

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

5
6 **Lakshminarayan M. Iyer, Vivek Anantharaman, Maxim Y. Wolf & L. Aravind***

7
8 National Center for Biotechnology Information, National Library of Medicine, National
9 Institutes of Health, Bethesda, MD 20894, USA

10
11 *Address for correspondence

12 LA (aravind@mail.nih.gov)

13 National Center for Biotechnology Information,

14 National Library of Medicine,

15 National Institutes of Health,

16 Bethesda, MD 20894,

17 United States of America

18 Phone: 301-594-2445

1		
2	Abstract	3
3	1. Introduction	4
4	2. Eukaryotic phylogeny and genomics	7
5	2.1 Repeated evolution of parasitism in protists	7
6	2.2 Key eukaryotic features revealed by comparative genomics	9
7	2.3 Demographic patterns in the distribution of transcription factors and	
8	chromatin proteins	10
9	3. Diversity of eukaryotic specific transcription factors	12
10	3.1 Identification of novel specific transcription factors in protist lineages	
11	12
12	3.2 Major trends in the evolution of TFs	14
13	4. Conserved domains in eukaryotic chromatin proteins	17
14	4.1 Definition and detection of chromatin protein domains	17
15	4.2 DNA-binding domains in chromatin proteins	18
16	5. The evolution of major functional guilds of chromatin proteins	19
17	5.1 Evolutionary history of histone acetylation-based regulatory systems	20
18	5.2 Natural history of histone-methylation-based regulation	24
19	5.3 Evolution of chromatin remodeling and assembling systems	28
20	5.4 Other chromatin protein modifications, potential histone tail	
21	interaction domains and histone chaperones	30
22	5.5 Natural history of epigenetic DNA modification enzymes	32
23	6. Domain architectures of chromatin proteins	34
24	6.1 Syntactical features in domain architectures of chromatin proteins:	
25	nature of interactions between different regulatory systems	34
26	6.2 Relationship between phylogeny, organizational complexity and	
27	domain architectures of chromatin proteins	36
28	7. Interactions between RNA-based regulatory systems and chromatin	
29	factors	38
30	8. General considerations and conclusions	40

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 are different even between sister lineages,
11 and have been shaped by gene loss and lineage-specific expansions. Recognition of
12 this principle has helped in identifying the hitherto unknown, dominant specific
13 transcription factors of several protists, 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
22 to 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 are
29 considerably variable across eukaryotes. This indicates that the histone code might
30 be “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 over eukaryotic evolution. Thus, *Trichomonas*, *Giardia*, *Naegleria* and
33 kinetoplastids have relatively simple domain architectures, whereas apicomplexans
34 and oomycetes have more complex architectures. RNA-dependent post-
35 transcriptional silencing systems, which interact with chromatin-level regulatory
36 systems, show considerable variability across parasitic protists, with complete loss in
37 many apicomplexans and partial loss in *T. vaginalis*. This evolutionary synthesis offers
38 a robust scaffold for future investigation of transcription and chromatin structure in
39 parasitic protists.

40
41 **Key words:** transcription factors, MYB, histones, methylation demethylation,
42 acetylation, deacetylation, domain architectures, evolution, PHD, chromo, bromo

1 **1. Introduction**

2 The unique configuration of the eukaryotic transcription apparatus sets it apart from
3 its counterparts in the archaeal and bacterial superkingdoms (Best et al., 2004;
4 Conaway and Conaway, 2004; Latchman, 2005). The basal or general transcription
5 apparatus of eukaryotes and archaea share several unique features. These include:
6 (1) structure of the RNA polymerase catalytic subunit (the 3 largest subunits
7 equivalent to the bacterial β' , β and α subunits); (2) specific accessory RNA
8 polymerase subunits (e.g. RPB10); (3) proteins constituting the basal transcription
9 initiation apparatus (general or global transcription factors), like TATA box-binding
10 protein (TBP), TFIIB, TFIIE and MBF (Reeve, 2003; Conaway and Conaway, 2004).
11 In contrast, certain components of the eukaryotic transcription elongation complex,
12 like the Spt6p-type of RNA-binding proteins, are shared with bacteria rather than
13 archaea (Anantharaman et al., 2002). Thus, during the endosymbiotic origin of
14 eukaryotes, the archaeal precursor appears to have contributed the core
15 transcription apparatus, including bulk of the basal or general transcription factors,
16 with a few additional elements being supplied by the bacterial partner (Dacks and
17 Doolittle, 2001; Reeve, 2003; Best et al., 2004; Conaway and Conaway, 2004;
18 Aravind et al., 2005; Aravind et al., 2006). Like the two prokaryotic superkingdoms,
19 several eukaryotes possess specific transcription factors (TFs) that are required for
20 transcriptional regulation of particular sets of genes (Latchman, 2005). In both
21 prokaryotic superkingdoms majority of specific TFs are members of a relatively small
22 group of protein families containing the helix-turn-helix (HTH) DNA-binding domain
23 (Aravind et al., 2005; Pellegrini-Calace and Thornton, 2005). Remaining prokaryote-
24 specific TFs mostly belong to two other superfamilies, the MetJ/Arc (Ribbons-helix-
25 helix) and the AbrB superfamily. Several families of eukaryote-specific TFs, like
26 homeodomain and Myb domain TFs, also bind DNA via the HTH domain, but TFs of
27 the AbrB and MetJ/Arc superfamilies are absent in eukaryotes (Aravind et al., 2005;
28 Latchman, 2005). However, almost all eukaryotic HTH-containing specific TFs do not
29 belong to any of the prokaryotic HTH families, and are only very distantly related to
30 them in sequence (Aravind et al., 2005; Pellegrini-Calace and Thornton, 2005).
31 Additionally, eukaryotes possess numerous large families of specific TFs containing
32 an astonishing array of DNA-binding domains (DBDs) that span the entire spectrum
33 of protein folds (Babu et al., 2004; Latchman, 2005). This deployment of specific TFs
34 with an immense structural diversity of DBDs is a dramatic difference in the
35 transcription apparatus of eukaryotes vis-à-vis the prokaryotic superkingdoms.

1
2 The nucleus, the defining feature of eukaryotes, along with their linear chromosomes
3 and highly dynamic chromatin also profoundly affect transcription regulation. This
4 cytological feature, in contrast to the prokaryotic situation, decoupled transcription
5 from translation, and necessitated transport of RNA from the nucleus to the
6 cytoplasm for translation (Mans et al., 2004; Denhardt et al., 2005). In terms of
7 chromosomal organization, eukaryotes share histones as the basic DNA-packaging
8 protein complex with several archaea (White and Bell, 2002; Reeve et al., 2004).
9 However, eukaryotic histones possess long positively charged tails, which are targets
10 of several post-translational modifications like acetylation, methylation,
11 phosphorylation and ubiquitination (Martens and Winston, 2003; Denhardt et al.,
12 2005; Allis et al., 2006; Kouzarides, 2007). Enzymes mediating these modifications
13 are a universal feature of eukaryotes and regulate transcription both globally and
14 locally by dynamically remodeling chromatin to allow or restrict access to general
15 and specific TFs (Collins et al., 2007; Kouzarides, 2007). In certain eukaryotes, the
16 dynamics of chromatin structure and transcription are also affected by modification
17 of bases in DNA (e.g. methylation) (Goll and Bestor, 2005; Allis et al., 2006).
18 Another aspect of chromatin remodeling in eukaryotes is the use of several distinct
19 types of ATP-dependent engines that alter chromatin structure both on a
20 chromosomal scale and locally. The extent of deployment of these ATP-dependent
21 chromatin remodeling engines in eukaryotes is vastly greater in magnitude than in
22 prokaryotes (Martens and Winston, 2003; Denhardt et al., 2005; Allis et al., 2006).
23 Also associated with chromatin are protein complexes of the nuclear envelope and
24 nuclear pores that mediate local interaction with chromosomes via telomeres and
25 matrix attachment regions (Mans et al., 2004). Post-transcriptional RNA-based
26 regulatory mechanisms that deploy small interfering RNAs and microRNAs (siRNAs
27 and miRNAs) interface with chromatin proteins and the transcription regulation
28 apparatus to effect specific transcriptional silencing, to direct modification of DNA
29 and chromatin proteins, and to initiate chromatin condensation (Anantharaman et
30 al., 2002; Grewal and Rice, 2004; Ullu et al., 2004; Allis et al., 2006; Vaucheret,
31 2006). The RNA component of eukaryotic chromatin also contains various pre-mRNA
32 processing complexes, and other poorly understood large non-coding RNAs
33 (Denhardt et al., 2005).
34

1 The unifying features of the transcription and chromatin dynamics apparatus across
2 eukaryotic model organisms notwithstanding, several studies have hinted at an
3 enormous lineage-specific diversity in the types of specific TFs and domain
4 architectures of chromatin proteins (Koonin et al., 2000; Coulson et al., 2001;
5 Lander et al., 2001; Lespinet et al., 2002; Sullivan et al., 2006). A potential corollary
6 to this observation was that the variety in specific TFs and chromatin-protein
7 architectures might provide the regulatory basis for the emergence of enormous bio-
8 diversity in terms of structure, life-styles and life-cycles across the eukaryotic
9 evolutionary tree (Coulson et al., 2001; Lander et al., 2001; Lespinet et al., 2002).
10 Phylogenetic investigations have shown that model organisms represent only a small
11 portion of the vast eukaryotic tree, with most of the bewildering diversity found in
12 the unicellular microbial eukaryotes or '**protists**' (Moon-van der Staay et al., 2001;
13 Baptiste et al., 2002; Simpson et al., 2006). Thus, a proper understanding of
14 transcription regulation and chromatin dynamics across the wide variety of protists is
15 critical to approach anywhere close to a complete picture of the natural history of
16 these systems in eukaryotes. Furthermore, given that several lineages of protists
17 have spawned human, livestock and crop parasites with an extraordinary range of
18 adaptations, this understanding will be critical in any future attempts to tackle
19 parasitic diseases. Fortunately, recent large-scale genome sequencing efforts have
20 generated complete or near-complete genome sequences of several protists, which
21 are either agents of major parasitic diseases or key players in world-wide
22 ecosystems (Fig. 1).

23
24 Traditional approaches to study protist parasitism have been greatly hampered by
25 practical difficulties relating to their complex multi-host lifecycles, *in vitro* culturing
26 and maintenance, as well as lack of proper animal models in certain cases (Kreier,
27 1977). Hence, experimental analyses on protist regulatory systems, especially
28 transcription and chromatin dynamics, are far from the levels that have been
29 achieved in eukaryotic model organisms. However, recent successes of comparative
30 genomics and its resonance with new technologies are vastly improving the situation.
31 In this article we use the treasure-trove of data from recently published protist
32 genome sequences to reconstruct and review chief aspects of the transcription
33 regulatory and chromatin reorganization apparatus in protists as well as multicellular
34 forms. Placing parasitic protists in the appropriate evolutionary context with their
35 free-living relatives, and other eukaryotes, helps us to highlight the multiple means

1 by which these regulatory systems have diversified across eukaryotes. Thus, this
2 review grounded in comparative genomics, attempts to fill the lacunae in the
3 evolutionary framework of our understanding of these systems, and tries to develop
4 a stage for future experimental forays in protists.

6 **2. Eukaryotic phylogeny and genomics**

7 **2.1 Repeated evolution of parasitism in protists**

8 Despite availability of genome-scale data, reconstruction of eukaryotic phylogeny has
9 not been straight-forward (Baptiste et al., 2002; Templeton et al., 2004; Arisue et
10 al., 2005; Walsh and Doolittle, 2005; Simpson et al., 2006). Some principal
11 problems that confound determination of higher order relationships amongst
12 eukaryotes are: 1) Rampant gene loss. This is common throughout the fungal
13 kingdom and especially pronounced in the microsporidian lineage (Aravind et al.,
14 2000; Katinka et al., 2001). *Entamoeba* amongst amoebozoans, *Cryptosporidium*
15 amongst apicomplexans and *Giardia* amongst basal eukaryotes also display extreme
16 gene loss relative to their sister lineages (Templeton et al., 2004; Loftus et al.,
17 2005; Carlton et al., 2007). 2) Gene loss also spurs concomitant rapid sequence
18 divergence of the proteins that have been retained on account of release from
19 selective constraints due to lost interacting partners (Aravind et al., 2000). 3) Lateral
20 gene transfer. Some eukaryotic lineages like chromists (stramenopiles) and
21 apicomplexans have emerged via secondary or tertiary endosymbiosis involving
22 engulfment of other eukaryotic cells from the plant lineage (Bhattacharya et al.,
23 2004). As a result their proteins show chimeric affinities to either those of the
24 original lineage or to those of the endosymbiont's lineage. In addition to these
25 issues, there are controversies concerning the rooting of the eukaryotic tree and the
26 nature of the last eukaryotic common ancestor (Arisue et al., 2005; Walsh and
27 Doolittle, 2005). Nevertheless, multiple independent recent studies using large multi-
28 protein datasets and algorithms to correct for differential evolutionary rates have
29 been robustly reproducing several higher order groupings (Fig. 1) (Baptiste et al.,
30 2002; Templeton et al., 2004; Walsh and Doolittle, 2005; Simpson et al., 2006).
31 This phylogenetic framework, combined with comparative genomics, provides a
32 reasonable model for eukaryotic evolution.

33
34 In this framework, animals and fungi form a strongly monophyletic lineage, with
35 amoebozoans as their immediate sister group. The plant lineage forms the sister

1 group to animals, fungi and amoebozoans, and together this assembly is referred to
2 here as the crown group (Fig. 1). Both unicellular (*protist*) as well as multicellular
3 forms spanning an entire organizational range are seen in each of the crown-group
4 lineages: for example, the plant lineage contains unicellular free-swimming algae
5 such as *Chlamydomonas*, as well as multicellular forms like the terrestrial vascular
6 plants, while amongst amoebozoans we encounter facultative multicellularity in
7 social amoebae like *Dictyostelium*. Likewise, parasitism has repeatedly emerged in
8 crown-group lineages (Fig. 1). The fungal lineage in particular has spawned several
9 parasites, including the human parasite *Cryptococcus* and plant parasites like
10 *Ustilago*. Most unusual of these are the structurally highly derived microsporidians,
11 which possess amongst the most reduced of eukaryotic genomes (Katinka et al.,
12 2001). Recent analyses suggest that they might be derived from within chytrids, the
13 basal-most lineage of fungi (James et al., 2006). The animal lineage too has given
14 rise to microbial parasites, namely the enigmatic myxozoa, which were previously
15 classified with microsporidians (Smothers et al., 1994). Amongst amoebozoans the
16 best-studied parasite is the human gut parasite *Entamoeba histolytica* (Loftus et al.,
17 2005). Even in the predominantly auxotrophic plant lineage microbial parasites have
18 emerged amongst rhodophytes, which deliver their nucleus into host cells belonging
19 to other rhodophyte species (Goff and Coleman, 1995).

20

21 The chromalveolate assemblage forms the next major monophyletic group that
22 includes the diverse stramenopiles (chromists) and alveolate lineages. Alveolates in
23 turn include apicomplexans, dinoflagellates (and *Perkinsus*) and ciliates, while
24 stramenopiles include an extraordinary range of predominantly photosynthetic forms
25 like diatoms, phaeophytes (brown algae, like kelp), chrysophytes (golden algae) and
26 non-photosynthetic oomycetes (Bhattacharya et al., 2004). Among alveolates,
27 apicomplexans are striking in being one of the few wholly parasitic lineages of
28 eukaryotes and include major animal parasites like the malarial parasite
29 *Plasmodium*, *Theileria*, *Toxoplasma* and *Cryptosporidium* (Kreier, 1977; Leander and
30 Keeling, 2003). Among stramenopiles, oomycetes, like *Phytophthora* are amongst
31 the most destructive of crop parasites (Tyler et al., 2006). The chromalveolate clade
32 forms a sister group to the crown group to the exclusion of other eukaryotes (Fig. 1).
33 Remaining "basal" eukaryotes mainly include numerous poorly characterized forms,
34 but some major monophyletic lineages are prominent amongst them. Of these the
35 euglenozoans, *Jakoba* and *Naegleria* form a well-supported lineage with diverse life-

1 styles and cycles (Fig. 1) (Simpson et al., 2006). Trypanosomes being major human
2 and livestock parasites are the best studied of euglenozoans, and more recently
3 there has been developing interest in *Naegleria*, an amoeboflagellate causing a rare
4 meningoencephalitis (Schuster and Visvesvara, 2004; El-Sayed et al., 2005). The
5 basal-most eukaryotic clades are believed to include the parabasalids and
6 diplomonads, which are respectively prototyped by the parasites *Trichomonas* and
7 *Giardia* (Best et al., 2004; Carlton et al., 2007).

8 9 **2.2 Key eukaryotic features revealed by comparative genomics**

10 Burgeoning genome sequencing projects have generated complete sequences of
11 major representatives of most of the above-discussed eukaryotic lineages (Fig. 1).
12 Results of comparative genomics have forcefully brought home certain large-scale
13 trends in eukaryotic evolution. Firstly, the hybrid evolutionary origins of most major
14 eukaryotic regulatory systems, especially those related to transcription and post-
15 transcriptional control, have been reinforced—different components have been
16 derived from either the archaeal precursor as well as the primary bacterial
17 endosymbiont (the mitochondrial), or, on multiple occasions, from various other
18 bacterial lineages, which might be associated as symbionts or consumed as food
19 (Koonin et al., 2000; Dacks and Doolittle, 2001; Walsh and Doolittle, 2005; Aravind
20 et al., 2006). Comparative genomics has also revealed the enormous plasticity of
21 eukaryotic genomes and rampant reorganization by lineage-specific expansions
22 (LSE) of genes and gene loss (Aravind et al., 2000; Katinka et al., 2001; Lespinet et
23 al., 2002). Massive gene loss relative to free-living forms is a prevalent feature of
24 most parasitic lineages. One exception is the basal eukaryote *Trichomonas*, which
25 possesses gene numbers comparable or greater than animals, plants and ciliates
26 (Carlton et al., 2007). The most parsimonious reconstruction considering the above
27 phylogenetic scenario suggests that the last eukaryotic common ancestor (LECA)
28 already possessed a distinctly larger gene complement (at least ~10,000 genes)
29 than its prokaryotic precursors. This complement coded numerous families of
30 proteins with multiple paralogous members, and several novel regulatory systems
31 with no direct prokaryotic equivalents (Aravind et al., 2006).

32
33 Availability of complete genome sequences also allows us to estimate the gross
34 differences in effects of natural selection on completely conserved orthologous
35 proteins belonging to different functional categories (Baptiste et al., 2002).

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

19

20 ***2.3 Demographic patterns in the distribution of transcription factors and*** 21 ***chromatin proteins***

22 Generation of sensitive sequence profiles and hidden Markov models for conserved
23 domains found in TFs (typically their DNA-binding domain) and chromatin proteins
24 allows their exhaustive and systematic detection across all complete eukaryotic
25 proteomes (Coulson et al., 2001; Babu et al., 2004; Finn et al., 2006). As a result,
26 reasonably robust counts or demography of potential TFs and
27 chromosomal/chromatin proteins (CPs) encoded by a given organism can be
28 obtained. These results show reasonably strong positive correlations between the
29 number of CPs or TFs coded by an organism and its proteome size (Fig. 2B, C).
30 These trends are best approximated by linear or mildly non-linear fits (weak
31 quadratic fit for TFs or weak power-law in chromatin factors), suggesting that, in
32 general, there is a proportional increase in the number of TFs for increasing number
33 of protein-coding genes. The trend observed in TFs is in contrast to that seen in
34 prokaryotes wherein a fit to a much stronger power-law trend is observed (Babu et
35 al., 2004; Aravind et al., 2005). However, in prokaryotes there appear to very few

1 dedicated CPs, and their number does not vary dramatically with proteome size. This
2 suggests that in general eukaryotes might optimize their transcription regulatory
3 potential by increasing numbers of both TFs and chromosomal proteins as their gene
4 numbers increase. As a result the scaling behavior of their TF counts is apparently
5 different from prokaryotes.

6
7 Parasites belonging to fungal, apicomplexan and stramenopile lineages show greater
8 or lesser degrees of gene loss in comparisons to their free-living sister clades, but
9 typically counts of their TFs and CPs do not deviate to a large extent from the
10 general trend observed across eukaryotes. This suggests that despite a degree of
11 genomic reduction, the overall regulatory input per protein-coding gene is roughly
12 comparable to other eukaryotes. Significant exceptions to the general eukaryotic
13 trend in TFs were seen in trypanosomes, while *Trichomonas vaginalis* and ciliates
14 displayed significant deviations in counts of both their TFs and CPs (Fig. 2B,C). The
15 notably lower TF count in trypanosomes relative to their proteome size might imply
16 that they possess a unique family of TFs that are unrelated to any previously
17 characterized variety and have eluded detection thus far. In *T. vaginalis* and ciliates
18 the absolute counts of TFs and CPs exceed those seen in other parasites or free-
19 living protists. However, their proteome size is similar to that of multicellular animals
20 and plants, and as result they have relatively far fewer TFs and CPs for their
21 proteome sizes compared to the multicellular forms (Fig. 2B, C). This might be due
22 to different parallel reasons: 1) Multicellular forms show both temporal
23 transcriptional changes during development and spatially differentiated cell-types
24 with diverse gene-expression states. In contrast, a parasite like *T. vaginalis* shows
25 relatively simple temporal development and has no equivalent of differentiated cell
26 fates. Likewise, though ciliates have amongst the most complex cell-architectures
27 seen in eukaryotes, they possess a relatively simple development and no
28 differentiated cell-types. Consequently, lower normalized counts of TFs in these
29 organisms might reflect differences in the amount of transcriptional control required
30 to regulate similarly sized genomes in the unicellular context (*T. vaginalis* or ciliates)
31 as opposed to multicellular forms with differentiation. 2) These protists also show
32 tremendous genetic redundancy with several closely related or near-identical gene
33 copies that, rather than being differentially regulated, might merely provide higher
34 effective concentrations of particular gene products (Aury et al., 2006; Carlton et al.,

1 2007). The gene counts, especially in *T. vaginalis*, are also exaggerated by numerous
2 transposable elements of diverse types (Carlton et al., 2007).

3 4 **3. Diversity of eukaryotic specific transcription factors**

5 6 **3.1 Identification of novel specific transcription factors in protist lineages**

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

24
25 On the practical side, this feature of eukaryotic TFs often makes their prediction in
26 poorly-studied lineages, especially parasites, a difficult task. This was poignantly
27 illustrated by the case of the apicomplexans, where multiple studies had initially
28 failed to recover *bona fide* specific TFs (Gardner et al., 2002; Templeton et al.,
29 2004). However, analysis of stage-specific gene expression in *Plasmodium*
30 *falciparum* revealed a complex pattern of changing gene expression that resulted in
31 genes with increasing functional specialization being expressed as intra-erythrocytic
32 development cycle (IDC) progressed (Bozdech et al., 2003; Le Roch et al., 2003).
33 This was also supported by expression studies in *Theileria* (Bishop et al., 2005), and
34 pointed to a specialized transcription regulatory program similar to that seen in
35 model organisms from the crown-group. Sensitive sequence profile analysis revealed

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

23

24 This discovery of the dominant specific TFs of apicomplexans serves as a model for
25 the identification of uncharacterized TFs in other protist lineages. Another
26 noteworthy example of this is provided by *T.vaginalis*. Transcription initiation in this
27 organism is primarily dependent on the protein IBP39, which binds the initiator
28 element (Inr) by means of a specialized winged HTH (wHTH) domain, termed the
29 IBD, and recruits the RNA polymerase via its C-terminal tail (Schumacher et al.,
30 2003; Lau et al., 2006). The recognition helix of the wHTH binds the major groove of
31 DNA, while a distinctive positively charged loop from a bi-helical hairpin at the N-
32 terminal contacts the adjacent minor groove. This novel DNA-binding domain, while
33 containing an ancient fold, has no close relatives in any other organism studied to
34 date (Schumacher et al., 2003; Lau et al., 2006). Given the generally low ratios of
35 specific TFs to proteome size in *T.vaginalis* and the elusive origins of the IBD of

1 IBP39, we investigated it using sequence profile searches to determine if it might
2 define a novel family of lineage-specific TFs. As a result we were able to identify a
3 family of at least 100 proteins in the *T.vaginalis* proteome, containing single IBDs
4 and congruent architectures as IBP39 (see Supplementary material). This suggests
5 that the IBD indeed defines a lineage-specific DNA-binding domain that is utilized by
6 a large family of specific TFs in this organism. Sequence divergence in the
7 recognition helix as well as the N-terminal positively charged loop across the IBD
8 family suggests that different versions of the domain have specialized to contact a
9 range of target sites, other than the *T.vaginalis* INR. Hence, this organism might
10 have utilized the lineage-specific expansion of a single family of TFs in wide range of
11 transcription regulatory contexts, including possibly global regulation as suggested
12 by the case of IBP39.

13

14 **3.2 Major trends in the evolution of TFs**

15 A survey of DBDs found in eukaryotic specific TFs shows that there are about 55
16 distinct superfamilies spanning all structural classes, with some of them present in
17 almost all eukaryotes studied to date (Fig. 3). This latter group contains at least 7
18 DBDs, namely the Basic-zipper (bZIP), C2H2 ZnF, HMG BOX, AT-hook, MYB,
19 CBF/NFYA, and E2F/DP1 DNA-binding domains. These, along with DBDs of general
20 transcription factors like that of the TATA-binding protein (TBP), TFIIB, TFIIE and
21 MBF which were inherited from the archaeal ancestor, and the BRIGHT/ARID which
22 emerged in eukaryotes, comprise the set of DBDs in TFs that can be confidently
23 traced to LECA (Best et al., 2004; Aravind et al., 2005). While majority of DBDs in
24 the ancient set shared with archaea contain the HTH fold, only the BRIGHT and MYB
25 domains amongst the early eukaryotic innovations possess this fold (Aravind et al.,
26 2005). This suggests that the recruitment of a structurally diverse set of DBDs to TFs
27 had already begun early in eukaryotic evolution. The wide distribution of specific TFs
28 with several other DBDs, like the MADS, GATA and Forkhead (FKH) domains, in
29 early-branching eukaryotes also suggests a relatively ancient origin for these
30 proteins in eukaryotic evolution (Fig. 3). Another major round innovation of TFs, with
31 new DBDs such as the CENPB, HSF and bHLH domains, appears to have happened
32 prior to divergence of the crown group and the chromalveolate clade. Finally, there
33 were extensive innovations of several other DBDs within the crown group. A striking
34 example of this are the DBDs of the fast-evolving p53-like fold. The earliest
35 representatives of this fold were present in the ancestor of the crown group and

1 typified by the DBD of the STAT proteins (Fig. 3) (Soler-Lopez et al., 2004). We
2 identified TFs of the STAT family in *Entamoeba histolytica* where they could
3 potentially function downstream of receptor kinases in processes related to its
4 pathogenesis (Fig. 3). The p53-like fold subsequently appears to have diversified
5 greatly in animals and fungi giving rise to 4 distinct families, including the animal
6 p53 proper. Finally, there are some TFs that appear to be found in a single lineage of
7 eukaryotes; some striking examples being the above-mentioned IBDs of *T. vaginalis*,
8 the APSES family of fungi, and a previously uncharacterized family of Zn-chelating
9 TFs (often also containing additional AT-hook motifs (Aravind and Landsman, 1998))
10 that are found in the stramenopiles like *Phytophthora* (Fig. 3).

11
12 Irrespective of their point of origin, individual eukaryote-specific TFs show highly
13 variable demographic patterns (Babu et al., 2004). Thus, the same family of TF
14 might be independently expanded across several distantly related taxa, whereas
15 sister taxa might drastically differ from each other in terms of their principal TFs (Fig.
16 3). The case of the ApiAP2 proteins suggests that the AP2 domain has been
17 independently expanded in both multicellular plants and apicomplexa but are present
18 in very low numbers in their respective immediate sister groups namely, the
19 chlorophyte algae (*Chlamydomonas* and *Ostreococcus*) and ciliates. Likewise the MYB
20 domain shows enormous LSEs in multicellular plants, the free-living ciliate
21 *Paramecium*, and phylogenetically distant parasites like *T. vaginalis*, *E. histolytica* and
22 *Naegleria gruberi*. In *E. histolytica* the expanded MYB proteins appear to constitute
23 the predominant specific TFs of this species (Fig. 3). Other examples of major
24 independent LSEs of TFs observed both in diverse parasites and free-living protist
25 groups include the bZIP domain in *Phytophthora* and *Paramecium*, and the heat-
26 shock factor (HSF) in most stramenopiles and *Paramecium*. While the C2H2 Zn-
27 finger (ZnF) is prevalent in most eukaryotic lineages, in each lineage its rise in
28 numbers appears to be a result of independent LSEs (Fig. 3) (Coulson et al., 2001;
29 Lespinet et al., 2002; Babu et al., 2004; Babu et al., 2006). For example in ciliates
30 like *Tetrahymena*, a LSE comprising of proteins combining the C2H2-ZnF with AT-
31 hooks appear to constitute the dominant TFs of this organism (Fig. 2). Interestingly,
32 ciliates (especially *Paramecium*) show an expansion of the DNA-binding CXC domain
33 that is normally found as a general DBD in chromosomal proteins rather than specific
34 TFs (Hauser et al., 2000). Its unusual expansion and presence in standalone form,

1 unlike chromosomal proteins, where it is fused to other domains, suggest that these
2 proteins possibly functions as specific TFs in ciliates (Fig. 3).

3
4 Beyond LSEs, other major forces in the evolution of TFs appear to be gene losses
5 and lateral transfers, as suggested sporadic phyletic patterns of several
6 superfamilies. Several families of TFs are shared by animals and plants or
7 amoebozoans to the exclusion of the fungi. However, phylogenetic analysis strongly
8 supports the exclusive grouping of animals and fungi, suggesting loss in the latter
9 (Fig. 3). One striking example is furnished by the dimeric E2F and DP1 transcription
10 factors (Templeton et al., 2004), which is present in animals, amoebozoans, plants,
11 chromalveolates and basal eukaryotes like *Trichomonas* and *Giardia*, while being
12 absent in all fungal lineages except the highly reduced parasite *Encephalitozoon*. This
13 pattern is highly suggestive of secondary loss of this ancient TF in the other fungi
14 after their separation from microsporidians. In contrast, some TFs like PBF2/TIF1,
15 exclusively shared by plants and chromalveolates might have been acquired by the
16 latter during the endosymbiotic association with the plant lineage. A specific version
17 of the WRKY TF is shared by plants, the plant parasite *Phytophthora* (shows a
18 notable expansion of over 20 copies) and *Giardia* (Babu et al., 2006). The C6 finger
19 was believed to be exclusively found in the fungal lineage, but has recently been
20 found in *Dictyostelium*, the stramenopile alga *Thalassiosira* and *Naegleria* with a
21 prominent lineage-specific expansion in the latter (Fig. 3). The sporadic phyletic
22 patterns of the WRKY and C6 domains in the protists are possibly the consequence of
23 lateral transfer respectively from the plant and fungal lineages. In some cases,
24 differentiating between the alternative explanations of gene loss and lateral transfer
25 is much harder with the current state of the data. For example, the homeodomain is
26 found in all crown group lineages in multiple copies. But amongst other protists the
27 atypical TALE subfamily of homeodomains (Burglin, 1997) are sporadically found in
28 ciliates, stramenopiles, *Naegleria* and *Trichomonas* pointing to a possible earlier
29 origin with frequent losses. However, in stramenopiles, certain homeodomains are
30 clearly closer to their plant counterparts, opening the possibility of lateral transfer
31 from the photosynthetic endosymbiont. Beyond the major endosymbiotic events and
32 close host-parasite interactions, the phagotrophic mode of nutrition of several
33 flagellate as well as amoeboid protists might have allowed lateral transfer between
34 distantly related lineages (Doolittle, 1998).

35

1 This pattern points to a universal situation in eukaryotes, where existing TFs are
2 being constantly lost, and new ones emerging through a variety of processes like LSE
3 and lateral transfer, and suggests a rapid turnover of regulatory DNA-protein
4 interactions. Hence, as previously postulated for crown-group model systems
5 (Lander et al., 2001; Lespinet et al., 2002), extensive differentiation of transcription
6 factors might be a major determinant that shapes adaptations of protists. This leads
7 to the question regarding the ultimate origin of eukaryotic TFs. Several families, like
8 the BRIGHT, homeo, POU, paired, HSF, IBD, MYB, TEA, FKH and pipsqueak domains
9 contain the HTH fold, albeit only distantly related to that seen in prokaryotic TFs.
10 Hence, they could have potentially emerged through rapid diversification of older
11 HTH domains inherited from prokaryotes (Aravind et al., 2005). Likewise, certain
12 other ancient folds like the C2H2 Znf and the immunoglobulin folds are found in the
13 DBDs of other eukaryotic TFs (Babu et al., 2004). Again these DBDs of eukaryotic
14 TFs might have been derived from the more ancient representatives of the respective
15 folds. Finally, as in the case of many other functional classes, eukaryotes have also
16 innovated transcription factor DBDs with entirely new folds. These are almost
17 entirely α -helical or metal-chelation supported structures, consistent with the greater
18 “ease” with which such structures are innovated *de novo* (Aravind et al., 2006). In
19 more immediate evolutionary terms, several specific TFs appear to have been
20 derived from DBDs of transposases and allied mobile elements. Examples of major
21 eukaryotic DBDs that appear to have had such an origin are the WRKY, AP2, PBF2,
22 VP1, paired, pipsqueak, CENPBP, APSES, BED-finger and GCR1 domains (Smit and
23 Riggs, 1996; Balaji et al., 2005; Babu et al., 2006). Typically, inactive mobile
24 elements that have lost catalytic activity of their transposase domain, but retain their
25 DBD appear to be “exapted” as new TFs.

26

27 **4. Conserved domains in eukaryotic chromatin proteins**

28

29 ***4.1 Definition and detection of chromatin protein domains***

30 Evidence from model systems suggests that various histone modifications comprise
31 an “extra-genetic” code termed the histone code (Dutnall, 2003; Peterson and
32 Laniel, 2004; Allis et al., 2006; Villar-Garea and Imhof, 2006; Kouzarides, 2007).
33 Critical to this regulatory process are enzymatic domains catalyzing covalent
34 histone/chromatin protein modification. These catalytic domains, along with those
35 which remove the covalent modifications, are a prominent class of regulatory

1 proteins found in eukaryotic chromatin. Reading of this histone code, in conjunction
2 with recognition of covalently modified bases in DNA, is central to the expression and
3 action of numerous epigenetic effects. An important class of protein domains includes
4 those mediating specific interactions with unmodified or variously covalently modified
5 histone side chains. These interactions are central to the recruitment of enzymatic
6 activities to lay out the histone code, read the code by specifically recognizing
7 modified histones, and allow energy-driven chromatin remodeling by recruiting
8 enzymes that catalyze these processes (de la Cruz et al., 2005; Allis et al., 2006;
9 Kim et al., 2006; Sullivan et al., 2006; Villar-Garea and Imhof, 2006; Kouzarides,
10 2007). It is impossible to precisely compartmentalize these disparate regulatory
11 complexes in chromatin from various complexes carrying out essential housekeeping
12 processes such as replication, recombination, DNA-repair and transcription.
13 Nevertheless, herein we adopt a strict definition for CPs and focus chiefly on
14 regulatory components. The distinctness of this set of proteins being defined here as
15 CPs is primarily suggested by the observation that they are mostly comprised of a
16 relatively small set of conserved protein domains (about 70-80), majority of which
17 are found nearly exclusively in eukaryotic CPs (Letunic et al., 2006) (Table 1). This
18 allows for relatively robust prediction of the complement of CPs through
19 computational analysis using sensitive sequence profile methods and HMMs (Finn et
20 al., 2006). Most of these domains can be classified under two broad functional
21 categories: **1) non-catalytic interaction or adaptor domains and 2) enzymatic**
22 **regulatory domains**. The former category can again be further sub-divided into
23 DNA-binding and protein-protein interaction domains (Table 1). We first briefly
24 discuss the DNA-binding domains, and then consider the remaining domains in
25 course of reconstructing the natural history of the major regulatory systems in
26 eukaryotic chromatin.

27

28 **4.2 DNA-binding domains in chromatin proteins**

29 The most basic DNA-protein interaction in eukaryotic chromatin is mediated by the 4
30 core histones that are universally conserved in all eukaryotes (Allis et al., 2006;
31 Woodcock, 2006). In addition to the core histones there are other related histone-
32 fold proteins, namely the smaller TAFs and general transcription factors like NFYB
33 and NFYC that appear to form octamer-like structures in the context of transcription
34 initiation complexes (Gangloff et al., 2001). The four core histones, NFYB, NFYC and
35 at least 3 of the histone fold TAFs (TAF6, TAF8 and TAF12) had diverged from each

1 other by the time of the last eukaryotic common ancestor (LECA). Interestingly,
2 these TAFs, and also the slightly later derived paralog TAF9, were independently,
3 repeatedly lost in most or all apicomplexans and all kinetoplastids. Histone H1, which
4 binds inter-nucleosomal linkers, is found in the crown group, stramenopiles and
5 *Naegleria*. Its distribution is suggestive of an origin in the crown group from the FKH
6 domain (Carlsson and Mahlapuu, 2002; Aravind et al., 2005) followed by lateral
7 transfers to stramenopiles during endosymbiosis with the plant lineage and
8 independently to *Naegleria*. The HMG box and AT-hook proteins can mediate
9 bending of the helical axis of DNA and play an important role in altering
10 chromosomal structure (Aravind and Landsman, 1998). The AT-hook also appears to
11 be frequently used as a supplementary DNA-binding interface in larger proteins to
12 form extended contacts by specifically interacting with the minor-groove of DNA.
13 DNA-binding domains of CPs such as the HMG box, CXXC, CXC domains, BRIGHT,
14 SAND (KDWK), C2H2-Znf and the AT-hook motif are shared with specific TFs.
15 However, excluding the C2H2 Zn fingers, these DBDs are predominantly found in CPs
16 and, unlike in TFs, they are typically found in the context of multi-domain proteins in
17 the CPs. The TAM (MBD) and SAD (SRA) domains specifically bind methylated DNA
18 and thereby allow recruitment of regulatory complexes to modified DNA (Aravind and
19 Landsman, 1998; Goll and Bestor, 2005; Johnson et al., 2007; Woo et al., 2007).
20 Yet others like the HIRAN, PARP-finger and Rad18 finger domains appear to
21 specifically recruit chromatin remodeling activities to damaged DNA (Iyer et al.,
22 2006).

23
24 Some DBDs, such as Ku have a critical role in chromosome structure and dynamics.
25 They bind matrix attachment regions of chromosomes, are part of the telomere
26 binding complex and are associated with the perinuclear localization of telomeres
27 (Riha et al., 2006). The ancestral Ku protein appears to have been acquired by the
28 eukaryotes from bacteria, where they are coded by a mobile DNA-repair operon
29 (Aravind and Koonin, 2001a), after the divergence of parabasalids and diplomonads.
30 On being acquired, a duplication gave rise to two paralogous subunits, Ku70 and
31 Ku80, which were vertically inherited ever since in eukaryotes. Interestingly, Ku was
32 lost independently in all studied apicomplexan lineages, with the exception of
33 *Toxoplasma*.

34

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

1 The opportunity offered by advances in genomics to reconstruct the evolutionary
2 history of the eukaryotic CPs allows us to answer certain, previously inaccessible
3 questions more robustly: 1) what was the complement of CPs functioning in LECA?
4 2) What were the lineage-specific innovations in CPs and how often they occur? 3)
5 What implications do differences in complements of CPs have for the epigenetic
6 regulation (e.g. generation and "interpretation" histone code) in different organisms?
7 4) Do differences in the domain organization of CPs have implication for eukaryotic
8 diversity? 5) Finally, with respect to parasites, we can now examine the degree to
9 which different regulatory systems are maintained and modified as parasitism
10 convergently evolved in different eukaryotic lineages (Sullivan et al., 2006). To
11 address these questions we discuss below the evolutionary history of the major
12 functional guilds amongst CPs, with a focus on new data from protists, and try to
13 infer the functional implications of this evolutionary history. It should be kept in mind
14 that protists with few notable exceptions are relatively poorly-studied eukaryotes and
15 the reconstruction presented here is necessarily speculative. However, basic trends
16 discussed here are likely to hold good even with new data from future experimental
17 investigations.

18

19 **5.1 Evolutionary history of histone acetylation-based regulatory systems**

20 Majority of confirmed histone lysine acetyltransferases (HATs) belong to the ancient
21 superfamily of N-acetyltransferases typified by the universally found eukaryotic
22 histone acetyltransferase GCN5 (also called GNAT acetyltransferases)(Neuwald and
23 Landsman, 1997). Recently a fungal specific class of HATs, the Rtt109p family, which
24 is also found in the reduced parasite *Encephalitozoon*, has been reported as being
25 unrelated to the GNAT enzymes (Schneider et al., 2006; Collins et al., 2007; Driscoll
26 et al., 2007; Han et al., 2007). However, analysis of the secondary structure
27 predictions suggests that it is a highly divergent derivative of the GNAT fold probably
28 derived from the bacterial acyl-homoserine lactone synthase family (Neuwald and
29 Landsman, 1997). At least 14 distinct families of the GNAT fold appear to be
30 dedicated acetylases and appear to have specialized to perform numerous specific
31 roles in eukaryotic chromatin (Fig. 4). Of these, at least 4 can be traced back to
32 LECA, and are multi-domain proteins fused to peptide-binding domains like bromo
33 (Gcn5p) and chromo (Esa1p), or other catalytic domains like an ATPase domain
34 related to the N-terminal domain of the SFI helicase module (Kre33p), and a radical
35 SAM enzyme domain (Elp3p). Of these, Gcn5p is critical for histone acetylation in

1 connection to transcriptional activation by specific TFs, Elp3p is required for
2 transcription elongation, and Esa1p appear to have a negative regulatory role by
3 favoring transcriptional silencing (Wittschieben et al., 1999; Durant and Pugh, 2006;
4 Paraskevopoulou et al., 2006). The radical SAM domain of Elp3p cleaves SAM and
5 might play a role in an as yet unknown modification or in interfering with histone
6 methylation that requires SAM as a substrate (Paraskevopoulou et al., 2006).

7

8 Of the remaining families of acetylases the Eco1p orthologs (implicated in
9 chromosome segregation (Bellows et al., 2003)) have been present at least prior to
10 the branching-off of kinetoplastids. Others like Hat1p, and CSRP2BP and some
11 paralogs of the Esa1p, which form the MYST family (Thomas and Voss, 2007),
12 emerged in the crown group or the common ancestor of the crown group and
13 chromalveolates. *T. vaginalis* shows independent expansions of the MYST (Esa1p
14 orthologs) and Gcn5p acetylases. Several families are restricted to a particular
15 lineage (Neuwald and Landsman, 1997). For example, fungi appear to have at least
16 4 lineage specific families (orthologs of Spt10p, Hpa2p, Rtt109p and *Neurospora*
17 NCU05993.1), while plants have a lineage-specific family of their own with fusion of
18 the acetylase domain with PHD fingers or AT-hook motifs (Fig. 4). Amongst parasitic
19 protists, an unusual lineage-specific representative is seen in *Phytophthora* and
20 related stramenopiles, where the acetylase domain is fused to a
21 carboxymethyltransferase domain (Fig. 4). It is possible that these enzymes might
22 carry out a second covalent modification, perhaps of acidic side-chains. In
23 evolutionary terms, the Elp3p and Kre33p acetylases are shared by eukaryotes and
24 archaea suggesting an inheritance from the archaeal precursor, whereas Esa1p and
25 Gcn5p orthologs appear to be innovations specific to eukaryotes, which were derived
26 through rapid divergence from a pre-existing version of the fold. In contrast,
27 affinities of the lineage-specific versions suggest that they were acquired repeatedly
28 by eukaryotes from the diverse bacterial radiation of NH₂ group acetylases (Fig. 4).

29

30 Histone deacetylases belong to two structurally distinct superfamilies, namely the
31 RPD3/HDAC superfamily and the Sir2 superfamily, both of which are universally
32 present in eukaryotes. Prokaryotic members of both superfamilies appear to have
33 played predominantly metabolic roles, respectively participating in acetoin and
34 nicotinamide metabolism, as opposed to a regulatory role in chromatin (Leipe and
35 Landsman, 1997; Sandmeier et al., 2002; Avalos et al., 2004). The RPD3

1 superfamily uses metal-dependent catalysis, whereas the Sir2, superfamily, which
2 resembles the classical Rossmann fold enzymes, uses a NAD cofactor (Leipe and
3 Landsman, 1997; Avalos et al., 2004). At least a single deacetylase of the
4 HDAC/Rpd3 superfamily was present in LECA, and appears to have been derived
5 from bacterial acetoin-hydrolyzing enzymes (Fig. 4). There have been several
6 lineage-specific innovations within this superfamily amongst eukaryotes. Consistent
7 with the expansion of the acetylases, *T.vaginalis* also shows an expansion of HDAC
8 deacetylases, while kinetoplastids show a unique family typified by LmjF21.1870
9 from *Leishmania*. The chromalveolate clade, including the apicomplexans
10 *Cryptosporidium* and *T.gondii* share with plants a distinctive version of this family
11 that contains N-terminal ankyrin repeats. The fungal-specific HDA1p deacetylases
12 combine the HDAC domain with a C-terminal inactive α/β hydrolase domain that
13 might be utilized for specific peptide-interactions. *Phytophthora* and *Naegleria* also
14 possess lineage-specific architectures that respectively combine the HDAC domain
15 with AP2 and PHD finger domains and the BRCT domain (Fig. 4).

16
17 Sir2 superfamily deacetylases can be traced back to the common ancestor of
18 eukaryotes and archaea. In LECA there were at least 2 members of this superfamily,
19 respectively corresponding to the classical Sir2 and the precursor of Sirtuin 4, 5, 6
20 and 7 (Fig. 4). Sirtuin 4, 5 and 7 split up into separate lineages prior to divergence of
21 *Naegleria* and kinetoplastids from rest of the eukaryotes. Several parasitic protists
22 like *Giardia* and *Cryptosporidium* additionally possess one or more Sir2 superfamily
23 proteins distinct from the above eukaryotic families. This appear to have been
24 transferred from bacteria relatively early in eukaryotic evolution. Like the HDAC
25 superfamily, members of this family also show parallel domain fusions in protists:
26 *Dictyostelium* and *Tetrahymena* show fusions to tetratricopeptide and ankyrin
27 repeats. A Sir2 deacetylase from ciliates, amoebozoans (including *E.histolytica*) and
28 *Naegleria*, contains a fusion to the ubiquitin-binding Zn finger domain, which,
29 interestingly, parallels an equivalent fusion of this domain to a HDAC deacetylase in
30 animal HDAC6 enzymes (Fig. 4). These fusions point to several unique interactions
31 being used to recruit enzymes containing deacetylase domains of either superfamily
32 to specific contexts. In particular, the AP2 domain could recruit the deacetylase to
33 specific DNA sequences, ankyrin repeats to large proteins complexes and the BRCT
34 domain to complexes associated with DNA repair. The Ubp-ZnF could on the other

1 hand specifically recruit deacetylases to regions of chromatin containing
2 ubiquitinated histones or other ubiquitinated proteins.

3
4 Members of the Sir2 superfamily have also been shown to carry out NAD dependent
5 mono-ADP ribosylation of proteins and generate ADP-ribose as a by-product of the
6 deacetylation reaction (Frye, 1999; Avalos et al., 2004). Versions of the Macro
7 domain, prototyped by the vertebrate macrohistone 2A have been shown to bind O-
8 acetyl-ADP-ribose or hydrolyze ADP-ribose-1''-phosphate (Aravind, 2001; Karras et
9 al., 2005; Shull et al., 2005). In *E.histolytica*, certain fungi and *Phytophthora*, the
10 Sir2 domain is fused to the Macro domain (Fig. 4). Versions of the Macro domain are
11 also found in other CPs, for instance, fused to the SWI2/SNF2 ATPase module. These
12 occurrences suggest that the O-acetyl-ADP-ribose generated by Sir2 action might
13 elicit additional regulatory roles on CP dynamics (Karras et al., 2005). It is possible
14 that the Macro domain might recognize mono-ADP-ribosylated proteins and catalyze
15 the removal of this modification. This is also supported by their fusion to classical
16 protein ADP-ribosyl transferases in animals (Aravind, 2001). By binding or
17 hydrolyzing O-acetyl-ADP-ribose it might elicit a regulatory effect on Sir2 action by
18 potentially favoring the forward (deacetylation) reaction by removing ADP ribose. A
19 representative of the Macro domain appears to have been acquired from bacteria
20 prior to LECA itself. It is possible that these versions have a role in RNA metabolism
21 rather than chromatin dynamics (Shull et al., 2005). Versions involved in chromatin
22 dynamics appear to represent independent transfers from bacteria on multiple
23 occasions in evolution. One potential example, typified by the *Plasmodium* protein
24 MAL13P1.74, is conserved throughout alveolates and expanded in certain ciliates
25 suggesting a major role for ADP-ribose metabolites in these organisms.

26
27 Acetylated peptides are chiefly recognized by the tetrahelical bromo domain that
28 appears to be a unique eukaryotic innovation, specifically geared towards recognition
29 of the acetylation aspect of the histone code (Zeng and Zhou, 2002; de la Cruz et
30 al., 2005; Kouzarides, 2007). Bromo domains are found in all eukaryotes and had at
31 least 4 representatives in the LECA (Fig. 4). Two of the ancient and highly conserved
32 versions of the bromo domain are fused to enzymatic domains (see below). The
33 presence of a bromo domain in TAF1, which goes back to LECA, indicates an
34 ancestral role for this modification (potentially catalyzed by GCN5) in the context of
35 transcription initiation. Another ancestral bromo domain is represented by orthologs

1 of the *Drosophila* Fsh protein that interacts with acetylated H4. These proteins
2 appear to interact with the TFIID transcription initiation complex, and probably
3 recognize acetylation by Esa1p orthologs (Durant and Pugh, 2006). It combines 1-2
4 bromo domains with another conserved C-terminal α -helical domain also found in
5 TAF14. In *T. vaginalis*, consistent with the LSE of acetylases and deacetylases, this
6 version shows an extraordinary expansion with at least 100 representatives. Several
7 new multi-domain architectures involving the bromo domain emerged in crown-
8 group eukaryotes and appear to have been acquired by stramenopiles during the
9 endosymbiotic association with the plant lineage. Additional sporadic, lineage-specific
10 architectures also appear to have emerged in the alveolates, stramenopiles and
11 kinetoplastids (Fig. 4).

12

13 **5.2 Natural history of histone-methylation-based regulation**

14 Methylation of histones on lysines (both mono and trimethylation) is mediated
15 predominantly by methyltransferases of the SET domain superfamily, which are
16 universally present in eukaryotes (Allis et al., 2006; Sullivan et al., 2006;
17 Kouzarides, 2007). They are unrelated to classical Rossmann fold methylases and
18 contain a β -clip fold (Iyer and Aravind, 2004). All eukaryotes encode SET domain
19 methylases, and at least 4 distinct versions, namely Skm/Bop2-like, trithorax-like,
20 E(z)-like and Ash1-like SET domains can be traced back to LECA (Fig. 5). Versions
21 found in the basal eukaryotes, *Giardia* and *Trichomonas*, despite being orthologous
22 to their counterparts from other eukaryotes, do not display complex multidomain
23 architectures. Most major domain accretion resulting in these complex architectures
24 appears to have happened in the crown group, and few of these might have been
25 sporadically transferred to the chromalveolate clade from the plant lineage. One
26 example of this is a protein typified by *P. falciparum* PF08_0012, contains a fusion of
27 the DNA-binding SAD domain (Makarova et al., 2001; Johnson et al., 2007) to the
28 SET domain, and seems to have been acquired from the apicoplast precursor.
29 However, occasional lineage-specific domain fusions do appear to have emerged in
30 parasitic protists. *T. gondii* shows a fusion to HMG domain, which has also
31 independently occurred in animals and the alga *Ostreococcus*. Apicomplexans also
32 display another unique lineage-specific methylase combining the SET domain with
33 ankyrin repeats (Fig. 5). Basidiomycete fungi, including the parasitic protist
34 *Cryptococcus*, contain an unusual fusion of a SET domain with a nucleic acid
35 deaminase related to Tad3p (Gerber and Keller, 1999). It remains to be seen if these

1 proteins, in addition to catalyzing histone methylation, also catalyze DNA
2 modification via deamination.
3
4 The SET domain shows massive LSEs in kinetoplastids (at least 25 copies) and
5 *Phytophthora* (up to 60 copies). The former organisms contain proteins with up to 9
6 tandem SET domains, and others with the SET domain fused to enzymatic domains
7 homologous to bacterial D-Ala-D-Ala ligases, which might catalyze additional protein
8 modifications. These domain architectures suggest that in addition to the conserved
9 methylation events, the SET superfamily appear to have expanded and specialized to
10 mediate several lineage-specific regulatory processes, which might involve
11 recruitment of CP methylation to specific contexts. Rossmann fold
12 methyltransferases also play a role in CP methylation and are predominantly typified
13 by Dot1p-type H3 K79 methyltransferases (Sawada et al., 2004; Janzen et al., 2006)
14 and CARM1-like histone arginine methyltransferases (Cheng et al., 2007). The
15 former family is conserved throughout the crown group, kinetoplastids and
16 stramenopiles, but is absent in alveolates and basal eukaryotes. The latter family
17 appears to be absent in the basal eukaryotes *Giardia* and *Trichomonas*, but is
18 observed in all other eukaryotes, barring the highly reduced microsporidians.
19
20 Demethylation in majority of eukaryotes is carried out by the Jumonji-related
21 (JOR/JmjC) domain, which contains a double-stranded β -helix domain catalyzing a
22 metal and 2-oxo acid dependent oxidative demethylation of modified histones
23 (Anantharaman et al., 2001; Aravind and Koonin, 2001b; Chen et al., 2006; Cloos et
24 al., 2006; Klose et al., 2006). These enzymes appear to be ultimately of bacterial
25 origin, because numerous related as well as more divergent versions of double-
26 stranded β -helix enzymes are found throughout bacteria (Aravind and Koonin,
27 2001b). This demethylase, as well as other known demethylase domains (see
28 below), are absent in the two basal eukaryotic lineages, as well as several other
29 degenerate parasites like microsporidians and *E. histolytica*. This implies that certain
30 organisms can effectively function apparently without demethylation, though it is
31 theoretically possible that some other enzyme catalyzes this reaction in them.
32 Nevertheless, prior to the divergence of the kinetoplastid-*Naegleria* clade around 9
33 distinct versions of demethylases had emerged. As in the case of the SET domain
34 these demethylase domains typically show relatively simple domain architectures in
35 most early-branching eukaryotic groups, but have accreted multiple protein-protein

1 interaction and DNA-binding domains in crown-group eukaryotes. Kinetoplastids,
2 certain fungi and choanoflagellates show a fusion between the demethylase domain
3 and a carboxymethyltransferase domain (also fused to acetylases), and these
4 proteins might catalyze additional uncharacterized protein modifications (Fig. 5).

5
6 Another histone demethylase with a more limited distribution is the LSD1-like
7 demethylase containing a classical dinucleotide cofactor-binding Rossmann fold
8 domain related to the amino oxidases that oxidize the primary NH₂ groups of
9 polyamines (Aravind and Iyer, 2002; Shi et al., 2004b; Metzger et al., 2005;
10 Stavropoulos et al., 2006). These enzymes are present throughout the crown group,
11 in apicomplexans, stramenopiles and *Naegleria*. Their evolutionary affinities suggest
12 an origin in the crown group followed by secondary transfer to certain protist
13 lineages. Almost all of these demethylases are fused to the SWIRM domain, and
14 additionally show some lineage-specific fusions, like to the HMG box domain in fungi,
15 PHD finger in apicomplexans and PHDX/ZF-CW in vertebrates. Given that their
16 closest relatives, the amino oxidases, oxidize polyamines which are present in
17 chromatin, it needs to be seen if these enzymes might additionally catalyze oxidation
18 of NH₂ groups of histone side-chains or of polyamines, as part of a second regulatory
19 mechanism. Crystal structures of these enzymes indicate that, in addition to DNA-
20 binding, the SWIRM domain in histone demethylases might also help in the
21 recognition of methylated target peptides (Stavropoulos et al., 2006).

22
23 An assemblage of structurally related domains that contain modified versions of the
24 SH3-like fold such as the chromo (including AGENET and MBT), tudor, BMB (PWWP),
25 and the bromo-associated motif/homology(BAM/BAH) domain are predominantly
26 found in CPs (Maurer-Stroh et al., 2003). Recent direct experimental results, as well
27 as circumstantial evidence from different sources show many, if not all
28 representatives of these domains, are the primary binders of methylated histone tails
29 (Bannister et al., 2001; Lachner et al., 2001; Sathyamurthy et al., 2003; Brehm et
30 al., 2004; Flanagan et al., 2005; Bernstein et al., 2006; Kim et al., 2006). The
31 classical SH3 domain is itself an ancient peptide-binding domain that appears to
32 have been acquired by eukaryotes from bacterial precursors. Bacterial homologs of
33 these chromo-related domains are found in secreted or periplasmic proteins
34 associated with peptidoglycan, like the bacterial SH3 and SHD1 (Slap homology
35 domain 1; a eukaryotic peptide binding domain) (Ponting et al., 1999). In

1 eukaryotes, members of the SH3 fold underwent an explosive radiation especially in
2 connection to CPs, and might coincide with key adaptations related to reading of the
3 histone code in the context of methylation. This is paralleled by the radiation of SH3-
4 fold domains in eukaryotic cytoplasmic peptide-interacting proteins (Finn et al.,
5 2006; Letunic et al., 2006). Thus, the ancestral SH3 fold domains acquired from
6 bacteria appear to have specialized for nuclear and cytoskeletal peptide interactions,
7 probably concomitant with the origin of the eukaryotic nucleo-cytoplasmic
8 compartmentalization.

9
10 Comparisons of protist genomes indicates that the distinct versions of this fold
11 namely chromo, tudor, BAM/BAH had already separated from each other in LECA
12 itself, and BMB (PWWP) emerged prior to the divergence of the kinetoplastid-
13 *Naegleria* clade (Fig. 5). At least 3 distinct versions of the chromo domain (including
14 a HP1-like protein), one BAM/BAH domain and one version of the chromatin-
15 associated tudor domain can be extrapolated as being present in LECA. As with the
16 bromo domain, the ancient representatives of these domains include both forms that
17 are fused to other enzymatic domains, as well as those in non-catalytic proteins.
18 Though most parasites, such as apicomplexans show a relatively low number of
19 these domains, with some domains like the BMB (PWWP) being entirely absent in
20 them, *T. vaginalis* shows a LSE of proteins containing chromo domains. In the free-
21 living ciliate *Paramecium*, but none of the other chromalveolates, we observe an
22 unusual expansion of proteins containing fusions of the BAM (BAH) and PHD finger
23 domains. Interestingly, chromalveolates show several unique architectures
24 combining a version of chromodomain related to those found in the *Drosophila*
25 *malignant brain tumor* (MBT) protein (Maurer-Stroh et al., 2003; Sathyamurthy et
26 al., 2003) with several domains related to ubiquitin signaling, such as different
27 deubiquitinating peptidases of the Otu and UBCH families, the RING finger E3-ligase,
28 and ubiquitin-like domains (Fig. 5). These architectures point to the development of
29 a functional association between histone methylation and chromatin-protein
30 ubiquitination in these protists. Most of these proteins have been lost in the
31 apicomplexan parasites, but are retained in the plant parasite *Phytophthora*, along
32 with several additional lineage-specific architectures involving the chromodomain. In
33 this context it is of interest to note that a transposon encoding a chromodomain
34 protein has extensively proliferated in the genome of *Phytophthora*.

35

1 Recent studies have also shown that certain versions of the binuclear, zinc chelating
2 treble-clef fold domain, the PHD finger to bind all nucleosomal histones (Eberharter
3 et al., 2004). Versions of this domain also interact specifically with trimethylated
4 lysines on histone H3 (Li et al., 2006b; Pena et al., 2006; Shi et al., 2006). Some
5 versions of the PHD finger have been claimed to bind to phosphoinositides, but
6 recent experiments suggest a downstream basic sequence rather than the PHD
7 finger is directly involved in this interaction (Kaadige and Ayer, 2006). Given the
8 exclusive prevalence of this domain in CPs and its sequence diversity (Aasland et al.,
9 1995), it is possible that different versions of the PHD finger mediate distinct
10 interactions with trimethylated histones, or other modified and unmodified histones.
11 At least a single copy of the PHD finger was present in LECA, and the domain showed
12 considerable evolutionary mobility, beginning prior to the separation of the crown
13 group and chromalveolate clades, and again within the crown group. In general, on
14 account of its evolutionary mobility, the PHD finger tends to form several lineage-
15 specific architectures in the two above clades (Fig. 5).

16

17 **5.3 Evolution of chromatin remodeling and assembling systems**

18 Enzymes mediating dynamics of eukaryotic chromatin on local and global scales
19 typically do so by utilizing the free-energy of NTP hydrolysis. Not surprisingly, most
20 of these enzymes contain motor domains of the P-loop NTPase fold; two major
21 classes of which are the SWI2/SNF2 ATPases and the SMC ATPases (Bork and
22 Koonin, 1993; Hirano, 2005). SWI2/SNF2 ATPases are primarily involved in local
23 chromatin remodeling events by affecting nucleosome positioning and assembly.
24 They are usually core subunits of large functional complexes that include other
25 chromatin modifying activities like acetylases, methylases or ubiquitinating enzymes
26 (Martens and Winston, 2003; Mohrmann and Verrijzer, 2005; Durr and Hopfner,
27 2006; Gangavarapu et al., 2006). SWI2/SNF2 ATPases had their origins in
28 bacteriophage replication systems and restriction-modification systems distributed
29 throughout the prokaryotic superkingdoms (Iyer et al., 2006). They appear to have
30 been recruited from such a source, in the earliest stages of eukaryotic evolution, and
31 expanded to give rise to at least 6 representatives by the time of LECA (Fig. 6). A
32 comparable count of these ATPases is found in the degraded genomes of *Giardia* and
33 *Encephalitozoon*, and includes most versions traceable to LECA. Thus, this ancient
34 set of SWI2/SNF2 ATPases is likely to comprise the most essential group of
35 chromatin remodeling enzymes required by any eukaryote. Domain architectures of

1 these predicted ancestral versions show that the ATPase module was already fused
2 to different peptide-binding domains like chromo, bromo and MYB (SANT) that
3 allowed them to specifically interact with modified or unmodified nucleosomes (Fig.
4 6).

5
6 Prior to divergence of the kinetoplastid-*Naegleria* clade the number of SWI2/SNF2
7 ATPases had gone up to at least 13 representatives, and at least 19-20
8 representatives can be extrapolated to the common ancestor of chromalveolates and
9 the crown group (Fig. 6). Consistent with this, even the most reduced parasitic
10 genomes amongst apicomplexans and kinetoplastids have similar numbers of these
11 ATPases as extrapolated for their respective common ancestors with other
12 eukaryotes. By the time of the former radiation, new architectures combining the
13 SWI2/SNF2 ATPase module with different DNA-binding domains, a HNH
14 (endonuclease VII) nuclease domain, a MACRO domain and the RING finger, had
15 occurred. This implies that their functional roles were expanding, with the new
16 versions sensing and repairing DNA damage or performing additional protein
17 modifications through ubiquitination. In subsequent radiations of SWI2/SNF2
18 ATPases several lineage-specific architectures appear to have arisen. Examples of
19 these include convergent fusions to PHD fingers in apicomplexans and the crown
20 group, and fusions to different DNA-modifying enzyme domains in kinetoplastids and
21 fungi (see below). In light of these associations with DNA metabolism, it remains to
22 be seen if at least some SWI2/SNF2 ATPases act as DNA helicases, like other
23 Superfamily-II helicases (Bork and Koonin, 1993). Other than in the crown group, a
24 striking lineage specific expansion of a SWI2/SNF2 ATPase fused to the SJA domain
25 (Lander et al., 2001) is encountered in the parasitic protist, *T. vaginalis*. A distinctive
26 version of the SWI2/SNF2 ATPase, typified by the *Drosophila* protein *Strawberry*
27 *notch* appears to have independently laterally transferred from bacteria or
28 bacteriophages to the crown group eukaryotes, but lost in amebozoans and fungi
29 (Fig. 6).

30
31 SMC ATPases belong to the ABC superfamily, and contain a coiled-coil domain and a
32 hinge domain inserted within the P-loop ATPase domain (Hirano, 2005). Working as
33 dimers along with other accessory proteins like kleisins they are primarily responsible
34 for the large-scale organizational dynamics of chromatin, including chromosome
35 condensation (Hirano, 2006; Uhlmann and Hopfner, 2006). SMC ATPases might have

1 been present in the common ancestor of all life forms, and by the time of LECA had
2 proliferated into at least 6 distinct versions, along with the more distantly related
3 form Rad50 (Fig. 6). These six SMC ATPases have been vertically conserved in
4 practically all eukaryotes, with apparent loss of SMC5 and SMC6 in kinetoplastids and
5 ciliates. Another catalytic domain found in CPs is the MORC domain, which is a
6 unique version of the Hsp90-type ATPase domain, related to those found in
7 topoisomerase II ATPase subunits and DNA repair proteins of the MutL family (Inoue
8 et al., 1999). It is likely that these proteins are also involved in poorly-known ATP-
9 dependent remodeling events throughout eukaryotes. MORC domains appear to be of
10 bacterial origin, and were perhaps acquired by crown group eukaryotes. Within the
11 crown group there are two distinct lineages of MORC proteins (Fig. 6). One of them is
12 interestingly fused to the hinge and coiled-coil domains found in the SMC ATPases
13 and a BAM domain (Fig. 6). These latter proteins might effectively function as
14 analogs of SMC ATPases, with the MORC domain playing a role equivalent to the ABC
15 ATPase domain of the former enzymes. Apicomplexans have a unique version of the
16 MORC ATPase fused to kelch-type β -propellers (Fig. 6). The MORC ATPase domain of
17 this animal is closer to the animal versions, and equivalents are absent in all other
18 members of the chromalveolate clade. These observations suggest that it could have
19 possibly been laterally transferred from the animal host early in apicomplexan
20 evolution. Similarly, *Naegleria* might have acquired the version of MORC ATPase with
21 the SMC-related domains from the crown group eukaryotes.

22

23 Beyond these ATP-dependent remodeling enzymes there might be other enzymatic
24 activities that have roles in assembly of CPs but remain poorly characterized. One
25 such example is suggested by a protein displaying a fusion of a peptidyl prolyl
26 isomerase of the FKBP family domain to the nucleoplasmin/HD2 domain (Aravind and
27 Koonin, 1998). Orthologs of this protein are seen in several eukaryotes, including
28 *Giardia*, and might play a role in the folding and assembly of histones by facilitating
29 conformational isomerization of proline.

30

31 ***5.4 Other chromatin protein modifications, potential histone tail interaction*** 32 ***domains and histone chaperones***

33 A less-understood covalent modification of CPs is the conjugation of ubiquitin (Ub)
34 and other related modifiers (Ubls; e.g. Nedd8 and SUMO) (Shilatifard, 2006; Collins
35 et al., 2007; Kouzarides, 2007). This process involves a 3-step reaction that

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

24

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

1 the SWIB domain, recruits ubiquitinating activities via the fused E3-ligase RING
2 finger domain to transcription factors like p53 (Bennett-Lovsey et al., 2002). The
3 standalone pan-eukaryotic version of this domain might be critical for recruitment of
4 SET domain methyltransferases to SWI2/SNF2 dependent remodeling enzymes to
5 chromatin (Stephens et al., 1998). Three unrelated ancient families of histone-
6 binding domains, namely the nucleoplasmin, ASF1 and NAP1 appear to be primarily
7 involved in the chaperoning and assembly of histones (Namboodiri et al., 2003; Park
8 and Luger, 2006; Tang et al., 2006). The HD2 domain related to nucleoplasmin was
9 originally claimed to be a histone deacetylase, but appears to be more likely to a
10 histone-binding domain (Aravind and Koonin, 1998). Presence of the
11 nucleoplasmin/HD2 and ASF1 domains in all eukaryotic including early branching
12 forms like *Giardia* and *Trichomonas*, points to the presence of at least two distinct
13 histone chaperones in LECA. NAP1 is absent in the basal eukaryotic taxa, and
14 appears to have emerged before the divergence of *Naegleria* and kinetoplastids from
15 other eukaryotes. In contrast, another class of histone chaperones, the Chz1p family,
16 has a more restricted distribution, being present only in animals and fungi (Luk et
17 al., 2007). Assembly of histone octamer complexes using multiple chaperones
18 appears to be an ancestral feature of eukaryotes distinguishing them from archaea,
19 and might be correlated with origin of low-complexity tails.

20

21 **5.5 Natural history of epigenetic DNA modification enzymes**

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

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

22

23 Distribution of these methylases suggests that DNA methylation might not be a
24 major regulatory factor in most parasitic protists, with exception of fungi and
25 possibly kinetoplastids and *Naegleria*. Consistent with this, the TAM (methylated
26 DNA-binding) domain is not observed in any of the lineages of parasitic protists
27 studied to date. However, the SAD (SRA) domain, which has also been shown to
28 interact with methylated DNA (Johnson et al., 2007; Woo et al., 2007), is found in
29 *Plasmodium*. A careful analysis of the conservation pattern of this domain suggests
30 that it contains a set of conserved polar residues suggestive of it being an enzyme
31 (Makarova et al., 2001), and might catalyze as yet unknown DNA modifications.
32 Another potentially important regulatory DNA modification, which is thus far
33 restricted to trypanosomes, is β -D-glucosyl hydroxyl methyl uracil (the J-base), a
34 modified thymine. The recently characterized, unique biosynthetic apparatus for this
35 base includes the JBP1/2 proteins, which share a double-stranded β -helix

1 dioxygenase domain distantly related to the JOR protein demethylase and AlkB-type
2 DNA demethylases (Yu et al., 2007). In JBP2 this domain is fused to a C-terminal
3 SWI2/SNF2 module, suggesting that DNA modification is coupled with chromatin
4 remodeling (DiPaolo et al., 2005). Dioxygenase domains specifically related to the
5 version found in JBP1/2 are found in animals (e.g. human CXXC6; translocated in
6 acute myeloid leukemia (Ono et al., 2002)), some actinomycete bacteria,
7 mycobacteriophages and in an expanded family of proteins in the fungus *Coprinopsis*
8 *cinerea*. While there is no evidence for modified bases like J in these organisms, it
9 remains to be seen if these enzymes could catalyze any other DNA modifications
10 such as DNA demethylation. Consistent with a chromatin related role the animal
11 versions like CXXC6 are fused to the DNA binding CxxC domains. The possibility of
12 other unusual regulatory DNA modifications linked to chromatin dynamics is
13 suggested by the presence, in certain fungi, of nucleic acid deaminases fused to the
14 SET histone methylases (see above, Fig. 5).

15

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

17

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

20 Information regarding domain architectures obtained from protist genomes helps in
21 understanding several aspects of chromosomal regulatory systems. One of these is
22 the degree of functional interaction between different regulatory systems, and the
23 evolutionary sequence of development of these interactions. Analysis of CPs reveals
24 certain strong “syntactical” patterns in their domain architectures (Fig.7, 8). Histone
25 methylase and acetylase domains never co-occur in the same polypeptide in any
26 eukaryote. Likewise, demethylases and deacetylases tend not to co-occur with each
27 other or respectively with methylases and acetylases (Fig. 7). This suggests that
28 acetylation and methylation are relatively stable modifications, and that their
29 removal is not temporally coupled or combined with re-modification. This is
30 consistent with methylation and acetylation being epigenetic markers and them
31 being independent, but potentially complementary in action (Peterson and Laniel,
32 2004; Shilatifard, 2006; Villar-Garea and Imhof, 2006; Kouzarides, 2007). Two of
33 the four acetyltransferases that can be traced to LECA are closely associated with the
34 basal transcription apparatus (GCN5, Elp3 families). Hence, the earliest roles of
35 acetylation were probably in the context of modulating histone-DNA interaction to

1 facilitate transcription. On the other hand, methylation appears to have emerged in
2 the more general context of organizing chromosomal structure by altering histone
3 properties. Whereas, acetylases show fusions to specific histone-tail-binding domains
4 even in the basal eukaryotes (e.g GCN5 with a bromo domain), histone methylases
5 develop such fusions only later in eukaryotic evolution (Fig. 5, 8). However,
6 eventually methylases developed greater domain architectural diversity than
7 acetylases (Fig. 4, 8), suggesting that they were probably utilized for modifying
8 histones in many more specific contexts than the latter. Interestingly, histone
9 demethylases show a clearly greater architectural complexity than deacetylases (Fig.
10 7). This might again suggest that while deacetylases tend to remove acetyl groups in
11 a generalized fashion, demethylases might select specific contexts via their
12 associated domains for reversing methylation. These observations are consistent
13 with results suggesting distinct roles for these two major components of the “histone
14 code” (Peterson and Laniel, 2004; Shilatifard, 2006; Villar-Garea and Imhof, 2006;
15 Kouzarides, 2007). Evidence from domain architectures suggests that both systems
16 interact to a certain degree with the ubiquitin system, and such associations began
17 emerging in the chromalveolate and crown-group clades.

18

19 Acetylases and methylases show preferential associations with certain peptide-
20 binding domains—acetylases most frequently combine with bromo domains and
21 methylases with PHD fingers (Fig. 7). Given the binding preferences of these
22 peptide-binding domains, it is possible that respectively recognizing previously
23 methylated or acetylated histones might be an important functional feature of some
24 versions of these enzymes. Conversely, methylases also come fused to acetylated-
25 peptide-binding domains and acetylases are fused to methylated-peptide-binding
26 domains (Fig. 7, 8), suggesting that each is recruited via the other modification.
27 Peptide-binding domains recognizing different forms of histone modifications might
28 also be combined with each other in the same polypeptide (Fig. 4, 5, 7). Thus, at
29 least in certain cases, these adaptor proteins might parallelly or serially recognize
30 different histone modifications. Often, such architectures have arisen in a lineage-
31 specific manner, including in several parasitic protists (Fig. 4, 5). For example
32 *Phytophthora* shows proteins respectively with 6 tandem bromo domains and serial
33 bromo, PHD finger and chromo domains, trypanosomes possess a protein with
34 bromo and ZF-CW(PHDX) domains, and *Giardia* a protein combining the bromo
35 domain and a WD-type β -propeller. This suggests that while histone modifications

1 might be universal in eukaryotes, their “interpretation” by peptide-binding adaptors
2 shows lineage-specific differences. Future experimental analyses of these lineage-
3 specific adaptors might be critical to understand the diversity of regulatory roles of
4 histone modifications in particular organisms. SWI2/SNF2 ATPases have been shown
5 to work with different histone-modifying enzymes in eukaryotic model systems
6 (Martens and Winston, 2003; Mohrmann and Verrijzer, 2005). However, their
7 domain architectures across eukaryotes show that there are no known fusions
8 between these ATPases and histone acetylase or methylase domains (or the
9 corresponding de-modifying enzymes) (Fig. 7). Hence, though their actions are
10 cooperative, they are not mechanistically closely coupled. However, SWI2/SNF2
11 ATPases are combined with Ub-conjugating E3 domains in the same polypeptide
12 (Gangavarapu et al., 2006), suggesting possible coupled action between these
13 activities.

14

15 ***6.2 Relationship between phylogeny, organizational complexity and domain*** 16 ***architectures of chromatin proteins***

17 Domain architectures can be depicted as an ordered graph or a network, in which
18 domains form the nodes and their linkages with other domains within a given
19 polypeptide (adjacent co-occurrence in polypeptide) are depicted as edges
20 connecting nodes (Fig. 7). These domain-architecture networks have been extremely
21 useful in extracting different features of domain associations (as discussed above),
22 and can also be used to assess complexity of domain architectures. Complexity of
23 domain architectures of proteins in a given functional system can also be
24 independently assessed using the complexity quotient that measures both the
25 variety and the number of domains in them (Fig. 2D). Anecdotal studies had
26 indicated that domain architectural complexity correlated with increased
27 organizational complexity of the organism- i.e. emergence of multicellularity and
28 increased cellular differentiation (Gibson and Spring, 1998; Lander et al., 2001).
29 Availability of protist proteomes allows a more objective evaluation of the correlation
30 between domain architecture complexity of CPs and organizational complexity across
31 eukaryotes. In functional terms, greater domain architectural complexity of CPs
32 would imply a greater variety and number of interactions made by them with
33 proteins, nucleic acids and small molecules.

34

1 Direct examination of the domain architecture networks points to a trend of
2 increasing domain architectural complexity in CPs in course of eukaryotic evolution
3 (Fig. 8). Diplomonads and parabasalids have the least complex domain architectures.
4 The *Naegleria*-kinetoplastid clade, apicomplexans and ciliates have higher
5 architectural complexity than these and chromists have even higher values.
6 However, the highest architectural complexity is observed in certain crown group
7 clades, and amongst them the animals are unparalleled in the complexity of their
8 domain architecture networks (Fig. 8). When complexity quotient of CPs is plotted
9 against the total number of predicted CPs encoded by an organism, we observe a
10 steady positively correlated rise in these values. Thus, in eukaryotes as the number
11 of CPs rises there is also a general increase in their architectural complexity. In many
12 cases this increase in architectural complexity occurs via "domain accretion" (Gibson
13 and Spring, 1998; Koonin et al., 2000; Lander et al., 2001). In this phenomenon,
14 new domains are added around an ancient orthologous core of the polypeptide. This
15 tendency is particularly prominent in histone methylases and SWI2/SNF2 ATPases
16 (Fig. 5, 6, 8). Despite having large absolute numbers of CPs, ciliates and
17 *Trichomonas* tend to have much lower architectural complexity. Mere increase in
18 proteome size without increase in architectural complexity of CPs, as seen in ciliates
19 and *T. vaginalis*, might be sufficient to achieve relatively complex organization within
20 a *single cell*. In contrast, the high complexity of animal proteins points to some
21 correlation between architectural complexity and number of CPs, and emergence of
22 numerous differentiated cell-types (Fig.2D, 8). Most major protist parasites, like
23 apicomplexans, kinetoplastids and diplomonads have relatively fewer and
24 architecturally less-complex CPs, as compared to their hosts (Fig.2D, 8). As a
25 consequence, relatively lesser effort might be needed to completely unravel their
26 regulatory interaction networks.

27

28 Unique domain architectures and phyletic patterns of the CPs can also be compared
29 for consistency with the inferred eukaryotic phylogenetic tree (See
30 above)(Templeton et al., 2004). In general, the most parsimonious explanation for
31 the observed architectures and phyletic patterns is consistent with the phylogenetic
32 tree (Fig. 1), albeit obscured by extensive losses in several parasites. Nevertheless,
33 certain clades are strongly supported by shared architectures and phyletic patterns:
34 1) The animal-fungi clade 2) the crown group clade 3) apicomplexans, alveolates
35 and, to certain extent, the chromalveolate clade. 4) A clade comprised of all

1 eukaryotes, excluding the diplomonad and parabasalid lineages. These points appear
2 to coincide with notable innovations amongst CPs and transcription factors. Plants
3 and stramenopiles exclusively share several transcription factors or CP domain
4 architectures, compared to the plants and alveolates (Armbrust et al., 2004; Tyler et
5 al., 2006). This is particularly intriguing given that the secondary endosymbiotic
6 event is believed to have occurred in the common ancestor of the chromalveolate
7 lineage (Bhattacharya et al., 2004). This might either imply selective loss of more
8 plant-derived genes in alveolates or a more recent further endosymbiotic event in
9 the ancestor of stramenopiles that delivered a new load of plant-derived genes
10 (Armbrust et al., 2004; Bhattacharya et al., 2004). Given that such tertiary
11 endosymbiosis events are also likely to have happened in dinoflagellates
12 (Bhattacharya et al., 2004), this latter alternative remains a distinct possibility. It is
13 also possible that these plant-derived transcription factors and CPs contributed to the
14 rise of organizational complexity and multicellularity in stramenopiles.

15

16 **7. Interactions between RNA-based regulatory systems and chromatin** 17 **factors**

18 Number of lines of evidence points to a functional link between RNA-based regulatory
19 systems, including post-transcriptional gene silencing or RNA interference (RNAi),
20 and chromatin-level regulatory events. Studies in plants have revealed a role for
21 siRNAs in directing DNA methylation and heterochromatin formation (Chan et al.,
22 2006; Li et al., 2006a; Pontes et al., 2006; Vaucheret, 2006). RNAi-like systems
23 have also been implicated in the epigenetic phenomenon like paramutation in plants
24 and meiotic silencing by unpaired DNA in *Neurospora* (Shiu et al., 2001; Alleman et
25 al., 2006). Comparative genomic analysis of fungi predicted a functional link between
26 the siRNA/miRNA biogenesis pathway and several CPs (Aravind et al., 2000).

27 Accumulating recent experimental evidence has confirmed this, and points to a
28 major role of small RNAs in directing histone methylation and heterochromatinization
29 in fungi like *Schizosaccharomyces* (Grewal and Moazed, 2003; Grewal and Rice,
30 2004). In ciliates, a similar small RNA-based pathway has been implicated in histone
31 H3 methylation, heterochromatin formation, and subsequent rearrangements and
32 elimination of DNA sequences during the development of the macronucleus
33 (Mochizuki et al., 2002; Mochizuki and Gorovsky, 2004; Malone et al., 2005). The
34 key conserved players in generation of these small regulatory RNAs are the dicer
35 nuclease and the RNA-dependent RNA polymerase (RDRP) which is involved in

1 amplifying them. Silencing action of these RNAs is mediated by the PIWI domain
2 RNases (the slicer nucleases), which might localize to chromatin to specifically
3 degrade transcripts right at the source (Grewal and Moazed, 2003; Grewal and Rice,
4 2004; Ullu et al., 2004; Li et al., 2006a; Pontes et al., 2006). Presence of PIWI
5 domains and RDRPs in representatives of all major eukaryotic clades studied to date
6 indicates that a minimal RNAi system comprising of these two proteins had already
7 emerged in LECA. Both the RDRP and the PIWI domain nucleases of this ancestral
8 system appear to have been acquired by the eukaryotic progenitor from bacterial
9 sources (Aravind et al., 2006). However, the system was repeatedly lost, partially or
10 entirely, in several eukaryotes. Vertebrate apicomplexan parasites, with exception of
11 the *Toxoplasma* lineage, have lost both the PIWI nuclease and the RDRP suggesting
12 that they are unlikely to possess a *bona fide* RNAi system (Ullu et al., 2004). Some
13 parasites like kinetoplastids and *Trichomonas* appear to have lost the RDRP, but
14 retain PIWI nucleases, and as a consequence display certain RNAi effects (Shi et al.,
15 2004a). Other parasites like *Giardia*, *Entamoeba* and the fungus *Cryptococcus*
16 possess both these enzymes suggesting the presence of both small RNA amplification
17 and degradation systems in these organisms. Interestingly, *Entamoeba* encodes an
18 inactive version of the RDRP (26.t00065), which might have a novel non-catalytical
19 regulatory role. With exception of HP1-like chromodomain proteins and some
20 conserved SET domain histone methylases, many CPs that appear to interact with
21 the RNAi machinery are largely limited to the crown group eukaryotes (Fig. 5). The
22 general architectural simplicity of protist CPs in comparison to their crown group
23 counterparts (Fig. 8) also raises questions regarding the degree of coordination
24 between CPs and the RNAi system in these organisms (Aravind et al., 2000).
25 Nevertheless, a core interacting regulatory network combining HP1-like
26 chromodomain proteins, histone methylases and the RNAi machinery could have
27 emerged very early in eukaryotic evolution with further elaboration in the crown
28 group.

29

30 Several studies in crown-group eukaryotes have implicated large non-coding RNAs in
31 heterochromatin formation and chromosome dosage compensation. Some
32 chromodomains have been shown to interact with these RNAs (Brehm et al., 2004;
33 Bernstein et al., 2006). Likewise, SAM domain proteins of the polycomb complex in
34 animals have also been shown to interact with large RNAs in chromatin (Zhang et al.,
35 2004). These suggest that there might be other RNA-based pathways, distinct from

1 RNAi pathways, which might have a direct role in chromatin level regulation.
2 Expression of the variant surface antigen Pfemp1 encoded by the *var* genes in
3 *P.falciparum* involves silencing of all of the copies of this gene except an active
4 version (Ralph and Scherf, 2005). Antigenic variation proceeds via silencing of the
5 currently active copy, and activation of a previously inactive copy. This silencing
6 process has been shown to resemble heterochromatin formation, and is mediated by
7 changes in histone modification, including the action of the PfSir2 deacetylase
8 (Duraisingh et al., 2005; Freitas-Junior et al., 2005). The transition between the
9 active and silenced state in *var* gene expression appears to depend on the
10 generation of a non-coding or “sterile” transcript from a promoter located in the
11 intron of the gene (Deitsch et al., 2001; Frank et al., 2006). This raises the
12 possibility of larger transcripts mediating chromatin dynamics in *P.falciparum*. These
13 tantalizing leads hint that there is likely to be a whole “world” of RNA-based
14 chromatin reorganizing processes that remains unexplored in different protists.

15

16 **8. General considerations and conclusions**

17 Most studies to date have examined the evolutionary history of eukaryotic
18 transcription and chromatin level regulation based on data coming from model
19 systems in the eukaryotic crown group (Koonin et al., 2000; Lander et al., 2001).
20 Given that these taxa only comprise the tip of the proverbial iceberg, the situation in
21 protist eukaryotes, and hence early stages of eukaryotic evolution remained unclear.
22 Recent advances in genomics have finally allowed exploration of this previously un-
23 navigated territory (Sullivan et al., 2006). These studies showed that there are many
24 aspects of transcription regulation and chromatin dynamics that have a rich diversity
25 in eukaryotes, beyond what is observed in the crown group. Most importantly, the
26 new data also enables an objective reconstruction of these systems in the last
27 eukaryotic common ancestor and their subsequent evolution. The prevalence of
28 histones in the euryarchaea along with certain other features are suggestive of the
29 archaeal precursor of the eukaryotes being an euryarchaeon (Reeve, 2003; Reeve et
30 al., 2004; Aravind et al., 2006). It is striking to note that several key players in
31 chromatin and eukaryotic transcription regulation, which were present in LECA, were
32 possibly derived from mobile elements and prophages, probably of bacterial origin.
33 These include the SWI2/SNF2 ATPases, the HEH domain, which helps in tethering
34 chromosomes to the nuclear membrane, and the RDRP (Mans et al., 2004; Aravind
35 et al., 2006; Iyer et al., 2006). Furthermore, lateral transfers from bacteria to the

1 eukaryotic progenitor appear to have contributed some of the acetylases, and
2 peptide-binding domains of the SH3 fold. Some of these events might correspond to
3 the primary endosymbiotic association between the euryarchaeon and the α -
4 proteobacterial mitochondrial precursor. It is possible that, even at this stage, some
5 of the lateral acquisitions were sporadic transfers from other bacterial sources. This
6 acquisition of various regulatory domains from bacteria continued even in the later
7 phases of eukaryotic evolution, as suggested by the history of the MORC domain,
8 DNA methylases, histone demethylases and several histone acetylases.

9
10 In any case, the main feature that defined the origin of eukaryotes was an early
11 spurt of drastic evolutionary innovation that accompanied the melding of the
12 archaeal and bacterial inheritances to give rise to a distinctive eukaryotic system
13 (Koonin et al., 2000; Dacks and Doolittle, 2001; Walsh and Doolittle, 2005; Aravind
14 et al., 2006). This appears to have happened between the point of emergence of the
15 first eukaryotic progenitor and the LECA from which all extant eukaryotes have
16 emerged. Chief evolutionary innovations in this phase were: 1) Multiple rounds of
17 duplications giving rise to various paralogous protein families, which diversified into
18 distinct functional niches (e.g. SWI2/SNF2). Emergence of families with multiple
19 paralogs increased the number of specific interactions made by proteins, and
20 appears to have played the primary role rise in organizational complexity of the
21 eukaryotes. 2) "Invention" of new α -helical domains (E.g. the bromodomain) and
22 diversification of metal-chelation supported structures resulted in the provenance of
23 whole new sets of protein-protein interactions (Aravind et al., 2006). For example,
24 the PHD and RING finger emerged from an ancestral Zn-chelating treble-clef fold
25 domain that recognized lysine-containing peptides, and subsequently diversified to
26 mediate specific interactions with methylated peptides and ubiquitination targets
27 respectively. 3) Emergence of proteins with long non-globular or low-complexity
28 stretches accreted to the ancient globular domains (E.g. tails of eukaryotic histones)
29 allowed for a greater degree of regulation of proteins through variety of post-
30 translational modifications (Liu et al., 2002). 4) Origin of nucleo-cytoplasmic
31 compartmentalization accompanied by diversification of several families of ancient
32 domains into versions with specific cytoplasmic or nuclear roles. This resulted in
33 paralogous domains that specifically functioned in regulatory events in either the
34 cytoplasm or the nucleus (Mans et al., 2004; Aravind et al., 2006). This might have

1 been central to the radiation of several SH3-fold peptide-binding domains in relation
2 to recognition of modified and unmodified histones.

3
4 The genomes of various early-branching eukaryotes (e.g. *Trichomonas* and *Giardia*)
5 suggest that recruitment of novel classes of DNA-binding domains had begun early in
6 eukaryotic evolution, with repeated emergence of new TFs in different lineages. In
7 particular, specific TFs in various parasitic protists remained unknown until recently.
8 However, this principle of lineage-specific expansions allowed us to identify the
9 major specific TFs of several parasitic lineages like apicomplexans, *T.vaginalis*,
10 *Entamoeba*, oomycetes and *Naegleria* (Fig. 3). Various demographic trends of TFs
11 and CPs show general positive correlations to proteome size and, to certain extent,
12 the degree of cellular differentiation associated with multicellularity. Typically,
13 parasitic protists, irrespective of their phylogeny, possess fewer specific TFs and less
14 complex CPs. The transcription regulation apparatus of protist parasites have taken
15 very different courses during adaptation to such a life-style. Microsporidians,
16 kinetoplastids and *Giardia* have highly reduced complements of specific transcription
17 regulators and CPs. Others like *Entamoeba* and apicomplexans and have lost most
18 TFs relative to their free-living counterparts, but have expanded single DBD families
19 to derive majority of their specific TFs. Differences can even be observed within
20 apicomplexans in the complements of specific TFs. For instance, *Cryptosporidium*
21 retains certain specific TFs like E2F/DP1 that are lost in other apicomplexans, and
22 *Toxoplasma* display a distinctly higher number of ApiAP2 TFs than other
23 apicomplexans, perhaps indicating a higher degree of specific transcriptional
24 regulation. Oomycetes, *Naegleria* and *T.vaginalis* have large numbers of TFs,
25 comparable in numbers to any free-living organism of a similar organizational grade.
26 Thus, the degree of transcriptional regulation in eukaryotic parasites appears to have
27 been shaped by a combination of factors, such as metabolic capabilities, degree of
28 obligate host-dependence, complexity of life cycles and effective coding capacity of
29 the genome. There appears to be no strong correlation between number of TFs and
30 CPs and general cellular morphology – an aspect so strikingly illustrated by the gross
31 demographic differences in these proteins between *Giardia* and *Trichomonas* despite
32 their comparable morphology.

33
34 In conclusion, we hope that the survey presented here provides a framework for the
35 functional analysis of transcription factors and CPs in protists. The most obvious lines

1 of future investigation would be to combine this information with high-throughput
2 methods such as expression studies, CHIP-chip methods, large-scale interaction
3 mapping, immuno-precipitation of complexes, fluorescence-tagged localization
4 studies and biochemical genomics to glean basic cell-biological information (Bozdech
5 et al., 2003; Le Roch et al., 2003; Dunn et al., 2005; LaCount et al., 2005; Collins et
6 al., 2007). In particular, this might be useful to obtain a handle on the biology of
7 parasites, where information on upstream regulators of genes implicated in
8 pathogenesis and progression of disease is limited. We also hope that these studies
9 would go hand-in-hand with more involved lines of investigation such as gene-
10 knockouts, phenotypic analysis and thorough biochemical characterization. Given the
11 presence of certain unique predicted enzymatic activities in protists, we believe that
12 such studies might also provide direct leads regarding novel biochemistries that have
13 been ignored in eukaryotic model systems. These studies might also provide new
14 targets for therapeutic and diagnostic applications. Specifically, the distinctness of
15 many protist regulatory enzymes from their animal and plant counterparts might
16 furnish targets for conventional drug development. The identification of specific TFs
17 might alternatively allow revisiting the relatively less-explored direction of
18 transcription-factor-targeting drugs (Ghosh and Papavassiliou, 2005; Visser et al.,
19 2006). Irrespective of the ultimate applications, we appear poised to reach new
20 levels of understanding in terms of eukaryotic transcription and chromatin dynamics
21 in the near future.
22

1 **Acknowledgements**

2 The authors are supported by the intramural program of the National Center for
3 Biotechnology Information. As field under consideration is vast and extremely active,
4 there are an enormous number of primary papers. We apologize to all colleagues
5 whose important contributions could not be cited to keep the article within
6 reasonable limits. Supplementary information comprising of a comprehensive
7 collection of Genbank identifiers for all chromatin proteins and transcription factors
8 included in this study will be provided as text files.

10 **Figure legends**

11
12 **Figure 1.** Phylogenetic relationships, genome sequencing efforts and major
13 distinguishing features of eukaryotes. The displayed tree is a maximum likelihood
14 (ML) tree derived from a concatenated alignment of 82 universally conserved
15 eukaryotic proteins spanning 19603 positions. The among site variation of rates for
16 the alignment was modeled as a distribution with 8 discrete rate categories and the
17 positions belonging to each rate category, rates and the α -parameters of the
18 distribution were estimated using the TreePuzzle 5.1 program with JTT matrix
19 (Schmidt et al., 2002). This was used to infer the ML tree with PROML (Felsenstein,
20 1989) and bootstrap support was estimated using 500 replicates with the PHYML
21 program (Guindon and Gascuel, 2003). All monophyletic nodes discussed in the text
22 were supported with >85% bootstrap support and are completely consistent with
23 previously published results using representatives of the same taxa. Rooting with
24 archaeal orthologs suggests a basal position for the diplomonads and parabasalids.
25 The approximate non-redundant protein count for a given genome was used to
26 calculate proteome size. For *Trichomonas vaginalis* (asterisk), the proteome size was
27 further reduced by removing fragmentary proteins that were identical to full-length
28 versions. There is extensive representation of multicellular animals, plants, and
29 fungi, including basidiomycete, ascomycetes and the microsporidian
30 *Encephalitozoon*. Recently, the diversity of genomes from the plant lineage has been
31 extended by the publication of the sequence of the minimalist unicellular chlorophyte
32 alga *Ostreococcus* and the complete sequencing of another chlorophyte
33 *Chlamydomonas*. Amongst amoebozoans, sequences of *Dictyostelium* and
34 *Entamoeba* have been published. Alveolates are represented by complete genome
35 sequences of at least 4 apicomplexan genera and two ciliates (*Tetrahymena* and
36 *Paramecium*). In the stramenopile clade the genomes of the abundant marine diatom
37 *Thalassiosira* and the oomycete *Phytophthora* have been reported. Of the
38 kinetoplastids we have genome sequences of human parasites of the genera
39 *Leishmania* and *Trypanosoma*, while that of *Naegleria* has also been recently
40 published. The genome sequences of the basal eukaryotes are represented by that of
41 *Giardia lamblia* and *Trichomonas*. In addition to these protist genome sequencing
42 projects which have been published or completed, there are several others which are
43 in different stages of completion.

44 **Figure 2**

45 **A)** Among site rate variation for different functional classes of eukaryotic proteins.
46 These were calculated using multiple alignments of highly conserved proteins that
47 are present in all eukaryotes in each functional category shown in the graph. The
48 total number of positions in each category were- translation: 6357; Transcription:
49 2275; Replication:5436; Histones: 381; Chaperones: 5154. The among site rate
50 variations for each functional class were calculated as described in Figure 1 but in
51 this case a Whelan and Goldman (WAG) substitution matrix was used as it confers
52 higher likelihood on the data. The fraction of the positions in each rate category is
53 plotted for each functional class – the categories to the left are slower evolving with

1 respect to those in the right. Note that the distribution for transcription and
2 replication proteins is U-shaped indicating an over-representation of extremes-
3 slowest-evolving and fastest-evolving positions.

4 **B)** Scaling of transcription factors with proteome size. The names of organisms used
5 for the plot and their abbreviations are indicated below. Organisms with significantly
6 lower than expected fraction of chromatin proteins are labeled.

7 **C)** Scaling of chromatin proteins with proteome size. The organisms are the same as
8 in A. Organisms with lower fraction of chromatin proteins than expected are marked.

9 **D)** Complexity quotient plot for chromatin proteins. The "complexity quotient" for an
10 organism is defined as the product of two values: the number of different types of
11 domains which co-occurs in signaling proteins, and the average number of domains
12 detected in these proteins. The complexity quotient is plotted against the total
13 number of chromatin proteins in a given organism. A polynomial curve fitting the
14 general trend of majority of organisms is shown. Crown group members are shown
15 in red and the non-crown group members are in green. Some organisms with much
16 lower complexity than those along the general trend are marked. Each protein has at
17 least a single known or predicted domain with a chromatin/transcription related
18 function. A total of 363 domains were considered, among which 121 were domains
19 specifically found in chromatin and transcription factors and the rest were other
20 domains with wider distributions encompassing other functional systems.

21 The organisms included in all these plots are the following: Crown group: *Aspergillus*
22 *fumigatus* – Afum, *Candida glabrata* – Cgla, *Debaryomyces hansenii* – Dhan, *Ashbya*
23 *gossypii* – Egos, *Gibberella zeae* – Gzea, *Kluyveromyces lactis* – Klac, *Neurospora*
24 *crassa* – Ncra, *Saccharomyces cerevisiae* – Scer, *Schizosaccharomyces pombe* –
25 Spom, *Yarrowia lipolytica* – Ylip, *Cryptococcus neoformans* – Cneo, *Ustilago maydis* –
26 Umay, *Encephalitozoon cuniculi* – Ecun, *Anopheles gambiae* – Agam, *Apis mellifera* -
27 Amel, *Branchiostoma floridae* – Bflor, *Caenorhabditis elegans* – Cele, *Ciona*
28 *intestinalis* - Cint, *Danio rerio* - Drer, *Drosophila melanogaster* – Dmel, *Homo*
29 *sapiens* – Hsap, *Mus musculus* – Mmus, *Pan troglodytes* – Ptro, *Rattus norvegicus* –
30 Rnor, *Strongylocentrotus purpuratus* - Spur, *Tetraodon nigroviridis* – Tnig, *Tribolium*
31 *castaneum* – Tcas, *Monosiga brevicollis* – Mbre, *Nematostella vectensis* – Nvec,
32 *Entamoeba histolytica* – Ehis, *Dictyostelium discoideum* – Ddis, *Chlamydomonas*
33 *reinhardtii* – Crei, *Ostreococcus tauri* – Otau, *Arabidopsis thaliana* - Atha,
34 *Phaeodactylum tricornutum* – Ptri, *Phytophthora sojae* – Psoj, *Phytophthora*
35 *ramorum* - Pram, *Thalassiosira pseudonana* – Tpse, *Tetrahymena thermophila* –
36 Tthe, *Paramecium tetraurelia* - Ptet, *Toxoplasma gondii* – Tgon, *Theileria parva* –
37 Tpar, *Theileria annulata* – Tann, *Cryptosporidium parvum* - Cpar, *Plasmodium*
38 *falciparum* – Pfal, *Trypanosoma cruzi* – Tcru, *Trypanosoma brucei* – Tbru, *Leishmania*
39 *major* - Lmaj, *Naegleria gruberi* – Ngru, *Giardia lamblia* – Glam, *Trichomonas*
40 *vaginalis* – Tvag, *Guillardia theta* – Gthe. The genomes were obtained from NCBI
41 Genbank. The *Toxoplasma gondii* sequence was the current release from Toxodb
42 (www.toxodb.org), while the Stramenopile, *Ciona intestinalis*, *Monosiga*
43 *brevicollis*, *Nematostella vectensis*, *Naegleria gruberi* and *Chlamydomonas reinhardtii*
44 genomes were obtained from Department of Energy's Joint Genome Institute
45 (<http://www.jgi.doe.gov/>).

46
47 **Figure 3.** Lineage-specific expansions and phyletic distributions of specific TFs. Only
48 those specific TFs that are present in protists and have LSEs or notable sporadic
49 phyletic patterns are shown. The distribution of the TFs across eukaryotic species is
50 shown below the eukaryotic tree. The key below the distribution gives the notations
51 used to describe presence, absence or LSEs. A "P" or a "Ps" next to the number of
52 TFs in the ciliate and oomycete columns represents LSE in *Paramecium* and

1 *Phytophthora sojae* respectively. Novel ZnBD denotes the novel zinc chelating TF
2 present in stramenopiles.
3
4

5 **Figure 4.** Ancient and lineage-specific domain architectures in acetylation-based
6 regulatory systems. Evolution of acetylation-based systems are shown using various
7 domain architectures that evolved either at different early stages in the evolution of
8 eukaryotes or more recently in different lineages. The number of ancient conserved
9 acetylases, deacetylases and acetyl-peptide detecting adaptors that were present in
10 the different temporal epochs are also shown to the right. Architectures are denoted
11 by their gene name and species abbreviations separated by underscores. If an
12 architecture is only restricted to a subset of species or lineages in a group then the
13 species or lineage abbreviations in which they are present are listed in brackets
14 below the architecture. Domain architectures of well known proteins are only
15 denoted by the protein names. For species abbreviations consult Fig. 2.
16 Abbreviations of lineages include: Amoe: Amoebozoans, Api: Apicomplexans, Cil:
17 Ciliates, FF: Filamentous fungi, Kin: Kinetoplastids, Oomy: Oomycetes, Pl: Plants,
18 Stram: Stramenopiles. Domains are denoted by their standard names and
19 abbreviations. For a comprehensive list of domain names and functions refer Table 1.
20 Atypical domain abbreviations include: A: Ankyrin repeat, B: B-box, BM: BMB/PWWP,
21 BrC: Brd2/TAF14 C-terminal domain, UBP: Bro: Bromo, C6: C6 fungal finger, Ch:
22 Chromo, Deam: Nucleotide deaminase, ECH: Enoyl-coA hydratase, FB: Fbox, Ing1N:
23 Ing1-like N-terminal domain, JN: JOR/JmjC N-terminal domain, K: Kelch repeats,
24 LCM: Leucine carboxymethyltransferase, MYND: MYND finger, OB nuclease:
25 Staphylococcal nuclease-like domain of the OB fold, OTU: OTU-like thiol protease, P:
26 PHD finger, PARPf: Zinc-chelating finger associated with Poly ADP ribose
27 polymerases, PX: PHDX/ZfCW, RAD16f: Zinc-chelating finger found in all RAD16
28 proteins, RAD18: Zinc-chelating finger associated with RAD18, R: RING finger,
29 TF2S2: The second domain of the TFIIS-like proteins, SnoC: Strawberry notch C-
30 terminal domain, T: TPR repeat, TopC: Zinc ribbon found at the C-terminii of
31 Topoisomerases, Tu: Tudor, WD: WD repeats, wH: winged HTH, Ubhyd: Ubiquitin
32 carboxy-terminal hydrolase of the papain-like thiol protease fold.
33

34 **Figure 5.** Ancient and lineage-specific domain architectures in the methylation-
35 dependent regulatory systems. Evolution of the methylation-based regulation is
36 shown using various domain architectures that evolved either at different early
37 stages in the evolution of eukaryotes or more recently in different lineages. The
38 number of ancient conserved protein methylases, demethylases and methylated-
39 peptide detecting adaptors that were present in the different temporal epochs are
40 also shown to the right. The scheme of labeling domain architectures, species and
41 lineages abbreviations are as in Fig. 4.
42

43 **Figure 6.** Evolution of ATP-dependent remodeling and DNA methylation systems.
44 The evolutionary history and inter-familial relationships of four different remodeling
45 ATPases, Sno ATPases, SWI2/SNF2 ATPases, MORC ATPases and SMC ATPases are
46 shown in addition to DNA methylases. Horizontal lines represent temporal epochs
47 that correspond to the major transitions of eukaryote evolution; the Last Eukaryotic
48 Common Ancestor, the divergence of kinetoplastids and heteroloboseans, the
49 divergence of the chromalveolates and crown group eukaryotes, and the divergence
50 of crown group eukaryotes. Solid lines show the maximum depth to which a
51 particular family can be traced. Solid triangles are used to club multiple families. The
52 ellipses encompass potential families from which a new family with a limited phyletic
53 distribution could have emerged. Domain architectures common to all members are

1 shown along the line depicting the family. Domain architectures limited to a few
2 members of the family are shown to the right with their phyletic distribution or
3 species abbreviations in brackets. Phyletic distribution of families with a limited
4 distribution is also shown next to the family name. For a full expansion of species
5 abbreviations, please refer to the Fig 2. For a correct expansion of atypical domain
6 names, refer to the Fig.4 legend.

7 8 **Figure 7**

9 **A)** A hypothetical example showing how domain architecture networks are
10 constructed.

11 **B)** The domain architecture network for eukaryotic chromatin proteins with a focus
12 on the primary catalytic regulatory systems, namely acetylation, methylation and
13 ATP-dependent chromatin remodeling. Within acetylases, deacetylases, methylases
14 and demethylases are included all enzymes known or predicted to catalyze the
15 respective activity irrespective of the superfamily to which they belong. The links
16 made by demethylase domains are shown in aquamarine, those by acetylases in red,
17 by SWI2/SNF2 ATPases in purple and by MORC ATPases in orange. Different
18 functional categories of domains and their labels are colored in the same way and
19 spatially grouped together. The thickness of the edges is approximately proportional
20 to the relative frequency with which linkages between two domains re-occur in
21 distinct polypeptides in all eukaryotes. The graphs were rendered using PAJEK
22 (<http://vlado.fmf.uni-lj.si/pub/networks/pajek/>).

23 24 **Figure 8**

25 **A)** Domain architecture networks of proteins known or predicted to be involved in the
26 chromatin protein methylation system are shown for representative eukaryotes. The
27 proteins belonging to the methylation system includes all proteins containing
28 methylase, demethylase and methylated-peptide binding domains. Their connections
29 with each other and all the other domains occurring in their respective polypeptides
30 proteins and the domain among themselves are shown. Certain key domains of the
31 system are marked with colored shapes as indicated in the right panel of the figure.
32 Note the increasing architectural complexity as indicated by the increasing density of
33 the network over the eukaryotic evolution, especially in several crown group
34 lineages.

35 **B)** The domain architecture network for the chromatin protein acetylation-based
36 system across all eukaryotes. This set includes proteins containing acetylase,
37 deacetylase, ADP-ribose metabolite binding and acetylated peptide-binding domains.
38 Architecture network was constructed exactly as illustrated in Fig. 7A and for the
39 methylation system, except that it includes all eukaryotes. Several key chromatin
40 protein domains have colored shapes and are labeled. Red edges denote domain
41 connections that can be traced back to LECA, green shows those emerging prior to
42 the divergence of the kinetoplastid-heterolobosean clade and cyan connections can
43 be traced back to the common ancestor of the crown group and chromalveolates.
44 Note the proliferation of lineage-specific architectures in course of eukaryotic
45 evolution.

46 **C)** A similar network as Fig.8B for the ATP-dependent chromatin remodeling system
47 across all eukaryotes. This includes all proteins containing SWI2/SNF2,
48 MORC and SMC domains and various notable domains are colored and labeled.
49 Certain edges have been colored based on their point of origin as described above.
50 The thickness of the edges is approximately proportional to the frequency with which
51 linkages between two domains appear in multiple polypeptides (thickness is relative
52 within a given figure). The graphs were rendered using PAJEK (<http://vlado.fmf.uni-lj.si/pub/networks/pajek/>).

1 **References**

2
3 Aasland, R., Gibson, T.J., Stewart, A.F., 1995. The PHD finger: implications for
4 chromatin-mediated transcriptional regulation. *Trends Biochem Sci* 20, 56-59.
5 Alleman, M., Sidorenko, L., McGinnis, K., Seshadri, V., Dorweiler, J.E., White, J.,
6 Sikkink, K., Chandler, V.L., 2006. An RNA-dependent RNA polymerase is
7 required for paramutation in maize. *Nature* 442, 295-298.
8 Allis, C.D., Jenuwein, T., Reinberg, D., Caparros, M., 2006. *Epigenetics* Cold Spring
9 Harbor Laboratory Press, New York.
10 Anantharaman, V., Koonin, E.V., Aravind, L., 2001. Regulatory potential, phyletic
11 distribution and evolution of ancient, intracellular small-molecule-binding
12 domains. *J Mol Biol* 307, 1271-1292.
13 Anantharaman, V., Koonin, E.V., Aravind, L., 2002. Comparative genomics and
14 evolution of proteins involved in RNA metabolism. *Nucleic Acids Res* 30, 1427-
15 1464.
16 Aravind, L., Koonin, E.V., 1998. Second Family of Histone Deacetylases. *Science* 280,
17 1167a.
18 Aravind, L., Landsman, D., 1998. AT-hook motifs identified in a wide variety of DNA-
19 binding proteins. *Nucleic Acids Res* 26, 4413-4421.
20 Aravind, L., Koonin, E.V., 2000. SAP - a putative DNA-binding motif involved in
21 chromosomal organization. *Trends Biochem Sci* 25, 112-114.
22 Aravind, L., Watanabe, H., Lipman, D.J., Koonin, E.V., 2000. Lineage-specific loss and
23 divergence of functionally linked genes in eukaryotes. *Proc Natl Acad Sci U S A*
24 97, 11319-11324.
25 Aravind, L., 2001. The WWE domain: a common interaction module in protein
26 ubiquitination and ADP ribosylation. *Trends Biochem Sci* 26, 273-275.
27 Aravind, L., Koonin, E.V., 2001a. Prokaryotic homologs of the eukaryotic DNA-end-
28 binding protein Ku, novel domains in the Ku protein and prediction of a
29 prokaryotic double-strand break repair system. *Genome Res* 11, 1365-1374.
30 Aravind, L., Koonin, E.V., 2001b. The DNA-repair protein AlkB, EGL-9, and leprecan
31 define new families of 2-oxoglutarate- and iron-dependent dioxygenases. *Genome*
32 *Biol* 2, RESEARCH0007.
33 Aravind, L., Iyer, L.M., 2002. The SWIRM domain: a conserved module found in
34 chromosomal proteins points to novel chromatin-modifying activities. *Genome*
35 *Biol* 3, RESEARCH0039.
36 Aravind, L., Anantharaman, V., Balaji, S., Babu, M.M., Iyer, L.M., 2005. The many
37 faces of the helix-turn-helix domain: transcription regulation and beyond. *FEMS*
38 *Microbiol Rev* 29, 231-262.
39 Aravind, L., Iyer, L.M., Koonin, E.V., 2006. Comparative genomics and structural
40 biology of the molecular innovations of eukaryotes. *Curr Opin Struct Biol* 16,
41 409-419.
42 Arisue, N., Hasegawa, M., Hashimoto, T., 2005. Root of the Eukaryota tree as inferred
43 from combined maximum likelihood analyses of multiple molecular sequence
44 data. *Mol Biol Evol* 22, 409-420.
45 Armbrust, E.V., Berges, J.A., Bowler, C., Green, B.R., Martinez, D., Putnam, N.H.,
46 Zhou, S., Allen, A.E., Apt, K.E., Bechner, M., et al., 2004. The genome of the

1 diatom *Thalassiosira pseudonana*: ecology, evolution, and metabolism. *Science*
2 306, 79-86.

3 Aury, J.M., Jaillon, O., Duret, L., Noel, B., Jubin, C., Porcel, B.M., Segurens, B., Daubin,
4 V., Anthouard, V., Aiach, N., et al., 2006. Global trends of whole-genome
5 duplications revealed by the ciliate *Paramecium tetraurelia*. *Nature* 444, 171-178.

6 Avalos, J.L., Boeke, J.D., Wolberger, C., 2004. Structural basis for the mechanism and
7 regulation of Sir2 enzymes. *Mol Cell* 13, 639-648.

8 Babu, M.M., Luscombe, N.M., Aravind, L., Gerstein, M., Teichmann, S.A., 2004.
9 Structure and evolution of transcriptional regulatory networks. *Curr Opin Struct*
10 *Biol* 14, 283-291.

11 Babu, M.M., Iyer, L.M., Balaji, S., Aravind, L., 2006. The natural history of the WRKY-
12 GCM1 zinc fingers and the relationship between transcription factors and
13 transposons. *Nucleic Acids Res* 34, 6505-6520.

14 Balaji, S., Babu, M.M., Iyer, L.M., Aravind, L., 2005. Discovery of the principal specific
15 transcription factors of Apicomplexa and their implication for the evolution of the
16 AP2-integrase DNA binding domains. *Nucleic Acids Res* 33, 3994-4006.

17 Bannister, A.J., Zegerman, P., Partridge, J.F., Miska, E.A., Thomas, J.O., Allshire, R.C.,
18 Kouzarides, T., 2001. Selective recognition of methylated lysine 9 on histone H3
19 by the HP1 chromo domain. *Nature* 410, 120-124.

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

24 Bellows, A.M., Kenna, M.A., Cassimeris, L., Skibbens, R.V., 2003. Human EFO1p
25 exhibits acetyltransferase activity and is a unique combination of linker histone
26 and Ctf7p/Eco1p chromatid cohesion establishment domains. *Nucleic Acids Res*
27 31, 6334-6343.

28 Bennett-Lovsey, R., Hart, S.E., Shirai, H., Mizuguchi, K., 2002. The SWIB and the
29 MDM2 domains are homologous and share a common fold. *Bioinformatics* 18,
30 626-630.

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

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

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

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

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

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

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

7 Brehm, A., Tufteland, K.R., Aasland, R., Becker, P.B., 2004. The many colours of
8 chromodomains. *Bioessays* 26, 133-140.

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

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

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

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

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

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

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

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

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

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

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

39 de la Cruz, X., Lois, S., Sanchez-Molina, S., Martinez-Balbas, M.A., 2005. Do protein
40 motifs read the histone code? *Bioessays* 27, 164-175.

41 Deitsch, K.W., Calderwood, M.S., Wellems, T.E., 2001. Malaria. Cooperative silencing
42 elements in var genes. *Nature* 412, 875-876.

43 Denhardt, D.T., Chaly, N., Walden, D.B., 2005. The eukaryotic nucleus: A thematic
44 issue. <http://www3.interscience.wiley.com/cgi-bin/jhome/109911273>
45 *BioEssays* 9, 43.

1 DiPaolo, C., Kieft, R., Cross, M., Sabatini, R., 2005. Regulation of trypanosome DNA
2 glycosylation by a SWI2/SNF2-like protein. *Mol Cell* 17, 441-451.

3 Doolittle, W.F., 1998. You are what you eat: a gene transfer ratchet could account for
4 bacterial genes in eukaryotic nuclear genomes. *Trends Genet* 14, 307-311.

5 Driscoll, R., Hudson, A., Jackson, S.P., 2007. Yeast Rtt109 promotes genome stability by
6 acetylating histone H3 on lysine 56. *Science* 315, 649-652.

7 Dunn, M.J., Jorde, L.B., Little, P.F., Subramanian, S., 2005. *Encyclopedia of Genetics,*
8 *Genomics, Proteomics and Bioinformatics.* John Wiley & Sons, Inc. , London.

9 Duraisingh, M.T., Voss, T.S., Marty, A.J., Duffy, M.F., Good, R.T., Thompson, J.K.,
10 Freitas-Junior, L.H., Scherf, A., Crabb, B.S., Cowman, A.F., 2005.
11 Heterochromatin silencing and locus repositioning linked to regulation of
12 virulence genes in *Plasmodium falciparum*. *Cell* 121, 13-24.

13 Durant, M., Pugh, B.F., 2006. Genome-wide relationships between TAF1 and histone
14 acetyltransferases in *Saccharomyces cerevisiae*. *Mol Cell Biol* 26, 2791-2802.

15 Durr, H., Hopfner, K.P., 2006. Structure-function analysis of SWI2/SNF2 enzymes.
16 *Methods Enzymol* 409, 375-388.

17 Dutnall, R.N., 2003. Cracking the histone code: one, two, three methyls, you're out! *Mol*
18 *Cell* 12, 3-4.

19 Eberharter, A., Vetter, I., Ferreira, R., Becker, P.B., 2004. ACF1 improves the
20 effectiveness of nucleosome mobilization by ISWI through PHD-histone contacts.
21 *Embo J* 23, 4029-4039.

22 El-Sayed, N.M., Myler, P.J., Blandin, G., Berriman, M., Crabtree, J., Aggarwal, G.,
23 Caler, E., Renauld, H., Worthey, E.A., Hertz-Fowler, C., et al., 2005.
24 Comparative genomics of trypanosomatid parasitic protozoa. *Science* 309, 404-
25 409.

26 Felsenstein, J., 1989. PHYLIP -- Phylogeny Inference Package (Version 3.2). *Cladistics*
27 5, 164-166.

28 Finn, R.D., Mistry, J., Schuster-Bockler, B., Griffiths-Jones, S., Hollich, V., Lassmann,
29 T., Moxon, S., Marshall, M., Khanna, A., Durbin, R., et al., 2006. Pfam: clans,
30 web tools and services. *Nucleic Acids Res* 34, D247-251.

31 Flanagan, J.F., Mi, L.Z., Chruszcz, M., Cymborowski, M., Clines, K.L., Kim, Y., Minor,
32 W., Rastinejad, F., Khorasanizadeh, S., 2005. Double chromodomains cooperate
33 to recognize the methylated histone H3 tail. *Nature* 438, 1181-1185.

34 Frank, M., Dzikowski, R., Costantini, D., Amulic, B., Berdugo, E., Deitsch, K., 2006.
35 Strict pairing of var promoters and introns is required for var gene silencing in the
36 malaria parasite *Plasmodium falciparum*. *J Biol Chem* 281, 9942-9952.

37 Freitag, M., Williams, R.L., Kothe, G.O., Selker, E.U., 2002. A cytosine
38 methyltransferase homologue is essential for repeat-induced point mutation in
39 *Neurospora crassa*. *Proc Natl Acad Sci U S A* 99, 8802-8807.

40 Freitas-Junior, L.H., Hernandez-Rivas, R., Ralph, S.A., Montiel-Condado, D.,
41 Ruvalcaba-Salazar, O.K., Rojas-Meza, A.P., Mancio-Silva, L., Leal-Silvestre,
42 R.J., Gontijo, A.M., Shorte, S., et al., 2005. Telomeric heterochromatin
43 propagation and histone acetylation control mutually exclusive expression of
44 antigenic variation genes in malaria parasites. *Cell* 121, 25-36.

- 1 Frye, R.A., 1999. Characterization of five human cDNAs with homology to the yeast
2 SIR2 gene: Sir2-like proteins (sirtuins) metabolize NAD and may have protein
3 ADP-ribosyltransferase activity. *Biochem Biophys Res Commun* 260, 273-279.
- 4 Gangavarapu, V., Haracska, L., Unk, I., Johnson, R.E., Prakash, S., Prakash, L., 2006.
5 Mms2-Ubc13-dependent and -independent roles of Rad5 ubiquitin ligase in
6 postreplication repair and translesion DNA synthesis in *Saccharomyces*
7 *cerevisiae*. *Mol Cell Biol* 26, 7783-7790.
- 8 Gangloff, Y.G., Romier, C., Thuault, S., Werten, S., Davidson, I., 2001. The histone fold
9 is a key structural motif of transcription factor TFIID. *Trends Biochem Sci* 26,
10 250-257.
- 11 Gardner, M.J., Hall, N., Fung, E., White, O., Berriman, M., Hyman, R.W., Carlton, J.M.,
12 Pain, A., Nelson, K.E., Bowman, S., et al., 2002. Genome sequence of the human
13 malaria parasite *Plasmodium falciparum*. *Nature* 419, 498-511.
- 14 Gearhart, M.D., Corcoran, C.M., Wamstad, J.A., Bardwell, V.J., 2006. Polycomb group
15 and SCF ubiquitin ligases are found in a novel BCOR complex that is recruited to
16 BCL6 targets. *Mol Cell Biol* 26, 6880-6889.
- 17 Gerber, A.P., Keller, W., 1999. An adenosine deaminase that generates inosine at the
18 wobble position of tRNAs. *Science* 286, 1146-1149.
- 19 Ghosh, D., Papavassiliou, A.G., 2005. Transcription factor therapeutics: long-shot or
20 lodestone. *Curr Med Chem* 12, 691-701.
- 21 Gibson, T.J., Spring, J., 1998. Genetic redundancy in vertebrates: polyploidy and
22 persistence of genes encoding multidomain proteins. *Trends Genet* 14, 46-49;
23 discussion 49-50.
- 24 Glickman, M.H., Ciechanover, A., 2002. The ubiquitin-proteasome proteolytic pathway:
25 destruction for the sake of construction. *Physiol Rev* 82, 373-428.
- 26 Goff, L.J., Coleman, A.W., 1995. Fate of Parasite and Host Organelle DNA during
27 Cellular Transformation of Red Algae by Their Parasites. *Plant Cell* 7, 1899-
28 1911.
- 29 Goll, M.G., Bestor, T.H., 2005. Eukaryotic cytosine methyltransferases. *Annu Rev*
30 *Biochem* 74, 481-514.
- 31 Goll, M.G., Kirpekar, F., Maggert, K.A., Yoder, J.A., Hsieh, C.L., Zhang, X., Golic,
32 K.G., Jacobsen, S.E., Bestor, T.H., 2006. Methylation of tRNA^{Asp} by the DNA
33 methyltransferase homolog Dnmt2. *Science* 311, 395-398.
- 34 Grewal, S.I., Moazed, D., 2003. Heterochromatin and epigenetic control of gene
35 expression. *Science* 301, 798-802.
- 36 Grewal, S.I., Rice, J.C., 2004. Regulation of heterochromatin by histone methylation and
37 small RNAs. *Curr Opin Cell Biol* 16, 230-238.
- 38 Guindon, S., Gascuel, O., 2003. A simple, fast, and accurate algorithm to estimate large
39 phylogenies by maximum likelihood. *Syst Biol* 52, 696-704.
- 40 Han, J., Zhou, H., Horazdovsky, B., Zhang, K., Xu, R.M., Zhang, Z., 2007. Rtt109
41 acetylates histone H3 lysine 56 and functions in DNA replication. *Science* 315,
42 653-655.
- 43 Hauser, B.A., He, J.Q., Park, S.O., Gasser, C.S., 2000. TSO1 is a novel protein that
44 modulates cytokinesis and cell expansion in *Arabidopsis*. *Development* 127,
45 2219-2226.

- 1 Hirano, T., 2005. SMC proteins and chromosome mechanics: from bacteria to humans.
2 *Philos Trans R Soc Lond B Biol Sci* 360, 507-514.
- 3 Hirano, T., 2006. At the heart of the chromosome: SMC proteins in action. *Nat Rev Mol*
4 *Cell Biol* 7, 311-322.
- 5 Inoue, N., Hess, K.D., Moreadith, R.W., Richardson, L.L., Handel, M.A., Watson, M.L.,
6 Zinn, A.R., 1999. New gene family defined by MORC, a nuclear protein required
7 for mouse spermatogenesis. *Hum Mol Genet* 8, 1201-1207.
- 8 Iyer, L.M., Aravind, L., 2004. The emergence of catalytic and structural diversity within
9 the beta-clip fold. *Proteins* 55, 977-991.
- 10 Iyer, L.M., Babu, M.M., Aravind, L., 2006. The HIRAN domain and recruitment of
11 chromatin remodeling and repair activities to damaged DNA. *Cell Cycle* 5, 775-
12 782.
- 13 James, T.Y., Kauff, F., Schoch, C.L., Matheny, P.B., Hofstetter, V., Cox, C.J., Celio, G.,
14 Gueidan, C., Fraker, E., Miadlikowska, J., et al., 2006. Reconstructing the early
15 evolution of Fungi using a six-gene phylogeny. *Nature* 443, 818-822.
- 16 Janzen, C.J., Hake, S.B., Lowell, J.E., Cross, G.A., 2006. Selective di- or trimethylation
17 of histone H3 lysine 76 by two DOT1 homologs is important for cell cycle
18 regulation in *Trypanosoma brucei*. *Mol Cell* 23, 497-507.
- 19 Johnson, L.M., Bostick, M., Zhang, X., Kraft, E., Henderson, I., Callis, J., Jacobsen, S.E.,
20 2007. The SRA methyl-cytosine-binding domain links DNA and histone
21 methylation. *Curr Biol* 17, 379-384.
- 22 Kaadige, M.R., Ayer, D.E., 2006. The polybasic region that follows the plant
23 homeodomain zinc finger 1 of Pfl is necessary and sufficient for specific
24 phosphoinositide binding. *J Biol Chem* 281, 28831-28836.
- 25 Karras, G.I., Kustatscher, G., Buhecha, H.R., Allen, M.D., Pugieux, C., Sait, F., Bycroft,
26 M., Ladurner, A.G., 2005. The macro domain is an ADP-ribose binding module.
27 *Embo J* 24, 1911-1920.
- 28 Katinka, M.D., Duprat, S., Cornillot, E., Metenier, G., Thomarat, F., Prensier, G., Barbe,
29 V., Peyretailade, E., Brottier, P., Wincker, P., et al., 2001. Genome sequence and
30 gene compaction of the eukaryote parasite *Encephalitozoon cuniculi*. *Nature* 414,
31 450-453.
- 32 Kim, J., Daniel, J., Espejo, A., Lake, A., Krishna, M., Xia, L., Zhang, Y., Bedford, M.T.,
33 2006. Tudor, MBT and chromo domains gauge the degree of lysine methylation.
34 *EMBO Rep* 7, 397-403.
- 35 Klose, R.J., Yamane, K., Bae, Y., Zhang, D., Erdjument-Bromage, H., Tempst, P., Wong,
36 J., Zhang, Y., 2006. The transcriptional repressor JHDM3A demethylates
37 trimethyl histone H3 lysine 9 and lysine 36. *Nature* 442, 312-316.
- 38 Koonin, E.V., Aravind, L., Kondrashov, A.S., 2000. The impact of comparative
39 genomics on our understanding of evolution. *Cell* 101, 573-576.
- 40 Kouzarides, T., 2007. Chromatin modifications and their function. *Cell* 128, 693-705.
- 41 Kreier, J., 1977. Parasitic Protozoa. Academic Press, New York.
- 42 Lachner, M., O'Carroll, D., Rea, S., Mechtler, K., Jenuwein, T., 2001. Methylation of
43 histone H3 lysine 9 creates a binding site for HP1 proteins. *Nature* 410, 116-120.
- 44 LaCount, D.J., Vignali, M., Chettier, R., Phansalkar, A., Bell, R., Hesselberth, J.R.,
45 Schoenfeld, L.W., Ota, I., Sahasrabudhe, S., Kurschner, C., et al., 2005. A protein

1 interaction network of the malaria parasite *Plasmodium falciparum*. *Nature* 438,
2 103-107.

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

6 Latchman, D., 2005. *Gene Regulation*. Taylor & Francis, New York.

7 Lau, A.O., Smith, A.J., Brown, M.T., Johnson, P.J., 2006. *Trichomonas vaginalis* initiator
8 binding protein (IBP39) and RNA polymerase II large subunit carboxy terminal
9 domain interaction. *Mol Biochem Parasitol* 150, 56-62.

10 Le Roch, K.G., Zhou, Y., Blair, P.L., Grainger, M., Moch, J.K., Haynes, J.D., De La
11 Vega, P., Holder, A.A., Batalov, S., Carucci, D.J., et al., 2003. Discovery of gene
12 function by expression profiling of the malaria parasite life cycle. *Science* 301,
13 1503-1508.

14 Leander, B.S., Keeling, P.J., 2003. Morphostasis in alveolate evolution. *Trends in*
15 *Ecology and Evolution* 18, 395-402.

16 Leipe, D.D., Landsman, D., 1997. Histone deacetylases, acetoin utilization proteins and
17 acetylpolymine amidohydrolases are members of an ancient protein superfamily.
18 *Nucleic Acids Res* 25, 3693-3697.

19 Lespinet, O., Wolf, Y.I., Koonin, E.V., Aravind, L., 2002. The role of lineage-specific
20 gene family expansion in the evolution of eukaryotes. *Genome Res* 12, 1048-
21 1059.

22 Letunic, I., Copley, R.R., Pils, B., Pinkert, S., Schultz, J., Bork, P., 2006. SMART 5:
23 domains in the context of genomes and networks. *Nucleic Acids Res* 34, D257-
24 260.

25 Li, C.F., Pontes, O., El-Shami, M., Henderson, I.R., Bernatavichute, Y.V., Chan, S.W.,
26 Lagrange, T., Pikaard, C.S., Jacobsen, S.E., 2006a. An ARGONAUTE4-
27 containing nuclear processing center colocalized with Cajal bodies in *Arabidopsis*
28 *thaliana*. *Cell* 126, 93-106.

29 Li, H., Ilin, S., Wang, W., Duncan, E.M., Wysocka, J., Allis, C.D., Patel, D.J., 2006b.
30 Molecular basis for site-specific read-out of histone H3K4me3 by the BPTF PHD
31 finger of NURF. *Nature* 442, 91-95.

32 Liu, J., Tan, H., Rost, B., 2002. Loopy proteins appear conserved in evolution. *J Mol Biol*
33 322, 53-64.

34 Loftus, B., Anderson, I., Davies, R., Alsmark, U.C., Samuelson, J., Amedeo, P.,
35 Roncaglia, P., Berriman, M., Hirt, R.P., Mann, B.J., et al., 2005. The genome of
36 the protist parasite *Entamoeba histolytica*. *Nature* 433, 865-868.

37 Luk, E., Vu, N.D., Patteson, K., Mizuguchi, G., Wu, W.H., Ranjan, A., Backus, J., Sen,
38 S., Lewis, M., Bai, Y., et al., 2007. Chz1, a nuclear chaperone for histone H2AZ.
39 *Mol Cell* 25, 357-368.

40 Makarova, K.S., Aravind, L., Wolf, Y.I., Tatusov, R.L., Minton, K.W., Koonin, E.V.,
41 Daly, M.J., 2001. Genome of the extremely radiation-resistant bacterium
42 *Deinococcus radiodurans* viewed from the perspective of comparative genomics.
43 *Microbiol Mol Biol Rev* 65, 44-79.

44 Malagnac, F., Wendel, B., Goyon, C., Faugeron, G., Zickler, D., Rossignol, J.L., Noyer-
45 Weidner, M., Vollmayr, P., Trautner, T.A., Walter, J., 1997. A gene essential for

1 de novo methylation and development in *Ascobolus* reveals a novel type of
2 eukaryotic DNA methyltransferase structure. *Cell* 91, 281-290.

3 Malone, C.D., Anderson, A.M., Motl, J.A., Rexer, C.H., Chalker, D.L., 2005. Germ line
4 transcripts are processed by a Dicer-like protein that is essential for
5 developmentally programmed genome rearrangements of *Tetrahymena*
6 *thermophila*. *Mol Cell Biol* 25, 9151-9164.

7 Manning, G., Plowman, G.D., Hunter, T., Sudarsanam, S., 2002. Evolution of protein
8 kinase signaling from yeast to man. *Trends Biochem Sci* 27, 514-520.

9 Mans, B.J., Anantharaman, V., Aravind, L., Koonin, E.V., 2004. Comparative genomics,
10 evolution and origins of the nuclear envelope and nuclear pore complex. *Cell*
11 *Cycle* 3, 1612-1637.

12 Martens, J.A., Winston, F., 2003. Recent advances in understanding chromatin
13 remodeling by Swi/Snf complexes. *Curr Opin Genet Dev* 13, 136-142.

14 Maurer-Stroh, S., Dickens, N.J., Hughes-Davies, L., Kouzarides, T., Eisenhaber, F.,
15 Ponting, C.P., 2003. The Tudor domain 'Royal Family': Tudor, plant Agenet,
16 Chromo, PWWP and MBT domains. *Trends Biochem Sci* 28, 69-74.

17 Metzger, E., Wissmann, M., Yin, N., Muller, J.M., Schneider, R., Peters, A.H., Gunther,
18 T., Buettner, R., Schule, R., 2005. LSD1 demethylates repressive histone marks to
19 promote androgen-receptor-dependent transcription. *Nature* 437, 436-439.

20 Mo, X., Kowenz-Leutz, E., Laumonier, Y., Xu, H., Leutz, A., 2005. Histone H3 tail
21 positioning and acetylation by the c-Myb but not the v-Myb DNA-binding SANT
22 domain. *Genes Dev* 19, 2447-2457.

23 Mochizuki, K., Fine, N.A., Fujisawa, T., Gorovsky, M.A., 2002. Analysis of a piwi-
24 related gene implicates small RNAs in genome rearrangement in tetrahymena.
25 *Cell* 110, 689-699.

26 Mochizuki, K., Gorovsky, M.A., 2004. Small RNAs in genome rearrangement in
27 *Tetrahymena*. *Curr Opin Genet Dev* 14, 181-187.

28 Mohrmann, L., Verrijzer, C.P., 2005. Composition and functional specificity of
29 SWI2/SNF2 class chromatin remodeling complexes. *Biochim Biophys Acta* 1681,
30 59-73.

31 Moon-van der Staay, S.Y., De Wachter, R., Vaultot, D., 2001. Oceanic 18S rDNA
32 sequences from picoplankton reveal unsuspected eukaryotic diversity. *Nature* 409,
33 607-610.

34 Namboodiri, V.M., Dutta, S., Akey, I.V., Head, J.F., Akey, C.W., 2003. The crystal
35 structure of *Drosophila* NLP-core provides insight into pentamer formation and
36 histone binding. *Structure* 11, 175-186.

37 Neuwald, A.F., Landsman, D., 1997. GCN5-related histone N-acetyltransferases belong
38 to a diverse superfamily that includes the yeast SPT10 protein. *Trends Biochem*
39 *Sci* 22, 154-155.

40 Ono, R., Taki, T., Taketani, T., Taniwaki, M., Kobayashi, H., Hayashi, Y., 2002. LCX,
41 leukemia-associated protein with a CXXC domain, is fused to MLL in acute
42 myeloid leukemia with trilineage dysplasia having t(10;11)(q22;q23). *Cancer Res*
43 62, 4075-4080.

44 Paraskevopoulou, C., Fairhurst, S.A., Lowe, D.J., Brick, P., Onesti, S., 2006. The
45 Elongator subunit Elp3 contains a Fe4S4 cluster and binds S-adenosylmethionine.
46 *Mol Microbiol* 59, 795-806.

1 Park, S.W., Hu, X., Gupta, P., Lin, Y.P., Ha, S.G., Wei, L.N., 2007. SUMOylation of Tr2
2 orphan receptor involves Pml and fine-tunes Oct4 expression in stem cells. *Nat*
3 *Struct Mol Biol* 14, 68-75.

4 Park, Y.J., Luger, K., 2006. The structure of nucleosome assembly protein 1. *Proc Natl*
5 *Acad Sci U S A* 103, 1248-1253.

6 Pellegrini-Calace, M., Thornton, J.M., 2005. Detecting DNA-binding helix-turn-helix
7 structural motifs using sequence and structure information. *Nucleic Acids Res* 33,
8 2129-2140.

9 Pena, P.V., Davrazou, F., Shi, X., Walter, K.L., Verkhusha, V.V., Gozani, O., Zhao, R.,
10 Kutateladze, T.G., 2006. Molecular mechanism of histone H3K4me3 recognition
11 by plant homeodomain of ING2. *Nature* 442, 100-103.

12 Peterson, C.L., Laniel, M.A., 2004. Histones and histone modifications. *Curr Biol* 14,
13 R546-551.

14 Pontes, O., Li, C.F., Nunes, P.C., Haag, J., Ream, T., Vitins, A., Jacobsen, S.E., Pikaard,
15 C.S., 2006. The Arabidopsis chromatin-modifying nuclear siRNA pathway
16 involves a nucleolar RNA processing center. *Cell* 126, 79-92.

17 Ponting, C.P., Aravind, L., Schultz, J., Bork, P., Koonin, E.V., 1999. Eukaryotic
18 signalling domain homologues in archaea and bacteria. *Ancient ancestry and*
19 *horizontal gene transfer. J Mol Biol* 289, 729-745.

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

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

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

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

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

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

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

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

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

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

- 1 Schumacher, M.A., Lau, A.O., Johnson, P.J., 2003. Structural basis of core promoter
2 recognition in a primitive eukaryote. *Cell* 115, 413-424.
- 3 Schuster, F.L., Visvesvara, G.S., 2004. Free-living amoebae as opportunistic and non-
4 opportunistic pathogens of humans and animals. *Int J Parasitol* 34, 1001-1027.
- 5 Shi, H., Chamond, N., Tschudi, C., Ullu, E., 2004a. Selection and characterization of
6 RNA interference-deficient trypanosomes impaired in target mRNA degradation.
7 *Eukaryot Cell* 3, 1445-1453.
- 8 Shi, X., Hong, T., Walter, K.L., Ewalt, M., Michishita, E., Hung, T., Carney, D., Pena, P.,
9 Lan, F., Kaadige, M.R., et al., 2006. ING2 PHD domain links histone H3 lysine 4
10 methylation to active gene repression. *Nature* 442, 96-99.
- 11 Shi, Y., Lan, F., Matson, C., Mulligan, P., Whetstone, J.R., Cole, P.A., Casero, R.A., Shi,
12 Y., 2004b. Histone demethylation mediated by the nuclear amine oxidase
13 homolog LSD1. *Cell* 119, 941-953.
- 14 Shilatifard, A., 2006. Chromatin modifications by methylation and ubiquitination:
15 implications in the regulation of gene expression. *Annu Rev Biochem* 75, 243-
16 269.
- 17 Shiu, P.K., Raju, N.B., Zickler, D., Metznerberg, R.L., 2001. Meiotic silencing by
18 unpaired DNA. *Cell* 107, 905-916.
- 19 Shull, N.P., Spinelli, S.L., Phizicky, E.M., 2005. A highly specific phosphatase that acts
20 on ADP-ribose 1"-phosphate, a metabolite of tRNA splicing in *Saccharomyces*
21 *cerevisiae*. *Nucleic Acids Res* 33, 650-660.
- 22 Simpson, A.G., Inagaki, Y., Roger, A.J., 2006. Comprehensive multigene phylogenies of
23 excavate protists reveal the evolutionary positions of "primitive" eukaryotes. *Mol*
24 *Biol Evol* 23, 615-625.
- 25 Smit, A.F., Riggs, A.D., 1996. Tiggers and DNA transposon fossils in the human
26 genome. *Proc Natl Acad Sci U S A* 93, 1443-1448.
- 27 Smothers, J.F., von Dohlen, C.D., Smith, L.H., Jr., Spall, R.D., 1994. Molecular evidence
28 that the myxozoan protists are metazoans. *Science* 265, 1719-1721.
- 29 Soler-Lopez, M., Petosa, C., Fukuzawa, M., Ravelli, R., Williams, J.G., Muller, C.W.,
30 2004. Structure of an activated *Dictyostelium* STAT in its DNA-unbound form.
31 *Mol Cell* 13, 791-804.
- 32 Stavropoulos, P., Blobel, G., Hoelz, A., 2006. Crystal structure and mechanism of human
33 lysine-specific demethylase-1. *Nat Struct Mol Biol* 13, 626-632.
- 34 Stephens, R.S., Kalman, S., Lammel, C., Fan, J., Marathe, R., Aravind, L., Mitchell, W.,
35 Olinger, L., Tatusov, R.L., Zhao, Q., et al., 1998. Genome sequence of an obligate
36 intracellular pathogen of humans: *Chlamydia trachomatis*. *Science* 282, 754-759.
- 37 Sullivan, W.J., Jr., Naguleswaran, A., Angel, S.O., 2006. Histones and histone
38 modifications in protozoan parasites. *Cell Microbiol* 8, 1850-1861.
- 39 Tang, Y., Poustovoitov, M.V., Zhao, K., Garfinkel, M., Canutescu, A., Dunbrack, R.,
40 Adams, P.D., Marmorstein, R., 2006. Structure of a human ASF1a-HIRA
41 complex and insights into specificity of histone chaperone complex assembly. *Nat*
42 *Struct Mol Biol* 13, 921-929.
- 43 Templeton, T.J., Iyer, L.M., Anantharaman, V., Enomoto, S., Abrahante, J.E.,
44 Subramanian, G.M., Hoffman, S.L., Abrahamsen, M.S., Aravind, L., 2004.
45 Comparative analysis of apicomplexa and genomic diversity in eukaryotes.
46 *Genome Res* 14, 1686-1695.

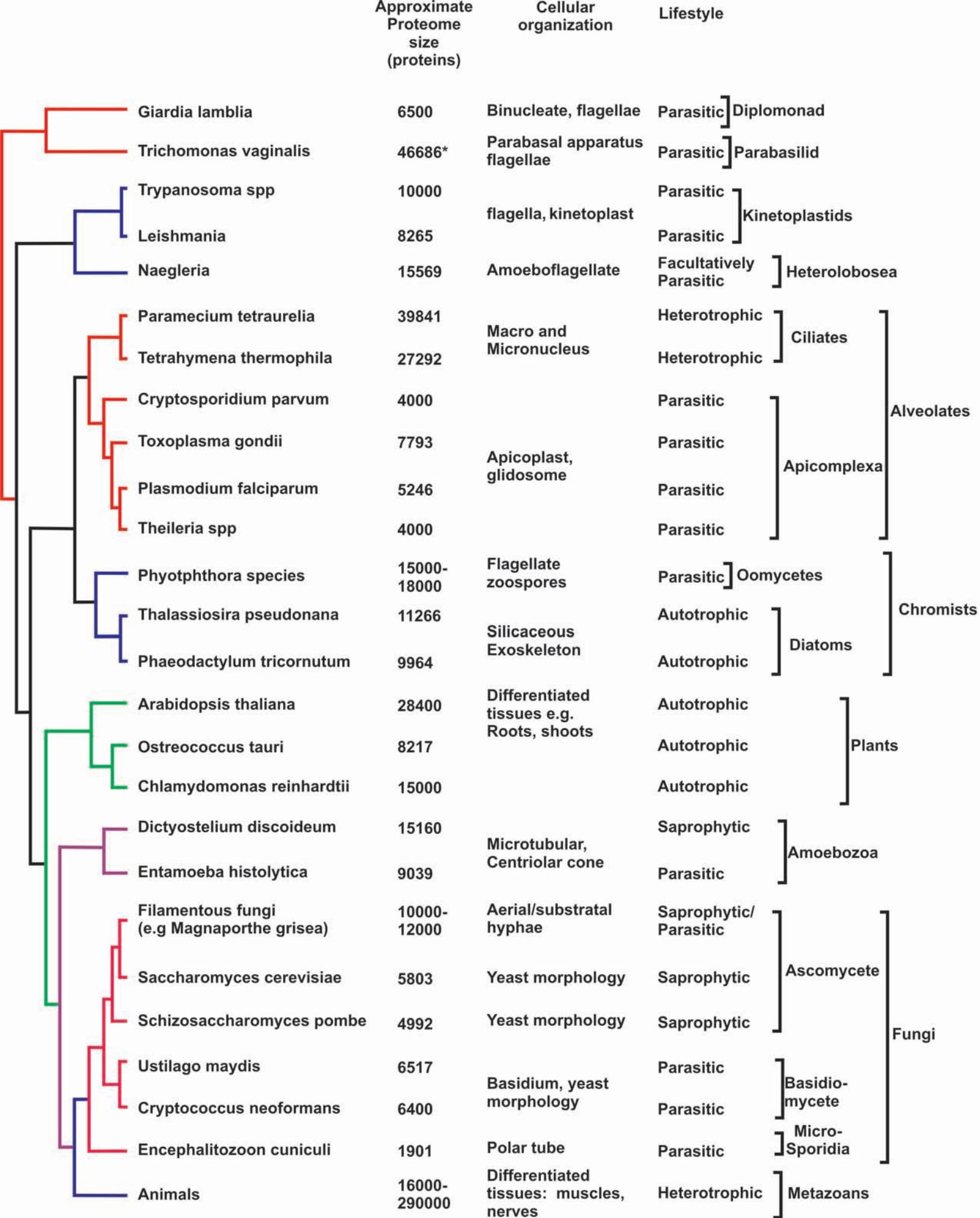
1 Thomas, T., Voss, A.K., 2007. The Diverse Biological Roles of MYST Histone
2 Acetyltransferase Family Proteins. *Cell Cycle* 6.
3 Tyler, B.M., Tripathy, S., Zhang, X., Dehal, P., Jiang, R.H., Aerts, A., Arredondo, F.D.,
4 Baxter, L., Bensasson, D., Beynon, J.L., et al., 2006. *Phytophthora* genome
5 sequences uncover evolutionary origins and mechanisms of pathogenesis. *Science*
6 313, 1261-1266.
7 Uhlmann, F., Hopfner, K.P., 2006. Chromosome biology: the crux of the ring. *Curr Biol*
8 16, R102-105.
9 Ullu, E., Tschudi, C., Chakraborty, T., 2004. RNA interference in protozoan parasites.
10 *Cell Microbiol* 6, 509-519.
11 Vaucheret, H., 2006. Post-transcriptional small RNA pathways in plants: mechanisms
12 and regulations. *Genes Dev* 20, 759-771.
13 Villar-Garea, A., Imhof, A., 2006. The analysis of histone modifications. *Biochim*
14 *Biophys Acta* 1764, 1932-1939.
15 Visser, A.E., Verschure, P.J., Gommans, W.M., Haisma, H.J., Rots, M.G., 2006. Step
16 into the groove: engineered transcription factors as modulators of gene
17 expression. *Adv Genet* 56, 131-161.
18 Walsh, D.A., Doolittle, W.F., 2005. The real 'domains' of life. *Curr Biol* 15, R237-240.
19 White, M.F., Bell, S.D., 2002. Holding it together: chromatin in the Archaea. *Trends*
20 *Genet* 18, 621-626.
21 Wittschieben, B.O., Otero, G., de Bizemont, T., Fellows, J., Erdjument-Bromage, H.,
22 Ohba, R., Li, Y., Allis, C.D., Tempst, P., Svejstrup, J.Q., 1999. A novel histone
23 acetyltransferase is an integral subunit of elongating RNA polymerase II
24 holoenzyme. *Mol Cell* 4, 123-128.
25 Woo, H.R., Pontes, O., Pikaard, C.S., Richards, E.J., 2007. VIM1, a methylcytosine-
26 binding protein required for centromeric heterochromatinization. *Genes Dev* 21,
27 267-277.
28 Woodcock, C.L., 2006. Chromatin architecture. *Curr Opin Struct Biol* 16, 213-220.
29 Yu, Z., Genest, P.A., Riet, B.T., Sweeney, K., Dipaolo, C., Kieft, R., Christodoulou, E.,
30 Perrakis, A., Simmons, J.M., Hausinger, R.P., et al., 2007. The protein that binds
31 to DNA base J in trypanosomatids has features of a thymidine hydroxylase.
32 *Nucleic Acids Res.*
33 Zeng, L., Zhou, M.M., 2002. Bromodomain: an acetyl-lysine binding domain. *FEBS Lett*
34 513, 124-128.
35 Zhang, H., Christoforou, A., Aravind, L., Emmons, S.W., van den Heuvel, S., Haber,
36 D.A., 2004. The *C. elegans* Polycomb gene SOP-2 encodes an RNA binding
37 protein. *Mol Cell* 14, 841-847.
38 Zhang, J., 2003. Are poly(ADP-ribosylation) by PARP-1 and deacetylation by Sir2
39 linked? *Bioessays* 25, 808-814.
40
41

Table1. Domains commonly found in chromatin proteins

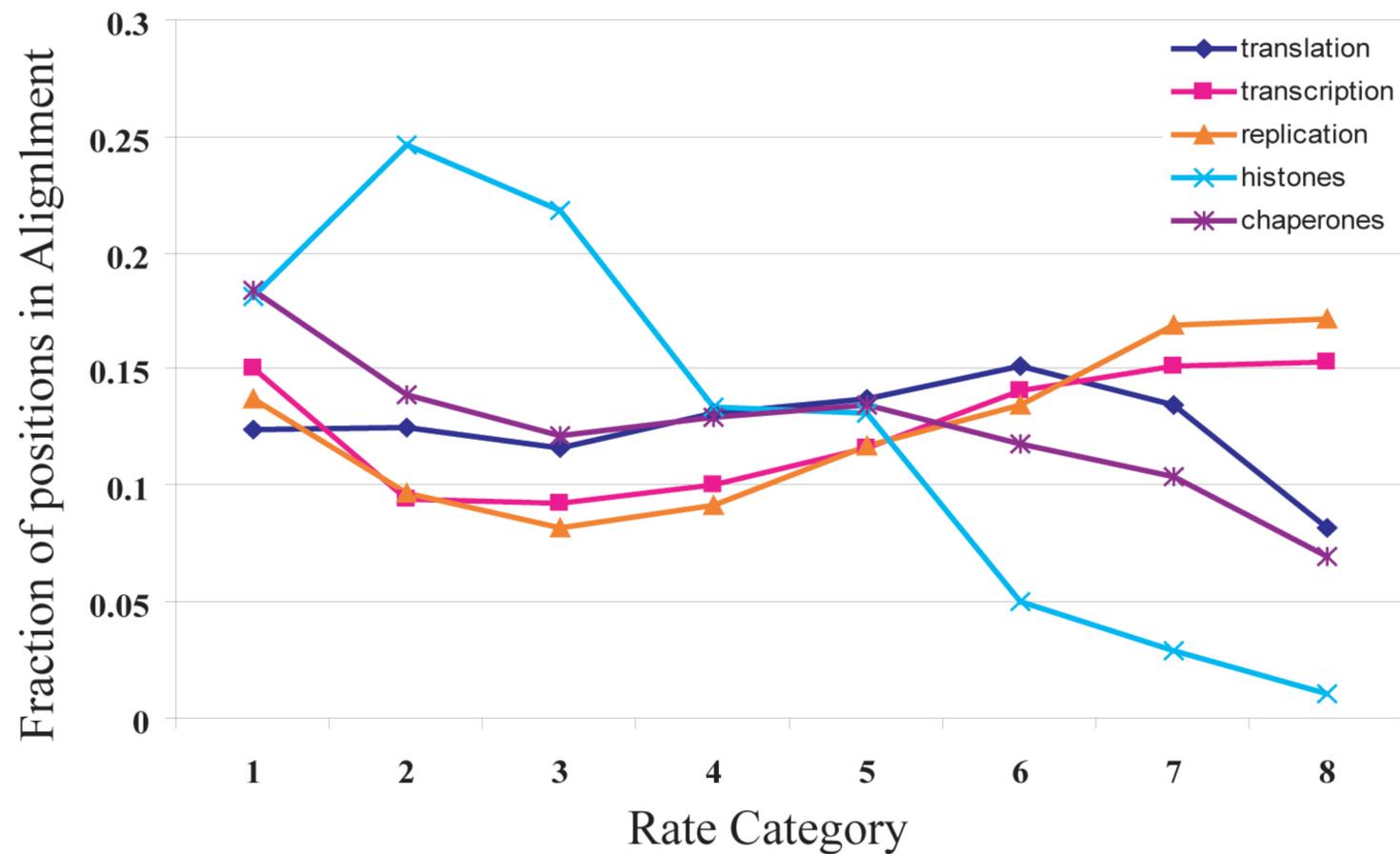
Domain	Structure	Comments
Enzymatic domains		
Acetyltransferases (GNAT)	$\alpha + \beta$ fold with 6 core strands	No particular universally conserved active site residues but a structurally conserved acetyl coA binding loop
RPD3/HDAC-like deacetylases	Haloacid dehalogenase class of Rossmannoid folds	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)	$\alpha + \beta$ 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.

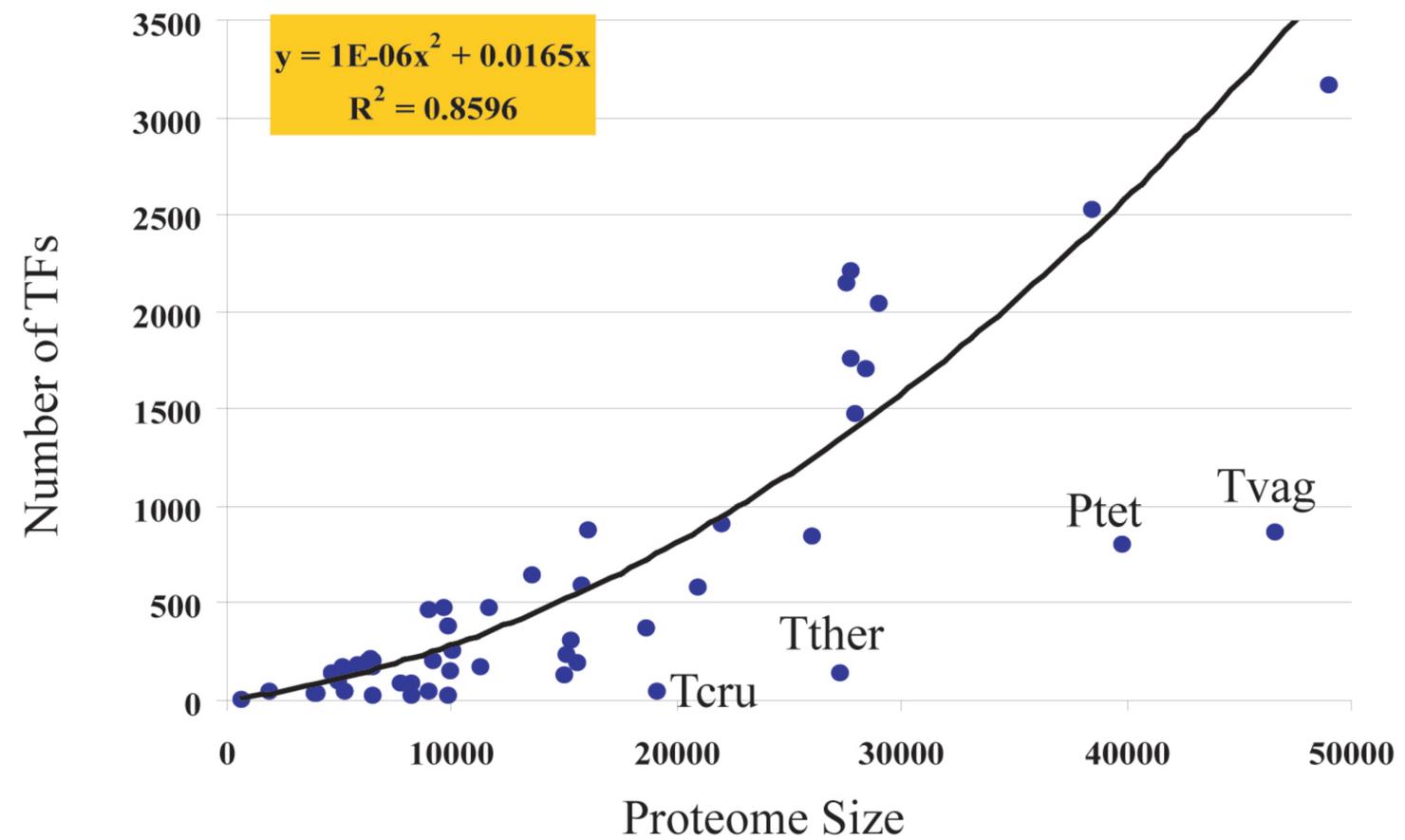
Other chromatin associated domains		
ZfCW/PHDX	Treble clef fold with a mononuclear Zn-chelation site	The earliest versions of this domain are traceable to the kinetoplastids.
EP1	α -helical	The version traceable to LECA is present in the enhancer of polycomb-like proteins and is a component of the NuA4 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.



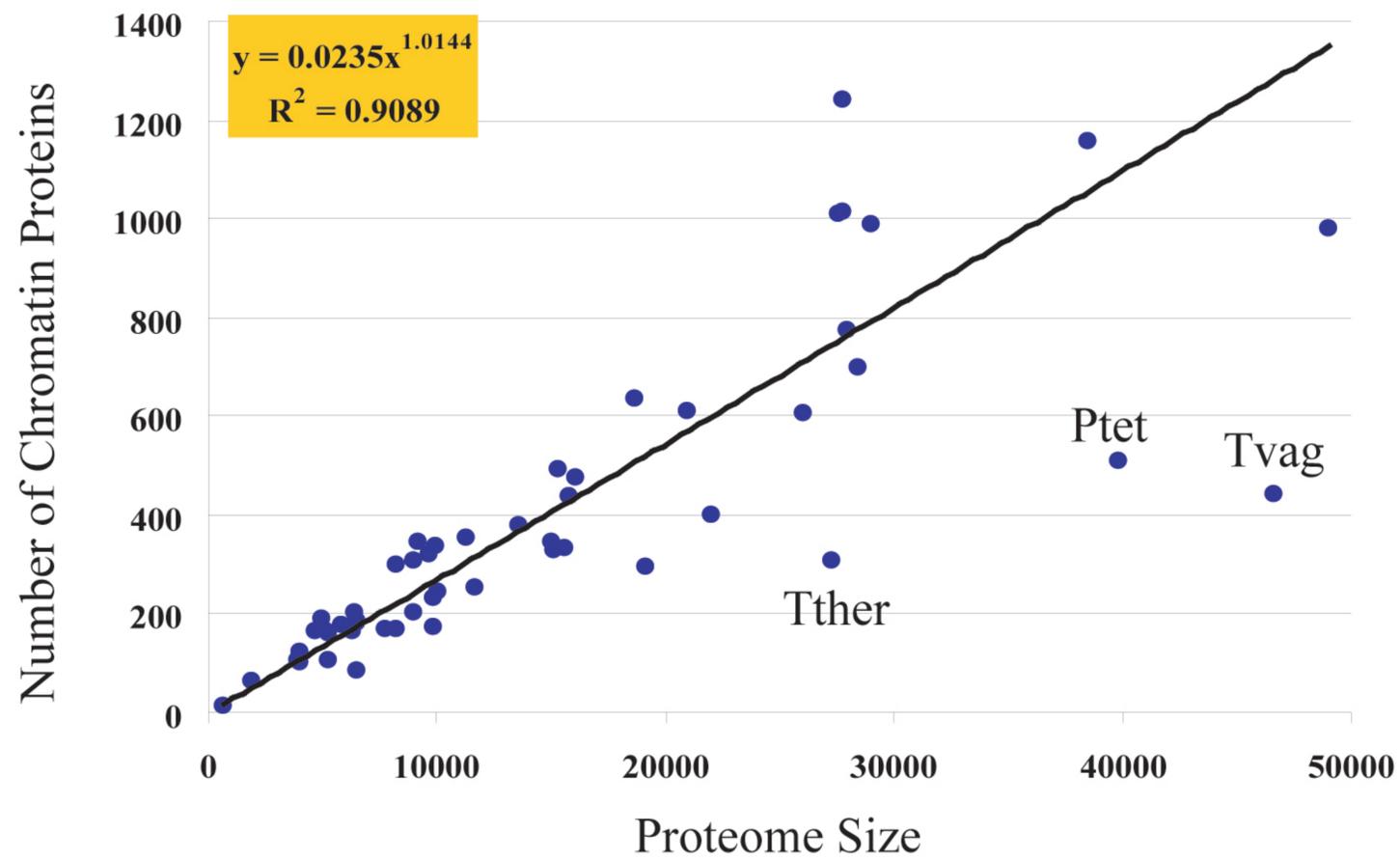
A



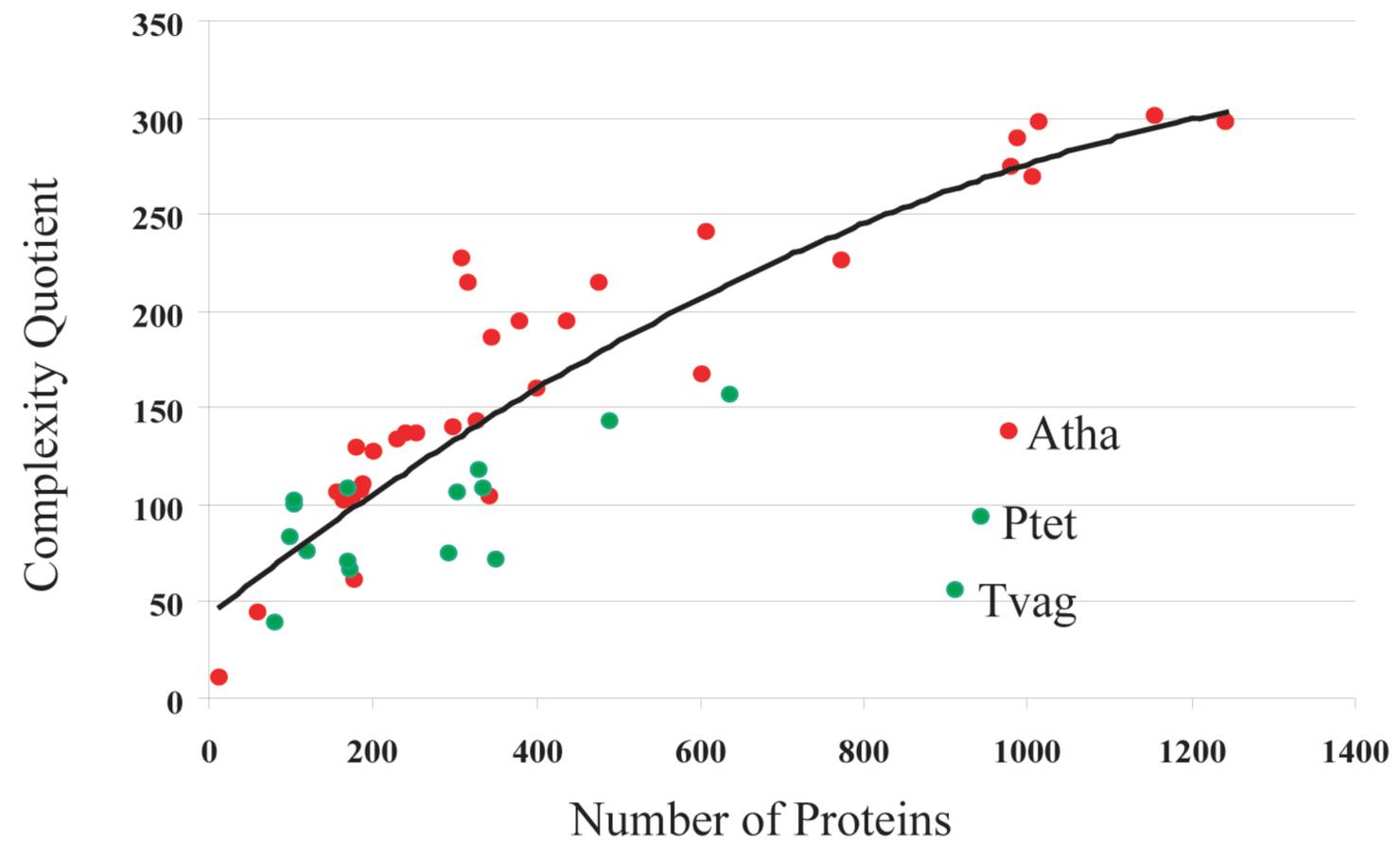
B

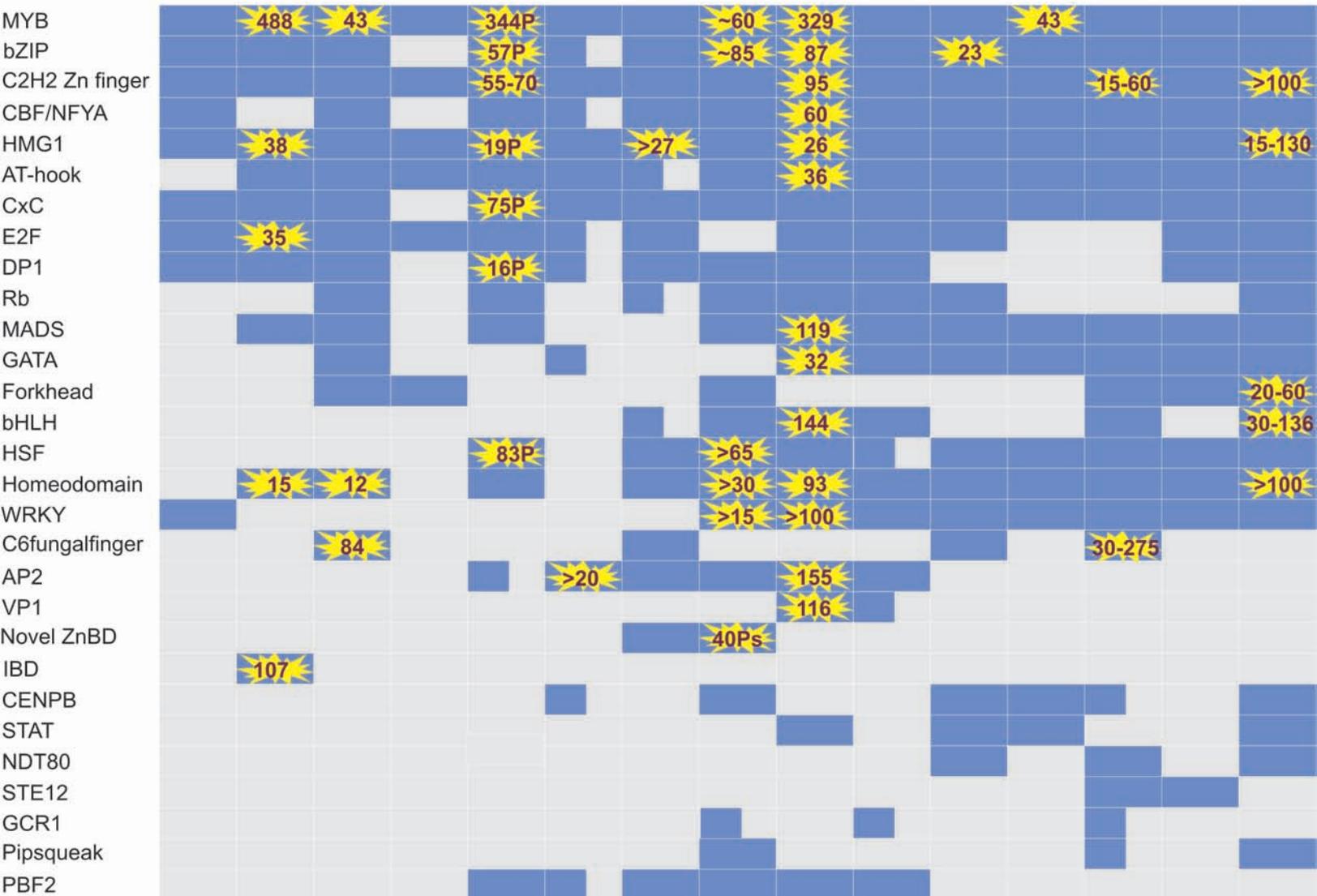
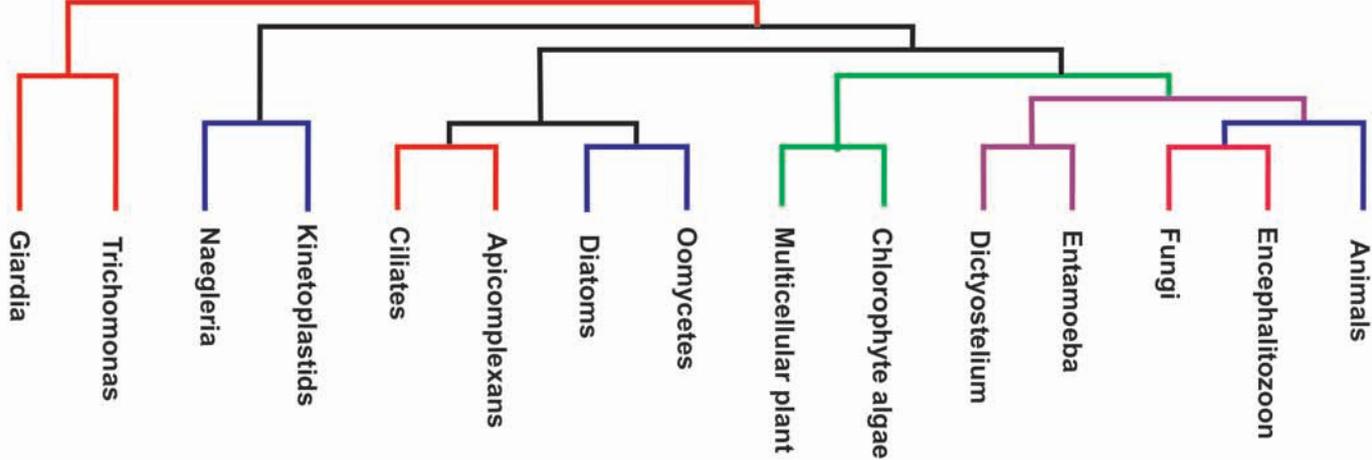


C

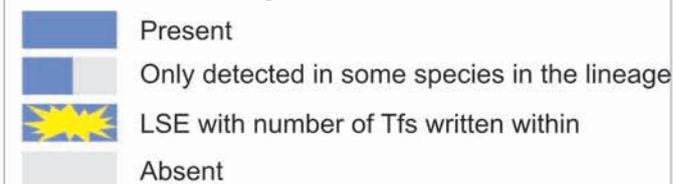


D



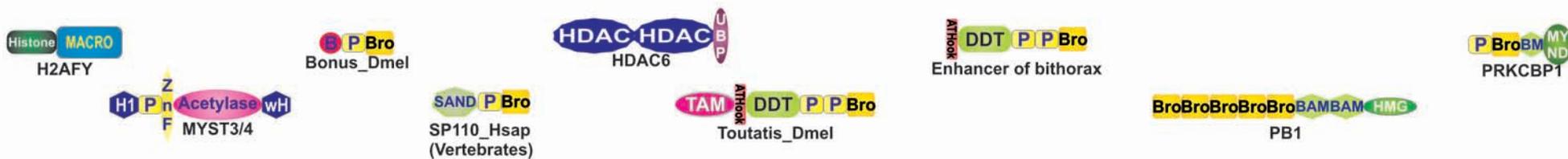


Key



Lineage-specific architectures

Animal



Fungi



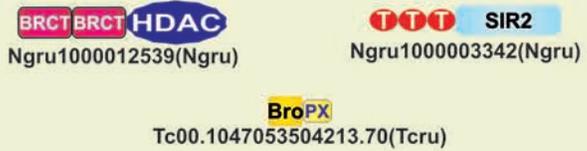
Plant



Amoebozoa



Kinetoplastid-Heterolobosea



Alveolate



Stramenopile



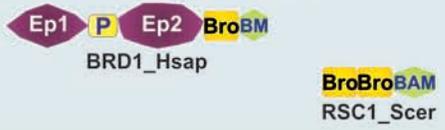
Basal



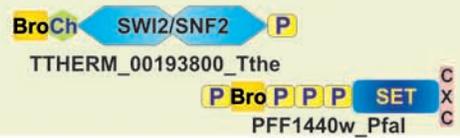
Plant >chrom-alveolate



Animal-fungi



Chromalveolates



Animal-fungi-amoebozoa



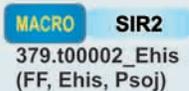
Ancient architectures

Divergence of crown group eukaryotes



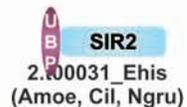
Acetylases: 7
Deacetylases: 7
Adaptors: 7

Divergence of chromalveolates and crown group eukaryotes



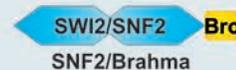
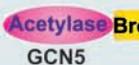
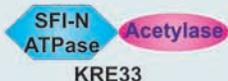
Acetylases: 6
Deacetylases: 7
Adaptors: 6

Divergence of kinetoplastids and heteroloboseans

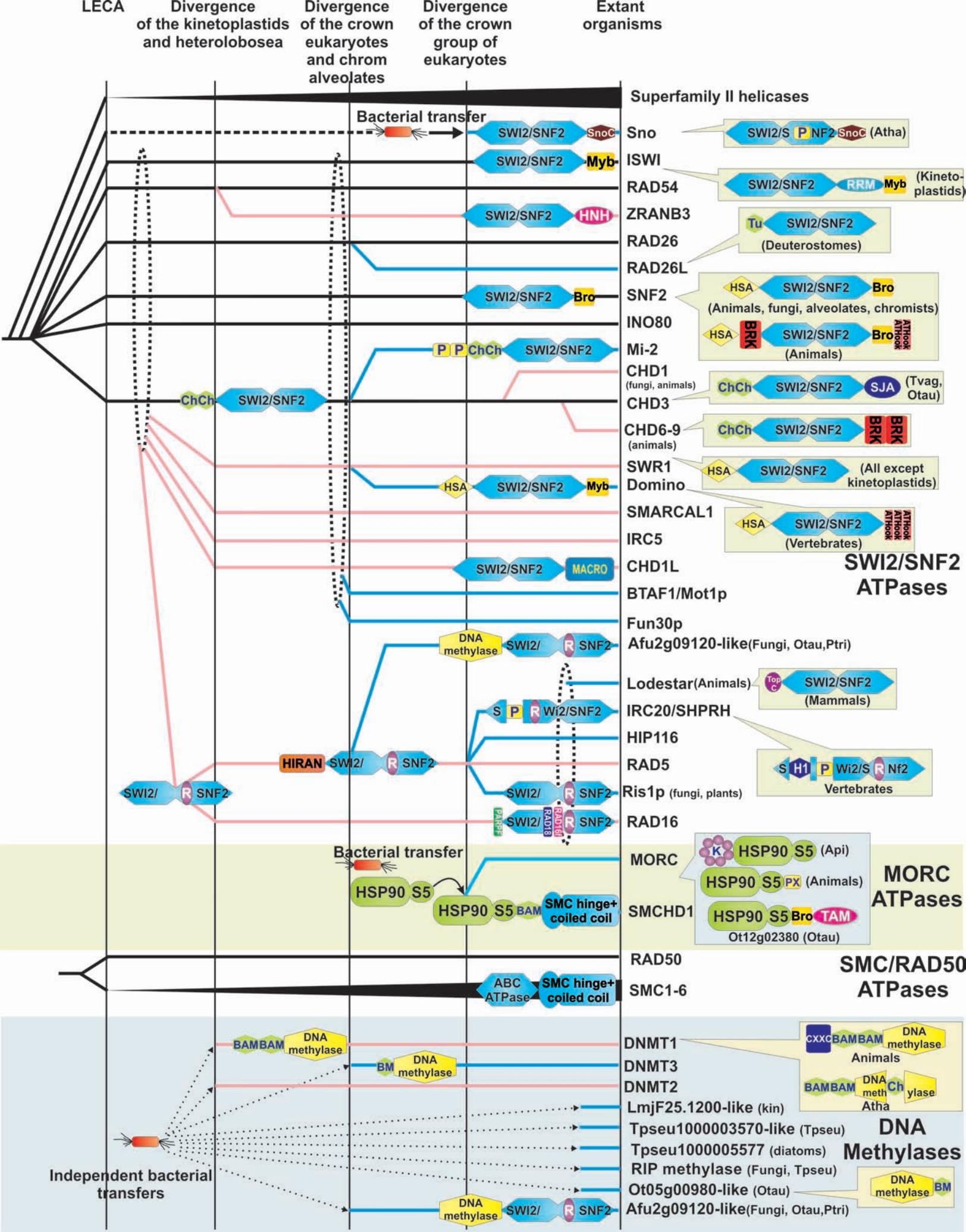


Acetylases: 6
Deacetylases: 7
Adaptors: 5

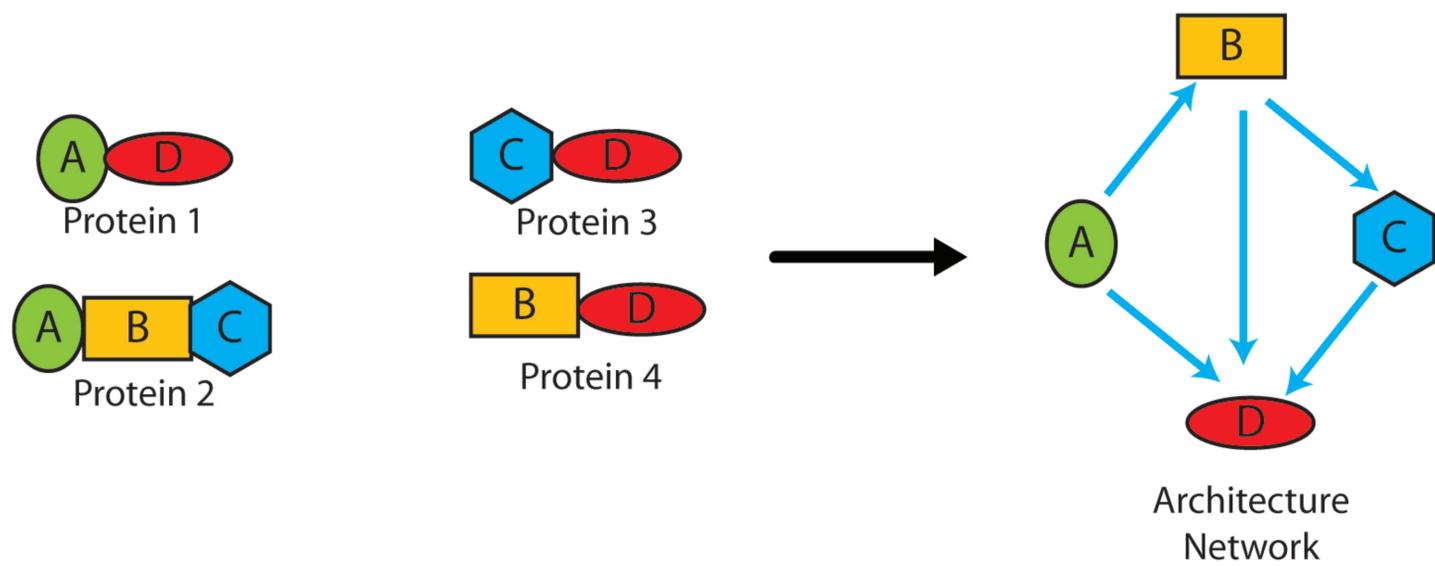
Last Eukaryotic Common ancestor



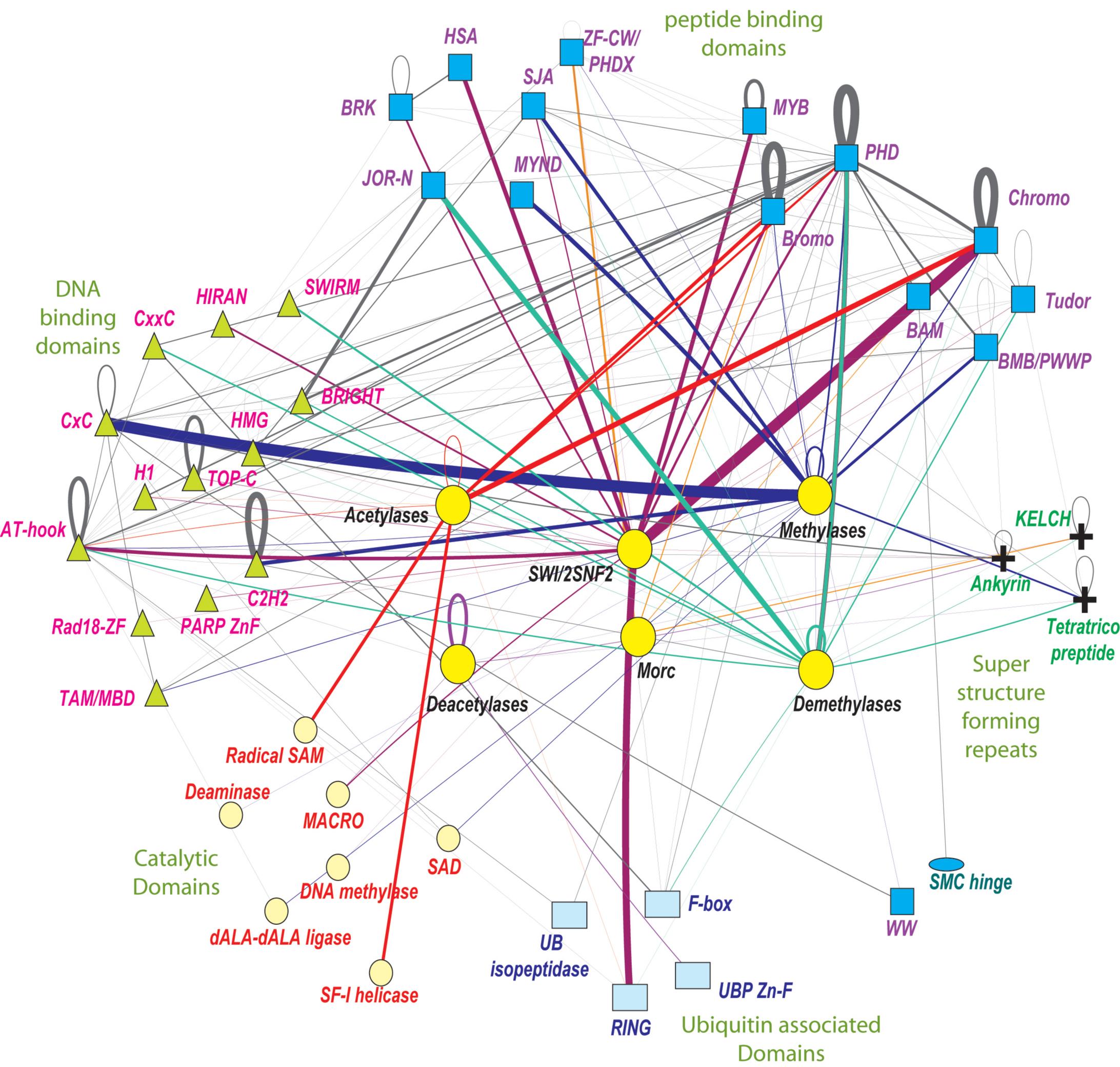
Acetylases: 4
Deacetylases: 3
Adaptors: 4



A



B



A Methylation/Demethylation

