Available online at www.sciencedirect.com
ScienceDirect
International Journal for Parasitology 38 (2008) 1–31
www.elsevier.com/locate/ijpara

Invited Review

Comparative genomics of transcription factors and chromatin proteins in parasitic protists and other eukaryotes

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

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

Received 18 April 2007; received in revised form 26 July 2007; accepted 30 July 2007

Abstract

Comparative genomics of parasitic protists and their free-living relatives are profoundly impacting our understanding of the regulatory systems involved in transcription and chromatin dynamics. While some parts of these systems are highly conserved, other parts are rapidly evolving, thereby providing the molecular basis for the variety in the regulatory adaptations of eukaryotes. The gross number of specific transcription factors and chromatin proteins are positively correlated with proteome size in eukaryotes. However, the individual types of specific transcription factors show an enormous variety across different eukaryotic lineages. The dominant families of specific transcription factors even differ between sister lineages, and have been shaped by gene loss and lineage-specific expansions. Recognition of this principle has helped in identifying the hitherto unknown, major specific transcription factors of several parasites, such as apicomplexans, *Entamoeba histolytica*, *Trichomonas vaginalis*, *Phytophthora* and ciliates. Comparative analysis of predicted chromatin proteins from protists allows reconstruction of the early evolutionary history of histone and DNA modification, nucleosome assembly and chromatin-remodeling systems. Many key catalytic, peptide-binding and DNA-binding domains in these systems ultimately had bacterial precursors, but were put together into distinctive regulatory complexes that are unique to the eukaryotes. In the case of histone methylases, histone demethylases and SWI2/SNF2 ATPases, proliferation of paralogous families followed by acquisition of novel domain architectures, seem to have played a major role in producing a diverse set of enzymes that create and respond to an epigenetic code of modified histones. The diversification of histone acetylases and DNA methylases appears to have proceeded via repeated emergence of new versions, most probably via transfers from bacteria to different eukaryotic lineages, again resulting in lineage-specific diversity in epigenetic signals. Even though the key histone modifications are universal to eukaryotes, domain architectures of proteins binding post-translationally modified-histones vary considerably across eukaryotes. This indicates that the histone code might be ''interpreted'' differently from model organisms in parasitic protists and their relatives. The complexity of domain architectures of chromatin proteins appears to have increased during eukaryotic evolution. Thus, *Trichomonas*, *Giardia*, *Naegleria* and kinetoplastids have relatively simple domain architectures, whereas apicomplexans and oomycetes have more complex architectures. RNA-dependent post-transcriptional silencing systems, which interact with chromatin-level regulatory systems, show considerable variability across parasitic protists, with complete loss in many apicomplexans and partial loss in *Trichomonas vaginalis*. This evolutionary synthesis offers a robust scaffold for future investigation of transcription and chromatin structure in parasitic protists.
Published by Elsevier Ltd on behalf of Australian Society for Parasitology Inc.

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

1. Introduction

The unique configuration of the eukaryotic transcription apparatus sets it apart from its counterparts in the archaeal and bacterial superkingdoms (Best et al., 2004; Conaway and Conaway, 2004; Latchman, 2005). On one hand, the basal or general transcription apparatus of eukaryotes and archaea share several unique features. These include: (i) structure of the RNA polymerase catalytic subunit (the three subunits equivalent to the bacterial β′, β and α subunits); (ii) specific accessory RNA polymerase subunits

* Corresponding author. Tel.: +1 301 594 2445; fax: +1 301 435 7793.
E-mail address: aravind@mail.nih.gov (L. Aravind).

0020-7519 Published by Elsevier Ltd on behalf of Australian Society for Parasitology Inc.
doi:10.1016/j.ijpara.2007.07.018

(e.g. RPB10); (iii) proteins constituting the basal transcription initiation apparatus (general or global transcription factors (TFs)), such as TATA box-binding protein (TBP), TFIIB, TFIIE and MBF (Reeve, 2003; Conaway and Conaway, 2004). On the other hand, certain components of the eukaryotic transcription elongation complex, such as the Spt6p-type RNA-binding proteins, are shared with bacteria rather than archaea (Anantharaman et al., 2002). Thus, the eukaryotic systems appear to have a chimeric pattern – the archaea-like elements contribute to the core transcription apparatus, including the bulk of the basal or general TFs, and the bacteria-like elements supply some additional factors of the basal transcription apparatus (Dacks and Doolittle, 2001; Reeve, 2003; Best et al., 2004; Conaway and Conaway, 2004; Aravind et al., 2005, 2006). Like the two prokaryotic superkingdoms, several eukaryotes possess specific TFs that are required for transcriptional regulation of particular sets of genes (Latchman, 2005). In both prokaryotic superkingdoms, the majority of specific TFs are members of a relatively small group of protein families containing the helix-turn-helix (HTH) DNA-binding domain (DBD) (Aravind et al., 2005; Pellegrini-Calace and Thornton, 2005). Several families of eukaryote-specific TFs, such as the homeodomain and Myb domain proteins, also bind DNA via the HTH domain (Aravind et al., 2005; Latchman, 2005). However, almost all eukaryotic HTH-containing specific TFs do not belong to any of the prokaryotic HTH families, and are only very distantly related to them in sequence (Aravind et al., 2005; Pellegrini-Calace and Thornton, 2005). Additionally, eukaryotes possess numerous large families of specific TFs containing an astonishing array of DBDs that span the entire spectrum of protein folds (Babu et al., 2004; Latchman, 2005). This deployment of specific TFs with an immense structural diversity of DBDs is a dramatic difference in the transcription apparatus of eukaryotes vis-à-vis the prokaryotic superkingdoms.

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

The unifying features of the transcription and chromatin dynamics apparatus across eukaryotic model organisms notwithstanding, several studies have hinted at an enormous lineage-specific diversity in the types of specific TFs and domain architectures of chromatin proteins (Koonin et al., 2000; Coulson et al., 2001; Lander et al., 2001; Lespinet et al., 2002; Sullivan et al., 2006). A potential corollary to this observation was that the variety in specific TFs and chromatin–protein architectures might provide the regulatory basis for the emergence of enormous bio-diversity in terms of structure, life-styles and life-cycles across the eukaryotic evolutionary tree (Coulson et al., 2001; Lander et al., 2001; Lespinet et al., 2002). Phylogenetic investigations have shown that model organisms represent only a small portion of the vast eukaryotic tree, with most of the bewildering diversity found in the unicellular microbial eukaryotes or 'protists' (Moon-van der Staay et al., 2001; Bapteste et al., 2002; Simpson et al., 2006). Several lineages of protists have spawned human, livestock and crop parasites with an extraordinary range of adaptations (Fig. 1). Hence, a proper understanding of the diversity of eukaryotic transcription regulation and chromatin dynamics will be critical in any future attempts to tackle parasitic diseases. A major boost for these studies has come from the recent large-scale genome sequencing efforts that have generated complete or near-complete genome sequences of several protists, which are either agents of major parasitic diseases or key players in world-wide ecosystems.

Traditional approaches to study protist parasitism have been greatly hampered by practical difficulties relating to their complex multi-host lifecycles, in vitro culturing and maintenance, as well as a lack of proper animal models in certain cases (Kreier, 1977). Hence, experimental analyses on protist regulatory systems, especially transcription and chromatin dynamics, are far from the levels that have been achieved in eukaryotic model organisms. However, recent successes of comparative genomics and its resonance

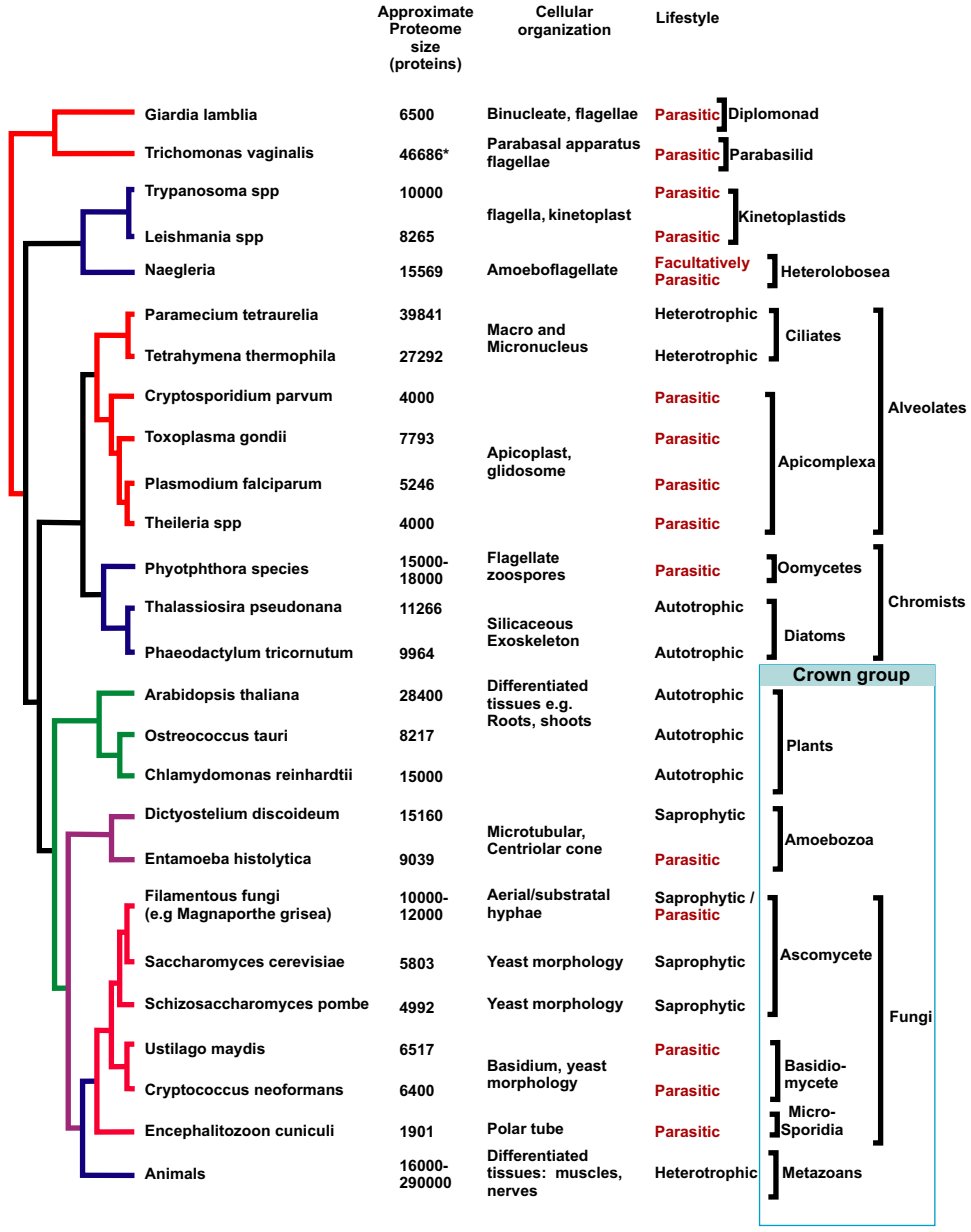

Fig. 1. Phylogenetic relationships, genome sequencing efforts and major distinguishing features of eukaryotes. The displayed tree is a maximum likelihood (ML) tree derived from a concatenated alignment of 82 universally conserved eukaryotic proteins spanning 19,603 positions. The among-site variation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

with new technologies are vastly improving the situation. In this article, we use the vast array of data from recently published protist genome sequences to present a comparative genomic overview of chief aspects of the transcription regulatory and chromatin remodeling apparatus in eukaryotes. Thus, this survey seeks to provide the larger framework and appropriate evolutionary context within which the biochemistry of transcription and chromatin can be explored in parasitic protists. In must be emphasized that the objective of this work is not to review, in the conventional sense, the literature on these regulatory processes in protists, but to synthesize the data from genomics to provide a base for future experimental forays on these protists.

2. Eukaryotic phylogeny and genomics

2.1. Repeated evolution of parasitism in protists

Despite availability of genome-scale data, reconstruction of eukaryotic phylogeny has not been straight-forward (Bapteste et al., 2002; Templeton et al., 2004; Arisue et al., 2005; Walsh and Doolittle, 2005; Simpson et al., 2006). Some principal problems that confound determination of higher order relationships amongst eukaryotes are: (i) Rampant gene loss is common throughout the fungal kingdom and especially pronounced in the microsporidian lineage (Aravind et al., 2000; Katinka et al., 2001). *Entamoeba* amongst amoebozoans, *Cryptosporidium* amongst apicomplexans and *Giardia* amongst basal eukaryotes also display extreme gene loss relative to their sister lineages (Templeton et al., 2004; Loftus et al., 2005; Carlton et al., 2007). (ii) Gene loss also spurs concomitant rapid sequence divergence of the proteins that have been retained on account of release from selective constraints due to lost interacting partners (Aravind et al., 2000). (iii) Lateral gene transfer occurs in some eukaryotic lineages like chromists (stramenopiles) and apicomplexans which have emerged via secondary or tertiary endosymbiosis involving engulfment of other eukaryotic cells from the plant lineage (Bhattacharya et al., 2004). As a result their proteins show chimeric affinities to either those of the original lineage or to those of the endosymbiont's lineage. In addition to these issues, there are controversies concerning the rooting of the eukaryotic tree and the nature of the last eukaryotic common ancestor (LECA) (Arisue et al., 2005; Walsh and Doolittle, 2005). Nevertheless, multiple independent recent studies using large multi-protein datasets and algorithms to correct for differential evolutionary rates have been robustly reproducing several higher order groupings (Fig. 1) (Bapteste et al., 2002; Templeton et al., 2004; Walsh and Doolittle, 2005; Simpson et al., 2006).

Animals and fungi are observed to form a monophyletic lineage, with amoebozoans as their immediate sister group. The plant lineage forms the sister group to animals, fungi and amoebozoans, and together this assembly is referred to here as the crown group (Fig. 1). Both unicellular (protist) as well as multicellular forms spanning an entire range of morphologies are seen in each of the crown-group lineages. Likewise, parasitism has repeatedly emerged in crown-group lineages (Fig. 1). The fungal lineage in particular has spawned several parasites, including the human parasite *Cryptococcus* and plant parasites such as *Ustilago*. The most unusual of these are the structurally highly-derived microsporidians, which possess some of the most reduced of eukaryotic genomes (Katinka et al., 2001). Recent analyses suggest that they might be derived from within chytrids, the basal-most lineage of fungi (James et al., 2006). The animal lineage has also given rise to microbial parasites, namely the enigmatic myxozoa, which were previously classified with microsporidians (Smothers et al., 1994). Amongst amoebozoans the best-studied parasite is the human gut parasite *Entamoeba histolytica* (Loftus et al., 2005). Even in the predominantly auxotrophic plant lineage microbial parasites have emerged amongst rhodophytes, which deliver their nucleus into host cells belonging to other rhodophyte species (Goff and Coleman, 1995).

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

2.2. Key eukaryotic features revealed by comparative genomics

Burgeoning genome sequencing projects have generated complete sequences of major representatives of most of the above-discussed eukaryotic lineages (Fig. 1). Results of comparative genomics have brought home certain large-scale trends in eukaryotic evolution. First and foremost, they have revealed the enormous plasticity of eukaryotic genomes and rampant reorganization by lineage-specific expansions (LSE) of genes and gene loss (Aravind et al., 2000; Katinka et al., 2001; Lespinet et al., 2002). Massive gene loss relative to free-living forms is a prevalent feature of most parasitic lineages. One exception is the basal eukaryote *Trichomonas*, which possesses gene numbers comparable to or greater than animals, plants and ciliates (Carlton et al., 2007). The most parsimonious reconstruction considering the above phylogenetic scenario suggests that the LECA already possessed a distinctly larger gene

complement (at least ~10,000 genes) than its prokaryotic precursors. This complement coded numerous families of proteins with multiple paralogous members and several novel regulatory systems with no direct prokaryotic equivalents (Aravind et al., 2006; Anantharaman et al., 2007).

The availability of complete genome sequences also allows us to estimate the gross differences in effects of natural selection on completely conserved orthologous proteins belonging to different functional categories (Bapteste et al., 2002). Examination of residues evolving at different rates in individual functional classes reveals certain interesting features (Fig. 2a). The machinery related to protein stability, namely chaperones and proteasomal subunits, comprise one of the most conserved groups of eukaryotic

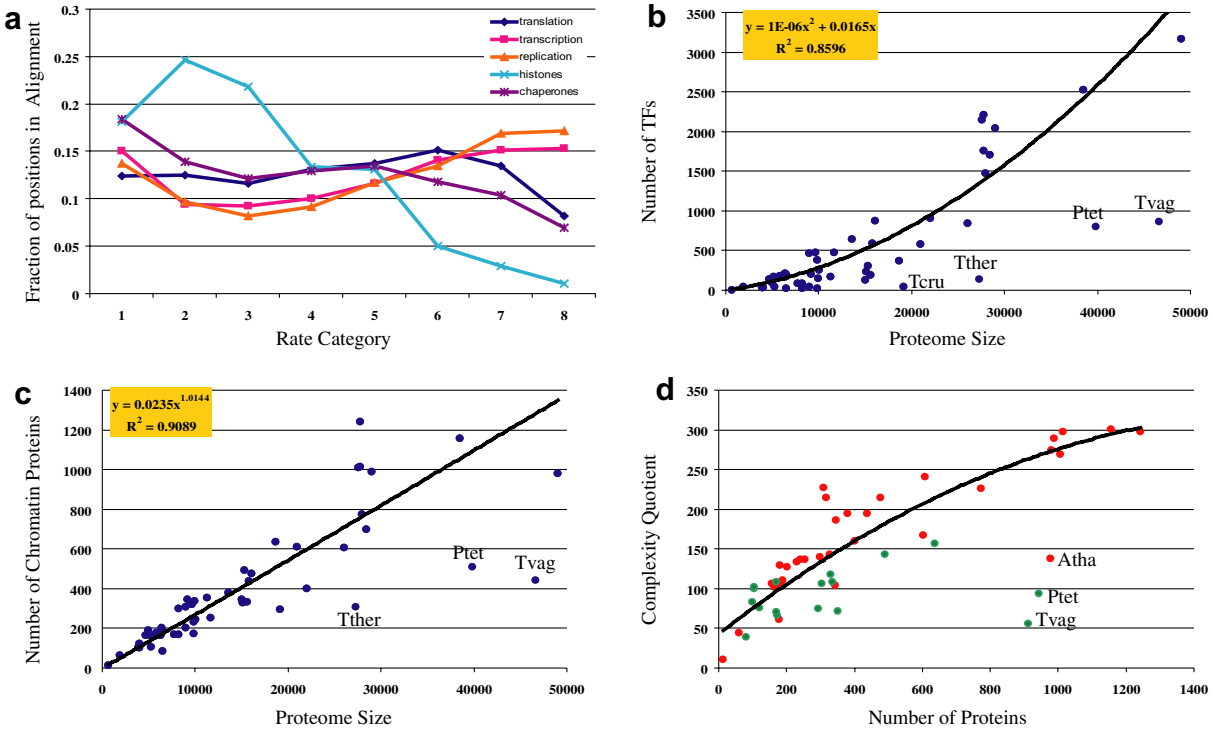

Fig. 2. Differences in rate categories in different functional classes, scaling of TFs and CPs and complexity quotient plots. (a) Among-site rate variation for different functional classes of eukaryotic proteins. These were calculated using multiple alignments of highly conserved proteins that are present in all eukaryotes in each functional category shown in the graph. The total numbers of positions in each category were- translation: 6357; transcription: 2275; replication: 5436; histones: 381; chaperones: 5154. The among-site rate variations for each functional class were calculated as described in Fig. 1 but in this case a Whelan and Goldman (WAG) substitution matrix was used as it confers higher likelihood on the data. The fraction of the positions in each rate category is plotted for each functional class – the categories on the left evolve slower than those on the right. Note that the distribution for transcription and replication proteins is U-shaped, indicating an over-representation of extremes – slowest-evolving and fastest-evolving positions. (b) Scaling of transcription factors with proteome size. The names of organisms used for the plot and their abbreviations are indicated below. Organisms with a significantly lower-than-expected fraction of chromatin proteins are labeled. (c) Scaling of chromatin proteins with proteome size. The organisms are the same as in (a). Organisms with a lower-than-expected fraction of chromatin proteins are marked. (d) Complexity quotient plot for chromatin proteins. The "complexity quotient" for an organism is defined as the product of two values: the number of different types of domains which co-occurs in signaling proteins, and the average number of domains detected in these proteins. The complexity quotient is plotted against the total number of chromatin proteins in a given organism. A polynomial curve fitting the general trend of the majority of organisms is shown. Crown group members are shown in red and the non-crown group members are in green. Some organisms with much lower complexity than those along the general trend are marked. Each protein has at least a single known or predicted domain with a chromatin/transcription-related function. A total of 363 domains were considered, among which 121 were domains specifically found in chromatin and transcription factors, and the rest were other domains with wider distributions encompassing other functional systems. The organisms included in all these plots are the following: Crown group: *Aspergillus fumigatus* – Afum, *Candida glabrata* – Cgla, *Debaryomyces hansenii* – Dhan, *Ashbya gossypii* – Egos, *Gibberella zeae* – Gzea, *Kluyveromyces lactis* – Klac, *Neurospora crassa* – Ncra, *Saccharomyces cerevisiae* – Scer, *Schizosaccharomyces pombe* – Spom, *Yarrowia lipolytica* – Ylip, *Cryptococcus neoformans* – Cneo, *Ustilago maydis* – Umay, *Encephalitozoon cuniculi* – Ecun, *Anopheles gambiae* – Agam, *Apis mellifera* – Amel, *Branchiostoma floridae* – Bflo, *Caenorhabditis elegans* – Cele, *Ciona intestinalis* – Cint, *Danio rerio* – Drer, *Drosophila melanogaster* – Dmel, *Homo sapiens* – Hsap, *Mus musculus* – Mmus, *Pan troglodytes* – Ptro, *Rattus norvegicus* – Rnor, *Strongylocentrotus purpuratus* – Spur, *Tetraodon nigroviridis* – Tnig, *Tribolium castaneum* – Tcas, *Monosiga brevicollis* – Mbre, *Nematostella vectensis* – Nvec, *Entamoeba histolytica* – Ehis, *Dictyostelium discoideum* – Ddis, *Chlamydomonas reinhardtii* – Crei, *Ostreococcus tauri* – Otau, *Arabidopsis thaliana* – Atha, *Phaeodactylum tricornutum* – Ptri, *Phytophthora sojae* – Psoj, *Phytophthora ramorum* – Pram, *Thalassiosira pseudonana* – Tpse, *Tetrahymena thermophila* – Tthe, *Paramecium tetraurelia* – Ptet, *Toxoplasma gondii* – Tgon, *Theileria parva* – Tpar, *Theileria annulata* – Tann, *Cryptosporidium parvum* – Cpar, *Plasmodium falciparum* – Pfal, *Trypanosoma cruzi* – Tcru, *Trypanosoma brucei* – Tbru, *Leishmania major* - Lmaj, *Naegleria gruberi* – Ngru, *Giardia lamblia* – Glam, *Trichomonas vaginalis* – Tvag, *Guillardia theta* – Gthe. The genomes were obtained from the NCBI Genbank (http://www.ncbi.nlm.nih.gov/genomes/static/euk_g.html and the NR database). The *T. gondii* sequence was the current release from Toxodb (www.toxodb.org), while the Stramenopile, *C. intestinalis*, *C. reinhardtii*, *M. brevicollis*, *N. vectensis*, *N. gruberi*, *Phytophthora* and *Thalassiosira* genomes were obtained from Department of Energy's Joint Genome Institute (http://www.jgi.doe.gov/).

proteins with the majority of their residues evolving slowly. In contrast nuclear proteins, especially those related to transcription and chromatin structure and dynamics, display a bimodality of evolutionary rates – a subset of the residues belong to the most slowly evolving category amongst all eukaryotic proteins, whereas another subset is rapidly evolving. Specifically, all core histones which comprise the nucleosomal octamer and parts of the RNA-polymerase catalytic subunits belong to the most slowly evolving categories (Fig. 2a). However, there are other parts of the same RNA-polymerase subunits that exhibit amongst the most rapid evolutionary rates of all the universally conserved orthologous proteins. A similar pattern of apparently bimodal evolutionary rates is observed amongst proteins comprising the replication apparatus. These observations suggest that while a subset or parts of chromosomal proteins have settled into highly conserved roles since the beginning of eukaryotic evolution, the remainder or remaining parts are rapidly diverging, indicating lineage-specific adaptations in these proteins (Fig. 2a).

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

Generation of sensitive sequence profiles and hidden Markov models for conserved domains found in TFs (typically their DBD) and chromatin proteins (CPs) allows their exhaustive and systematic detection across all complete eukaryotic proteomes (Coulson et al., 2001; Babu et al., 2004; Finn et al., 2006) (see Supplementary material file 1 for details on methods). As a result, reasonably robust counts or demography of potential TFs and CPs encoded by a given organism can be obtained. These results show positive correlations between the number of CPs or TFs coded by an organism and its proteome size (Fig. 2b and c; Supplementary material files 2 and 3). These trends are best approximated by linear or mildly non-linear fits (weak quadratic fit for TFs or weak power-law in chromatin factors) suggesting that, in general, there is a proportional increase in the number of TFs for an increasing number of protein-coding genes. The trend observed in TFs is in contrast to that seen in prokaryotes wherein a fit to a much stronger power-law trend is observed (Babu et al., 2004; Aravind et al., 2005). However, in prokaryotes there appear to be very few dedicated CPs, and their number does not vary dramatically with proteome size. This suggests that eukaryotes might optimize their transcription regulatory potential by increasing numbers of both TFs and chromosomal proteins as their gene numbers increase. As a result the scaling behavior of their TF counts is different from prokaryotes.

Parasites belonging to fungal, apicomplexan and stramenopile lineages show greater or lesser degrees of gene loss in comparisons with their free-living sister clades, but typically counts of their TFs and CPs do not deviate to a large extent from the general trend observed across eukaryotes. Hence, despite a degree of genomic reduction, the overall regulatory input per protein-coding gene in these parasites is roughly comparable with other eukaryotes. Significant exceptions to the general eukaryotic trend in TFs were seen in trypanosomes, while *T. vaginalis* and ciliates displayed significant deviations in counts of both their TFs and CPs (Fig. 2b and c). The notably lower TF count in trypanosomes relative to their proteome size might imply that they possess a unique family of TFs that are unrelated to any previously characterized variety and have eluded detection thus far. In *T. vaginalis* and ciliates the absolute counts of TFs and CPs exceed those seen in other parasites or free-living protists. However, their proteome size is similar to that of multicellular animals and plants, and as result they have relatively fewer TFs and CPs for their proteome sizes compared with the multicellular forms (Fig. 2b and c). This might be due to different parallel causes: (i) Multicellular forms show both temporal transcriptional changes during development and spatially differentiated cell-types with diverse gene-expression states. In contrast, a parasite like *T. vaginalis* shows relatively simple temporal development and has no equivalent of differentiated cell fates. Likewise, though ciliates have amongst the most complex cell-architectures seen in eukaryotes, they possess a relatively simple development and no differentiated cell-types. Consequently, lower normalized counts of TFs in these organisms might reflect differences in the amount of transcriptional control required to regulate similarly sized genomes in the unicellular context (*T. vaginalis* or ciliates) as opposed to multicellular forms with differentiation. (ii) These protists also show tremendous genetic redundancy with several closely related or near-identical gene copies that, rather than being differentially regulated, might merely provide higher effective concentrations of particular gene products (Aury et al., 2006; Carlton et al., 2007). The gene counts, especially in *T. vaginalis*, are also exaggerated by numerous transposable elements of diverse types (Carlton et al., 2007).

3. Diversity of eukaryotic-specific transcription factors

3.1. Identification of novel-specific transcription factors in protist lineages

Eukaryotes are distinguished by the extreme diversity of their specific TFs, both in terms of superfamilies of the DBDs they contain and the lineage-specific differences in their distributions (Coulson et al., 2001; Lespinet et al., 2002; Babu et al., 2004). Thus, the most utilized TFs differ widely across major eukaryotic lineages: for example, in multicellular plants TFs with the MADS, VP1 and Apetala2 (AP2) DBDs are most prevalent, whereas in animals TFs containing homeodomains and C_2H_2 Zn fingers are dominant, and in fungi the C6-binuclear Zn fingers are dominant (Fig. 3). Until recently, no examples of the C6-binuclear finger were found outside the fungi, suggesting that some DBDs of these TFs can have extremely restricted

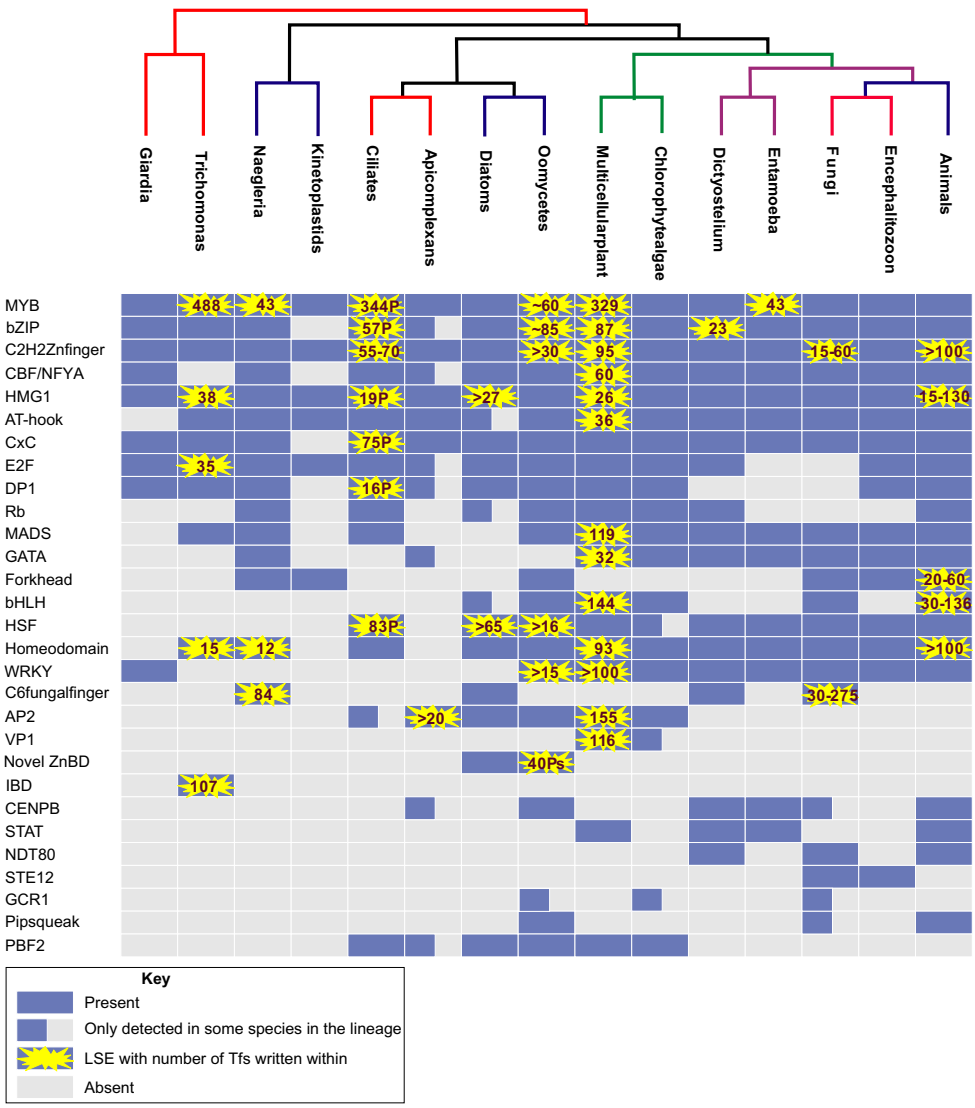

Fig. 3. Lineage-specific expansions and phyletic distributions of specific transcription factors (TFs). Only those specific TFs that are present in protists and have lineage-specific expansions (LSEs) or notable sporadic phyletic patterns are shown. The distribution of the TFs across eukaryotic species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

phyletic patterns (Babu et al., 2004). It is notable that this lineage-specific diversity of specific TFs exists, despite a fairly strong global trend in TF demography across eukaryotes (Fig. 2b). This implies a general constraint in terms of the number of TFs required to regulate a proteome of a given size, even though there appears to be no major constraint on the actual type of TF being deployed (i.e. their evolutionary origin). A corollary is that different superfamilies of TFs have independently expanded in each major lineage to convergently produce overall counts corresponding to that dictated by the general constraint (Figs. 2b and 3).

On the practical side, this feature of eukaryotic TFs often makes their prediction in poorly-studied lineages, especially parasites, a difficult task. This was poignantly illustrated by the apicomplexans, where multiple studies had initially failed to recover bonafide-specific TFs (Gardner et al., 2002; Templeton et al., 2004). However, analysis of stage-specific gene expression in *Plasmodium falciparum* revealed a complex pattern of changing gene expression that resulted in genes with increasing functional specialization being expressed as the intra-erythrocytic development cycle (IDC) progressed (Bozdech et al., 2003; Le Roch et al., 2003). This was also supported by expression studies in *Theileria* (Bishop et al., 2005) and pointed to a specialized transcription regulatory program similar to that seen in model organisms from the crown group. Sensitive sequence profile analysis revealed a major lineage-specific expanded family of proteins (ApiAP2 family) with one or more copies of the AP2 DBD, similar to those found in plant AP2 TFs, to be present in all studied apicomplexan clades from *Cryptosporidium* to *Plasmodium* (Balaji et al.,

2005). Further analysis of expression of the ApiAP2 genes in the course of the IDC showed that they clustered into specific co-expression guilds that notably corresponded to the major development stages namely the ring, trophozoite, early schizont and schizogony/merozoite. Analysis of physical interactions of ApiAP2 proteins based on recently published large-scale protein interaction data (LaCount et al., 2005) revealed homo- and hetero-dimeric interaction with other ApiAP2 proteins, as well as interaction with various CPs such as the GCN5 histone acetyltransferase, CHD1 and Rad5/16-type SWI2/SNF2 ATPases and the HMG1 ortholog (MAL8P1.72). These observations suggested that the ApiAP2 proteins are indeed the predominant specific TFs of apicomplexans, and are likely to function similar to their counterparts from crown-group model organisms by recruiting histone-modifying and chromatin remodeling factors to their target sites. The types of factors recruited by them are suggestive of being involved in both transcription activation (e.g. GCN5) and repression (e.g. CHD1) (Allis et al., 2006). Studies on altered gene expression patterns in response to febrile temperatures in *P. falciparum* revealed that in addition to the ApiAp2 proteins a small set of specific TFs with other types of DBDs might play important regulatory roles in apicomplexans (Oakley et al., 2007). They include a C2H2 Zn finger protein (PFL0455c) and a plant PBF2/TIF1 ortholog (PFE1025c) which, as in ciliates, might regulate expression of rRNA (Saha et al., 2001).

This discovery of the dominant specific TFs of apicomplexans serves as a model for the identification of uncharacterized TFs in other protists, such as *T. vaginalis*. Transcription initiation in this organism is primarily dependent on the protein IBP39, which binds the initiator element (Inr) by means of a specialized winged HTH (wHTH) domain, termed the IBD, and recruits the RNA polymerase via its C-terminal tail (Schumacher et al., 2003; Lau et al., 2006). The recognition helix of the wHTH binds the major grove of DNA, while a distinctive positively-charged loop from a bi-helical hairpin at the N-terminal contacts the adjacent minor grove. This novel DBD, while containing an ancient protein fold, has no close relatives in any organism studied to date (Schumacher et al., 2003; Lau et al., 2006). Given the generally low ratios of specific TFs to proteome size in *T. vaginalis* and the elusive origins of the IBD of IBP39, we investigated it using sequence profile searches to determine if it might define a novel family of lineage-specific TFs. As a result we were able to identify a family of at least 100 proteins in the *T. vaginalis* proteome, containing single IBDs and congruent architectures as IBP39 (see Supplementary material file 2). This suggests that the IBD indeed defines a lineage-specific DBD that is utilized by a large family of specific TFs in this organism. Sequence divergence in the recognition helix as well as the N-terminal positively charged loop across the IBD family suggests that different versions of the domain have potentially specialized to contact a range of target sites, other than the *T. vaginalis* Inr.

3.2. Major trends in the evolution of TFs

A survey of DBDs found in specific TFs shows that there are about 55 distinct superfamilies spanning all structural classes, with some of those present in almost all eukaryotes studied to date (Fig. 3). This latter group contains at least seven distinct DBDs, namely the Basic-zipper (bZIP), C2H2 ZnF, HMG box, AT-hook, MYB, CBF/NFYA and E2F/DP1 DBDs. These, along with DBDs of general TFs such as TBP, TFIIB, TFIIE and MBF which are shared with archaea, and the BRIGHT/ARID which emerged in eukaryotes, comprise the set of DBDs in TFs that can be confidently traced to the LECA (Best et al., 2004; Aravind et al., 2005). While the majority of DBDs in the ancient set shared with archaea contain the HTH fold, amongst the early eukaryotic innovations only the BRIGHT and MYB domains possess this fold (Aravind et al., 2005). This suggests that recruitment of a structurally diverse set of DBDs in TFs had already begun early in eukaryotic evolution. The wide distribution of specific TFs with several other DBDs, such as the MADS, GATA and Forkhead (FKH) domains, in early-branching eukaryotes also suggests a relatively ancient origin for these proteins in eukaryotic evolution (Fig. 3). Another major round of innovation of TFs, with new DBDs such as the CENPB, HSF and bHLH domains, appears to have happened prior to divergence of the crown group and the chromalveolate clade. Finally, there were extensive innovations of several other DBDs within the crown group, for example DBDs of the fast-evolving p53-like fold. The earliest representatives of this fold were present in the ancestor of the crown group and typified by the DBD of the STAT proteins (Fig. 3) (Soler-Lopez et al., 2004). We identified TFs of the STAT family in *E. histolytica* (Fig. 3, e.g. *E. histolytica* 83.t00003), where they could potentially function downstream of receptor kinases in processes related to this organism's pathogenesis. The p53-like fold subsequently appears to have diversified greatly in animals and fungi giving rise to four distinct families, including the animal p53 proper. Finally, there are some TFs that appear to be found in a single lineage of eukaryotes; striking examples being the above-mentioned IBDs of *T. vaginalis*, the APSES family of fungi and a previously uncharacterized family of predicted Zn-chelating TFs (often also containing additional AT-hook motifs (Aravind and Landsman, 1998)) found in the plant parasite *Phytophthora* (Fig. 3; Supplementary material file 3).

Irrespective of their point of origin, individual eukaryote-specific TFs show highly variable demographic patterns (Babu et al., 2004). For example, the AP2 domain has been independently expanded in both multicellular plants and apicomplexa but is present in very low numbers in its respective immediate sister groups namely, the chlorophyte algae (*Chlamydomonas* and *Ostreococcus*) and ciliates (Balaji et al., 2005). Likewise, the MYB domain shows enormous LSEs in multicellular plants, the free-living ciliate *Paramecium*, and phylogenetically distant para-

sites such as *T. vaginalis*, *E. histolytica* and *Naegleria*. The expanded MYB proteins are predicted to constitute the predominant-specific TFs in *E. histolytica* (Fig. 3). Other examples of major independent LSEs of TFs observed both in diverse parasites and free-living protist groups include the bZIP domain in *Phytophthora* and *Paramecium*, and the heat-shock factor (HSF) in most stramenopiles and *Paramecium*. While the C2H2 Zn-finger (ZnF) is prevalent in most eukaryotic lineages, its rise in each lineage appears to be a result of independent LSEs (Fig. 3) (Coulson et al., 2001; Lespinet et al., 2002; Babu et al., 2004, 2006). For example, a LSE comprised of proteins combining the C2H2-ZnF with AT-hooks appears to constitute the dominant TFs in ciliates such as *Tetrahymena* (Fig. 3). Interestingly, ciliates (especially *Paramecium*) show an expansion of the DNA-binding CXC domain that is normally found as a general DBD in chromosomal proteins rather than specific TFs (Hauser et al., 2000) (Fig. 3). Its unusual expansion and presence in standalone form, unlike chromosomal proteins where it is fused to other domains, suggest that these proteins possibly functions as specific TFs in ciliates.

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

This extensive lineage-specific diversification seen in eukaryotic TFs might be a major determinant that shapes the adaptations of protists. This leads to the question regarding the ultimate origin of eukaryotic TFs. Several families, such as the BRIGHT, homeo, POU, paired, HSF, IBD, MYB, TEA, FKH and pipsqueak domains contain the HTH fold, albeit only distantly related to that seen in prokaryotic TFs (Aravind et al., 2005). Hence, they could have potentially emerged through rapid diversification of older HTH domains inherited from prokaryotes (Aravind et al., 2005). Likewise, certain other ancient folds such as the C2H2 Znf and the immunoglobulin folds are found in the DBDs of eukaryotic TFs (Babu et al., 2004). These DBDs might also have been derived from more ancient representatives of their respective folds. Finally, as in the case of many other functional classes, eukaryotes have innovated TFs with DBDs containing entirely new folds. These are almost entirely α-helical or metal-chelation supported structures, consistent with the greater "ease" with which such structures are innovated de novo (Aravind et al., 2006). In more immediate evolutionary terms, several specific TFs appear to have been derived from DBDs of transposases and allied mobile elements. Examples of major eukaryotic DBDs that appear to have had such an origin are the WRKY, AP2, PBF2, VP1, paired, pipsqueak, CENPBP, APSES, BED-finger and GCR1 domains (Smit and Riggs, 1996; Balaji et al., 2005; Babu et al., 2006). Typically, inactive mobile elements that have lost the catalytic activity of their transposase domain but retain their DBD appear to be "re-cycled" as new TFs (Smit and Riggs, 1996; Babu et al., 2006).

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

4.1. Definition and detection of chromatin protein domains

It is impossible to precisely compartmentalize the disparate regulatory complexes in chromatin from the complexes responsible for essential housekeeping processes such as replication, recombination, DNA-repair and transcription. Nevertheless, herein we adopt a restricted definition for CPs by focusing chiefly on "regulatory" components. These regulatory components chiefly include enzymes catalyzing histone modifications that comprise an "extragenetic" code termed the histone code (Dutnall, 2003; Peterson and Laniel, 2004; Allis et al., 2006; Villar-Garea and Imhof, 2006; Kouzarides, 2007). These enzymes typically function in conjunction with energy-driven chromatin-remodeling enzymes. The "reading" of this histone code and recognition of covalently modified bases in DNA is mediated by another important class of regulatory proteins that bind unmodified or various covalently

modified histone side chains (de la Cruz et al., 2005; Allis et al., 2006; Kim et al., 2006; Sullivan et al., 2006; Villar-Garea and Imhof, 2006; Kouzarides, 2007). The distinctness of this set of proteins being defined here as CPs is primarily supported by the observation that they are mostly comprised of a relatively small set of conserved protein domains (about 70–80), the majority of which are found nearly exclusively in eukaryotic CPs (Letunic et al., 2006) (Table 1). This allows for relatively robust prediction of the complement of CPs through computational analysis using sensitive sequence profile methods and HMMs (Finn et al., 2006) (Supplementary material file 1). Most of these domains can be classified under two broad biochemical categories: (i) non-catalytic interaction or adaptor domains and (ii) enzymatic regulatory domains. The former category can again be further sub-divided into DNA-binding and protein–protein interaction domains (see Table 1 for summary). We first briefly discuss the DBDs and then consider the remaining domains in the course of reconstructing the natural history of the major regulatory systems in chromatin.

4.2. DNA-binding domains in chromatin proteins

The most basic DNA–protein interaction in eukaryotic chromatin is mediated by the four core histones that are universally conserved in all eukaryotes (Allis et al., 2006; Woodcock, 2006). In addition to the core histones there are other homologous histone-fold proteins, namely the smaller TATA-binding protein associated factors (TAFs) and general TFs such as NFYB and NFYC that appear to form octamer-like structures in the context of transcription initiation complexes (Gangloff et al., 2001). The four core histones, NFYB, NFYC and at least three of the TAFs with a histone fold (TAF6, TAF8 and TAF12) had diverged from each other by the time of the LECA. Interestingly, these TAFs and the slightly later derived paralog TAF9 were independently, repeatedly lost in most or all apicomplexans and all kinetoplastids. The four core nucleosomal histones often show variants which have been shown in model systems to specify ''specialized chromatin'' in the regions where they are deposited on DNA (Boulard et al., 2007; Kusch and Workman, 2007). For example, centromere-specific histone H3 is critical for the assembly of the kinetochore complexes. Amongst parasitic protists, an example of such a variant histone H3 is presented by the *Plasmodium* protein PF13_0185, which contains a distintictive N-terminal tail (Supplementary material file 3). The kinetoplastids on the other hand contain rapidly evolving histones such as H4, which might indicate adaptive evolution (Lukes and Maslov, 2000). Histone H1, which binds inter-nucleosomal linkers, is found in the crown-group stramenopiles (including the plant parasite *Phytophthora*) and *Naegleria*. Its distribution is suggestive of an origin in the crown group from the more widespread paralogous FKH domain (Carlsson and Mahlapuu, 2002; Aravind et al., 2005), followed by lateral transfers to stramenopiles during endosymbiosis with the plant lineage and independently to *Naegleria*.

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

5. The evolution of major functional guilds of chromatin proteins

The opportunity offered by advances in genomics to reconstruct the evolutionary history of the eukaryotic CPs allows us to answer certain previously inaccessible questions more robustly: (i) what was the complement of CPs functioning in LECA? (ii) What were the lineage-specific innovations in CPs of parasitic eukaryotes relative to other organisms? (iii) What implications do differences in complements of CPs have for epigenetic regulation (e.g. generation and ''interpretation'' of the histone code) in parasites when compared with other eukaryotes? With respect to parasites, we can now examine the degree to which different regulatory systems are maintained or modified as parasitism convergently evolves in different eukaryotic lineages. It should be kept in mind that parasitic protists, with few notable exceptions, are relatively poorly studied and the reconstruction presented here is necessarily speculative. Nevertheless, we hope that highlighting the major differences in the natural history of CPs will offer a starting point with material and a hypothesis for case by case experimental investigations.

Table 1
Domains commonly found in chromatin proteins

Domain	Structure	Comments
Enzymatic domains		
Acetyltransferases (GNAT)	α + β fold with six 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 Rossmanoid 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-dependent deacetylation
MACRO domain	Derived α/β fold with N-terminal β-hairpin in core sheet	There are at least eight 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 two histidine residues that might chelate an active metal, typically iron. The oxidative demethylation of proteins by these proteins resembles the oxidative demethylation of DNA by AlkB family enzymes
LSD1-like demethylase	Classical 6-stranded dehydrogenase-type Rossmann fold	This enzyme is believed to catalyze protein demethylation by an oxidative process by utilizing flavin dinucleotides 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 helicases
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/ diooxygenase domain	Double-stranded-β helix	Found in the kinetoplastid J-binding proteins. Distant homologs of AlkB and protein demethylases
DNA-binding domains		
Histone fold	trihelical fold with long central helix	At least nine 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 *Trichomonas* 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 CXXCXXCX(n) unit with the second unit inserted into the first
CXC	A trinuclear Zn cluster	Three 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 three strands and helix	Found only in animals, plants and stramenopiles. Apparently lost in fungi and amoebozoans

(*continued on next page*)

Table 1 (*continued*)

Domain	Structure	Comments
SAD (SRA)	α + β fold	Methylated DNA binding domain with conserved N-terminal histidine and C-terminal YDG signature suggesting possible catalytic activity. Of bacterial origin and fused to McrA-type HNH (Endonuclease VII) endonucleases in them
HIRAN	All β-fold	Typically fused to SWI2/SNF2 ATPases in eukaryotes. Found as a standalone domain in bacteria in conserved operons encoding a range of phage replication enzymes
PARP finger	Single Zn coordinated by three 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 three 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 2nd and 3rd helix	Two superfamilies, namely the SAP and LEM domains contain this fold and are 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 *Entamoeba*
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 *Tetrahymena* 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	Winged helix-turn-helix domain	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. *Phytophothora sojae* 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 Rossmanoid topology	Domain of bacterial origin in LECA. Several eukaryotic versions bind phosphorylated peptides in context of DNA repair

Table 1 (*continued*)

Domain	Structure	Comments
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

5.1. Evolutionary history of histone acetylation-based regulatory systems

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

Of the remaining families of HATs, the Eco1p orthologs (implicated in chromosome segregation (Bellows et al., 2003)) were present at least prior to the branching-off of kinetoplastids. Others such as Hat1p, CSRP2BP and some paralogs of the Esa1p, which form the MYST family (Thomas and Voss, 2007), emerged in the crown group or the common ancestor of the crown group and chromalveolates. *T. vaginalis* shows independent expansions of the MYST (Esa1p orthologs) and Gcn5p type HATs. Several families are restricted to a particular lineage (Neuwald and Landsman, 1997). For example, fungi appear to have at least four lineage-specific families (orthologs of Spt10p, Hpa2p, Rtt109p and *Neurospora* NCU05993.1), while plants have a lineage-specific family of their own with fusion of the acetylase domain with PHD fingers or AT-hook motifs (Fig. 4). Amongst parasitic protists, an unusual lineage-specific representative is seen in *Phytophthora* and related stramenopiles, where the acetylase domain is fused to a carboxymethyltransferase domain (Fig. 4). It is possible that these enzymes might carry out a second covalent protein modification, perhaps of acidic side-chains. The Elp3p and Kre33p acetylases are shared by eukaryotes and archaea, suggesting an inheritance from the archaeal precursor, whereas Esa1p and Gcn5p orthologs appear to be innovations specific to eukaryotes, which were derived through rapid divergence from a pre-existing version of the fold. In contrast, affinities of the lineage-specific versions suggest that they were acquired repeatedly by eukaryotes from the diverse bacterial radiation of NH_2 group acetylases (Fig. 4).

Histone deacetylases belong to two structurally distinct superfamilies, namely the RPD3/HDAC superfamily and the Sir2 superfamily, both of which are universally present in eukaryotes. Prokaryotic members of both superfamilies appear to have played predominantly metabolic roles, respectively participating in acetoin and nicotinamide metabolism, as opposed to a regulatory role in chromatin (Leipe and Landsman, 1997; Sandmeier et al., 2002; Avalos et al., 2004). The RPD3 superfamily uses metal-dependent catalysis, whereas the Sir2 superfamily, which resembles the classical Rossmann fold enzymes, uses a NAD cofactor (Leipe and Landsman, 1997; Avalos et al., 2004). At least one deacetylase of the HDAC/Rpd3 superfamily was present in LECA and appears to have been derived from bacterial acetoin-hydrolyzing enzymes (Fig. 4). There have been several lineage-specific innovations within this superfamily amongst eukaryotes. Consistent with the expansion of HATs,*T. vaginalis* also shows an expansion of HDAC deacetylases, while kinetoplastids show a unique family

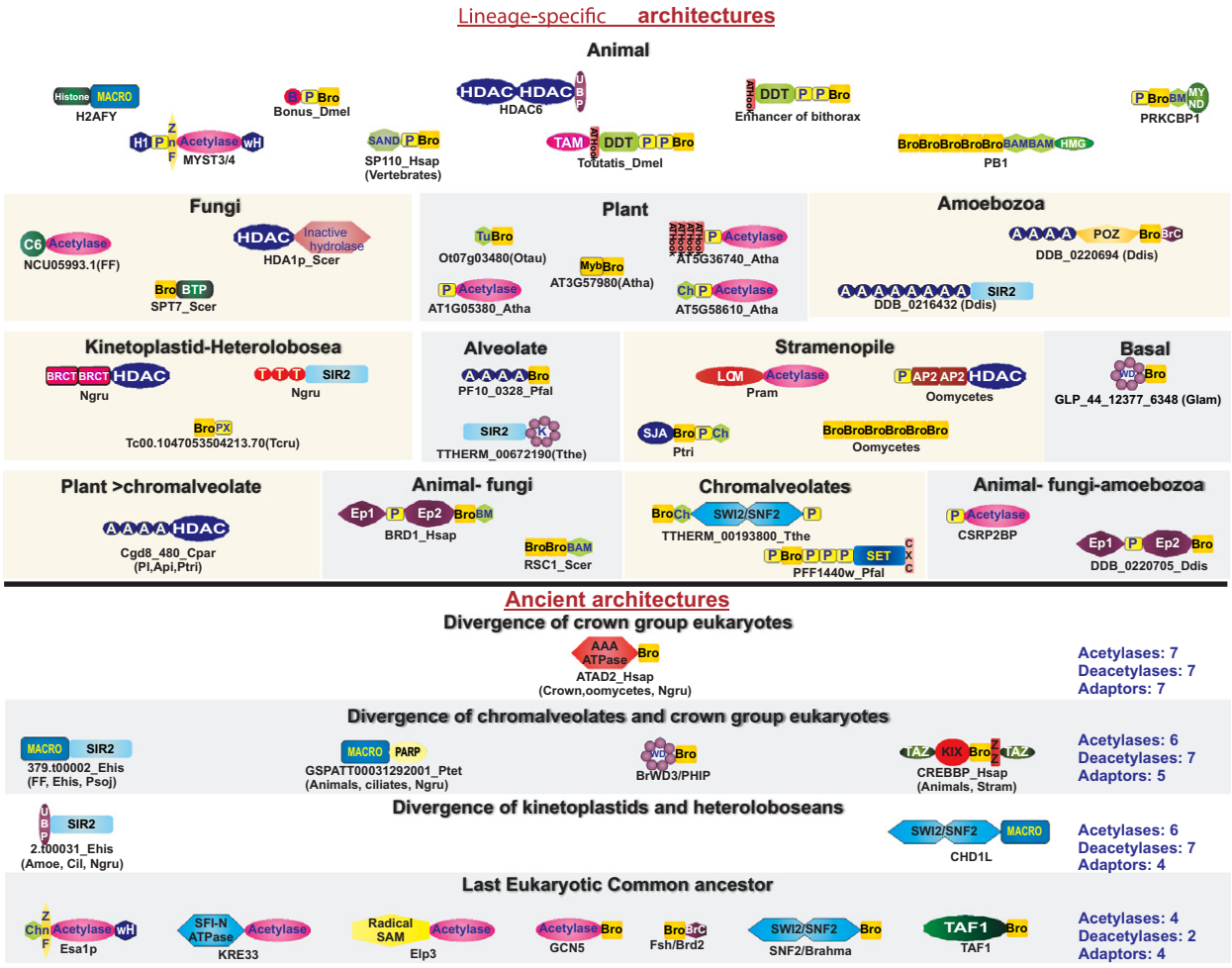

Fig. 4. Ancient and lineage-specific domain architectures in acetylation-based regulatory systems. Evolution of acetylation-based systems are shown using various domain architectures that evolved either at different early stages in the evolution of eukaryotes or more recently in different lineages. The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

typified by LmjF21.1870 from *Leishmania*. The chromalveolate clade, including the apicomplexans *Cryptosporidium* and *Toxoplasma gondii*, has a distinctive version of HDAC that contains N-terminal ankyrin repeats which is shared with plants. Fungal-specific HDA1p deacetylases combine the HDAC domain with a C-terminal inactive α/β hydrolase domain that might be utilized for specific peptide-interactions. The parasites *Phytophthora* and *Naegleria* possess lineage-specific architectures that, respectively, combine the HDAC domain with AP2 and PHD finger domains and the BRCT domain (Fig. 4).

At least one member of the Sir2 superfamily deacetylases, the classical SIR2, can be traced back to the common ancestor of eukaryotes and archaea. All other major families appear to have been acquired from bacteria much later in eukaryotic evolution: Sirtuin 4, 5 and 6 appear to have been independently acquired prior to the divergence of *Naegleria* and kinetoplastids from other eukaryotic lineages. Yet another sporadic lineage of Sir2-like proteins typified by *Cryptosporidium* cgd7_2030 is present in gut parasites such as *Giardia* and *Cryptosporidium* (Fig. 4). Like the HDAC superfamily, members of this family show

parallel domain fusions in various protists: *Dictyostelium* and *Tetrahymena* show fusions to tetratricopeptide and ankyrin repeats. A Sir2 deacetylase from ciliates, amoebozoans (including parasitic *E. histolytica*) and *Naegleria*, contains a fusion to the ubiquitin-binding Zn finger domain which, interestingly, parallels a similar fusion of the ubiquitin-binding Zn finger domain to a HDAC deacetylase in animal HDAC6 enzymes (Fig. 4) (Pandey et al., 2007). These fusions point to several unique interactions being used to recruit enzymes containing deacetylase domains of either superfamily to specific contexts. In particular, the AP2 domain could recruit the deacetylase to specific DNA sequences, ankyrin repeats to large proteins complexes and the BRCT domain to complexes associated with DNA repair. The Ubp-ZnF could, on the other hand, specifically recruit deacetylases to regions of chromatin containing ubiquitinated histones or other ubiquitinated proteins (Pandey et al., 2007).

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

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

5.2. Natural history of histone-methylation-based regulation

Methylation of histones on lysines (both mono and trimethylation) is mediated predominantly by methyltransferases of the Su(var)3–9, Enhancer-of-zeste, Trithorax (SET) domain superfamily (Table 1), which are universally present in eukaryotes (Allis et al., 2006; Sullivan et al., 2006; Kouzarides, 2007). They are unrelated to classical Rossmann fold methylases and contain a β-clip fold (Iyer and Aravind, 2004). All eukaryotes encode SET domain methylases, and at least five distinct versions, including Skm/Bop2-like, trithorax-like, E(z)-like and Ash1-like SET domains can be traced back to the LECA (Fig. 5). The one other SET domain protein traceable to the LECA combines the SET domain with an amino acid ligase domain homologous to polyglutamylases (van Dijk et al., 2007). This protein might catalyze ligation of amino acids, such as peptide polyglutamylation, in addition to lysine methylation. All other SET domain proteins from *Giardia* and *Trichomonas* do not display complex multidomain architectures, unlike orthologs from other eukaryotes. Most domain accretion resulting in complex architectures appears to have happened in the crown group, and few of these proteins have been sporadically transferred to chromalveolates from the plant lineage. One such example is a protein typified by *P. falciparum* PF08_0012, contains a fusion of the DNA-binding SET-associated DR1533 (SAD) domain (Makarova et al., 2001; Johnson et al., 2007) to the SET domain, and seems to have been acquired from the apicoplast precursor. However, occasional lineage-specific domain fusions do appear to have emerged in parasitic protists. *T. gondii* shows a fusion to the High mobility group (HMG) box domain, which has also independently occurred in animals and the alga *Ostreococcus*. Apicomplexans also display another unique lineage-specific methylase combining the SET domain with ankyrin repeats (Fig. 5). Basidiomycete fungi, such as the parasitic form *Cryptococcus,* contain an unusual fusion of a SET domain with a nucleic acid deaminase related to Tad3p (Gerber and Keller, 1999). It remains to

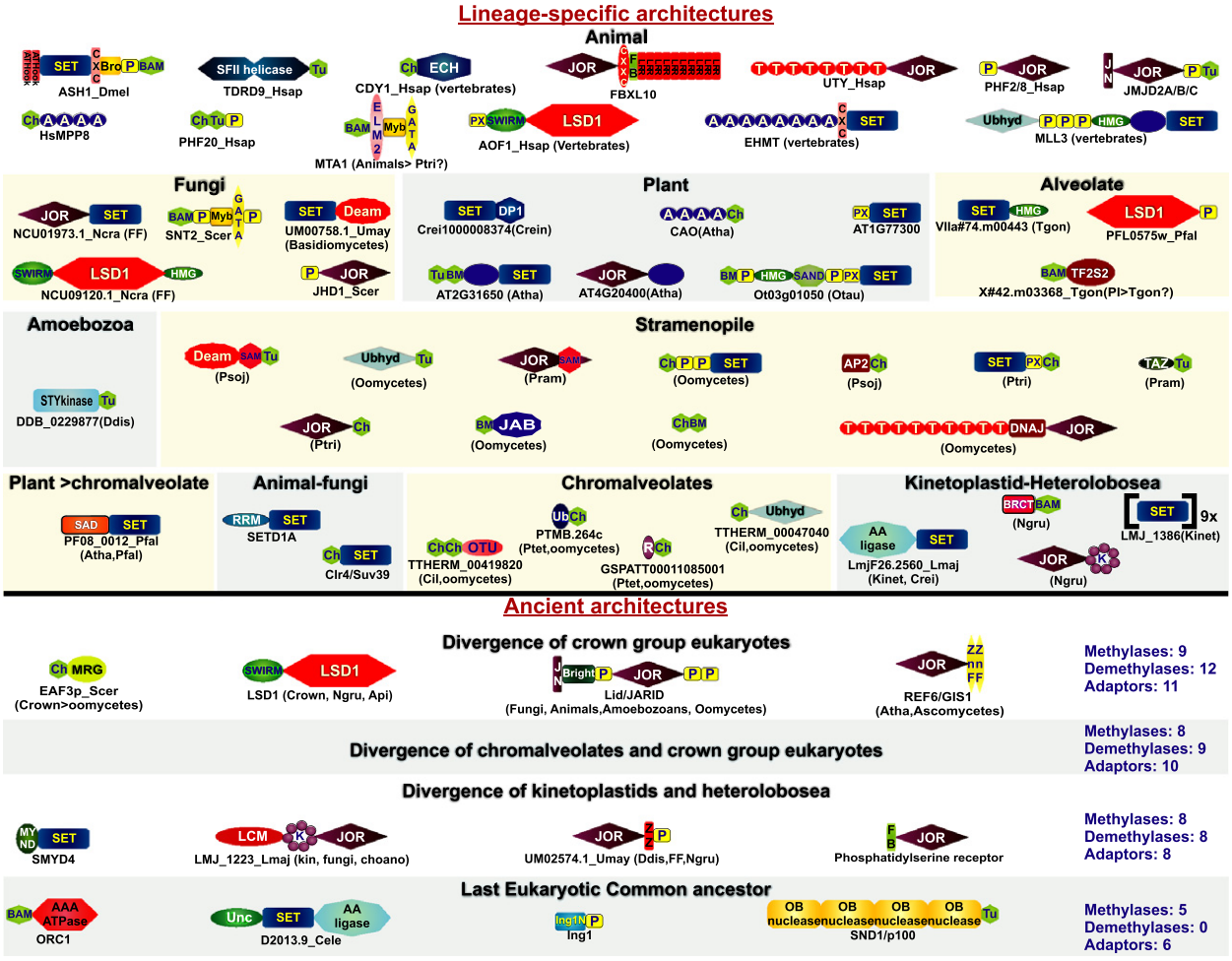

Fig. 5. Ancient and lineage-specific domain architectures in the methylation-dependent regulatory systems. Evolution of methylation-based regulation is shown using various domain architectures that evolved either at different early stages in the evolution of eukaryotes or more recently in different lineages.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

be seen if these proteins, in addition to catalyzing histone methylation, mediate DNA modification via deamination.

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

Demethylation in majority of eukaryotes is carried out by the Jumonji-related (JOR/JmjC) domain, which contains a double-stranded β-helix domain catalyzing a metal and 2-oxo acid dependent oxidative demethylation of modified histones (Anantharaman et al., 2001; Aravind and Koonin, 2001b; Chen et al., 2006; Cloos et al., 2006; Klose et al., 2006). These enzymes appear to be ultimately of bacterial origin, because numerous related as well as more divergent versions of double-stranded β-helix enzymes are found throughout bacteria (Aravind and Koonin, 2001b). This demethylase, as well as other known demethylase domains (see below), are absent in *Giardia* and *Trichomonas*, and other parasites like *E. histolytica* and microsporidians. This implies that certain organisms can apparently function without demethylation, though it is theoretically possible that they possess some unrelated enzyme for this purpose. Nevertheless, prior to the divergence of the kinetoplastid-*Naegleria* clade around nine distinct versions of

demethylases had emerged. As in the case of the SET domain these demethylase domains typically show relatively simple domain architectures in most early-branching eukaryotic groups, but have accreted multiple protein–protein interaction and DBDs in crown-group eukaryotes. Kinetoplastids, certain fungi and choanoflagellates show a fusion between the demethylase domain and a carboxymethyltransferase domain (also fused to acetylases) (Fig. 5).

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

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

Comparisons of protist genomes indicate that distinct versions of the SH3 fold, namely chromo, tudor and BAM/BAH domains, had already separated from each other in the LECA itself, and the BMB (PWWP) domain emerged just prior to the divergence of the kinetoplastid-*Naegleria* clade (Table 1 and Fig. 5). At least three distinct versions of the chromo domain (including a HP1-like protein), one BAM/BAH domain and one version of the chromatin-associated tudor domain, can be extrapolated as being present in the LECA. The ancient representatives of these domains include both forms that are fused to other enzymatic domains, as well as those in non-catalytic proteins. Most parasites such as apicomplexans show a relatively low number of these domains, with some domains such as the BMB (PWWP) being entirely absent. In contrast, *T. vaginalis* shows a LSE of proteins containing chromo domains. In the free-living ciliate *Paramecium*, but none of the other chromalveolates, we observe an unusual expansion of proteins containing fusions of the BAM (BAH) and PHD finger domains. Interestingly, chromoalveolates show several unique architectures combining a version of the chromodomain related to those found in the *Drosophila* malignant brain tumor (MBT) protein (Maurer-Stroh et al., 2003; Sathyamurthy et al., 2003) with several domains related to ubiquitin signaling, such as different deubiquitinating peptidases of the Otu and UBCH families, the RING finger E3-ligase and ubiquitin-like domains (Fig. 5). These architectures point to the development of a functional association between histone methylation and chromatin–protein ubiquitination in these protists. Most of these proteins have been lost in apicomplexan parasites, but are retained in the plant parasite *Phytophthora,* along with several additional lineage-specific architectures involving the chromodomain. In this context, it is of interest to note that a transposon encoding a chromodomain protein has proliferated extensively in the genome of *Phytophthora*.

Recent studies have also shown that certain versions of the binuclear, zinc chelating treble-clef fold domain protein, the PHD finger, bind all nucleosomal histones (Eberharter et al., 2004). Other versions of this domain also interact specifically with trimethylated lysines on histone H3 (Li et al., 2006b; Pena et al., 2006; Shi et al., 2006). Some versions of the PHD finger have been claimed to bind to phosphoinositides, but recent experiments suggest a downstream basic sequence, rather than the PHD finger, is directly involved in this interaction (Kaadige and Ayer, 2006). Given the exclusive prevalence of this domain in CPs and its sequence diversity (Aasland et al., 1995), it is possible that different versions of the PHD finger mediate distinct interactions with trimethylated histones, other modified and unmodified histones or peptides in other chromatin proteins. At least a single copy of the PHD finger was present in the LECA and the domain showed

considerable evolutionary mobility, beginning prior to the separation of the crown group and chromalveolate clades, and again within the crown group (Fig. 5).

5.3. Evolution of chromatin remodeling and assembling systems

Enzymes mediating dynamics of eukaryotic chromatin on local and global scales typically do so by utilizing the free-energy of NTP hydrolysis. Not surprisingly, most of these enzymes contain motor domains of the P-loop NTPase fold (Table 1); two major classes of which are the SWI2/SNF2 ATPases and the SMC ATPases (Bork and Koonin, 1993; Hirano, 2005). SWI2/SNF2 ATPases are primarily involved in local chromatin remodeling events by affecting nucleosome positioning and assembly. They are usually core subunits of large functional complexes that include other chromatin-modifying activities such as acetylases, methylases or ubiquitinating enzymes (Martens and Winston, 2003; Mohrmann and Verrijzer, 2005; Durr and Hopfner, 2006; Gangavarapu et al., 2006). SWI2/SNF2 ATPases had their origins in bacteriophage replication systems and restriction-modification systems found in the prokaryotic superkingdoms (Iyer et al., 2006). They appear to have been recruited from such a source in the earliest stages of eukaryotic evolution and expanded to give rise to at least six representatives by the time of the LECA (Fig. 6). A comparable count of these ATPases is found in the degraded genomes of *Giardia* and *Encephalitozoon* and includes most versions traceable to the LECA.Thus, this ancient set of SWI2/SNF2 ATPases is likely to comprise the most essential group of chromatin remodeling enzymes required by any eukaryote. Domain architectures of these predicted ancestral versions show that the ATPase module was already fused to different peptide-binding domains such as chromo, bromo and MYB (SANT) which allowed them to specifically interact with modified or unmodified nucleosomes (Fig. 6).

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

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

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

A less-understood covalent modification of CPs is the conjugation of ubiquitin (Ub) and other related modifiers

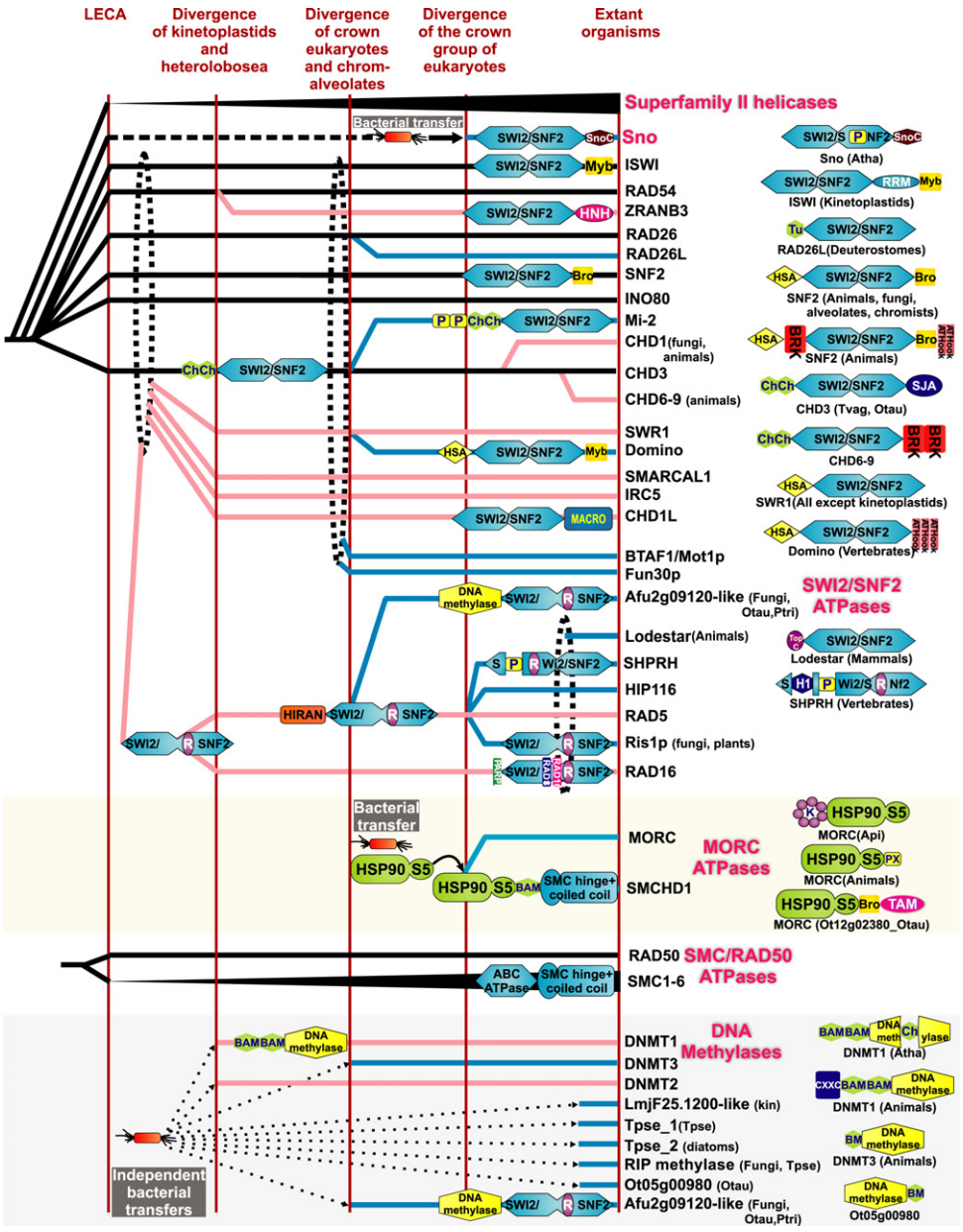

Fig. 6. Evolution of ATP-dependent remodeling and DNA methylation systems. The evolutionary history and inter-familial relationships of four different remodeling ATPases, Sno ATPases, SWI2/SNF2 ATPases, MORC ATPases and SMC ATPases, are shown in addition to DNA methylases. Horizontal lines represent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

(Ubls; e.g. Nedd8 and SUMO) (Shilatifard, 2006; Collins et al., 2007; Kouzarides, 2007). This process involves a three-step reaction that transfers the Ub/Ubl to its target protein. The substrate specificity for the transfer mainly lies in the third enzyme, the E3, which typically contains a RING finger domain (Glickman and Ciechanover, 2002). Several RING finger proteins are exclusive residents of eukaryotic chromatin: the PML family of SUMO-specific E3s, the RING finger containing Rad5/Rad8 family of SWI2/SNF2 ATPases and the Posterior Sex combs (PSC) family of proteins of the Polycomb group that combine a RING finger with a C-terminal Ub-like domain (Gangavarapu et al., 2006; Gearhart et al., 2006; Shilatifard, 2006; Collins et al., 2007; Park et al., 2007). The latter family is conserved in both the crown group and alveolates, including certain apicomplexans such as *Theileria* and *Cryptosporidium* and was shown to mono-ubiquitinate H2A (Gearhart et al., 2006). The presence of dedicated enzymes

for removal of Ub modifications from histones and other nuclear proteins is suggested by the predicted deubiquitinating enzymes which combine the JAB peptidase domain with the SWIRM domain in animals and *Dictyostelium* (Aravind and Iyer, 2002). An unusual set of proteins in *Trichomonas* combine MYB domains with Ub-binding UBA domains, suggesting that they might interact with ubiquitinated chromosomal proteins. Other less-known protein modifications in chromatin are suggested by the presence of nuclear poly-ADP ribosyltransferases. In plants these enzymes are fused to the DNA-binding SAP domain that is likely to tether the catalytic domain to chromosome scaffold attachment regions (Aravind and Koonin, 2000; Zhang, 2003). Interestingly, histone-modifying kinases do not appear to show any notable fusions to other chromatin-specific peptide-binding domains, and are drawn from several ancient families of eukaryotic protein kinases (Manning et al., 2002).

In addition to well-characterized modified-histone-interacting domains, there are numerous less-studied potential peptide-interaction domains in eukaryotic CPs that might also play analogous roles (Table 1). Several versions of the MYB domain found in CPs (often termed SANT domains), bind histone tails rather than DNA (Boyer et al., 2002; de la Cruz et al., 2005; Mo et al., 2005). This appears to represent a eukaryote-specific functional shift in the ancient DNA-binding HTH fold for peptide interaction. Contextual information from domain architecture suggests that domains such as the ELM2, SJA, EP1/2 and the PHDX/ZF-CW with a potential treble-clef fold domain (Finn et al., 2006; Letunic et al., 2006) might interact with histone tails and play a role in reading the histone code or in recruiting other activities to the nucleosome (Table 1; Figs. 7 and 8). One version of another peptide-binding domain, the SWIB domain, recruits ubiquitinating activities via the fused E3-ligase RING finger domain to TFs such as p53 (Bennett-Lovsey et al., 2002). The stand-alone pan-eukaryotic version of this domain might be critical for recruitment of SET domain methyltransferases to SWI2/SNF2-dependent remodeling enzymes to chromatin (Stephens et al., 1998).

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

5.5. Natural history of epigenetic DNA modification enzymes

Modification of DNA by cytosine methyltransferases with the AdoMet-binding Rossmann fold (Table 1) plays a central role in epigenetic regulation in several crown-group eukaryotes (Goll and Bestor, 2005). The common ancestor of crown-group eukaryotes had at least two cytosine methylases, the DNMT1 and DNMT3 families, which appear to have possessed both maintenance and de novo methylation activity (Fig. 6). They were repeatedly lost in many lineages of animals, fungi and amoebozoans. A third methylase, DNMT2, was found in the crown group as well as chromalveolates and *Naegleria*; however recent results suggest that this enzyme might be a $tRNA^{Asp}$ methylase (Goll et al., 2006). Interestingly, several filamentous fungi and *Ostreococcus* code for a novel DNA-methylase, related to the bacterial *dam* DNA adenine methylases fused to a RAD5-like SWI2/SNF2 ATPase and another uncharacterized enzymatic domain (Fig. 6). This might point to a hitherto unstudied adenine methylation in these organisms. *Ostreococcus* and diatoms possess other potential DNA methylases in addition to those conserved in the crown group. At least one of those is fused to a BAM domain, suggesting a chromatin-associated role (Fig. 6). Several filamentous fungi, including plant parasites, contain a distinct cytosine methylase that is involved in the point mutation of repetitive DNA sequences (RIP) and developmental gene regulation (Malagnac et al., 1997; Freitag et al., 2002). The new genome sequences suggest that an ortholog of this enzyme is also present in diatoms such as *Thalassiosira*. In this context, it is interesting to note that kinetoplastids also possess a distinct cytosine methylase (prototyped by *Leishmania* LmjF25.1200) related to bacterial restriction-modification enzymes, although no such DNA modification has been reported in these organisms (Yu et al., 2007). It remains to be seen if this enzyme catalyzes cryptic DNA methylation or is involved in a process similar to repeat-induced point mutation of the fungi. Evolutionary analysis of eukaryotic DNA methylases suggests that they are all related to methylases of different restriction-modification systems or the *dam* methylation system of prokaryotic provenance (Goll and Bestor, 2005) (Fig. 6). Thus, all eukaryotic DNA methylase families, including the DNMT1 and DNMT3 families, appear to have been derived from multiple independent transfers

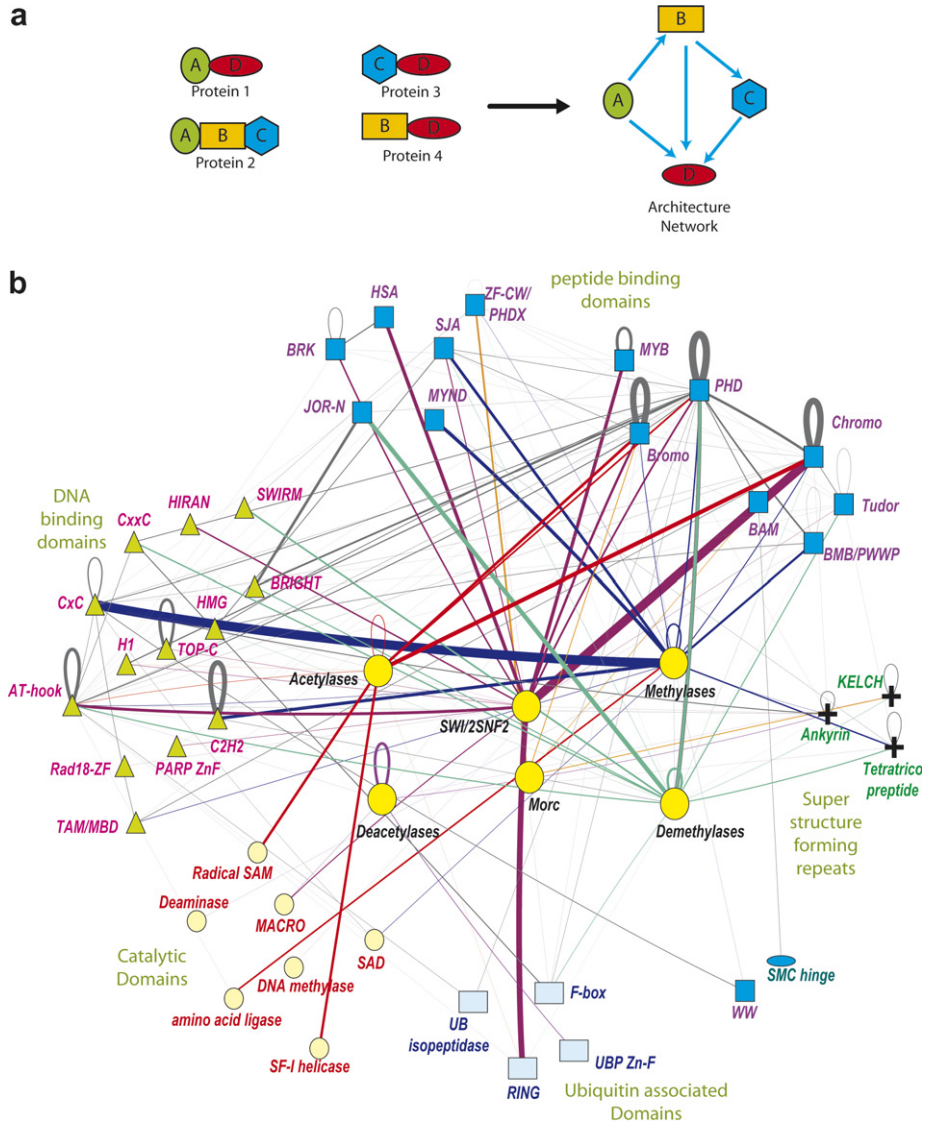

Fig. 7. Network representations of the domain architectures of eukaryotic chromatin proteins. (a) A hypothetical example showing how domain architecture networks are constructed. A, B, C and D are globular domains that occur in a range of combinations. These are combined into an architectural network where the globular domains are notes and the edges reflect their physical connectivity. (b) The domain architecture network for eukaryotic chromatin proteins with a focus on the primary catalytic regulatory systems, namely acetylation, methylation and ATP-dependent chromatin remodeling. Included within acetylases, deacetylases, methylases and demethylases are all enzymes known or predicted to catalyze the respective activity, irrespective of the superfamily to which they belong. The links made by demethylase domains are shown in aquamarine, those by acetylases in red, by SWI2/SNF2 ATPases in purple and by MORC ATPases in orange. Different functional categories of domains and their labels are colored in the same way and spatially grouped together. The thickness of the edges is approximately proportional to the relative frequency with which linkages between two domains re-occur in distinct polypeptides in all eukaryotes. The graphs were rendered using PAJEK (http://vlado.fmf.uni-lj.si/pub/networks/pajek/).

(around six to nine instances) from bacteria to different eukaryotic lineages. Subsequent to their transfer, they appear to have combined with a range of domains found in eukaryotic CPs (e.g. BMB/PWWP in DNMT3, CXXC and BAM/BAH in DNMT1, insertion of chromo domain into methylase domain in plants CMTs of the DNMT1 family (Chan et al., 2006)) that probably helped them to interact specifically with different chromosomal target sites.

Distribution of these methylases suggests that DNA methylation might not be a major regulatory factor in most parasitic protists, with the exception of fungi and possibly kinetoplastids and *Naegleria*. Consistent with this, the TAM (MBD) domain (Table 1) is not observed in any of the lineages of parasitic protists studied to date. However, the SAD (SRA) domain (Table 1), which has also been shown to interact with methylated DNA (Johnson et al., 2007; Woo et al., 2007), is found in *Plasmodium*. An analysis of the conservation pattern of this domain suggests that it contains a set of conserved polar residues suggestive of it being an enzyme (Makarova et al., 2001), and might catalyze as yet unknown DNA modifications. Another potentially important regulatory DNA modification, which

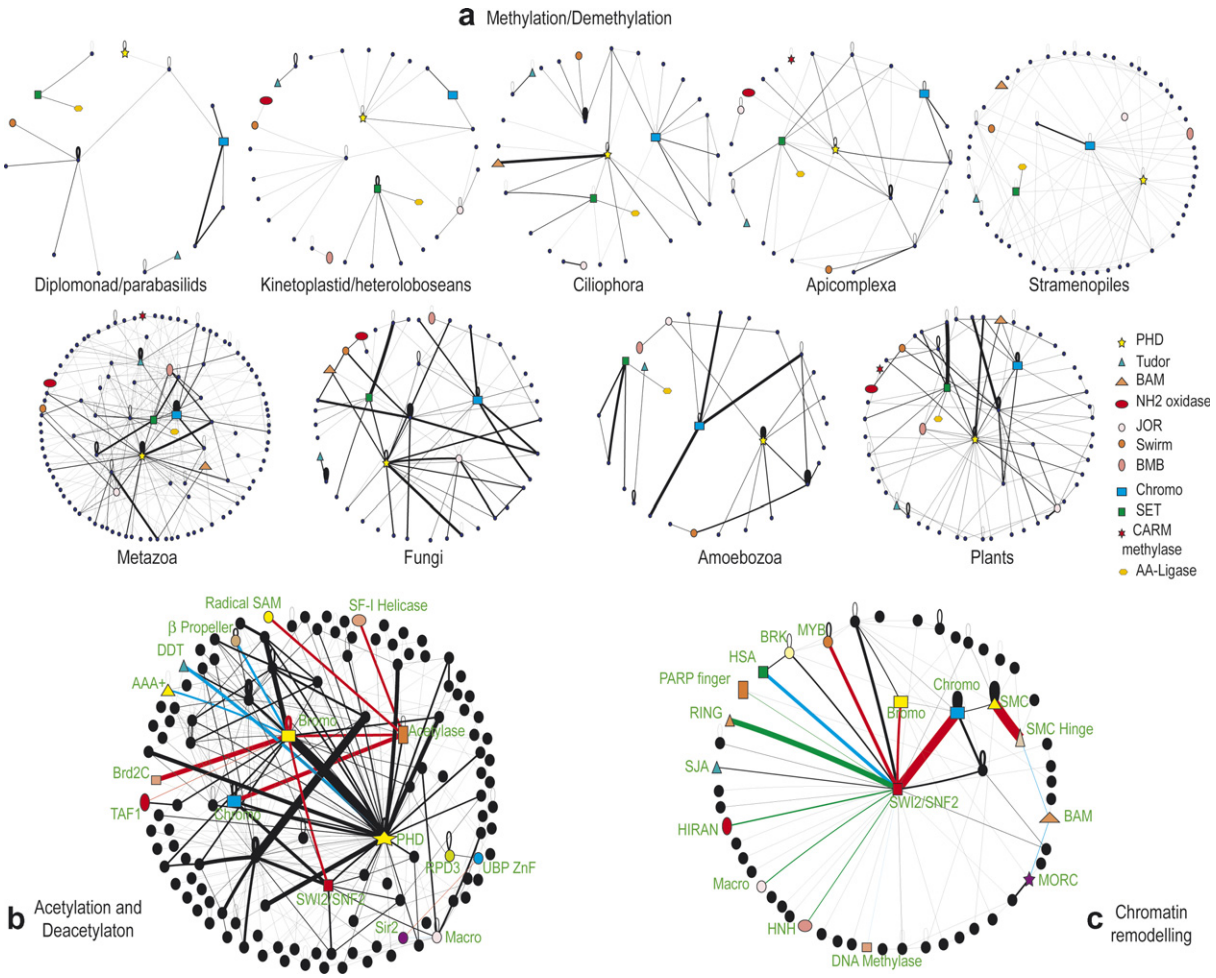

Fig. 8. Domain architecture networks of proteins involved in protein methylation, acetylation and ATP-dependent chromatin remodeling. (a) Domain architecture networks of proteins known or predicted to be involved in the chromatin protein methylation system are shown for representative eukaryotes. The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

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

6. Domain architectures of chromatin proteins

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

Domain architectures of CPs reveal certain strong "syntactical" patterns (Figs. 7, and 8). For example, histone methylase and acetylase domains never co-occur in the same polypeptide in any eukaryote. Likewise, demethylases and deacetylases tend not to co-occur with each other or, respectively, with methylases and acetylases (Fig. 7). This suggests that acetylation and methylation are relatively stable modifications, and that their removal is not temporally coupled or combined with re-modification. This is consistent with methylation and acetylation being epigenetic markers and being independent but potentially complementary in action (Peterson and Laniel, 2004; Shilatifard, 2006; Villar-Garea and Imhof, 2006; Kouzarides, 2007). Two of the four acetyltransferases that can be traced to the LECA are closely associated with the basal transcription apparatus (GCN5, Elp3 families). Hence, the earliest roles of acetylation were probably in the context of modulating histone–DNA interaction to facilitate transcription. On the other hand, methylation appears to have emerged in the more general context of organizing chromosomal structure by altering histone properties. Whereas acetylases show fusions to specific histone-tail-binding domains even in the basal eukaryotes (e.g. GCN5 with a bromo domain), histone methylases only develop such fusions later in eukaryotic evolution (Figs. 5, and 8). However, methylases eventually developed greater domain architectural diversity than acetylases (Figs. 4 and 8). Similarly, histone demethylases show a clearly greater architectural complexity than deacetylases (Fig. 7). These patterns could suggest that methylases and demethylases might have evolved a greater selectivity for the specific contexts (for example, other co-occuring modifications) of their target residues or respond to a larger range of inputs sensed by the fused domains. These observations are consistent with results suggesting distinct roles for these two major components of the "histone code" (Peterson and Laniel, 2004; Shilatifard, 2006; Villar-Garea and Imhof, 2006; Kouzarides, 2007).

Acetylases and methylases show preferential associations with certain peptide-binding domains – acetylases most frequently combine with bromo domains, and methylases with PHD fingers (Fig. 7). Given the binding preferences of these peptide-binding domains, it is possible that, respectively, recognizing previously methylated or acetylated histones might be an important functional feature of some versions of these enzymes, especially in the context of maintaining an epigenetic mark. Conversely, methylases are also fused to acetylated-peptide-binding domains and acetylases are fused to methylated-peptide-binding domains (Figs. 7 and 8), suggesting that a degree of cross-talk or interdependence developed between these modification processes in the course of eukaryote evolution. Likewise, evidence from domain architectures suggests that both systems interact to a certain degree with the ubiquitin system and such associations began emerging in the chromalveolate and crown-group clades. Peptide-binding domains recognizing different forms of histone modifications might also be combined with each other in the same polypeptide (Figs. 4, 5, and 7). Often, such architectures have arisen in a lineage-specific manner, including in several parasitic protists (Figs. 4 and 5). For example, *Phytophthora* shows proteins with six tandem bromo domains and serial bromo, PHD finger and chromo domains, trypanosomes possess a protein with bromo and ZF-CW(PHDX) domains, and *Giardia* possesses a protein combining the bromo domain and a WD-type β-propeller (Figs. 4 and 5). This suggests that while histone modifications might be universal in eukaryotes, their "interpretation" by peptide-binding adaptors shows lineage-specific differences. SWI2/SNF2 ATPases have been shown to work with different histone-modifying enzymes in eukaryotic model systems (Martens and Winston, 2003; Mohrmann and Verrijzer, 2005). However, their domain architectures across eukaryotes show that there are no known fusions between these ATPases and histone acetylase or methylase domains (or the corresponding demodifying enzymes) (Fig. 7). Hence, though their actions are cooperative, they are not closely coupled mechanistically. However, SWI2/SNF2 ATPases are combined with Ub-conjugating E3 domains in the same polypeptide, suggesting possible coupled action between these activities (Gangavarapu et al., 2006).

6.2. Relationship between phylogeny, organizational complexity and domain architectures of chromatin proteins

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

Domain architecture networks show a trend of increasing domain architectural complexity in CPs in the course of eukaryotic evolution (Fig. 8). Diplomonads and parabasalids have the least complex domain architectures. The

Naegleria-kinetoplastid clade, apicomplexans and ciliates have higher architectural complexity than these and chromists have even higher values. However, the highest architectural complexity is observed in certain crown-group clades, and amongst those the animals are unparalleled in the complexity of their domain architecture networks (Fig. 8). When the complexity quotient of CPs is plotted against the total number of predicted CPs encoded by an organism, we observe a steady positively-correlated rise in these values. In many cases, this increase in architectural complexity occurs via "domain accretion" or fusion of new domains around an ancient orthologous core of the polypeptide (Gibson and Spring, 1998; Koonin et al., 2000; Lander et al., 2001). This tendency is particularly prominent in histone methylases and SWI2/SNF2 ATPases (Figs. 5, 6, and 8). Despite having large absolute numbers of CPs, ciliates and *Trichomonas* tend to have much lower architectural complexity. Mere increase in proteome size without increase in architectural complexity of CPs, as seen in ciliates and *T. vaginalis*, might be sufficient to achieve relatively complex organization within a single cell. In contrast, the high complexity of animal proteins points to a possible relationship between architectural complexity and the number of CPs, and emergence of numerous differentiated cell-types (Figs. 2d and 8). Excluding *Naegleria* and *Trichomonas,* other protist parasites such as apicomplexans, kinetoplastids and diplomonads have relatively fewer and architecturally less complex CPs, compared with their hosts (Figs. 2d and 8). As a consequence, relatively less experimental effort might be needed to completely unravel their regulatory interaction networks.

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

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

A number of lines of evidence point to a functional link between RNA-based regulatory systems, including post-transcriptional gene silencing or RNA interference (RNAi) and chromatin-level regulatory events. Studies in plants have revealed a role for siRNAs in directing DNA methylation and heterochromatin formation (Chan et al., 2006; Li et al., 2006a; Pontes et al., 2006; Vaucheret, 2006). RNAi-like systems have also been implicated in epigenetic phenomenon such as paramutation in plants and meiotic silencing by unpaired DNA in *Neurospora* (Shiu et al., 2001; Alleman et al., 2006). Comparative genomic analysis of fungi predicted a functional link between the siRNA/miRNA biogenesis pathway and several CPs (Aravind et al., 2000). Accumulating recent experimental evidence has confirmed this, and points to a major role of small RNAs in directing histone methylation and heterochromatinization in fungi such as *Schizosaccharomyces* (Grewal and Moazed, 2003; Grewal and Rice, 2004). In ciliates, a similar small RNA-based pathway has been implicated in histone H3 methylation, heterochromatin formation and subsequent rearrangement and elimination of DNA sequences during the development of the macronucleus (Mochizuki et al., 2002; Mochizuki and Gorovsky, 2004; Malone et al., 2005). The key conserved players in the generation of these small regulatory RNAs are the dicer nuclease and the RNA-dependent RNA polymerase (RDRP), which is involved in amplifying those. The silencing action of these RNAs is mediated by the PIWI (after the *Drosophila* Piwi protein) domain RNAses (the slicer nucleases), which might localize to chromatin to specifically degrade transcripts at the source (Grewal and Moazed, 2003; Grewal and Rice, 2004; Ullu et al., 2004; Li et al., 2006a; Pontes et al., 2006). The presence of PIWI domains and RDRPs in representatives of all major eukaryotic clades studied to date indicates that a minimal RNAi system comprising these two proteins had already emerged in the LECA. Both the RDRP and the PIWI domain nucleases of this ancestral system appear to have been acquired by the eukaryotic progenitor from bacterial sources (Aravind et al., 2006). However, the system was repeatedly lost, either partially or entirely, in several eukaryotes. Vertebrate apicomplexan parasites, with exception of the *Toxoplasma* lineage, have lost both the PIWI nuclease and the RDRP, suggesting that they are unlikely to possess a bona fide RNAi system (Ullu et al., 2004). Some parasites such as kinetoplastids and *Trichomonas* appear to have lost the RDRP but retain PIWI nucleases, and as a consequence display certain RNAi effects (Shi et al., 2004a). Other parasites such as *Giardia*, *Entamoeba* and the fungus *Cryptococcus* possess both these enzymes, suggesting the presence of both small RNA amplification and degradation systems in these organisms. Interestingly, *Entamoeba* encodes an inactive version of the RDRP (26.t00065), which might have a novel non-catalytical regulatory role.

With the exception of HP1-like chromodomain proteins and some conserved SET domain histone methylases, many CPs that appear to interact with the RNAi machinery are largely limited to the crown-group eukaryotes (Fig. 5) (Aravind et al., 2000). Nevertheless, a core interacting regulatory network combining HP1-like chromodomain proteins, histone methylases and the RNAi machinery could have emerged very early in eukaryotic evolution.

Several studies in crown-group eukaryotes have implicated large non-coding RNAs in heterochromatin formation and chromosome dosage compensation. Some chromodomains have been shown to interact with these RNAs (Brehm et al., 2004; Bernstein et al., 2006). Likewise, SAM domain proteins of the polycomb complex in animals have also been shown to interact with large RNAs in chromatin (Zhang et al., 2004). These suggest that there might be other RNA-based pathways, distinct from RNAi pathways, which might have a direct role in chromatin level regulation. Expression of the variant surface antigen Pfemp1, encoded by the *var* genes in *P. falciparum,* involves silencing of all of the copies of this gene except an active version (Ralph and Scherf, 2005). Antigenic variation proceeds via silencing of the currently active copy and activation of a previously inactive copy. This silencing process has been shown to resemble heterochromatin formation and is mediated by changes in histone modification, including the action of the PfSir2 deacetylase (Duraisingh et al., 2005; Freitas-Junior et al., 2005). The transition between the active and silenced state in *var* gene expression appears to depend on the generation of a non-coding or "sterile" transcript from a promoter located in the intron of the gene (Deitsch et al., 2001; Frank et al., 2006). This raises the possibility of larger transcripts mediating chromatin dynamics in *P. falciparum*. These tantalizing leads hint that there is likely to be a whole "world" of RNA-based chromatin reorganizing processes that remain unexplored in different protists.

8. General considerations and conclusions

As seen from the above discussion, the new data enables an objective reconstruction of various transcription- and chromatin-related regulatory systems in the LECA (see Supplementary material files 2 and 3) and their subsequent evolution. Strikingly, several key players in chromatin and eukaryotic transcription regulation which were present in the LECA were possibly derived from mobile elements and prophages, probably of bacterial origin. These include the SWI2/SNF2 ATPases, the HEH domain which helps in tethering chromosomes to the nuclear membrane, and the RDRP (Mans et al., 2004; Aravind et al., 2006; Iyer et al., 2006). An important feature that defined the origin of eukaryotes was an early spurt of drastic evolutionary innovation that accompanied the melding of the archaeal and bacterial inheritances to give rise to a distinctive eukaryotic system (Koonin et al., 2000; Dacks and Doolittle, 2001; Walsh and Doolittle, 2005; Aravind et al., 2006). This appears to have happened between the point of emergence of the first eukaryotic progenitor and the LECA from which all extant eukaryotes have emerged.

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

Genomes of various early-branching eukaryotes (e.g. *Trichomonas* and *Giardia*) suggest that recruitment of novel classes of DBDs had begun early in eukaryotic evolution, with repeated emergence of new TFs in different lineages. In particular, specific TFs in various parasitic protists remained unknown until recently. However, this principle of lineage-specific expansions allowed us to identify the major specific TFs of several parasitic lineages such as apicomplexans, *T. vaginalis*, *Entamoeba*, oomycetes and heterolobosans (Fig. 3; Supplementary material file 2). Typically, parasitic protists, irrespective of their phylogeny, possess fewer specific TFs and less complex CPs. The transcription regulation apparatus of protist parasites have taken very different courses during adaptation to such a life-style. Microsporidians, kinetoplastids and *Giardia* have highly reduced complements of specific transcription regulators and CPs. Others such as *Entamoeba* and apicomplexans have lost most TFs relative to their free-living sister-groups, but have expanded single DBD families to derive the majority of their specific TFs. Differences can even be observed within apicomplexans in the complements of specific TFs: for instance, *Cryptosporidium* retains certain specific TFs such as E2F/DP1 that have been lost in other apicomplexans, and *Toxoplasma* displays a distinctly higher number of ApiAP2 TFs than all other apicomplexans, perhaps indicating a higher degree of specific transcriptional regulation. Oomycetes, *Naegleria* and *T. vaginalis* have large numbers of TFs, comparable in numbers to any free-living organism of a similar organizational grade (Babu et al., 2004). Thus, the degree of transcriptional regulation in eukaryotic parasites appears to have been shaped by a combination of factors such as metabolic capabilities, degree of obligate host-dependence,

complexity of life cycles and effective coding capacity of the genome. There also appears to be no strong correlation between the number of TFs and CPs and general cellular morphology – an aspect strikingly illustrated by the gross demographic differences in these proteins between *Giardia* and *Trichomonas* despite their comparable morphology.

Translating this information into experimental results leading to a new understanding of parasitic protists is a major challenge. However, a first level approximation can be obtained via a directed effort using the most obvious high-throughput methods such as expression studies, CHIP-chip methods, large-scale interaction mapping, immuno-precipitation of complexes, fluorescence-tagged localization studies and biochemical genomics to glean basic cell-biological information (Bozdech et al., 2003; Le Roch et al., 2003; Dunn et al., 2005; LaCount et al., 2005; Collins et al., 2007). In particular, these approaches might be useful to obtain insight into the upstream regulators of genes implicated in pathogenesis and the progression of parasitic disease. We also hope that these studies would go hand-in-hand with more involved lines of investigation such as gene-knockouts, phenotypic analysis and thorough biochemical characterization. Given the presence of certain unique predicted enzymatic activities in protists, we believe that such studies might also provide direct leads regarding novel biochemistries that have been ignored in eukaryotic model systems. These studies might also lead to new targets for therapeutic and diagnostic applications. Specifically, the distinctness of many protist regulatory enzymes from their animal and plant counterparts might furnish targets for conventional drug development. Identification of distinctive specific TFs in protists also raises the hope of revisiting the relatively less-explored direction of TF-targeting drugs (Ghosh and Papavassiliou, 2005; Visser et al., 2006). Irrespective of the ultimate applications, these explorations appear poised to deliver new information on eukaryotic transcription and chromatin dynamics in the near future.

Acknowledgements

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

We are grateful to US Department of Energy Joint Genome Institute for making available the draft genome sequence of Naegleria gruberi, an amoeboflagellate and representative of the genus Naegleria, which includes a facultative human parasite causing meningoencephalitis. We would like to specifically acknowledge the effort of Lillian Fritz-Laylin, Scott Dawson, Simon Prochnik, Michael Ginger, Alan Kuo, Igor Iev, Harris Shapiro, Joel Mancuso, Jonathan Pham, W.Zacheus Cande, Daniel Rokhsar and the Joint Genome Institute for sequencing and annotating the Naegleria gruberi genome. The Joint Genome Institute provides these data in good faith, but makes no warranty, expressed or implied, nor assumes any legal liability or responsibility for any purpose for which the data are used.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.ijpara.2007.07.018.

References

Aasland, R., Gibson, T.J., Stewart, A.F., 1995. The PHD finger: implications for chromatin-mediated transcriptional regulation. Trends Biochem. Sci. 20, 56–59.

Alleman, M., Sidorenko, L., McGinnis, K., Seshadri, V., Dorweiler, J.E., White, J., Sikkink, K., Chandler, V.L., 2006. An RNA-dependent RNA polymerase is required for paramutation in maize. Nature 442, 295–298.

Allis, C.D., Jenuwein, T., Reinberg, D., Caparros, M., 2006. Epigenetics. Cold Spring Harbor Laboratory Press, New York.

Anantharaman, V., Koonin, E.V., Aravind, L., 2001. Regulatory potential, phyletic distribution and evolution of ancient, intracellular small-molecule-binding domains. J. Mol. Biol. 307, 1271–1292.

Anantharaman, V., Koonin, E.V., Aravind, L., 2002. Comparative genomics and evolution of proteins involved in RNA metabolism. Nucleic Acids Res. 30, 1427–1464.

Anantharaman, V., Iyer, L.M., Aravind, L., 2007. Comparative genomics of protists: new insights on evolution of eukaryotic signal transduction and gene regulation. Annu. Rev. Microbiol.

Aravind, L., Koonin, E.V., 1998. Second family of histone deacetylases. Science 280, 1167a.

Aravind, L., Landsman, D., 1998. AT-hook motifs identified in a wide variety of DNA-binding proteins. Nucleic Acids Res. 26, 4413–4421.

Aravind, L., Koonin, E.V., 2000. SAP - a putative DNA-binding motif involved in chromosomal organization. Trends Biochem. Sci. 25, 112–114.

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

Aravind, L., 2001. The WWE domain: a common interaction module in protein ubiquitination and ADP ribosylation. Trends Biochem. Sci. 26, 273–275.

Aravind, L., Koonin, E.V., 2001a. Prokaryotic homologs of the eukaryotic DNA-end-binding protein Ku, novel domains in the Ku protein and prediction of a prokaryotic double-strand break repair system. Genome Res. 11, 1365–1374.

Aravind, L., Koonin, E.V., 2001b. The DNA-repair protein AlkB, EGL-9, and leprecan define new families of 2-oxoglutarate- and iron-dependent dioxygenases. Genome Biol. 2, RESEARCH0007.

Aravind, L., Iyer, L.M., 2002. The SWIRM domain: a conserved module found in chromosomal proteins points to novel chromatin-modifying activities. Genome Biol. 3, RESEARCH0039.

Aravind, L., Anantharaman, V., Balaji, S., Babu, M.M., Iyer, L.M., 2005. The many faces of the helix-turn-helix domain: transcription regulation and beyond. FEMS Microbiol. Rev. 29, 231–262.

Aravind, L., Iyer, L.M., Koonin, E.V., 2006. Comparative genomics and structural biology of the molecular innovations of eukaryotes. Curr. Opin. Struct. Biol. 16, 409–419.

Arisue, N., Hasegawa, M., Hashimoto, T., 2005. Root of the Eukaryota tree as inferred from combined maximum likelihood analyses of multiple molecular sequence data. Mol. Biol. Evol. 22, 409–420.

Armbrust, E.V., Berges, J.A., Bowler, C., Green, B.R., Martinez, D., Putnam, N.H., Zhou, S., Allen, A.E., Apt, K.E., Bechner, M., et al., 2004. The genome of the diatom *Thalassiosira pseudonana*: ecology, evolution, and metabolism. Science 306, 79–86.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Denhardt, D.T., Chaly, N., Walden, D.B., 2005. The eukaryotic nucleus: a thematic issue. BioEssays 9, 43.

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

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

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

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

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

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

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

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

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

Felsenstein, J., 1989. PHYLIP – phylogeny inference package (version 3.2). Cladistics 5, 164–166.

Finn, R.D., Mistry, J., Schuster-Bockler, B., Griffiths-Jones, S., Hollich, V., Lassmann, T., Moxon, S., Marshall, M., Khanna, A., Durbin, R., et al., 2006. Pfam: clans, web tools and services. Nucleic Acids Res. 34, D247–D251.

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

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

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

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

Frye, R.A., 1999. Characterization of five human cDNAs with homology to the yeast SIR2 gene: Sir2-like proteins (sirtuins) metabolize NAD and may have protein ADP-ribosyltransferase activity. Biochem. Biophys. Res. Commun. 260, 273–279.

Gangavarapu, V., Haracska, L., Unk, I., Johnson, R.E., Prakash, S., Prakash, L., 2006. Mms2-Ubc13-dependent and -independent roles of Rad5 ubiquitin ligase in postreplication repair and translesion DNA synthesis in *Saccharomyces cerevisiae*. Mol. Cell. Biol. 26, 7783–7790.

Gangloff, Y.G., Romier, C., Thuault, S., Werten, S., Davidson, I., 2001. The histone fold is a key structural motif of transcription factor TFIID. Trends Biochem. Sci. 26, 250–257.

Gardner, M.J., Hall, N., Fung, E., White, O., Berriman, M., Hyman, R.W., Carlton, J.M., Pain, A., Nelson, K.E., Bowman, S., et al., 2002. Genome sequence of the human malaria parasite *Plasmodium falciparum*. Nature 419, 498–511.

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

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

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

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

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

Goff, L.J., Coleman, A.W., 1995. Fate of parasite and host organelle DNA during cellular transformation of red algae by their parasites. Plant Cell 7, 1899–1911.

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

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

Grewal, S.I., Moazed, D., 2003. Heterochromatin and epigenetic control of gene expression. Science 301, 798–802.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Klose, R.J., Yamane, K., Bae, Y., Zhang, D., Erdjument-Bromage, H., Tempst, P., Wong, J., Zhang, Y., 2006. The transcriptional repressor JHDM3A demethylates trimethyl histone H3 lysine 9 and lysine 36. Nature 442, 312–316.

Koonin, E.V., Aravind, L., Kondrashov, A.S., 2000. The impact of comparative genomics on our understanding of evolution. Cell 101, 573–576.

Kouzarides, T., 2007. Chromatin modifications and their function. Cell 128, 693–705.

Kreier, J., 1977. Parasitic Protozoa. Academic Press, New York.

Kusch, T., Workman, J.L., 2007. Histone variants and complexes involved in their exchange. Subcell. Biochem. 41, 91–109.

Lachner, M., O'Carroll, D., Rea, S., Mechtler, K., Jenuwein, T., 2001. Methylation of histone H3 lysine 9 creates a binding site for HP1 proteins. Nature 410, 116–120.

LaCount, D.J., Vignali, M., Chettier, R., Phansalkar, A., Bell, R., Hesselberth, J.R., Schoenfeld, L.W., Ota, I., Sahasrabudhe, S., Kurschner, C., et al., 2005. A protein interaction network of the malaria parasite *Plasmodium falciparum*. Nature 438, 103–107.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke

R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J.P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann N, Stojanovic N, Subramanian A, Wyman D, Rogers J, Sulston J, Ainscough R, Beck S, Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J.C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R.H., Wilson, R.K., Hillier, L.W., McPherson, J.D., Marra, M.A., Mardis, E.R., Fulton, L.A., Chinwalla, A.T., Pepin, K.H., Gish, W.R., Chissoe, S.L., Wendl, M.C., Delehaunty, K.D., Miner, T.L., Delehaunty, A., Kramer, J.B., Cook, L.L., Fulton, R.S., Johnson, D.L., Minx, P.J., Clifton, S.W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J.F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., Gibbs, R.A., Muzny, D.M., Scherer, S.E., Bouck, J.B., Sodergren, E.J., Worley, K.C., Rives, C.M., Gorrell, J.H., Metzker, M.L., Naylor, S.L., Kucherlapati, R.S., Nelson, D.L., Weinstock, G.M., Sakaki, Y., Fujiyama, A., Hattori, M., Yada, T., Toyoda, A., Itoh, T., Kawagoe, C., Watanabe, H., Totoki, Y., Taylor, T., Weissenbach, J., Heilig, R., Saurin, W., Artiguenave, F., Brottier, P., Bruls, T., Pelletier, E., Robert, C., Wincker, P., Smith, D.R., Doucette-Stamm, L., Rubenfield, M., Weinstock, K., Lee, H.M., Dubois, J., Rosenthal, A., Platzer, M., Nyakatura, G., Taudien, S., Rump, A., Yang, H., Yu, J., Wang, J., Huang, G., Gu, J., Hood, L., Rowen, L., Madan, A., Qin, S., Davis, R.W., Federspiel, N.A., Abola, A.P., Proctor, M.J., Myers, R.M., Schmutz, J., Dickson, M., Grimwood, J., Cox, D.R., Olson, M.V., Kaul, R., Raymond, C., Shimizu, N., Kawasaki, K., Minoshima, S., Evans, G.A., Athanasiou, M., Schultz, R., Roe, B.A., Chen, F., Pan, H., Ramser, J., Lehrach, H., Reinhardt, R., McCombie, W.R., de la Bastide, M., Dedhia, N., Blocker, H., Hornischer, K., Nordsiek, G., Agarwala, R., Aravind, L., Bailey, J.A., Bateman, A., Batzoglou, S., Birney, E., Bork, P., Brown, D.G., Burge, C.B., Cerutti, L., Chen, H.C., Church, D., Clamp, M., Copley, R.R., Doerks, T., Eddy, S.R., Eichler, E.E., Furey, T.S., Galagan, J., Gilbert, J.G., Harmon, C., Hayashizaki, Y., Haussler, D., Hermjakob, H., Hokamp, K., Jang, W., Johnson, L.S., Jones, T.A., Kasif, S., Kaspryzk, A., Kennedy, S., Kent, W.J., Kitts, P., Koonin, E.V., Korf, I., Kulp, D., Lancet, D., Lowe, T.M., McLysaght, A., Mikkelsen, T., Moran, J.V., Mulder, N., Pollara, V.J., Ponting, C.P., Schuler, G., Schultz, J., Slater, G., Smit, A.F., Stupka, E., Szustakowski, J., Thierry-Mieg, D., Thierry-Mieg, J., Wagner, L., Wallis, J., Wheeler, R., Williams, A., Wolf, Y.I., Wolfe, K.H., Yang, S.P., Yeh, R.F., Collins, F., Guyer, M.S., Peterson, J., Felsenfeld, A., Wetterstrand, K.A., Patrinos, A., Morgan, M.J., de Jong, P., Catanese, J.J., Osoegawa, K., Shizuya, H., Choi, S., Chen, Y.J.; International Human Genome Sequencing Consortium. 2001. Initial sequencing and analysis of the human genome. Nature 409, 860–921.

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

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

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

Leander, B.S., Keeling, P.J., 2003. Morphostasis in alveolate evolution. Trends Ecol. Evol. 18, 395–402.

Leipe, D.D., Landsman, D., 1997. Histone deacetylases, acetoin utilization proteins and acetylpolyamine amidohydrolases are members of an ancient protein superfamily. Nucleic Acids Res. 25, 3693–3697.

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

Letunic, I., Copley, R.R., Pils, B., Pinkert, S., Schultz, J., Bork, P., 2006. SMART 5: domains in the context of genomes and networks. Nucleic Acids Res. 34, D257–D260.

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

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

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

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

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

Lukes, J., Maslov, D.A., 2000. Unexpectedly high variability of the histone H4 gene in *Leishmania*. Parasitol. Res. 86, 259–261.

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

Malagnac, F., Wendel, B., Goyon, C., Faugeron, G., Zickler, D., Rossignol, J.L., Noyer-Weidner, M., Vollmayr, P., Trautner, T.A., Walter, J., 1997. A gene essential for de novo methylation and development in *Ascobolus* reveals a novel type of eukaryotic DNA methyltransferase structure. Cell 91, 281–290.

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

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

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

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

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

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

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

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

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

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

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

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

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

Oakley, M.S., Kumar, S., Anantharaman, V., Zheng, H., Mahajan, B., Haynes, J.D., Moch, J.K., Fairhurst, R., McCutchan, T.F., Aravind, L., 2007. Molecular factors and biochemical pathways induced by febrile temperature in intraerythrocytic *Plasmodium falciparum* parasites. Infect. Immun. 75, 2012–2025.

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

Pandey, U.B., Nie, Z., Batlevi, Y., McCray, B.A., Ritson, G.P., Nedelsky, N.B., Schwartz, S.L., DiProspero, N.A., Knight, M.A., Schuldiner, O., et al., 2007. HDAC6 rescues neurodegeneration and provides an essential link between autophagy and the UPS. Nature 447, 859–863.

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

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

Park, Y.J., Luger, K., 2006. The structure of nucleosome assembly protein 1. Proc. Natl. Acad Sci. USA 103, 1248–1253.

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

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

Peterson, C.L., Laniel, M.A., 2004. Histones and histone modifications. Curr. Biol. 14, R546–R551.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Shiu, P.K., Raju, N.B., Zickler, D., Metzenberg, R.L., 2001. Meiotic silencing by unpaired DNA. Cell 107, 905–916.

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

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

Smit, A.F., Riggs, A.D., 1996. Tiggers and DNA transposon fossils in the human genome. Proc. Natl. Acad Sci. USA 93, 1443–1448.

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

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

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

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

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

Tang, Y., Poustovoitov, M.V., Zhao, K., Garfinkel, M., Canutescu, A., Dunbrack, R., Adams, P.D., Marmorstein, R., 2006. Structure of a human ASF1a-HIRA complex and insights into specificity of histone chaperone complex assembly. Nat. Struct. Mol. Biol. 13, 921–929.

Templeton, T.J., Iyer, L.M., Anantharaman, V., Enomoto, S., Abrahante, J.E., Subramanian, G.M., Hoffman, S.L., Abrahamsen, M.S., Aravind, L., 2004. Comparative analysis of apicomplexa and genomic diversity in eukaryotes. Genome Res. 14, 1686–1695.

Thomas, T., Voss, A.K., 2007. The diverse biological roles of MYST histone acetyltransferase family proteins. Cell Cycle 6, 704–969.

Tyler, B.M., Tripathy, S., Zhang, X., Dehal, P., Jiang, R.H., Aerts, A., Arredondo, F.D., Baxter, L., Bensasson, D., Beynon, J.L., et al., 2006. *Phytophthora* genome sequences uncover evolutionary origins and mechanisms of pathogenesis. Science 313, 1261–1266.

Uhlmann, F., Hopfner, K.P., 2006. Chromosome biology: the crux of the ring. Curr. Biol. 16, R102–R105.

Ullu, E., Tschudi, C., Chakraborty, T., 2004. RNA interference in protozoan parasites. Cell. Microbiol. 6, 509–519.

van Dijk, J., Rogowski, K., Miro, J., Lacroix, B., Edde, B., Janke, C., 2007. A targeted multienzyme mechanism for selective microtubule polyglutamylation. Mol. Cell 26, 437–448.

Vaucheret, H., 2006. Post-transcriptional small RNA pathways in plants: mechanisms and regulations. Genes Dev. 20, 759–771.

Villar-Garea, A., Imhof, A., 2006. The analysis of histone modifications. Biochim. Biophys. Acta 1764, 1932–1939.

Visser, A.E., Verschure, P.J., Gommans, W.M., Haisma, H.J., Rots, M.G., 2006. Step into the groove: engineered transcription factors as modulators of gene expression. Adv. Genet. 56, 131–161.

Walsh, D.A., Doolittle, W.F., 2005. The real 'domains' of life. Curr. Biol. 15, R237–R240.

White, M.F., Bell, S.D., 2002. Holding it together: chromatin in the archaea. Trends Genet. 18, 621–626.

Wittschieben, B.O., Otero, G., de Bizemont, T., Fellows, J., Erdjument-Bromage, H., Ohba, R., Li, Y., Allis, C.D., Tempst, P., Svejstrup, J.Q., 1999. A novel histone acetyltransferase is an integral subunit of elongating RNA polymerase II holoenzyme. Mol. Cell 4, 123–128.

Woo, H.R., Pontes, O., Pikaard, C.S., Richards, E.J., 2007. VIM1, a methylcytosine-binding protein required for centromeric heterochromatinization. Genes Dev. 21, 267–277.

Woodcock, C.L., 2006. Chromatin architecture. Curr. Opin. Struct. Biol. 16, 213–220.

Yu, Z., Genest, P.A., Riet, B.T., Sweeney, K., Dipaolo, C., Kieft, R., Christodoulou, E., Perrakis, A., Simmons, J.M., Hausinger, R.P., van Luenen, H.G., Rigden, D.J., Sabatini, R., Borst, P., 2007. The protein that binds to DNA base J in trypanosomatids has features of a thymidine hydroxylase. Nucleic Acids Res.

Zeng, L., Zhou, M.M., 2002. Bromodomain: an acetyl-lysine binding domain. FEBS Lett. 513, 124–128.

Zhang, H., Christoforou, A., Aravind, L., Emmons, S.W., van den Heuvel, S., Haber, D.A., 2004. The *C. elegans* Polycomb gene SOP-2 encodes an RNA binding protein. Mol. Cell 14, 841–847.

Zhang, J., 2003. Are poly(ADP-ribosyl)ation by PARP-1 and deacetylation by Sir2 linked? Bioessays 25, 808–814.