



# Innovation in gene regulation: The case of chromatin computation

Sonja J. Prohaska<sup>a,b,\*</sup>, Peter F. Stadler<sup>a,b,c,d,e</sup>, David C. Krakauer<sup>b</sup>

<sup>a</sup> Bioinformatics Group, Department of Computer Science, and Interdisciplinary Center for Bioinformatics, University of Leipzig, Härtelstraße 16–18, D-04107 Leipzig, Germany

<sup>b</sup> Santa Fe Institute, 1399 Hyde Park Rd., Santa Fe, NM 87501, USA

<sup>c</sup> Max-Planck Institute for Mathematics in the Sciences, Inselstraße 22, D-04103 Leipzig, Germany

<sup>d</sup> Fraunhofer Institut für Zelltherapie und Immunologie–IZI, Perlickstraße 1, D-04103 Leipzig, Germany

<sup>e</sup> Department of Theoretical Chemistry University of Vienna, Währingerstraße 17, A-1090 Wien, Austria

## ARTICLE INFO

### Article history:

Received 22 December 2009

Accepted 6 March 2010

Available online 18 March 2010

### Keywords:

Chromatin

Gene regulation

Evolution

## ABSTRACT

Chromatin regulation is understood to be one of the fundamental modes of gene regulation in eukaryotic cells. We argue that the basic proteins that determine the chromatin architecture constitute an evolutionary ancient layer of transcriptional regulation common to all three domains of life. We explore phylogenetically, sources of innovation in chromatin regulation, focusing on protein domains related to chromatin structure and function, demonstrating a step-wise increase of complexity in chromatin regulation. Building upon the highly conserved use of variants of chromosomal architectural proteins to distinguish chromosomal states, Eukarya secondarily acquired mechanisms for “writing” chemical modifications onto chromatin that constitute persistent signals. The acquisition of reader domains enabled decoding of these complex, signal combinations and a decoupling of the signal from immediate biochemical effects. We show how the coupling of reading and writing, which is most prevalent in crown-group Eukarya, could have converted chromatin into a powerful computational device capable of storing and processing more information than pure *cis*-regulatory networks.

© 2010 Elsevier Ltd. All rights reserved.

## 1. Summary of findings and evolutionary hypothesis

Gene regulation in extant organisms is a complex adaptive system making use of a diversity of mechanisms to ensure that proteins and complementary macromolecular components are produced and maintained at functional levels in variable environments.

In this contribution we focus on one key regulatory subsystem of the cell: chromatin regulation. We argue that chromatin functioned as general regulator of transcription in the primordial nucleus, and that a series of key molecular innovations has significantly expanded the regulatory scope of the cell. In this section we summarize the key empirical results of the paper. Sections 2–6 provide the empirical support for these claims. In Section 7, we formulate an evolutionary hypothesis that seeks to explain our findings in terms of the evolution of proto-genetic regulation. We then interpret this evolutionary sequence in terms of increasing computational power by locating key stages of the evolutionary sequence within a formal model of computation. Since Sections 2–6 are primarily concerned with presenting

evidence, those interested in the conceptual development of the argument can focus on Sections 7 and 8.

Two variants of *chromosomal architectural proteins* (ChAPs) are sufficient to define binary genomic, and potentially phenotypic, states. We show how extensions to this binary system lead to potential distinctions among an increasing number of genomic states, culminating in forms of control localized to specific sites in the genome, as illustrated for example by extant mammalian histone variants capable of differential expression and/or localization to restricted chromosomal regions (Hake and Allis, 2006).

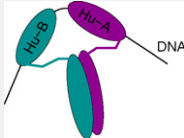
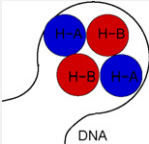
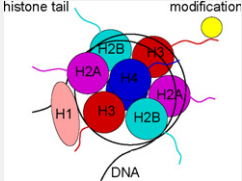
The recruitment of pre-existing modification enzymes able to modulate structural, or binding properties of ChAPs, represents a first significant regulatory innovation in chromatin. Modification of ChAPs has been widely reported across the Archaea and Eukarya, with many of the modification enzymes homologous in both sequence and function. Modification enzymes are also present in bacteria but it is not known to what extent bacteria can instruct chemical changes to ChAPs.

The second innovation in chromatin regulation is the appearance of a diverse set of protein structures that bind with high specificity chemical modifications of histones. These modification “readers” are wide spread, among eukaryotes. The emergence of “readers” qualitatively changes chromatin regulation. Chemical modifications function as chromosomal marks or signals. Modification enzymes thereby shift from a purely structural role to an informational one, taking on the role of “writers” and “erasers” of signals association with gene expression.

\* Corresponding author at: Bioinformatics Group, Department of Computer Science, and Interdisciplinary Center for Bioinformatics, University of Leipzig, Härtelstraße 16–18, D-04107 Leipzig, Germany.

E-mail addresses: [sonja@bioinf.uni-leipzig.de](mailto:sonja@bioinf.uni-leipzig.de) (S.J. Prohaska), [studla@bioinf.uni-leipzig.de](mailto:studla@bioinf.uni-leipzig.de) (P.F. Stadler), [krakauer@santafe.edu](mailto:krakauer@santafe.edu) (D.C. Krakauer).

**Table 1**  
Comparison of the main features of Chromosomal Architectural Proteins (ChAP) and their interaction with DNA in the three domains of life.

|                | Eubacteria  | Archaea   | Eukarya   |
|----------------|---|---|---|
| Histones       | –   | +   | +   |
| Histone tails  | –   | –   | +   |
| Other ChAPs    | HU  | HU, Alba  | –   |
| ChAP variants  | HU- $\alpha$ , HU- $\beta$  | H- $\alpha$ , H- $\beta$  | H3, H4, H2A, H2B  |
| ChAP dimers    | homo- (and hetero-) dimers  | (homo- and) hetero-tetramers  | hetero-octamers   |
| ChAP complexes |   |   | histone tail  |
| Structure      |  |  |  |
| Supercoiling   | negative  | positive or negative  | negative  |
| DNA contact    | 9 bp/complex  | 60 bp/complex   | 146 bp/complex  |
| Wrapping       | –   | ~ 1 turn  | ~ 2 turns   |

Writing and reading allow for signal propagation (along the genome) and epigenetic transmission (across generations). Chromatin level modification also supports structural memory, a partial independence from the underlying genomic layer of information, and a capability to interpret sets of signals as inputs to generate complex outputs of chromatin modification. The extant epigenetic system observed in higher eukaryotes allows for fast and flexible/reversible integration of environmental signals into the regulatory program of the genome. These mechanisms provide crucial support for cell differentiation in multicellular organisms.

The addition of these key functions can be seen to be associated with increasing computational power, shifting from a simple Markovian mechanism of regulation in the ancestral state, to a context sensitive system in several derived lineages, with capabilities perhaps exceeding those of sequence-mediated, *cis*-regulatory control.

2. A survey of extant chromatin regulation

2.1. The architecture of chromatin

DNA is never naked in a cell but associated with small, abundant, basic proteins, the ChAPs, that facilitate DNA organization and prevent DNA aggregation and tangling. This ensures efficient replication, segregation and gene expression. ChAPs are phylogenetically diverse.<sup>1</sup> The tight association of DNA and ChAPs is known as *chromatin*, and occurs in different compositions in all three domains of life (Travers and Muskhelishvili, 2005; Luijsterburg et al., 2008). The ChAPs with the widest phylogenetic distribution are HU and histones in the Eubacterial and Eukaryotic clades, Archaea favor two types of ChAPs from a set comprising HU, Alba, and histone (Luijsterburg et al., 2008) summarized in Table 1. In addition to these common proteins, many nucleoid-associated proteins with more restricted phylogenetic distributions are known from microbes (Sandman et al., 1998).

Histone proteins contain the histone fold, a 64 amino acid long helix–loop–helix–loop–helix motif stabilized through dimerization with a second histone fold. Functional characteristics

distinguish DNA packaging via histones from other architectural proteins. In histone containing species, DNA is wrapped around a histone complex, forming the *nucleosome* leading to a significant reduction of the contour length of DNA. First identified in Euryarchaeota, representatives of archaeal histones have subsequently been found in species from all major archaeal lineages (Cubonová et al., 2005; Slesarev et al., 1998). Most archaeal genomes encode 1–6 distinct histone proteins (Sandman et al., 1998; Bailey et al., 2002). Many Crenarchaea, however, completely lack histones. The only ChAP that is highly conserved in this phylogenetic group is Cren7 (Guo et al., 2008). Archaeal nucleosomes consist of a histone tetramer with about 60 bp of DNA wrapped around the histones approximately once. Eukaryotic nucleosomes consist of a histone octamer in which two H3–H4 heterodimers form a histone core homologous to the archaeal histone tetramer (Sandman and Reeve, 2000). In comparison to archaea, about twice as much DNA is wrapped around the octamer in two turns rather than one. Whereas archaeal histones introduce positive or negative supercoiling depending on salt concentration, temperature, and tetramer formation (Marc et al., 2002), eukaryotic nucleosomes always induce negative supercoiling (Table 1). Restricted to Eukaryotes are higher-order chromatin structures formed due to interactions between histone tails that protrude from the core and H1 (histone 1)—a non-histone fold linker protein. Linker histones in the protein configuration of Metazoans, appear first in late protists (Kasinsky et al., 2001). Dinoflagellates have secondarily lost their histones and do not form nucleosomes (Moreno Díaz de la Espina et al., 2005). Their major ChAP, HCC, is most closely related to bacterial HU proteins (Wong et al., 2003).

Eubacteria do not as a rule have proteins with histone-folds (Reeve et al., 2004). An exception are the homologs of the non-histone-fold protein histone H1 in Chlamydiae (Grieshaber et al., 2004; Murata et al., 2007) likely acquired by lateral transfer from a eukaryotic host. HU and its homologs are phylogenetically the most widespread ChAPs in Eubacteria. On average, three HU variants per genome are found, Table 1, which encode monomers that form homo- and heterodimers (Pinson et al., 1999). HU binds the DNA backbone in a sequence-unspecific manner, sharply bending DNA introducing negative supercoils. In *Escherichia coli*, HU dimers bind on average every 290 bp but form dense arrays with a dimer every 9 bp *in vitro* (Broyles and Pettijohn, 1986). Surprisingly, HU is not essential for viability (Dri et al., 1991), as it can be replaced by alternative ChAPs (Azam and Ishihama, 1999).

<sup>1</sup> Groups of unspecific DNA binding proteins, such as single-stranded DNA binding proteins or RecA are usually not counted as ChAPs.

To-date, HU homologs have been characterized in a few Archaea (Orfaniotou et al., 2009), and in a variety of eukaryotes including *Giardia lamblia* (Triana et al., 2001), Dinoflagellates (Wong et al., 2003) and Apicomplexa (Arenas et al., 2007).

Alba (acetylation lowers binding affinity) is a candidate non-histone ChAP (Bell et al., 2002; Wardleworth et al., 2002). It was first found in Archaea, where it is one of the most abundant proteins in thermophilic and hyperthermophilic species. Eukaryotic homologs in human, green plants, and protists have since been reported. Depending on its concentration and conformation, the Alba either bridges two DNA duplexes or cooperatively binds to a single DNA duplex (Noom et al., 2009). Alba also binds to ssDNA and RNA *in vivo* (Guo et al., 2008), leading some to question its role as a chromatin architectural protein.

In Eubacteria, binding by alternative ChAPs leads to different supercoiling patterns of DNA and to different spatial organizations of the nucleoid (Thanbichler et al., 2005). The binding affinities of ChAPs differ by at least an order of magnitude, exhibit varying levels of sequence specificity and can be sensitive to DNA curvature (Azam and Ishihama, 1999).

A fundamental difference between the chromatin organization of Eubacteria and the Eukarya/Archaea clade is that the latter two domains form multimeric nucleosomes that have DNA wrapped around themselves, whereas eubacterial DNA is organized into loops with dense regions in complex with ChAPs. The complexes of DNA and ChAPs are referred to as nucleoid in both Archaea and Eubacteria despite structural differences.

## 2.2. ChAP variants

ChAPs are frequently present in paralogous copies. The number of ChAP variants is limited in prokaryotes (on average 1–3), whereas eukaryotic genomes can encode a large number of differentially expressed paralogs (Malik and Henikoff, 2003).

Different cellular forms of microbes typically correlate with different nucleoid structures and differential abundance of the various nucleoid proteins (Travers and Muskhelishvili, 2005). As with HU, the abundance of certain histone proteins changes with cell states in Archaea and Eukarya. Paralogs can influence the degree of compaction introduced into DNA (Sandman et al., 1998). H2A.Z, for example, is less stable than H2A and inhibits spreading of silent chromatin. Several paralogs/variants exist for all histone types in Eukarya. Phylogenetic analysis suggests that H3 and H2A are more variable than H4 and H2B, respectively (Malik and Henikoff, 2003; Thatcher and Gorovsky, 1994).

Functional diversification has been demonstrated for all variants except H4. Deposition of these proteins is dependent on the cell cycle (replication-dependent and replication-independent histone variants, H3 and H3.3, respectively) and cell type (sperm- and testis specific variants of H1 and H3). Alternative variants localize to genomic regions (Ahmad and Henikoff, 2002; Brown, 2001; Kamakaka and Biggins, 2005; Malik and Henikoff, 2003; Wenkert and Allis, 1984), e.g. CENP-A, a H3 variant, and MacroH2A are deposited at centromeres and the inactive X chromosome, respectively. Hv1 and Hv2 are H2A and H3 variants exclusively used in the macronucleus of *Tetrahymena thermophila*. Trypanosomes use specific combinations of numerous histone variants to ensure demarcation of polycistronic transcription units (Siegel et al., 2009). In addition, ChAPs can deviate from non-specific DNA binding, and binding preferences of each variant can diverge. Nucleosomes, for example, have a slight sequence preference (Bailey et al., 2000; Lowary and Widom, 1998), favoring DNA that bends more easily. Furthermore, they tend to influence the positioning of neighboring nucleosomes (Rando and Ahmad, 2007). Paralogous ChAPs are known to modulate the

sequence preference for nucleosome positioning (Bailey et al., 2002; Marc et al., 2002).

A similar diversification of ChAPs is observed in Archaea and Eubacteria. Some Archaea have two histone variants, H- $\alpha$  and H- $\beta$ . H- $\beta$  has been shown to induce greater compaction (Grayling et al., 1996). Heterodimers occur predominantly in the stationary growth phase, whereas the (H- $\alpha$ )<sub>2</sub> homodimers are prevalent during exponential growth phase (Sandman et al., 1994). Alba often appears in two paralogs, with heterodimers promoting DNA compaction as with histone variants (Jelinska et al., 2005).

In *E. coli* and related enteric bacteria, the two HU variants  $\alpha$  and  $\beta$  form homo- and heterodimers.  $\alpha\beta$  and  $\alpha_2$  dominates the exponential growth phase. A shift to higher  $\beta$  concentrations leads to formation of mainly  $\alpha\beta$  and stationary phase (Claret and Rouviere-Yaniv, 1997).

## 3. The regulation of transcription by chromatin

Chromatin structure has an immediate effect on local transcriptional activity. In Eubacteria, distinct nucleoid structures are associated with different cellular forms, such as vegetative cells or spores. At the onset of the stationary phase, bacterial chromatin undergoes a massive reorganization into ordered toroidal structures through a process dictated by the intrinsic properties of DNA and the ubiquitous, starvation-induced DNA-binding protein Dps (Frenkiel-Krispin et al., 2004). Eubacteria appear to have evolved a hierarchy of nucleoid-associated factors with each spanning a different range of sequence specificity: less specific variants act as functional backups for the more specific variants (Pérez-Martín and de Lorenzo, 1997; Azam and Ishihama, 1999). Non-linear DNA structures promote signal integration paralleling the transduction cascades employed by higher organisms to control cell growth and differentiation (Pérez-Martín and de Lorenzo, 1997). Eubacterial ChAPs such as H-NS, FiS, IHF, and SLPa, therefore, act as general regulators of transcription (Dame et al., 2006; Perez et al., 2008; Lucchini et al., 2006).

DNA occupied by nucleosomes, and structured into chromatin blocks, slows down initiation and elongation of transcription compared to naked DNA (Armstrong, 2007; Izban and Luse, 1992). This effect is strongest in eukaryotic octamers. However, for Eukaryotic (H3–H4)<sub>2</sub> tetramers as well as archaeal nucleosomes, DNA wrapped around histone tetramers is comparatively easy to access (Sandman and Reeve, 2000). For initiation of transcription, activators usually bind the unoccupied “linker” DNA between nucleosomes (Rando and Ahmad, 2007). Binding sites covered by a nucleosome are largely inaccessible to binding factors unless nucleosomes are destabilized, moved or evicted. Chromatin is understood as a negative regulator of transcription. A major distinction is made between *open* and *closed*, *active* and *silent* chromatin. Open chromatin correlates with active gene expression (Ryan et al., 1998). Nevertheless, chromatin opening and gene activation are distinct processes (Schübeler et al., 2000). Closed chromatin, on the other hand, does not necessarily impair binding of transcription factors (Ryan et al., 1998). In fact, transcription factors are able to open chromatin (Lomvardas and Thanos, 2002; Cirillo et al., 2002). Reduced transcriptional activity is often associated with compacted “closed” nucleoid structures in Eubacteria (Frenkiel-Krispin et al., 2004).

Hence chromatin does not exclusively determine the onset or efficiency of transcription in most species. Only trypanosomatids rely largely on chromatin, and in particular, on variations in nucleosome compositions (Talbert and Henikoff, 2009; Siegel et al., 2009) to regulate gene expression. In nearly all organisms, the local chromatin composition sets the stage for further, more

specific mechanisms of regulation, such as sequence-specific transcription factors (Li et al., 2007). ChAPs therefore (nearly universally) act as general regulators of transcription.

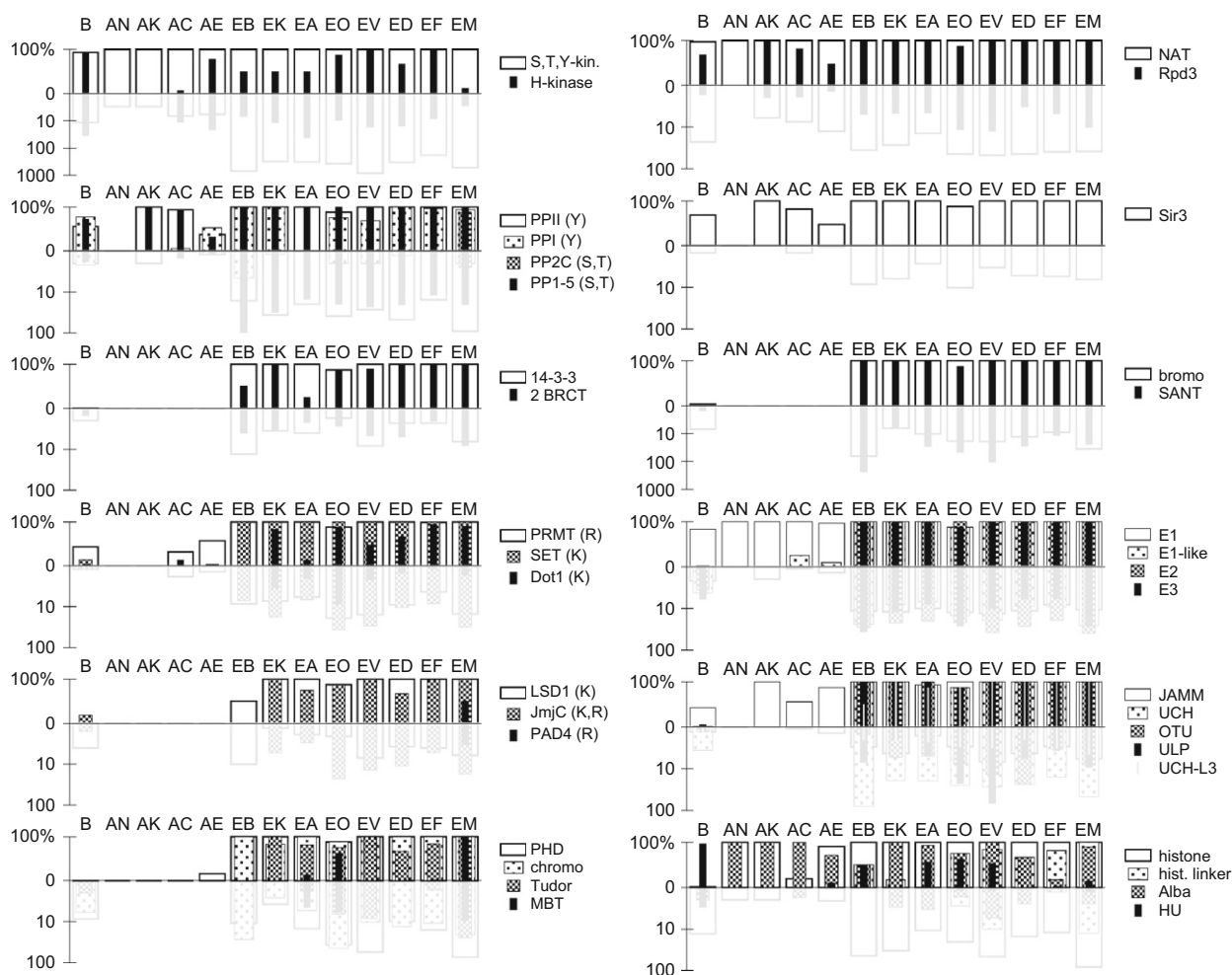
#### 4. Chemical (de-)modifications of ChAPs in the three domains of life

Chromatin destabilization and subsequent transcriptional activation can be achieved either by the exchange of ChAP variants or by chemical modifications of ChAPs. As we have described above, the first mechanism is present in all three domains of life. By contrast, chemical modifications of ChAPs do not appear uniformly. We have performed a comprehensive evaluation of the phylogenetic distribution of key modification enzymes making use of the SCOP and SuperFamily databases (Andreeva et al., 2008; Wilson et al., 2009), and complemented

these findings with an extensive literature survey of the topic (see Appendix A). The results are summarized in Fig. 1 (see also Supplementary Tables) and commented upon in the following subsections. Overall, chemical modifications of ChAPs are prevalent in Archaea and Eukarya. In Eubacteria, only a few reports assert that phosphorylation (Cozzone, 2009) and hydroxybutyrylation (Norris, 2005) are present in this domain.

Chemical modifications of ChAPs can function into two fundamentally different ways: (1) by means of direct influence on the thermodynamic stability of chromatin, and (2) by means of active disassembly, eviction, or mobilization performed by enzyme complexes (Henikoff, 2008; Lesne, 2006).

In Archaea, the first, thermodynamic mode is predominant. For example, Alba—which stands for “acetylation lowers binding affinity”—is acetylated and de-acetylated to decrease and increase the binding affinity of Alba to DNA. A regulatory role for direct chemical effects can also be observed in Eukarya.



**Fig. 1.** Phylogenetic distribution of functional domains involved in regulation of expression by chromatin. The upper panel of each individual graph shows the fraction of species per clade containing at least one protein with the functional domain specified to the right. The lower panel depicts the average over the absolute number of proteins for these species. While all domains of life utilize ChAPs to construct chromosomes, and modification writers (in particular phosphorylation, and acetylation) are present ubiquitously, we observe that modification readers are restricted to Eukarya. *Left panel from top to bottom:* Enzymes for phosphorylation: Serine/Threonine/Tyrosine-kinases and Histidine-kinases. Enzymes for de-phosphorylation: Tyrosine-phosphatases I and II, Serine/Threonine-phosphatases 1–5 and 2C. Reading of phosphorylation marks by 14-3-3 and tandem BRCT domains. Enzymes for methylation: protein arginine methyltransferases (PRMTs) and lysine methyltransferases SET and Dot1. De-methylation enzymes homologous to LSD1 or JmjC and de-amination enzymes homologous to PAD4. Reading of methylation marks by PHD fingers, chromo, Tudor, and MTB domains. *Right panel from top to bottom:* Enzyme for acetylation: N-acetyltransferase (NAT). De-acetylation enzymes homologous to Sir3. Reading of acetylation marks (bromo domain) and unmodified side chains (SANT domain). Enzymes E1, E2 and E3 required for ubiquitination. Enzymes for de-ubiquitination: JAMM metalloprotease, Ubiquitin carboxyl-terminal hydrolases Otubain and Ubiquitin-like protein-specific proteases. *Chromosomal architectural proteins (ChAPs):* histones, linker histones, and homologs of Alba and HU. Phylogenetic clades: B—Eubacteria; Archaea: AN—Nanoarchaeota; AK—Korarchaeota; AC—Crenarchaeota; AE—Euryarchaeota; Eukarya: EB—basal eukaryotes (e.g. Diplomonadida and Parabasilia); EK—Kinetoplastida (including Heterolobosea); EA—Alveolata (i.e. Ciliophora and Apicomplexa); EO—Chromista (i.e. Cryptophyta, Haptophyta, and Heterokonta) (Yoon et al., 2002); EV—Viridiplantae (incl. Chlorophyta); ED—Amoebozoa (e.g. Dictyostelium); EF—Fungi; EM—Metazoa. Note that the groups AN (Nanoarchaeota) and AK (Korarchaeota) each are represented only by one species.



Acetylation of histone H3 and H4 in Eukarya leads to chromatin opening/activation (Roh et al., 2005). As demonstrated for yeast, the level of gene expression is dependent on the total number of acetyl groups on H4K5, H4K8 and H4K12, whereas the sites themselves are interchangeable (Dion et al., 2005). This observation of degeneracy has been seen as a challenge to the idea of a complex, chromatin code (Henikoff, 2005).

Several unrelated protein families are responsible for each type of ChAP modification. In the following sections, we examine phosphorylation, acetylation, methylation, and ubiquitination in some detail. In the presence of modification “readers”, i.e. in the eukaryotes, we refer to modifiers functionally as “writers” and “erasers” describing the addition and removal of chemical modifications.

#### 4.1. (De-)phosphorylation

Protein phosphorylation involves chemical modifications to serine/threonine/tyrosine and histidine residues. The corresponding enzymes performing these reactions can be partitioned in histidine and serine/threonine/tyrosine protein kinases. The latter constitute a part of one protein superfamily and are the most common and abundant modification enzymes throughout all domains of life contributing to histone phosphorylation. Histidine phosphorylation on histones has also been reported for H4 in mammalian tissues (Besant and Attwood, 2000; Tan et al., 2004). Phosphate groups can induce conformational changes and are used to activate enzymes and serve as signals. Phosphorylation cascades provide rapid response kinetics (e.g. heat shock response) and reversible action. Determining which from Ser/Thr/Tyr kinases phosphorylate ChAPs is difficult to estimate since reported histone kinases (e.g. Aurora, Polo, Nek, or Haspin, Dai and Higgins, 2005) cannot be distinguished from other protein kinases based on family-level domain descriptions.

Phosphorylation of serine, threonine and tyrosine is reversed by phosphatases utilizing different protein domains and molecular mechanisms. In a phylogenetic respect, phosphatases co-occur with kinases. Nevertheless, genomes contain fewer phosphatases than kinases (in Eukarya by a factor of 10, Fig. 1) (Kennely, 2002, 2003). This suggests that a highly specific phosphorylation machinery might be counteracted by a fairly general de-phosphorylation machinery.

#### 4.2. (De-)acetylation

Protein lysine residues can be modified by acetylation or methylation. In general, acetylation of histones is known to correlate with transcriptional activity. However, certain sites, such as H4K16 in yeast, contradict this remark and seem to be strongly context dependent (Dion et al., 2005). Eukaryotic HATs (histone acetyltransferases) belong to the (super)family of NATs (acetyl-CoA N-acetyltransferases) among which are relatives of GCN5 and Elp3, MYST, p300/CBP, TAF<sub>II</sub>250 and nuclear hormone-related HATs (Pandey et al., 2002). NAT members do not only perform protein/histone acetylation. Bacterial members of this superfamily can acetylate antibiotics (Vetting et al., 2004). Pats (protein acetyltransferases) from Eubacteria, serve as regulators of metabolic enzymes (Starai and Escalante-Semerena, 2004) and might have been co-opted into the role of specific regulators of Alba within Archaea (Marsh et al., 2005). According to our analysis, NAT-domain containing acetyltransferases are phylogenetically as widespread as kinases. Prokaryotes have comparable numbers of kinases and NATs. In contrast, eukaryotes have on average 10 times more kinases than acetyltransferases. Taken

together, this suggests that acetyltransferases originated close the root of the three domains of life.

Two structurally different domains are commonly employed for protein de-acetylation in all domains of life. Sir2 family enzymes, sirtuins (silent information regulators), are found in all three domains of life (Frye, 2000; Smith et al., 2000; White and Bell, 2002; Buck et al., 2004) but are missing from genomes of many prokaryotes. In Eukarya and Archaea, these enzymes are NAD<sup>+</sup> dependent and suggest an ancient link between chromatin structure and the metabolic state of the cell. Evidence for this comes from the conserved functional role of Sir2 even though it de-acetylates different substrates in eukarya and archaea, namely, histones and Alba, respectively (Bell et al., 2002). Rpd3-like deacetylases have no similarity to Sir2 proteins. They are found in all eukaryotic genomes and have distant homologs in archaea and bacteria (Gregoret et al., 2004; Taunton et al., 1996; White and Bell, 2002; Rundlett et al., 1996; Pandey et al., 2002). Another structurally different class is HD2-type histone deacetylases which appear to be present only in plants (Pandey et al., 2002).

We find that only a small fraction of prokaryotic species have at least one deacetylase, and that the numbers of these enzymes are very small in each species. In contrast, deacetylases are ubiquitous in Eukarya. It is interesting to note that there are at least twice as many acetyl-transferases than deacetylases per eukaryotic genome.

#### 4.3. (De-)methylation

Methylation can be manifold: Lysine residues can be mono-, di- or tri-methylated and arginine residues are subject to mono-methylation and symmetric or asymmetric dimethylation. All of these have been observed to play a role in chromatin modification (Cheng et al., 2005). Usually, the substrate for the enzymatic reaction is either S-adenosyl-L-methionin (AdoMet) or S-adenosyl-L-homocysteine (AdoHcy) (Schubert et al., 2003). However, the structural class of SAM-MTases does not include all AdoMet utilizing methyltransferases. Methylation of lysine can be carried out by SET (abbreviation for “Su(var), enhancer of zeste, trithorax”) (Dillon et al., 2005) or Dot1 domain containing proteins. Less than one-third of prokaryotic organisms have such proteins. There are no SET domain containing proteins in Archaea (White and Bell, 2002), consistent with our analysis. The only exception is a SET domain protein in *Candidatus Methanoregula boonei* which is homologous to bacterial Histone-lysine N-methyltransferases, probably acquired horizontally. We observe 3–15 times more Set than Dot1 domain proteins in Eukarya. Only Stramenopiles have at least one Dot1 domain in each genome. Several lysine residues of the archaeal ChAP Sul7d are subject to mono-methylation by an hitherto unknown enzyme (White and Bell, 2002; Marsh et al., 2005). Furthermore, a SAM-MTase in the archaeon *Pyrococcus horikoshii* and its orthologs has been suggested to methylate rRNA and tRNA (Sun et al., 2005).

Methylation of arginine is carried out by PRMTs (protein arginine-methyltransferase) (Cheng et al., 2005; Bedford and Richard, 2005). More than 95% of prokaryotes have at least one PRMT and all Eukaryotes examined have at least two. The average number of paralogs per species is always lower than that of acetyltransferases. Histone modifying PRMTs do not cluster within the family of PRMTs (Bedford and Richard, 2005).

Eubacteria can have DNA, RNA as well as protein methyltransferases in their genomes. DNA and RNA methyltransferases are involved in defense against foreign DNA or antibiotics (Long et al., 2006). Furthermore, protein methylation is found to play a role in sensory adaptation, where methylation is responsible for

signal transduction and implementation of a short term memory (Taylor, 2004).

Histone methylation has been thought irreversible until an amine oxidase reaction, carried out by LSD1 (lysine-specific demethylase), was proposed to cause de-methylation of mono- and di-methylated lysine residues (Shi et al., 2004; Trewick et al., 2005). At about the same time a hydroxylase with a JmjC domain was discovered able to de-methylate also tri-methylated substrates in a radical reaction (Tsukada et al., 2006; Trewick et al., 2005). The cell pays a high price for de-methylation as both reactions produce formaldehyde, a substance toxic to living cells because of oxidative stress. Deimination of methyl-arginine, carried out by PAD4 (peptidylarginine deiminase 4), also removes methylation marks but does not recover the arginine residue (Wang et al., 2004). Although methylation is not irreversible, one expects massive reversal of methylation to be rare.

Fig. 3 shows that there is a global correlation between the number of modification and demodification enzymes in a given organism. With the exception of ubiquitination, “erasers” are substantially less frequent than “writers”. Somewhat surprisingly, this difference is only moderate for methylation/demethylation.

#### 4.4. (De-)ubiquitination

Protein ubiquitination denotes the attachment of one or more ubiquitin molecules to an amino acid side chain, commonly a lysine. Ubiquitin is itself a small protein (76 AA in human) specific to eukaryotic cells. Similar polypeptides like SUMO (small ubiquitin-like modifier) and RUB1 (related-to-ubiquitin 1) can be attached to proteins in similar enzymatic reactions, termed sumoylation and rubylation (Hochstrasser, 2000). The ubiquitination process requires a cascade of enzymatic reactions carried out by an activation enzyme (E1), a conjugation enzyme (E2), and sometimes a protein ligase (E3) (Hochstrasser, 2000). The functional consequences of the latter modifications are just beginning to be determined. In general, poly-ubiquitination is a marker for protein degradation, while mono-ubiquitination plays a role in signaling and has been found in all major histone families and several histone variants (Wang et al., 2006; Weake and Workman, 2008). Recent studies suggest that histone ubiquitination is a universal response to DNA damage and induces DNA damage repair (Zhou et al., 2009).

Deubiquitination enzymes and other proteases, which detach ubiquitin-like proteins, belong to either metalloproteases or cysteine proteases. At least seven protease lineages with different evolutionary origin contain proteases with a cysteine in the catalytic site (Barrett and Rawlings, 2001). Two contain the following protease domains relevant for de-ubiquitination: UCH and UCH-L3 (ubiquitin C-terminal hydrolases), OTU, and ULP (Ubiquitin-like protease). Other domains, e.g. Josephin, could not be found in the Superfamily database (see Appendix A for details).

Ubiquitination, de-ubiquitination and ubiquitin-like proteins have long thought to be completely absent in prokaryotes (Eichler and Adams, 2005). The discovery of small prokaryotic proteins with a  $\beta$ -grasp fold, as in ubiquitin, and C-terminal activation by E1-like enzymes have changed this picture and suggest derivation of the ubiquitin conjugation system from the more ancient sulfur transfer pathway (Hochstrasser, 2009). In addition, all canonical domains of life and viruses have members of both types of protease, cysteine and metallo-proteases. This strengthens the evidence that ubiquitin-like proteins, and the corresponding conjugation system and proteases (Nijman et al., 2005), trace back to the root of the phylogenetic tree.

The picture for the phylogenetic distribution of (de-)ubiquitination enzymes is very similar to other marks. Both relatives of

some modifiers and demodifiers can be found in nearly all domains of life. The minimal set of necessary modifiers (E1 and E2) is present but still rare in archaea. The ratio between modifiers and demodifiers is different from other marks on the side of the modifier. The number of de-modifiers is sometimes higher than that of the modifiers, Fig. 3.

Summarizing this subsection, we detect a few general patterns. Most importantly, there is strong evidence that the evolutionary origin of chemical modifications leading to functional diversification of ChAPs predates the origin of eukaryotes (Bell et al., 2002). These are common to (Eukarya+Archaea) but exclude Eubacteria (White and Bell, 2002). Furthermore, the repertoire of modification writers is much more elaborate than that of modification erasers. In particular, not all Archaea appear to make use of erasers. Even in Eukarya, the use of writers and erasers is far from being balanced, Fig. 3. This could be explained by the fact that modified ChAPs can also be removed by local replacement with unmodified proteins and subsequent degradation of the modified ones.

### 5. The evolution of molecular literacy: modification readers

A major innovation that sets Eukarya apart from the other two domains is the invention of an elaborate signaling system based on chemical modification of ChAPs. In contrast to prokaryotes, chemical modification can be both placed and removed by specific enzymes and be recognized specifically. Reading is chemically and logically distinct from erasing modification: chemically, eraser enzymes recognize modifications with their active center; logically, a modification can be recognized only once by an eraser. In contrast, readers recognize the chemical modification and its context with an enzymatically inert domain; since the modification remains unchanged and can be read arbitrarily often.

#### 5.1. The diversity of reader domains

It is an intriguing observation that each type of histone modification is recognized by several distinct, apparently unrelated, protein domains (de la Cruz et al., 2005; Taverna et al., 2007). We briefly summarize the major reader domains:

Bromodomains (Mujtaba et al., 2007) read acetylation marks on lysine residues and are found in chromatin-associated proteins such as modifiers and remodellers. In addition, non-histone acetylation marks can be targeted. Proteins containing a double bromodomain bind two acetylation marks in a critical topological configuration, increasing specificity and affinity.

Methylation marks can be found on lysine and arginine residues. Lysine can be found in mono-, di- and trimethylated states *in vivo*. Several domains (e.g. chromodomains, Eisenberg, 2001, chromo barrels, Xu et al., 2008, tudor, and MBT, Adams-Cioaba and Min, 2009) of the Royal superfamily can read lysine methylation marks and distinguish between the number of methyl groups. PHD-fingers also specifically recognize either methylated or unmethylated lysine residues. The key residues are often located at protein termini that can be inserted into the binding pocket.

Methylation marks can also be placed on arginine. Mono- as well as symmetric and asymmetric di-methylation can be found. Information on the corresponding readers is very limited, and no specific recognition proteins are known. WD40 repeats have evolved to detect many different signals, among them, histone methylation marks on lysine and arginine and histone phosphorylation marks. The PHD domains found in the archaeal genomes hit a hypothetical protein of *Methanococcus maripaludis* and the

RecJ-like exonuclease of a few *Pyrococcus* species suggesting that these matches are almost certainly false-positives.

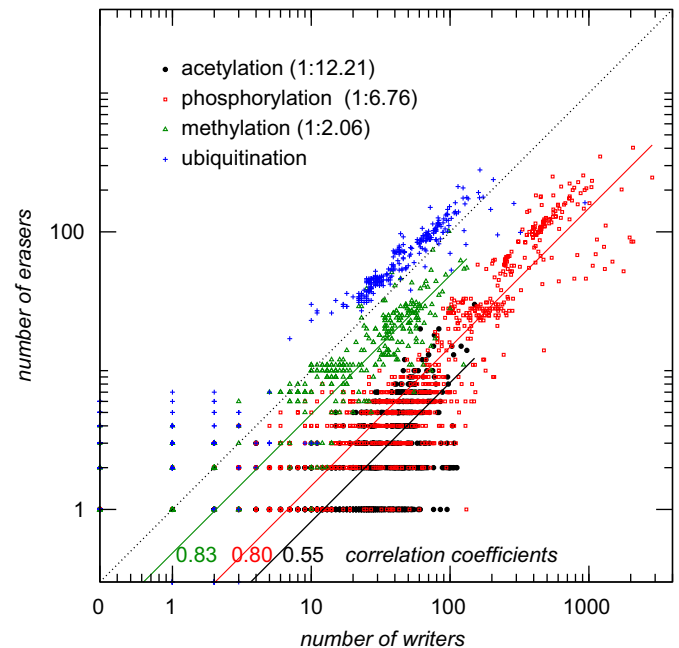
14-3-3 proteins (Morrison, 2009) are the most common readers of phosphoserine residues. Found in all eukaryotes, they are absent from all prokaryotes examined (Ferl et al., 2002). A tandem BRCT domain (Singh et al., 2008) can also recognize phosphorylation marks at the interface of domains. In budding yeast, readers of the latter type are, e.g. involved in cell-cycle arrest in response to DNA damages marked by H2AX phosphorylation (Hammet et al., 2007). Only single BRCT domains can be found in eubacteria and archaea.

As for the modifying enzymes, we evaluated the phylogenetic distribution of reader domains in detail, Fig. 1. The most striking observation is that modification readers are almost exclusively found in eukaryotes. The few exceptions, mostly in bacteria, are found in a handful of species, who likely acquired them via horizontal gene transfer from an eukaryotic host. Among the major Eukaryotic clades, there is little variation in the distribution of reader domains, and there appears to be no or at most very little (sub)kingdom-specific innovation of relevant protein domains within the Eukarya. The number of genomic copies of particular domains, on the other hand, can vary by more than an order of magnitude between different clades, indicating clade specific patterns of diversification. The repertoire of phosphorylation readers (14-3-3 and 2BRCT domains) is almost two orders of magnitude smaller than the typical number of kinase domains, since kinases have functions beyond the regulation of chromatin. For methylation and acetylation, the copy numbers of modifiers and readers are comparable.

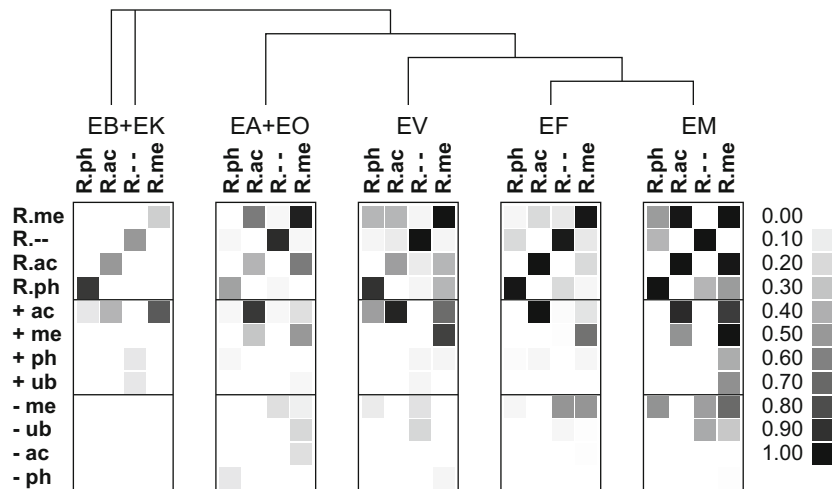
Many eukaryotic chromatin-associated enzymes combine two or more distinct domains, Fig. 2. The combination of reader domains is likely to increase specificity and allows for recognition of more diverse modification patterns. Our data show that the use of reader domain combinations increases towards the crown-group Eukarya. While combinations of domains recognizing the same modification are present in most species, combinations of methylation and acetylation readers, for instance, are most frequent in Chromalveolata and Metazoa.

Previous studies have reported an enrichment of reader domains in chromatin modifying enzymes (de la Cruz et al., 2005). The analysis of the co-occurrence of reader and writer

domains shows that acetylation is often coupled to reading of methylation marks. Animals in particular, appear to have an elaborate system of directing all four major types of modification, dependent on methylation and acetylation marks. Fungi, on the other hand, largely restrict themselves to writing methylation and acetylation marks depending on existing methylation and acetylation marks. Since marks may be set at nucleosomes adjacent to those that the enzymes read, the combination of reader and writer domains provides a means for the autonomous propagation of histone modification in response to



**Fig. 3.** Correlation between the number of modification (writer) and demodification (eraser) enzymes. The effect is largest for acetylation with a ratio of more than 12 and phosphorylation with a ratio of more than 6. For methylation the ratio is much smaller but still significant. In contrast, the bulk of species has slightly more ubiquitination erasers than writers. See text for more details.



**Fig. 2.** Phylogenetic distribution of protein domains co-occurring with reader domains. Reader domains are more often coupled with each other and with modification enzymes in crown-group Eukarya. Metazoa, in particular, have an extensive repertoire of such combinations that is highly conserved within the kingdom. The grayscale value indicates the fraction of species in a clade that has at least one protein containing the domain combination specified by the left and top index of the matrices. 'R.me'—methylation reader; '+ me'—methylation enzyme; '- me'—de-methylation enzyme. Indices are analogous for ac—acetylation; ph—phosphorylation; ub—ubiquitination; 'R.--'—reader of an unmodified side chain. EB+EK—basal eukaryotes and kinetoplastids; EA+EO—Chromalveolata (Martens et al., 2008); EV—Viridiplantae; EF—Fungi; EM—Metazoa.

histone-modification-dependent target-selection (Forneris et al., 2005). Combinations of readers and erasers are comparably rare. Only animals seem to use combinations of modification readers and demethylation domains in a systematic way.

Over all, we observe an increase in the diversity of combinations of reader, writer and eraser domains in “higher” eukaryotes. It is also evident that combinations are more uniformly adopted in animals, and to a lesser extent in plants.

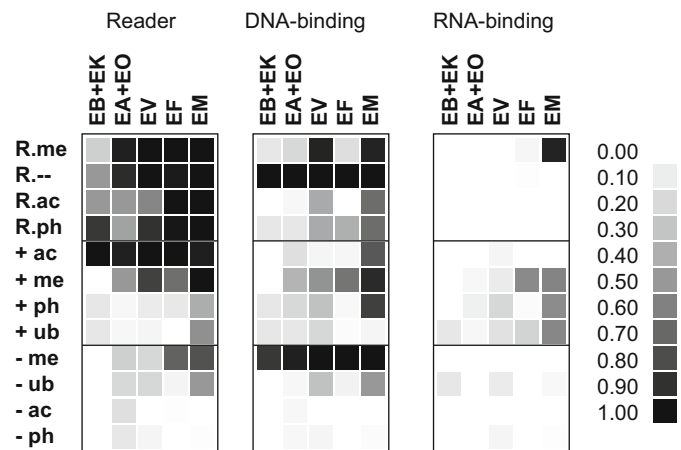
We emphasize that the analysis of domain co-occurrences does not exhaust the full complexity of the chromatin modification machinery. We anticipate protein complexes comprising several proteins with reader and writer domains permitting additional combinations. The domain co-occurrences nevertheless provide information on general evolutionary trends, and on functionalities that have preserved tight coupling over long evolutionary time scales.

## 5.2. Guiding chromatin-modification to DNA loci

The function of chromatin in the regulation of transcription requires a connection between chromatin, the placement of variant nucleosomes, chemical modification, and the underlying sequence of genomic DNA. How this connection is established in practice, however, is by no means well understood at present. This is in part due to the diversity of documented mechanisms, which involve *cis*-acting DNA elements, nascent transcripts, and trans-acting RNAs. A particular location may be recognized in principle by any combination of DNA sequence information, chromatin state, and local genome activities (e.g. transcription). Corresponding binding domains in chromatin modification complexes may give a hint on how the spatial positioning is achieved, Fig. 2. Nucleosome positioning further complicates the picture since it is influenced, but apparently not determined by DNA sequence features (Jiang and Pugh, 2009; Washietl et al., 2008; Segal and Widom, 2009).

Little is known about the detailed mechanisms that lead to the position-specific deposition of variant histones or other ChAPs. Histone chaperones and ATP-dependent remodeling complexes, which typically contain reader domains, are implicated in this process (Jin et al., 2005).

Similar to promoter elements or transcriptional enhancers, chromatin-opening elements (Schübeler et al., 2000), such as the HSFE involved in the regulation of  $\beta$ -globin expression (Nemeth and Lowrey, 2004), have been found that are involved in chromatin opening. Furthermore, interaction of sequence-specific DNA-binding transcription factors with chromatin modification enzymes has been proposed (Cirillo et al., 2002). The middle panel of Fig. 4 shows that reader, writer, and eraser domains for specific modifications are frequently associated with DNA binding domains, demonstrating that there are extensive protein families that bring together DNA sequence information and chromatin modifications. In particular, histone demethylation appears to be firmly linked to DNA binding in all eukaryotes, implying that this process is at least in part directed by information residing in the DNA sequence. Only fungi and animals have a systematic repertoire of demethylases that are tied to modification readers. There is a general trend towards a tighter association between protein domains operating on chemical modifications and DNA binding domains in animals and to a lesser extent in plants. Besides DNA sequence-dependent mechanisms and DNA-methylation dependent mechanisms (Tariq and Paszkowski, 2004; Freitag and Selker, 2005), process-dependent modifications have been described that store the recent local transcriptional history to chromatin states (Henikoff, 2008).



**Fig. 4.** Evolutionary trends in domain combination of chromatin (de-)modifiers and three potential target selectors. *L.h.s.*: Combinations of reader domains are prevalent in all major Eukaryotic clades, while the coupling of reader and modifier domains increases towards the animals. In particular acetylation and methylation is coupled with reader domains. *Middle*: DNA binding domains are frequently coupled with modifier domains. In particular, demethylation appears to be governed by DNA information. Interestingly, a large class of proteins, which again increases towards animals, combines DNA binding with the reading of specific modifications. *R.h.s.*: The combination of RNA binding domains with chromatin modification shows a clear increase towards the animal clade. Abbreviations as in Fig. 2.

Several studies have demonstrated that siRNAs can promote DNA methylation as well as specific chromatin modifications in a number of lineages (Chan, 2008; Klenov et al., 2007; Liu et al., 2007; Verdel et al., 2009). Small RNAs are involved in targeting by direct interaction with the DNA or with nascent RNAs. While heterochromatin formation in *Schizosaccharomyces pombe* is guided by RNA–RNA recognition (Bühler et al., 2006), plants appear to rely on direct RNA–DNA binding (Pélissier and Wassenegger, 2000). In ciliates, small scnRNAs are produced from the entire micronucleus, compared to the macronuclear DNA and degraded if the sequences match. After sexual reproduction, the retained scnRNAs target the DNA of the newly formed macronucleus to identify the “internal eliminated sequences” IES, trigger histone methylation, and eventually the excision of the corresponding DNA regions (Kurth and Mochizuki, 2009). Plants have evolved an elaborate transcriptional machinery dedicated to eliciting sequence-specific, chromatin-based gene silencing (Matzke et al., 2009).

Long, mRNA-like ncRNAs (lncRNAs) play a crucial role in imprinting and other chromatin-level regulation that contribute to cell fate (Pandey et al., 2008; Leeb et al., 2009). Recently lncRNAs have been identified as crucial components in the polycomb/trithorax regulation system. The interaction of PcG and TrxG proteins with their target sites, the PRE/TREs (polycomb/trithorax responsive elements) can be achieved by several distinct mechanisms (Hekimoglu and Ringrose, 2009) that are also employed by other modification enzymes. The modifier complex may bind directly to a nascent complex, also proposed e.g. for the histone acetylase ASH-1 in flies (Sanchez-Elsner et al., 2006). Non-coding RNA may associate with their protein partners independent of the chromatin and guide them back to a target *in cis* or *in trans* recognized by RNA–DNA binding. This mode of action likely guides *trans*-actions such as the silencing of the HOXD cluster by HOTAIR (Rinn et al., 2007). Finally, nascent anti-sense transcripts might form anchors for ncRNAs-modifier complexes as suggested by the frequent bi-directional transcription from PRE/TREs. Recent data show that almost a quarter of the human lncRNAs are physically associated with the repressive chromatin-modifying



complex PRC2 (Khalil et al., 2009), suggesting that mlncRNA play a crucial role in guiding chromatin modification.

In the r.h.s. panel of Fig. 4 we summarize co-occurrences of RNA binding domains with reader and modifier domains. There appears to be a strong increase in these combinations as we evolve towards the animal kingdom. In particular, the strong link between methylation readers and RNA binding is unique to animals. From these data it appears that chromatin modification in animals come to rely more heavily on specific RNA-binding than the other major groups of eukaryotes. This is consistent with the vast amount of mlncRNAs produced by animal genomes, which have not been reported, at least to this extent, for other clades. Plants, fungi and basal eukaryotes make extensive use of small RNAs in directing chromatin modifications. In this mode of action, the chromatin modifiers presumably interact with the RNP complexes of the RNAi machinery rather than directly with the small RNAs.

Fig. 4 appears to indicate an ancestral role for DNA binding domains and a prevalence of modifier-DNA interactions over modifier-RNA interactions. This could be the result of a strong annotation bias as nucleic acid binding domains are “by default” interpreted as DNA binders, and much less is known about RNA binding proteins in general. Zinc finger domains, here counted as DNA binding due to their function in transcription factors, for instance, are well known to also bind RNA (Brown, 2005; Hall, 2005) and DNA/RNA duplexes (Shi and Berg, 1995). The detailed distribution and relative importance of DNA- versus RNA-directed modifiers remains a question for future research. Irrespective of this outcome, the data show that there is a strong anchor that guides the chromatin modification machinery to sequence-specific loci. In particular (de)methylation is predominantly guided by nucleic acid sequence information.

## 6. Epigenetic inheritance of chromatin regulation

An intriguing feature of chromatin is that it can transmit patterns of gene expression across cell divisions. This *cellular memory* is of crucial importance in the development of multicellular organisms and underlies cell differentiation. There is no single mechanism responsible for copying epigenetic information from one cell generation to the next. Instead, various combinations of modification and targeting mechanisms are combined to achieve a more or less faithful propagation of information. Rather few cellular transmission mechanisms are well understood, beyond the replication of DNA methylation by Dnmt1 (Vilkaitis et al., 2005). Much less is known about mechanisms through which histone modifications are copied.

Histone modification does not prevent the transfer of parental histones to newly replicated DNA and thus are free to serve as a means of epigenetic inheritance. However, after replication, half of the nucleosomes must be assembled *de novo*, requiring mechanisms that “copy” chemical modifications forward, in order to establish faithful inheritance of the epigenetic marks (Benson et al., 2006). The deposition of histone H3.3, for example, has been implicated in the inheritance of the active chromatin state (Ng and Gurdon, 2008). Complex interactions of Polycomb and trithorax group proteins with several methylation and acetylation enzymes are necessary to maintain both inactive and active states and their boundaries (Schuettengruber et al., 2007; Schwartz and Pirrotta, 2008). Although critical details remain to be determined, it has become clear that some of the histone marks are regenerated after replication from partially transmitted information using the reader/modifier machinery (Probst et al., 2009).

For most chemical modifications it remains unknown whether they are epigenetically inherited. It is conceivable that the

transcriptional status of chromatin, rather than the specific pattern of chemical modification, is transmitted to the next generation. We do know that chromatin-based information can be actively erased or reset upon replication (Morgan et al., 2005; Reik, 2007).

It remains to be determined to what extent epigenetic inheritance is tied to underlying DNA sequences. It has been established that histone modifications are not solely determined by the underlying sequence (in which case the “epigenetic” information would be a one-to-one mapping from the sequence information under the action of the modification enzymes). On the other hand, the epigenetic system is not a “free-floating” system, completely detached from DNA sequence information. Studies in fission yeast and other multicellular organisms suggest that effector complexes target nascent chromatin-bound non-coding RNAs and recruit chromatin-modifying complexes (Amaral and Mattick, 2008). This has been suggested to contribute to the inheritance of chromatin states during the process of chromosome duplication (Moazed, 2009).

Chromatin structure also plays a key role in DNA damage repair (Escargueil et al., 2008). Nucleotide excision repair, a repair pathway conserved in animals and fungi, starts with histone modifications, in particular H2A mono-ubiquitinylation (Zhou et al., 2009) that marks the damaged region. A complex remodelling cascade then enables repair by excision of the affected DNA segment and the reconstitution of the chromatin structure (Zhang et al., 2009). H3K9 modifications furthermore play a role in targeting class switch recombination in the Ig heavy chain gene (Kuang et al., 2009).

The diversity and complexity of the mechanisms of epigenetic inheritance, as well as the apparently late evolutionary origin of several key components (Iyer et al., 2008; Schuettengruber et al., 2007), suggests that an elaborate heritable cellular memory was a late addition to the chromatin-based regulation machinery.

## 7. A hypothesis for the evolution of chromatin regulation

In this section we make use of our comparative, phylogenetic data, in combination with well known biochemical properties of DNA, RNA, and protein to propose an evolutionary series of events leading to complex forms of genetic regulation through chromatin-mediated mechanisms. The scenario includes a number of testable hypotheses.

### 7.1. Origins of DNA genomes

The early history of life presents us with many unsolved problems. Most researchers agree on the existence of an “RNA-Protein World” stage preceding the divergence of Eubacteria, Archaea, and Eukarya, in which genetic information was stored in RNA (Boussau et al., 2008; Glansdorff et al., 2008; Rodríguez-Trelles et al., 2006; Taylor, 2006; Wolf and Koonin, 2007). Two competing theories postulate either a last common ancestor (LUCA) of the three domains with a DNA genome (Becerra et al., 2007; Mat et al., 2008), or a LUCA with an RNA genome (Di Giulio, 2008; Forterre, 2006; Glansdorff et al., 2008; Koonin and Martin, 2005; Poole and Logan, 2005). In the latter scenario, the transition to a DNA genomes occurred twice (once in Eubacteria and once in Archaea+Eukarya) (Forterre, 2002) or even thrice (Forterre, 2006), possibly mediated by viral entities. The literature agrees, however, that complex cellular machineries were present in the ancestral RNA genomes before they were replaced by modern DNA-based information storage, and that this substitution was driven by the advantages of a chemically and thermodynamically more stable genomic material (Boussau et al., 2008).

A late and multiple transfer of a large amount of genetic information from RNA to DNA implies that the regulation of protein expression must have been predominantly translational in the ancestral RNA state. The logical alternative is specific differential regulation of the replication of specific RNA fragments<sup>2</sup> (Taylor, 2006). It is unlikely that such a mechanism will have been shifted over to a new DNA chromosome with the same RNA and protein factors in place. This is because RNA-binding proteins do not necessarily bind DNA, and even if they do, there is no chemical reason for them to recognize the same nucleotide sequences.

We propose that the transition from an RNA to a DNA genome was most likely from an ancestral state in which the protein-production was regulated predominantly at the level of translation. A generic transcription machinery, perhaps of viral origin, without complex fine-grained regulation would be capable of reproducing an RNA complement similar to that of the former RNA/protein world ancestor, so that the ancestral post-transcriptional mechanism could continue to function.

Gene-specific transcriptional regulation, we argue, must have been a later innovation. Hence this generates the hypothesis that specific transcription factors should have arisen towards the distal branches of the tree of life. This hypothesis is supported by the observation that specific regulation by transcription factors is evolutionary plastic, with global transcription factors poorly conserved across phylogeny (Lozada-Chávez et al., 2006).

## 7.2. ChAPs as early transcriptional regulators

ChAPs probably arose together with the DNA genome, presumably for one or more of the following reason:

1. In a scenario where DNA was imported from viruses, it may have been necessary to distinguish cellular from foreign DNA.
2. A relatively large DNA genome may have required some architectural stabilization to prevent agglutination.
3. ChAPs may have evolved as a simple mechanism for separating a transcriptionally active state from a replication phase. The ability to “decide” whether to replicate or not could provide substantial advantage in nutrient poor and highly variable environments.
4. The stable DNA genome may have enabled the formation of durable “spores” capable of surviving harsh temporary conditions.

Initially, only the global transcriptional activity would be regulated, presumably by means of the concentration of ChAPs. Differential protein expression, on the other hand, was still organized at the RNA-protein level—translational—just as in the ancestral RNA/protein organism. This leads us to hypothesize that ChAPs along with many other DNA binding proteins, such as transcription factors, originated from ancestral RNA binding domains. Support for this hypothesis comes from the fact that similar folds can bind both classes of nucleic acids (Brown, 2005; Guo et al., 2008; Yang et al., 2006). The diversity of ChAPs in the different domains is consistent with independent origins of DNA genomes in Eubacteria and the Archaea/Eukarya lineage.

Paralogous ChAPs with somewhat different sequence preferences have arisen by gene duplication, providing a potential means of distinguishing chromosomal regions in which transcriptional activity could be regulated differentially through modulation of the concentration and relative abundance of

paralogous ChAPs. Recall that such mechanism are still used in e.g. distinguishing exponential and stationary growth phases in Eubacteria and Archaea. Most plausibly, at this stage, the regulatory effect would be exerted directly by the physico-chemical properties of the ChAPs.

The earliest forms of transcriptional regulation were thus generalized and repressive, and entirely dependent upon a small collection of ChAPs. This defined a handful of chromatin states on a single genome that mapped onto multiple, clearly distinct phenotypic states. Sequence-specific transcription factors would have evolved to allow activating exceptions to the general repressed states refining/backing up the specific regulation of expression implemented on the level of translation. Eventually, this leads to the current solution, in which the expression of a typical protein is subject to both transcriptional and post-transcriptional regulation (Bailey-Serres et al., 2009; Izawa and Inoue, 2009).

Chemical modification of ChAPs offers a more economical alternative to the use of paralogs, and could have co-opted enzymes that originally modified other proteins in a host-defense or signaling context. Chemical modifications provide a faster means of responding to external signals than the expression of ChAP paralogs, and reduces the metabolic burden imposed by ChAP synthesis. The combination of a small set of non-specific DNA-binding ChAPs, and several types of modifications set by specifically targeted modification enzymes, allows for a fine-grained definition of chromatin states.

The imbalance in the prevalence of modification “writers” over modification “erasers” hints at a later addition of general demodification activities. There are at least three interpretations for this observation: (1) Harmful substances are produced during the de-modification reaction (e.g. demethylation) causing selection against the massive use of such a reaction pathway. (2) A very general de-modification mechanism is already in place since replication and ChAP turnover thins out modified ChAPs through re-synthesis of unmodified ChAPs. The controlled degradation of modified ChAPs could therefore quickly return the chromatin to the unmodified state, rendering specific de-modification an unnecessary step. Only when multiple modifications are written on the same nucleosome, the degradation pathway becomes impractical. (3) A cascade of specific, effectively irreversible modifications may have been evolutionarily advantageous as a means of implementing asymmetric differentiation of cell states, thereby setting the stage for organismal development. Unfortunately, we cannot entirely rule out that the under-representation of erasers is caused by a bias in the protein domain data.

## 7.3. ChAPs as a memory device

The invention of reader domains in the ancestral Eukarya turned chromatin into a cellular memory device. Chemical marks written onto the ChAPs in the cellular past can now be interpreted and modified in a context-sensitive fashion by proteins that combine reader domains and effector domains. It allows the cell to keep a record of former states—in particular, of the past transcriptional activity in any given genomic region. A major advantage of such a setup is that the transcriptional programs no longer needs to be activated or terminated using direct feedback e.g. through measurements of metabolites or environmental factors.

Interestingly, we found that the reader/effector system is quite variable among the various Eukaryotic lineages. Many of the protein families that implement specific combinations of reader and effector domains are exclusive to specific clades, with much

<sup>2</sup> Models for “transcription” in the RNA world typically envision this as a replication-like process mediated by an ancestor or relative of the ribosome.

of the combinatorial complexity of readers and effectors explored throughout the course of evolution.

Neither the advent of a reader domain nor the innovation of reader/effector combinations are particularly unlikely events. Fusion proteins and recombinations of protein domains are abundant throughout protein evolution. It is perhaps more surprising that Archaea and Eubacteria do not seem to have evolved a ChAP-associated memory system. We speculate that this is not for lack of opportunity but rather for the lack of immediate advantage or access.

The combination of reader domains with writer and eraser domains enables a network of histone modification in response to histone-modification-dependent target-selection (Forneris et al., 2005) and promotes an autonomous dynamical system resulting in epigenetic information cascades superimposed onto and uninfluenced by the underlying DNA (Benecke, 2006; Fischle et al., 2003; Hall et al., 2002; Sedighi and Sengupta, 2007). The potential detachment of the chromatin modification network from underlying DNA is prone to conflict with the adaptive requirements of transcriptional regulation. We hypothesize that this semi-independence could be a source of pathology, such as in cancer. It is no surprise, therefore, that transcription and histone modification are tightly linked. A variety of distinct mechanisms, from DNA binding enhancers (Koutroubas et al., 2008) to the employment of small and large ncRNAs have evolved to anchor the modification activities of the chromatin to the underlying DNA sequence.

#### 7.4. Epigenetics

Epigenetic inheritance is the most recent control layer in the chromatin regulation system. In fact, it does not consist of a single coherent mechanism but a collection of rather elaborate (and somewhat mysterious at this point) “tricks” to propagate *selective* information stored in the chromatin memory to daughter cells. To the extent that epigenetic inheritance is understood at all, it re-utilizes diverse components of the chromatin-regulation machinery to regenerate part of the cell state-information following stochastic assortment of the histone complexes to daughter cells. Very little is known to what extent the detailed mechanism are lineage-specific. Organismal development depends on this process to propagate epigenetic states across cell divisions and to implement a program of differentiation steps that are effectively irreversible.

### 8. Chromatin computation

#### 8.1. Gene regulation as computation

Gene regulation can be thought of as a form of computation. So far, this point has been made most explicit in the case of *cis*-regulatory networks (Bonn and Furlong, 2008; Istrail et al., 2007; Levine and Davidson, 2005). *Cis*-regulatory modules (CRMs) are abstracted as Boolean functions that combine the input—a pattern of currently present transcription factors—by means of conjunction, disjunction and complementation (AND, OR, and NOT gates). Complex circuits are formed since both “down-stream effectors” as well as transcription factors themselves are regulated by such CRMs. This *cis*-regulatory model of genomic computation represents one of the essential computational modalities of the cell, and has been shown to play a crucial role in development.

Here we suggest that the functional interpretation of our phylogenetic findings and evolutionary hypothesis is that chromatin

regulation adds a computational layer that, in Eukarya, is qualitatively different and potentially more powerful than the CRM networks. As we shall see, the difference is the explicit and extensive memory implemented in the ChAP modifications. In order to proceed, we need to introduce the idea of “computation” generally and more formally, which we can view essentially as statement about constraints on input–output functions. Stated in this way, we shall see that CRMs are but one class of important mechanisms for information processing. We start from a set of basic or atomic states  $S$  and a transition operation  $\rightarrow$  operating on these states, which is simply a relation  $\rightarrow \subseteq S \times S$  on the set of states that tells us which transitions are allowed. Formally, a computation is then simply a sequence  $(s_1, s_2, \dots, s_n)$  of states  $s_i \in S$  that are related by  $s_i \rightarrow s_{i+1}$ . The set of all computations w.r.t.  $\rightarrow$  will be denoted by  $M$ . For computations to be effective they need to both come to an end (or halt  $H$ ) and produce a result that is available or readable by some other components of a system,  $R$ . Hence we need to introduce a stop predicate  $H: M \rightarrow \{0,1\}$  and consider only the computations  $M_H$  that halt. For these we further require an “output mapping”  $\rho$  that takes the results of a complete computation and makes them available for evaluations:  $\rho: M_H \rightarrow R$ . A *computing system*  $\Gamma$  is then simply defined in terms of the quadruple:  $\Gamma = (S, \rightarrow, R, \rho)$ .

This formal specification is still too abstract for our purposes. Let us therefore think of each state  $s \in S$  as a particular realization of a biological structure that is composed of elementary objects that can be manipulated individually during the state transitions. The abstract transitions  $\rightarrow$  then become concrete rewriting operations on collections of these elementary objects/data structures. In the simplest case,  $s$  is a string over some alphabet  $\mathcal{A}$ . In this particular implementation, computer science defines a so-called Chomsky (1959) hierarchy that establishes a correspondence between the form of the rewrite operations, computational power, and the structure of the transition rules that implement a computation.

#### 8.2. Memory capacity

CRMs are naturally interpreted in terms of finite state machines. Each state  $s$  corresponds to a particular concentration profile of a few hundred to a few thousand regulators, mostly transcription factors, signalling proteins, and micro-RNAs. Each of these regulators encodes in its concentrations, a small number of states that are distinguishable at the transition level, and hence contributes a few bits to the overall storage capacity of the system. A crude upper bound can be obtained from the number of regulators,  $n_{\text{reg}}$ , and their number of copies  $N$  in the cell: even assuming that half the human genes are regulators,  $n_{\text{reg}} = 10^4$ , with 1000 copies each, there are no more than  $10^5$  bit of information in a CRM network.

In chromatin computation, each state  $s \in S$  is the particular arrangement of all the specifically modified ChAPs on the DNA. Since little is known about the feedback between rearrangements of ChAP/nucleosome positioning (chromatin remodeling) and chemical modifications, we simplify this picture by neglecting the details of nucleosome positioning. In this approximation, a state  $s \in S$  is then a particular pattern of ChAP modifications and variants in the linear array of nucleosomes. This arrangement is similar to the linear memory organization of a digital computer. A nucleosome then corresponds to a particular *page* of memory. The organization of each page deviates in detail from a simple string because different residues can sustain different numbers of modifications of different types.

In the absence of readers, chemically stored information is not persistent in that it cannot be utilized without changing it. Demethylation, therefore, is fundamentally different from reading

a particular methylation pattern. With the advent of reader domains, however, the same piece of information can be accessed repeatedly and in different contexts. As a consequence, eukaryotic chromatin can store information on a much longer time-scale than CRMs.

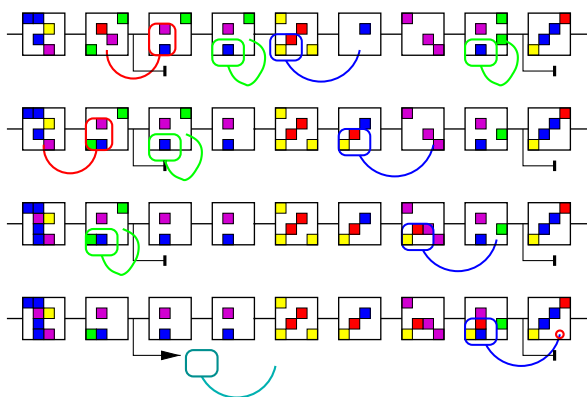
Eukaryotic nucleosomes have a sizable memory capacity owing to the different combinations of methylation, acetylation, phosphorylation, and ubiquitination that can be present or absent at multiple residues in each of several ChAPs forming a single nucleosome. Based on published histone modification maps, yeast can store up to about 70 bits of information in a single nucleosome, in human the capacity can be estimated as ~200 bits (see Methods section for details). Note that this value is comparable to a memory capacity of not more than 400 bits provided by the approximately 200 nt of DNA in the sphere of influence of a nucleosome.

Since the number of accessible states is still finite, we could formally map chromatin computation to a finite state machine, albeit with a state space that is 4–6 orders of magnitude larger than that of CRMs. A more natural interpretation is that chromatin computation implements a context-sensitive mechanism with a more modest state space.

We can extend this computational interpretation by recognizing that the transitions in eukaryotic chromatin computation are typically local and massively parallel at the same time. Each reader/writer recognizes a sub-pattern of just a few bits on each memory page and writes/modifies a few on the same or an adjacent nucleosome. The chromatin modification machinery is thus reminiscent of a vector or even parallel computer, Fig. 5. Each of the memory pages (nucleosomes) is subject to the transitions caused by a particular combination of reader and modifier enzymes active at any given time. If the chemical activity is uniform across the genome, the system behaves much like a vector processor. In conjunction with region specific targeting systems, different modifications processed can be active concurrently in different regions of the genome/memory, so that chromatin computation becomes a kind of imperfectly synchronous, parallel computing.

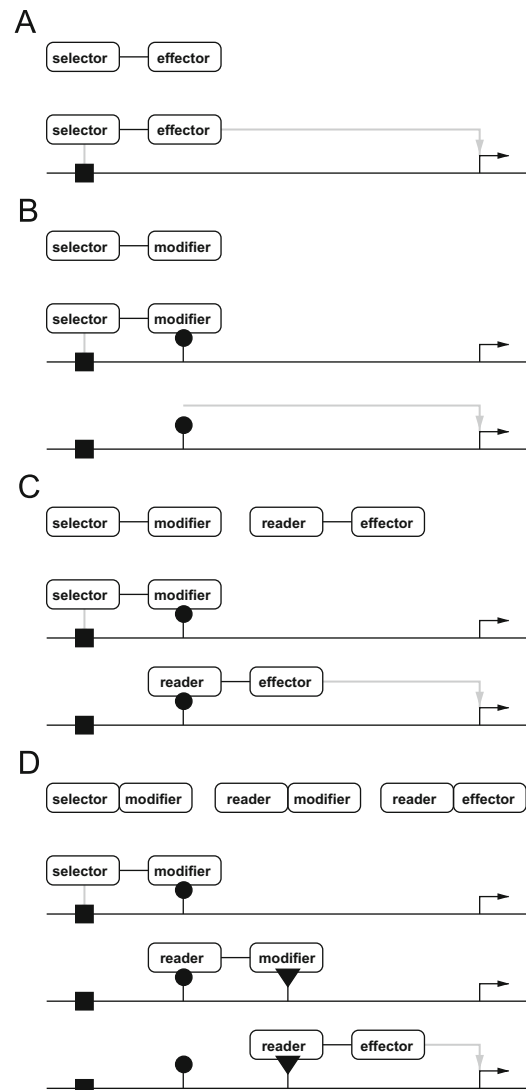
### 8.3. Biochemical support of chromatin computation

Transcriptional regulation, in this computational paradigm, is the result of associating a computational process to specific



**Fig. 5.** Computational model of Chromatin. Memory is organized in a linear sequence of pages (nucleosome) and all state changes are restricted in a cellular-automaton-like fashion to local neighborhoods. State transitions are determined by *rewrite rules* (reader/modifier complexes) that recognize (part of) the information stored on a particular memory page (nucleosome) and change (some of) the content of the same or an adjacent memory page. The set of available rules also depends on the state of particular local memory pages (indicated by the 'transcription start site symbols'), corresponding the transcriptional activity of particular reader/modifier complexes.

biochemical effectors. Here we discuss the detailed biochemical micro-structure of our apparatus in terms of symbolic operations. A *selector* guides an *effector* to a particular genomic position and induces or prevents transcription at a given locus. In CRMs, this selector is the DNA binding domain of a transcription factor binding to specific DNA locus, Fig. 6A. Here, the presence of the transcription factor is required to exert the desired effect. This appears to be the dominant mode of regulation in Eubacteria. Using a programming analogy, it corresponds to an *if then* statement. Logically related is the use of a *selector-modifier* combination. In the case of the deposition of variant nucleosomes and the writing or histone modifications in Archaea, the modification itself physically influences transcription; the effect persists as long as the modification is present, Fig. 6B. The important distinction from CRMs is that in the latter case



**Fig. 6.** Conceptual innovation in chromatin regulation. (A) Direct effects on transcription are carried out by specific or unspecific selector–effector pairs, such as transcription factors or variants of chromosomal architectural proteins (ChAPs). (B) Only Archaea and Eukarya are known to chemically modify their ChAPs. Initially, it is this modification itself that effects the binding affinity of the protein and therefore regulation. (C) Modification readers are unique to Eukarya. Now a signal can be set by a selector–modifier. This persistent signal is then interpreted by a reader–effector combination. The modification thus becomes a signal in an information theoretic sense and provides a means for separating the decision making process from the execution of the decision. (D) The coupling of reader and modifier allows signal propagation along the DNA and provides the physical basis for signal inheritance and complex computations.



modifications have only transient memory for the action of a modifier. From a programming point of view, modifications may be seen as *while* statements. The combination of *effectors* with modification *readers* expands the use of modifications as memory to include interpretation of the stored information in different ways, Fig. 6C. Chromatin in higher eukaryotes appears to be utilized in this way because most modification marks do not have unambiguous interpretations, but their function depends on the cellular context. *Reader-modifier* combinations, finally, allow the implementation of complex programs without the need to produce an immediate effect on transcription at all, Fig. 6D.

CRMs and chromatin differ in a second important property. Most of the states of CRMs are subject to selective constraints because the expression of regulatory gene products, almost by definition, has an impact on the production of downstream target genes. Transitions between regulatory states therefore can be expected to have direct physiological consequences. Changes in the regulatory program are typically immediately visible to the forces of natural selection, implying rather stringent constraints on the computational programs that “can be run” on a CRM.

In contrast, chromatin modifications have rather indirect, aggregate effects on transcriptional activity, such that a large fraction of its memory capacity can be utilized without direct physiological effect of the chemical modifications themselves. Hence it is feasible to implement rather elaborate computations in chromatin which are selectively (nearly) neutral (except for the resource consumption of the modification chemistry itself).

## 9. Concluding remarks

We have surveyed the phylogenetic distribution of mechanisms of chromatin modification and regulation. Chromosomal architectural proteins (ChAPs) and chromatin are present throughout extant organisms and have a demonstrable impact on gene expression. Multiple alternative ChAPs and paralogous variants of ChAPs influence chromatin structure. The mode and extent of chromatin regulation is quite different in Eubacteria, Archaea, and Eukaryotes. In contrast to Eubacteria, both Archaea and Eukarya regularly utilize chemical modification of ChAPs, which in Archaea is largely restricted to setting modifications with immediate biophysical impact. In contrast, Eukarya possess an elaborate system for managing chemical modifications comprising proteins that can write, read, and erase in a highly specific way post-translational modification such as phosphorylation, methylation, acetylation and ubiquitination. A phylogenetic analysis of domain-occurrences indicates increasingly complex interactions between reading and modifying (writer and eraser) domains in the crown-group Eukaryotes.

We view the regulation of gene expression through chromatin regulation as a parallel information processing system that is evolutionarily prior to the transcription factor based templating machinery stressed in *cis*-regulatory modules. It is only with the appearance of the Eukarya, that chromatin regulation transitions from a write–delete system to write–delete–read system. This suggests to us that with the Eukarya we move from a crude low-dimensional ancestral finite-state computing device, to a highly tuned parallel computing architecture with a significantly expanded memory capacity.

The regulatory complexity that can be achieved (in the form of CRMs) through DNA-binding proteins is somewhat limited by their binding properties (Mattick, 2007). As a consequence, small non-coding RNAs have recently been proposed to serve a better function in specific and selective DNA recognition with the additional benefit of genome compactness of ncRNAs and evolvability of the interaction networks (Mattick, 2007; Mattick

et al., 2009; Prohaska and Stadler, 2008). We know that many of these RNAs appear to be involved as selectors for action on chromatin. We suggest that chromatin regulation could provide significant increases in both the resolution of cellular information processing and contribute to fault tolerance by providing a redundant layer above CRM mediated regulation.

With the appearance of the Eukarya we see the emergence of complex forms of multicellular life, and it is tempting to speculate that this was correlated with the expansion of computational power. We know that differentiated multicellular organisms rely on epigenetic processes to retain cell states, and that these epigenetic marks are often carried by chromatin. Here we suggest that epigenetics in the form of chromatin modifying systems adds a layer of context-sensitivity and thereby regulatory flexibility during the developmental process.

Because chromatin regulation presents an additional, complementary, regulatory level to that provided by *cis*-regulation, the regulatory layers might diverge when they behave in ways that are fitness neutral. Hence extensive write–read–write operations at the chromatin layer that have a minimal impact at the level of *cis*-regulation (and hence translation) will not be inhibited by anything other than the potential cost of frequent chemical modification. Hence there is the possibility for extensive “free-wheeling” at the chromatin level, generating significant variation that could come to play an important adaptive role (both negative and positive) with a change in cellular context. The most obvious example is cancer as a chromatin disease (Esteller and Herman, 2002). Since DNA repair mechanisms appear to be programmed by chromatin, perhaps the increasing chromosomal aberrations in many cancers are not just incidental DNA damage, but the downstream effect of free-running chromatin dynamics loosely coupled to immediate selective consequences.

In the germline, the nearly complete erasure of chromatin marks could be a “precaution” against the accumulations of computational errors at the chromatin level: the propagation of epigenetic information is limited by Eigen’s error-threshold (Eigen, 1971; Eigen et al., 1989). This theory implies that extremely high accuracies on the order of a single copying error per replication are required for the long-term information maintenance. With the possible exception of DNA methylation, mechanisms of epigenetic inheritance appear to be much noisier, which could explain why they are restricted to somatic cells. With largely unmarked chromatin as a starting point, early embryonic development could hardly be governed by chromatin. Instead, CRMs have been tremendously successful in regulating these processes. The transcriptional activity in these early stages also leads to an accumulation of chromatin marks, which soon take control of the cell fate.

These findings, taken together, provide evidence for the evolution of complex, combinatorial forms of information regulation, starting from simple structural precursors. Moreover, multiple regulatory mechanisms act in parallel, and in all likelihood, redundantly. These facts attest to the inability of complex regulatory systems to suppress all sources of uncertainty.

## Acknowledgements

This work was supported in part by the VW Foundation and the Pathfinder initiative on Complexity through the projects EDEN (043251) and SYNLET (043312). And by a grant on Innovation in Natural, Experimental and Applied Evolution from the Packard Foundation to DCK. In order to preserve chances for future federal funding we refrained from using the alternative title “How to get histoned”.

## Appendix A. Methods

### A.1. Phylogenetic distribution of protein domains

Domain annotation for 751 species (1099 strains and/or different genomes) were extracted from the genome assignment data of the Superfamily database (Wilson et al., 2009) (version 14-Jun-2009) in flat file format.

Domains relevant for chromatin regulation were derived from known proteins involved in chromatin regulation taken from the literature, and their domain annotation as given by *scop* (Andreeva et al., 2008) and Superfamily.

**Phylogeny:** Species (with one to several strains each) were assigned to the following clades, B—Eubacteria; A—Archaea; E—Eukarya; AN—Nanoarchaeota; AK—Korarchaeota; AC—Crenarchaeota; AE—Euryarchaeota; EB—basal eukaryotes (e.g. Diplomonadida and Parabasilia); EK—Kinetoplastida (incl. Heterolobosea); EA—Alveolata (i.e. Ciliophora and Apicomplexa); EO—Chromista (i.e. Cryptophyta, Haptophyta, and Heterokonta) (Yoon et al., 2002); EV—Viridiplantae (incl. Chlorophyta); ED—Amoebozoa (e.g. Dictiostelium); EF—Fungi; EM—Metazoa, based on the deep phylogeny presented in Iyer et al. (2008) for critical splits and the “Tree of Life” or NCBI taxonomy otherwise. Monophyletic Chromalveolata (EA+EO) with Chromista and Alveolata as sister groups (Martens et al., 2008) were assumed here.

**Domain distribution:** For each domain, gene counts were maximized over all strains for each species and averaged over all species, with at least one gene, for each clade. This reflects the *average abundance* of a domain *X* within a clade. Furthermore, we computed the fraction of species within a specific clade that has at least one gene with domain *X*, indicating the *prevalence* of domain *X* within a clade.

**Computational tools:** The domain abundance and prevalence values for single domains or domain combinations are computed with a sequence of *awk* and *sort* commands together with simple *perl* scripts using the Superfamily data described above and phylogenetic information as input.

### A.2. Information on DNA versus nucleosomes

The amount of information that can be stored by a nucleosome was calculated based on the number of different sites and types of modifications under the assumption that all sites on the eight histones can be modified independently. A comprehensive list of possible modifications for mammalian histones was collected from the literature:

**Modifications of histone H 3:** H3R2me (Bhaumik et al., 2007; Bedford and Clarke, 2009; Torres-Padilla et al., 2007), H3R2cit (Cuthbert et al., 2004), H3T3ph (Bhaumik et al., 2007), H3K4me (Bhaumik et al., 2007; Garcia et al., 2007), H3K4ac (Garcia et al., 2007), H3K4bio (Kothapalli et al., 2005), H3R8me (Bhaumik et al., 2007), H3R8cit (Cuthbert et al., 2004), H3K9ac (Bhaumik et al., 2007; Garcia et al., 2007), H3K9bio (Kothapalli et al., 2005), H3K9me (Bhaumik et al., 2007; Garcia et al., 2007), H3S10ph (Bhaumik et al., 2007), H3T11ph (Bhaumik et al., 2007), H3K14ac (Bhaumik et al., 2007; Garcia et al., 2007), H3R17me (Bhaumik et al., 2007), H3R17cit (Cuthbert et al., 2004), H3K18me1 (Garcia et al., 2007), H3K18ac (Bhaumik et al., 2007; Garcia et al., 2007), H3K18bio (Kothapalli et al., 2005), H3K23me1 (Garcia et al., 2007), H3K23ac (Bhaumik et al., 2007; Garcia et al., 2007), H3R26me (Bhaumik et al., 2007), H3R26cit (Cuthbert et al., 2004), H3K27me (Bhaumik et al., 2007; Garcia et al., 2007), H3K27ac (Bhaumik et al., 2007; Garcia et al., 2007), H3S28ph (Bhaumik et al., 2007), H3P30iso (Nelson et al., 2006), H3K36me (Bhaumik et al., 2007; Garcia et al., 2007), H3K36ac (Bhaumik et al., 2007; Garcia et al., 2007), H3P38iso (Nelson et al., 2006), H3K56me (Garcia et al., 2007), H3K56ac (Bhaumik et al., 2007; Mersfelder and Parthun, 2006; Garcia et al., 2007), H3K79me (Bhaumik et al., 2007; Mersfelder and Parthun, 2006; Garcia et al., 2007), H3K79ac (Garcia et al., 2007), H3K115ac (Zhang et al., 2003), H3T118ph (Freitas et al., 2004), H3K122me (Freitas et al., 2004), H3K122ac (Zhang et al., 2003).

et al., 2007; Garcia et al., 2007), H3K36ac (Bhaumik et al., 2007; Garcia et al., 2007), H3P38iso (Nelson et al., 2006), H3K56me (Garcia et al., 2007), H3K56ac (Bhaumik et al., 2007; Mersfelder and Parthun, 2006; Garcia et al., 2007), H3K79me (Bhaumik et al., 2007; Mersfelder and Parthun, 2006; Garcia et al., 2007), H3K79ac (Garcia et al., 2007), H3K115ac (Zhang et al., 2003), H3T118ph (Freitas et al., 2004), H3K122me (Freitas et al., 2004), H3K122ac (Zhang et al., 2003).

| Group | Domain         | Superfamily IDs |
|-------|----------------|-----------------|
| ChAP  | HU             | 47730           |
|       | Alba           | 82704           |
|       | Histone        | 47114, 47129    |
|       | Linker histone | 46827           |
| Mub   | E1             | 46935, 69572    |
|       | E2             | 54496           |
|       | E3             | 56205           |
| Dub   | UCH            | 82568           |
|       | UCH-L3         | 54050           |
|       | ULP            | 54054           |
|       | OTU            | 110773          |
|       | JAMM           | 102712          |
| Mph   | H-kinase       | 55884           |
|       | STY-kinase     | 88854           |
| Dph   | PPI            | 52788           |
|       | PPII           | 52799           |
|       | PP1-5          | 56310           |
|       | PP2C-like      | 81601           |
| Rph   | 14-3-3         | 48446           |
|       | BRCT           | 52113           |
| Mac   | NAT            | 55730           |
|       | Rpd3           | 52773           |
| Dac   | Sir2           | 63984           |
| Rac   | bromo          | 47370           |
| R-    | SANT           | 46739           |
| Mme   | SET            | 82200           |
|       | PRMT           | 53351           |
|       | Dot1           | 89746           |
| Dme   | JmjC           | 82194           |
|       | LSD1           | 140222          |
|       | PAD4           | 110107          |
| Rme   | chromo         | 54165           |
|       | MBT            | 89299           |
|       | PHD            | 57911           |
|       | Tudor          | 63749           |

**Modifications of histone H 4:** H4S1ph (Bhaumik et al., 2007), H4R3me (Bhaumik et al., 2007), H4R3cit (Cuthbert et al., 2004), H4K5ac (Bhaumik et al., 2007), H4K8ac (Bhaumik et al., 2007),

H4K8bio (Kothapalli et al., 2005), H4K12ac (Bhaumik et al., 2007), H4K12me1 (Zhang et al., 2003), H4K12bio (Kothapalli et al., 2005), H4K16ac (Bhaumik et al., 2007), H4K20me (Bhaumik et al., 2007; Garcia et al., 2007), H4K20ac (Zhang et al., 2003), H4K31me (Beck et al., 2006; Garcia et al., 2007), H4K31ac (Garcia et al., 2007), H4S47ph (Freitas et al., 2004), H4R55me (Beck et al., 2006), H4K59me (Freitas et al., 2004), H4H75ph (Besant and Attwood, 2009), H4K77me (Beck et al., 2006), H4K77ac (Zhang et al., 2003), H4K79ac (Zhang et al., 2003), H4K91ac (Mersfelder and Parthun, 2006), H4R92me (Freitas et al., 2004).

**Modifications of histone H 2A:** H2AS1ph (Bhaumik et al., 2007), H2AK5ac (Bhaumik et al., 2007), H2AK9ac (Bhaumik et al., 2007; Basu et al., 2009), H2AK9bio (Kothapalli et al., 2005), H2AK13ac (Bhaumik et al., 2007; Basu et al., 2009), H2AK13bio (Kothapalli et al., 2005), H2AK15ac (Basu et al., 2009), H2AK36ac (Basu et al., 2009), H2AK99me (Freitas et al., 2004), H2AK119ub (Bhaumik et al., 2007), H2AT120ph (Bhaumik et al., 2007), H2AK125ac (Basu et al., 2009), H2AK127ac (Basu et al., 2009), H2AK129bio (Kothapalli et al., 2005).

**Modifications of H 2B:** H2BK5ac (Bhaumik et al., 2007), H2BK11ac (Basu et al., 2009), H2BK12ac (Bhaumik et al., 2007), H2BS14ph (Bhaumik et al., 2007), H2BK15ac (Bhaumik et al., 2007), H2BK16ac (Basu et al., 2009), H2BK20ac (Bhaumik et al., 2007), H2BK23me (Zhang et al., 2003), H2BK23ac (Basu et al., 2009), H2BK24ac (Basu et al., 2009), H2BK43me (Zhang et al., 2003), H2BK47me (Beck et al., 2006), H2BK57me (Beck et al., 2006), H2BK85ac (Zhang et al., 2003), H2BK99me (Zhang et al., 2003), H2BK108me (Beck et al., 2006), H2BK108ac (Zhang et al., 2003), H2BK116ac (Basu et al., 2009), H2BK120ub (Bhaumik et al., 2007), H2BK120ac (Zhang et al., 2003). Based on the sparsity of functional information on non-enzymatic histone biotinylation (Healy et al., 2009), we excluded these marks from further calculations.

The information content (in bits) is calculated by summing over the logarithm (to the base 2) of different states per site for a histone octamer (i.e.  $2 \times H3$ ,  $2 \times H4$ ,  $2 \times H2A$ ,  $2 \times H2B$ ). Notice that the unmodified state is a state and that lysine and arginine methylations each contribute three states (mono-, di-, or trimethylation and mono-methylation, symmetric or asymmetric di-methylation, respectively). The total information content is 205.84 bits ( $I_{H3} = 38.65$ ,  $I_{H4} = 25.87$ ,  $I_{H2A} = 14.17$ ,  $I_{H2B} = 24.23$ ).

The storage capacity of DNA can be estimated as  $I_{DNA} = 2L$  since each position contributes  $\log_2 4 = 2$  bits for the four states A, C, G, or T. Assuming one nucleosome per 200nts, the DNA has an information content of 400 bits per nucleosome position. Epigenetic information thus may constitute up to 1/3 of the total information stored on a chromosome.

## Appendix B. Supplementary data

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.jtbi.2010.03.011.

## References

- Adams-Cioaba, M.A., Min, J., 2009. Structure and function of histone methylation binding proteins. *Biochem. Cell Biol.* 87, 93–105.
- Ahmad, K., Henikoff, S., 2002. The histone variant H3.3 marks active chromatin by replication-independent nucleosome assembly. *Mol. Cell* 9, 1191–1200.
- Amaral, P.P., Mattick, J.S., 2008. Noncoding RNA in development. *Mamm. Genome* 19, 454–492.
- Andreeva, A., Howorth, D., Chandonia, J.-M., Brenner, S.E., Hubbard, T.J.P., Chothia, C., Murzin, A.G., 2008. Data growth and its impact on the SCOP database: new developments. *Nucleic Acids Res.* 36, D419–D425.
- Arenas, A.F., Gutierrez Escobar, A.J., Gómez-Marin, J.E., 2007. Evolutionary origin of the protozoan parasites histone-like proteins (HU). *In Silico Biol.* 8, 2.
- Armstrong, J.A., 2007. Negotiating the nucleosome: factors that allow RNA polymerase II to elongate through chromatin. *Biochem. Cell Biol.* 85, 426–434.
- Azam, T.A., Ishihama, A., 1999. Twelve species of the nucleoid-associated protein from *Escherichia coli* sequence recognition specificity and DNA binding affinity. *J. Biol. Chem.* 274, 33105–33113.
- Bühler, M., Verdel, A., Moazed, D., 2006. Tethering RITS to a nascent transcript initiates RNAi- and heterochromatin-dependent gene silencing. *Cell* 125, 873–886.
- Bailey, K.A., Pereira, S.L., Widom, J., Reeve, J.N., 2000. Archaeal histone selection of nucleosome positioning sequences and the prokaryotic origin of histone-dependent genome evolution. *J. Mol. Biol.* 303, 25–34.
- Bailey, K.A., Marc, F., Sandman, K., Reeve, J.N., 2002. Both DNA and histone fold sequences contribute to archaeal nucleosome stability. *J. Biol. Chem.* 277, 9293–9301.
- Bailey-Serres, J., Sorenson, R., Juntawong, P., 2009. Getting the message across: cytoplasmic ribonucleoprotein complexes. *Trends Plant Sci.* 14, 443–453.
- Barrett, A.J., Rawlings, N.D., 2001. Evolutionary lines of cysteine peptidases. *Biol. Chem.* 382, 727–733.
- Basu, A., Rose, K.L., Zhang, J., Beavis, R.C., Ueberheide, B., Garcia, B.A., Chait, B., Zhao, Y., Hunt, D.F., Segal, E., Allis, C.D., Hake, S.B., 2009. Proteome-wide prediction of acetylation substrates. *Proc. Natl. Acad. Sci. USA* 106, 13785–13790.
- Becerra, A., Delaye, L., Islas, S., Lazcano, A., 2007. The very early stages of biological evolution and the nature of the Last Common Ancestor of the three major cell domains. *Annu. Rev. Ecol. Evol. Syst.* 38, 361–379.
- Beck, H.C., Nielsen, E.C., Matthies, R., Jensen, L.H., Sehested, M., Finn, P., Grauslund, M., Hansen, A.M., Jensen, O.N., 2006. Quantitative proteomic analysis of post-translational modifications of human histones. *Mol. Cell. Proteomics* 5, 1314–1325.
- Bedford, M.T., Clarke, S.G., 2009. Protein arginine methylation in mammals: Who, what, and why. *Mol. Cell* 33, 1–13.
- Bedford, M.T., Richard, S., 2005. Arginine methylation: an emerging regulator of protein function. *Mol. Cell* 18, 263–272.
- Bell, S.D., Botting, C.H., Wardleworth, B.N., Jackson, S.P., White, M.F., 2002. The interaction of Alba, a conserved archaeal chromatin protein, with Sir2 and its regulation by acetylation. *Science* 296, 148–151.
- Benecke, A., 2006. Chromatin code local non-equilibrium dynamics, and the emergence of transcription regulatory programs. *Eur. Phys. J. E Soft Matter* 19, 353–366.
- Benson, L.J., Gu, Y., Yakovleva, T., Tong, K., Barrows, C., Strack, C.L., Cook, R.G., Mizzen, C.A., Annunziato, A.T., 2006. Modifications of H3 and H4 during chromatin replication nucleosome assembly and histone exchange. *J. Biol. Chem.* 281, 9287–9296.
- Besant, P.G., Attwood, P.V., 2000. Detection of a mammalian histone H4 kinase that has yeast histidine kinase-like enzymic activity. *Int. J. Biochem. Cell Biol.* 32, 243–253.
- Besant, P.G., Attwood, P.V., 2009. Detection and analysis of protein histidine phosphorylation. *Mol. Cell. Biochem.* 329, 93–106.
- Bhaumik, S.R., Smith, E., Shilatifard, A., 2007. Covalent modifications of histones during development and disease pathogenesis. *Nat. Struct. Mol. Biol.* 14, 1008–1016.
- Bonn, S., Furlong, E.E., 2008. cis-regulatory networks during development: a view of *Drosophila*. *Curr. Opin. Genet. Dev.* 18, 513–520.
- Boussau, B., Blanquart, S., Neculea, A., Lartillot, N., Gouy, M., 2008. Parallel adaptations to high temperatures in the Archaeal Eon. *Nature* 456, 942–946.
- Brown, D.T., 2001. Histone variants: Are they functionally heterogeneous? *Genome Biol.* 2.
- Brown, R.S., 2005. Zinc finger proteins: getting a grip on RNA. *Curr. Opin. Struct. Biol.* 15, 94–98.
- Broyles, S.S., Pettijohn, D.E., 1986. Interaction of the *Escherichia coli* HU protein with DNA. Evidence for formation of nucleosome-like structures with altered DNA helical pitch. *J. Mol. Biol.* 187, 47–60.
- Buck, S.W., Gallo, C.M., Smith, J.S., 2004. Diversity in the Sir2 family of protein deacetylases. *J. Leukoc. Biol.* 75, 939–950.
- Chan, S.W., 2008. Inputs and outputs for chromatin-targeted RNAi. *Trends Plant Sci.* 13, 383–389.
- Cheng, X., Collins, R.E., Zhang, X., 2005. Structural and sequence motifs of protein (histone) methylation enzymes. *Annu. Rev. Biophys. Biomol. Struct.* 34, 267–294.
- Chomsky, N., 1959. On certain formal properties of grammars. *Inf. Control* 2, 137–167.
- Cirillo, L.A., Lin, F.R., Cuesta, I., Friedman, D., Jarnik, M., Zaret, K.S., 2002. Opening of compacted chromatin by early developmental transcription factors HNF3 (FoxA) and GATA-4. *Mol. Cell* 9, 279–289.
- Claret, L., Rouviere-Yaniv, J., 1997. Variation in HU composition during growth of *Escherichia coli*: the heterodimer is required for long term survival. *J. Mol. Biol.* 273, 93–104.
- Cozzone, A.J., 2009. Bacterial tyrosine kinases: novel targets for antibacterial therapy? *Trends Microbiol.* 17, 536–543.
- Cubonová, L., Sandman, K., Hallam, S.J., Delong, E.F., Reeve, J.N., 2005. Histones in crenarchaea. *J. Bacteriol.* 187, 5482–5485.
- Cuthbert, G., Daujat, S., Snowden, A.W., Erdjument-Bromage, H., Hagiwara, T., Yamada, M., Schneider, R., Gregory, P.D., Tempst, P., Bannister, A.J., Kouzarides, T., 2004. Histone deimination antagonizes arginine methylation. *Cell* 118, 545–553.

- Dai, J., Higgins, J.M.G., 2005. Haspin: a mitotic histone kinase required for metaphase chromosome alignment. *Cell Cycle* 4, 665–668.
- Dame, R.T., Noom, M., Wuite, G.J.L., 2006. Bacterial chromatin organization by H-NS protein unravelled using dual DNA manipulation. *Nature* 444, 387–390.
- de la Cruz, X., Lois, S., Sánchez-Molina, S., Martínez-Balbás, M.A., 2005. Do protein motifs read the histone code? *Bioessays* 27, 164–175.
- Di Giulio, M., 2008. The origin of genes could be polyphyletic. *Gene* 426, 39–46.
- Dillon, S.C., Zhang, X., Trievel, R.C., Cheng, X., 2005. The SET-domain protein superfamily: protein lysine methyltransferases. *Genome Biol.* 6, 227.
- Dion, M.F., Altschuler, S.J., Wu, L.F., Rando, O.J., 2005. Genomic characterization reveals a simple histone H4 acetylation code. *Proc. Natl. Acad. Sci. USA* 102, 5501–5506.
- Dri, A.M., Rouviere-Yaniv, J., Moreau, P.L., 1991. Inhibition of cell division in hupA/hupB mutant bacteria lacking HU protein. *J. Bacteriol.* 173, 2852–2863.
- Eichler, J., Adams, M.W., 2005. Posttranslational protein modification in Archaea. *Microbiol. Mol. Biol. Rev.* 69, 393–425.
- Eigen, M., 1971. Selforganization of matter and the evolution of macromolecules. *Naturwiss.* 58, 465–523.
- Eigen, M., McCaskill, J., Schuster, P., 1989. The molecular quasispecies. *Adv. Chem. Phys.* 75, 149–263.
- Eissenberg, J.C., 2001. Molecular biology of the chromo domain: an ancient chromatin module comes of age. *Gene* 275, 19–29.
- Escargueil, A.E., Soares, D.G., Salvador, M., Larsen, A.K., Henriques, J.A., 2008. What histone code for DNA repair? *Mutat. Res.* 658, 259–270.
- Esteller, M., Herman, J.G., 2002. Cancer as an epigenetic disease: DNA methylation and chromatin alterations in human tumours. *J. Pathol.* 196, 1–7.
- Ferl, R.J., Manak, M.S., Reyes, M.F., 2002. The 14–3–3s. *Genome Biol.* 3, 7.
- Fischle, W., Wang, Y., Allis, C.D., 2003. Binary switches and modification cassettes in histone biology and beyond. *Nature* 425, 475–479.
- Forneris, F., Binda, C., Vanoni, M.A., Battaglioli, E., Mattevi, A., 2005. Human histone demethylase LSD1 reads the histone code. *J. Biol. Chem.* 280, 41360–41365.
- Forterre, P., 2002. The origin of DNA genomes and DNA replication proteins. *Curr. Opin. Microbiol.* 5, 525–532.
- Forterre, P., 2006. Three RNA cells for ribosomal lineages and three DNA viruses to replicate their genomes: a hypothesis for the origin of cellular domain. *Proc. Natl. Acad. Sci. USA* 103, 3669–3674.
- Freitag, M., Selker, E.U., 2005. Controlling DNA methylation: many roads to one modification. *Curr. Opin. Genet. Dev.* 15, 191–199.
- Freitas, M.A., Sklenar, A.R., Parthun, M.R., 2004. Application of mass spectrometry to the identification and quantification of histone post-translational modifications. *J. Cell. Biochem.* 92, 691–700.
- Frenkiel-Krispin, I., Ben-Avraham, D., Englander, J., Shimoni, E., Wolf, S.G., Minsky, A., 2004. Nucleoid restructuring in stationary-state bacteria. *Mol. Microbiol.* 51, 395–405.
- Frye, R.A., 2000. Phylogenetic classification of prokaryotic and eukaryotic Sir2-like proteins. *Biochem. Biophys. Res. Commun.* 273, 793–798.
- García, B.A., Hake, S.B., Diaz, R.L., Kauer, M., Morris, S.A., Recht, J., Shabanowitz, J., Mishra, N., Strahl, B.D., Allis, C.D., Hunt, D.F., 2007. Organismal differences in post-translational modifications in histones H3 and H4. *J. Biol. Chem.* 282, 7641–7655.
- Glansdorff, N., Xu, Y., Labedan, B., 2008. The last universal common ancestor: emergence, constitution and genetic legacy of an elusive forerunner. *Biol. Direct* 3, 29.
- Grayling, R.A., Sandman, K., Reeve, J.N., 1996. Histones and chromatin structure in hyperthermophilic Archaea. *FEMS Microbiol. Rev.* 18, 203–213.
- Gregoret, I.V., Lee, Y.M., Goodson, H.V., 2004. Molecular evolution of the histone deacetylase family: functional implications of phylogenetic analysis. *J. Mol. Biol.* 338, 17–31.
- Grieshaber, N.A., Fischer, E.R., Mead, D.J., Dooley, C.A., Hackstadt, T., 2004. Chlamydial histone-DNA interactions are disrupted by a metabolite in the methylerythritol phosphate pathway of isoprenoid biosynthesis. *Proc. Natl. Acad. Sci. USA* 101, 7451–7456.
- Guo, L., Feng, Y., Zhang, Z., Yao, H., Luo, Y., Wang, J., Huang, L., 2008. Biochemical and structural characterization of Cren7, a novel chromatin protein conserved among Crenarchaea. *Nucleic Acids Res.* 36, 1129–1137.
- Hake, S.B., Allis, C.D., 2006. Histone H3 variants and their potential role in indexing mammalian genomes: the “H3 barcode hypothesis”. *Proc. Natl. Acad. Sci. USA* 103, 6428–6435.
- Hall, I.M., Shankaranarayana, G.D., Noma, K., Ayoub, N., Cohen, A., Grewal, S.I., 2002. Establishment and maintenance of a heterochromatin domain. *Science* 297, 2232–2237.
- Hall, T.M., 2005. Multiple modes of RNA recognition by zinc finger proteins. *Curr. Opin. Struct. Biol.* 15, 367–373.
- Hammet, A., Magill, C., Heierhorst, J., Jackson, S.P., 2007. Rad9 BRCT domain interaction with phosphorylated H2AX regulates the G1 checkpoint in budding yeast. *EMBO Rep.* 8, 851–857.
- Healy, S., Heightman, T.D., Hohmann, L., Schriemer, D., Gravel, R.A., 2009. Nongenymatic biotinylation of histone H2A. *Protein Sci.* 18, 314–328.
- Hekimoglu, B., Ringrose, L., 2009. Non-coding RNAs in polycomb/trithorax regulation. *RNA Biol.* 6, 129–137.
- Henikoff, S., 2005. Histone modifications: combinatorial complexity or cumulative simplicity? *Proc. Natl. Acad. Sci. USA* 102, 5308–5309.
- Henikoff, S., 2008. Nucleosome destabilization in the epigenetic regulation of gene expression. *Nat. Rev. Genet.* 9, 15–26.
- Hochstrasser, M., 2000. Evolution and function of ubiquitin-like protein-conjugation systems. *Nat. Cell Biol.* 2, E153–E157.
- Hochstrasser, M., 2009. Origin and function of ubiquitin-like proteins. *Nature* 458, 422–429.
- Istrail, S., De-Leon, S.B., Davidson, E.H., 2007. The regulatory genome and the computer. *Dev. Biol.* 310, 187–195.
- Iyer, L.M., Anantharaman, V., Wolf, M.Y., Aravind, L., 2008. Comparative genomics of transcription factors and chromatin proteins in parasitic protists and other eukaryotes. *Int. J. Parasitol.* 38, 1–31.
- Izawa, S., Inoue, Y., 2009. Post-transcriptional regulation of gene expression in yeast under ethanol stress. *Biotechnol. Appl. Biochem.* 53, 93–99.
- Izban, M.G., Luse, D.S., 1992. Factor-stimulated RNA polymerase II transcribes at physiological elongation rates on naked DNA but very poorly on chromatin templates. *J. Biol. Chem.* 267, 13647–13655.
- Jelinska, C., Conroy, M.J., Craven, C.J., Hounslow, A.M., Bullough, P.A., Waltho, J.P., Taylor, G.L., White, M.F., 2005. Obligate heterodimerization of the archaeal Alba2 protein with Alba1 provides a mechanism for control of DNA packaging. *Structure* 13, 963–971.
- Jiang, C., Pugh, B.F., 2009. Nucleosome positioning and gene regulation: advances through genomics. *Nat. Rev. Genet.* 10, 161–172.
- Jin, J., Cai, Y., Li, B., Conaway, R.C., Workman, J.L., Conaway, J.W., Kusch, T., 2005. In and out: histone variant exchange in chromatin. *Trends Biochem. Sci.* 30, 680–687.
- Kamakaka, R.T., Biggins, S., 2005. Histone variants: Deviants? *Genes Dev.* 19, 295–310.
- Kasinsky, H.E., Lewis, J.D., Dacks, J.B., Ausió, J., 2001. Origin of H1 linker histones. *FASEB J.* 15, 34–42.
- Kennely, P.J., 2002. Protein kinases and protein phosphatases in prokaryotes: a genomic perspective. *FEMS Microbiol. Lett.* 206, 1–8.
- Kennely, P.J., 2003. Archaeal protein kinases and protein phosphatases: insights from genomics and biochemistry. *Biochem. J.* 370, 373–389.
- Khalil, A.M., Guttman, M., Huarte, M., Garber, M., Raj, A.M., Rivea Morales, D., Thomas, K., Presser, A., Bernstein, B.E., van Oudenaarden, A., Regev, A., Lander, E.S., Rinn, J.L., 2009. Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. *Proc. Natl. Acad. Sci. USA* 106, 11675–11680.
- Klenov, M.S., Lavrov, S.A., Stolyarenko, A.D., Ryazansky, S.S., Aravin, A.A., Tuschl, T., Gvozdev, V.A., 2007. Repeat-associated siRNAs cause chromatin silencing of retrotransposons in the *Drosophila melanogaster* germline. *Nucleic Acids Res.* 35, 5430–5438.
- Koonin, E.V., Martin, W., 2005. On the origin of genomes and cells within inorganic compartments. *Trends Genet.* 21, 647–654.
- Kothapalli, N., Camporeale, G., Kueh, A., Chew, Y.C., Oommen, A.M., Griffin, J.B., Zempleni, J., 2005. Biological functions of biotinylated histones. *J. Nutr. Biochem.* 16, 446–448.
- Koutroubas, G., Merika, M., Thanos, D., 2008. Bypassing the requirements for epigenetic modifications in gene transcription by increasing enhancer strength. *Mol. Cell Biol.* 28, 926–938.
- Kuang, F.L., Luo, Z., Scharff, M.D., 2009. H3 trimethyl K9 and H3 acetyl K9 chromatin modifications are associated with class switch recombination. *Proc. Natl. Acad. Sci. USA* 106, 5288–5293.
- Kurth, H.M., Mochizuki, K., 2009. Non-coding RNA: a bridge between small RNA and DNA. *RNA Biol.* 6, 138–140.
- Leeb, M., Steffen, P.A., Wutz, A., 2009. X chromosome inactivation sparked by non-coding RNA. *RNA Biol.* 6, 94–99.
- Lesne, A., 2006. The chromatin regulatory code: beyond a histone code. *Eur. Phys. J. E Soft Matter* 19, 375–377.
- Levine, M., Davidson, E.H., 2005. Gene regulatory networks for development. *Proc. Natl. Acad. Sci. USA* 102, 4936–4942.
- Li, B., Carey, M., Workman, J.L., 2007. The role of chromatin during transcription. *Cell* 128, 707–719.
- Liu, Y., Taverna, S.D., Muratore, T.L., Shabanowitz, J., Hunt, D.F., Allis, C.D., 2007. RNAi-dependent H3K27 methylation is required for heterochromatin formation and DNA elimination in *Tetrahymena*. *Genes Dev.* 21, 1530–1545.
- Lomvardas, S., Thanos, D., 2002. Opening chromatin. *Mol. Cell* 9, 209–211.
- Long, K.S., Poehlsgaard, J., Kehrenberg, C., Schwarz, S., Vester, B., 2006. The Cfr rRNA methyltransferase confers resistance to phenicols, lincosamides, oxazolidinones, pleuromutilins, and streptogramin A antibiotics. *Antimicrob. Agents Chemother.* 50, 2500–2505.
- Lowary, P.T., Widom, J., 1998. New DNA sequence rules for high affinity binding to histone octamer and sequence-directed nucleosome positioning. *J. Mol. Biol.* 276, 19–42.
- Lozada-Chávez, I., Chandra Janga, S., Collado-Vides, J., 2006. Bacterial regulatory networks are extremely flexible in evolution. *Nucleic Acids Res.* 34, 3434–3445.
- Lucchini, S., Rowley, G., Goldberg, M.D., Hurd, D., Harrison, M., Hinton, J.C.D., 2006. H-NS mediates the silencing of laterally acquired genes in bacteria. *PLoS Pathogens* 2, e81.
- Luijsterburg, M.S., White, M.F., van Driel, R., Dame, R.T., 2008. The major architects of chromatin: architectural proteins in bacteria archaea and eukaryotes. *Crit. Rev. Biochem. Mol. Biol.* 43, 393–418.
- Malik, H.S., Henikoff, S., 2003. Phylogenomics of the nucleosome. *Nat. Struct. Biol.* 10, 882–891.
- Marc, F., Sandman, K., Lurz, R., Reeve, J.N., 2002. Archaeal histone tetramerization determines DNA affinity and the direction of DNA supercoiling. *J. Biol. Chem.* 277, 30879–30886.



- Marsh, V.L., Peak-Chew, S.Y., Bell, S.D., 2005. Sir2 and the acetyltransferase, Pat, regulate the archaeal chromatin protein, Alba. *J. Biol. Chem.* 280, 21122–21128.
- Martens, C., Vandepoele, K., Van de Peer, Y., 2008. Whole-genome analysis reveals molecular innovations and evolutionary transitions in chromalveolate species. *Proc. Natl. Acad. Sci. USA* 105, 3427–3432.
- Mat, W.K., Xue, H., Wong, J.T., 2008. The genomics of LUCA. *Front Biosci.* 13, 5605–5613.
- Mattick, J.S., 2007. A new paradigm for developmental biology. *J. Exp. Biol.* 210, 1526–1547.
- Mattick, J.S., Amaral, P.P., Dinger, M.E., Mercer, T.R., Mehler, M.F., 2009. RNA regulation of epigenetic processes. *Bioessays* 31, 51–59.
- Matzke, M., Kanno, T., Daxinger, L., Huettel, B., Matzke, A.J., 2009. RNA-mediated chromatin-based silencing in plants. *Curr. Opin. Cell Biol.* 21, 367–376.
- Mersfelder, E.L., Parthun, M.R., 2006. The tale beyond the tail: histone core domain modifications and the regulation of chromatin structure. *Nucleic Acids Res.* 34, 2653–2662.
- Moazed, D., 2009. Small RNAs in transcriptional gene silencing and genome defence. *Nature* 457, 413–420.
- Moreno Díaz de la Espina, S., Alverca, E., Cuadrado, A., Franca, S., 2005. Organization of the genome and gene expression in a nuclear environment lacking histones and nucleosomes: the amazing dinoflagellates. *Eur. J. Cell Biol.* 84, 137–149.
- Morgan, H.D., Santos, F., Green, K., Dean, W., Reik, W., 2005. Epigenetic reprogramming in mammals. *Hum. Mol. Genet.* 14 (Spec No 1), 47–58.
- Morrison, D.K., 2009. The 14-3-3 proteins: integrators of diverse signaling cues that impact cell fate and cancer development. *Trends Cell Biol.* 19, 16–23.
- Mujtaba, S., Zeng, L., Zhou, M.M., 2007. Structure and acetyl-lysine recognition of the bromodomain. *Oncogene* 26, 5521–5527.
- Murata, M., Azuma, Y., Miura, K., Rahman, M.A., Matsutani, M., Aoyama, M., Suzuki, H., Sugi, K., Shirai, M., 2007. Chlamydia SET domain protein functions as a histone methyltransferase. *Microbiology* 153, 585–592.
- Nelson, C.J., Santos-Rosa, H., Kouzarides, T., 2006. Proline isomerization of histone H3 regulates lysine methylation and gene expression. *Cell* 126, 905–916.
- Nemeth, M.J., Lowrey, C.H., 2004. An erythroid-specific chromatin opening element increases beta-globin gene expression from integrated retroviral gene transfer vectors. *Gene Ther. Mol. Biol.* 8, 475–486.
- Ng, R.K., Gurdon, J.B., 2008. Epigenetic inheritance of cell differentiation status. *Cell Cycle* 7, 1173–1177.
- Nijman, S.M., Luna-Vargas, M.P., Velds, A., Brummelkamp, T.R., Dirac, A.M., Sixma, T.K., Bernards, R., 2005. A genomic and functional inventory of deubiquitinating enzymes. *Cell* 123, 773–786.
- Noom, M.C., Hol, F.J.H., Laurens, N., While, M.F., Dame, R.T., Wuite, G.J.L., 2009. Unravelling the role of Alba in the organization of the archaeal nucleoid. *Biophys. J.* 96 (Suppl. 1), 61a (Poster Abstract).
- Norris, V., 2005. Poly-(R)-3-hydroxybutyrate and the pioneering work of Rosetta Natoli Reusch. *Cell Mol. Biol.* 51, 629–634.
- Orfanotiou, F., Tzamalís, P., Thanassoulas, A., Stefanidi, E., Zees, A., Boutou, E., Vlassi, M., Nounesis, G., Vorgias, C.E., 2009. The stability of the archaeal HU histone-like DNA-binding protein from *Thermoplasma volcanium*. *Extremophiles* 13, 1–10.
- Pandey, R., Müller, A., Napoli, C.A., Selinger, D.A., Pikaard, C.S., Richards, E.J., Bender, J., Mount, D.W., Jorgensen, R.A., 2002. Analysis of histone acetyltransferase and histone deacetylase families of *Arabidopsis thaliana* suggests functional diversification of chromatin modification among multicellular eukaryotes. *Nucleic Acids Res.* 30, 5036–5055.
- Pandey, R.R., Mondal, T., Mohammad, F., Enroth, S., Redrup, L., Komorowski, J., Nagano, T., Mancini-Dinardo, D., Kanduri, C., 2008. Kcnq1ot1 antisense noncoding RNA mediates lineage-specific transcriptional silencing through chromatin-level regulation. *Mol. Cell* 32, 232–246.
- Péllissier, T., Wassenegger, M., 2000. A DNA target of 30 bp is sufficient for RNA-directed DNA methylation. *RNA* 6, 55–65.
- Perez, J.C., Latifi, T., Groisman, E.A., 2008. Overcoming H-NS-mediated transcriptional silencing of horizontally acquired genes by the PhoP and SlyA proteins in *Salmonella enterica*. *J. Biol. Chem.* 283, 10773–10783.
- Pérez-Martín, J., de Lorenzo, V., 1997. Clues and consequences of DNA bending in transcription. *Annu. Rev. Microbiol.* 51, 593–628.
- Pinson, V., Takahashi, M., Rouviere-Yaniv, J., 1999. Differential binding of the *Escherichia coli* HU, homodimeric forms and heterodimeric form to linear, gapped and cruciform DNA. *J. Mol. Biol.* 267, 485–497.
- Poole, A.M., Logan, D.T., 2005. Modern mRNA proofreading and repair: clues that the last universal ancestor possessed an RNA genome? *Mol. Biol. Evol.* 22, 1444–1455.
- Probst, A.V., Dunleavy, E., Almouzni, G., 2009. Epigenetic inheritance during the cell cycle. *Nat. Rev. Mol. Cell Biol.* 10, 192–206.
- Prohaska, S.J., P.F., Stadler, P.F., 2007. A story of growing confusion: genes and their regulation. In: Mondaini, R.P., Dilão, R. (Eds.), *BIOMAT-2007: International Symposium on Mathematical and Computational Biology*. World Scientific, Singapore, pp. 325–345 (Armação dos Búzios, RJ, Brazil, 24–29 November 2008).
- Rando, O.J., Ahmad, K., 2007. Rules and regulation in the primary structure of chromatin. *Curr. Opin. Cell Biol.* 19, 250–256.
- 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.
- Reik, W., 2007. Stability and flexibility of epigenetic gene regulation in mammalian development. *Nature* 447, 425–432.
- Rinn, J.L., Kertesz, M., Wang, J.K., Squazzo, S.L., Xu, X., Bruggmann, S.A., Goodnough, L.H., Helms, J.A., Farnham, P.J., Segal, E., Chang, H.Y., 2007. Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. *Cell* 129, 1311–1323.
- Rodríguez-Trelles, F., Rosa Tarrío, R., Ayala, F.J., 2006. Origins and evolution of spliceosomal introns. *Annu. Rev. Genet.* 40, 47–76.
- Roh, T.Y., Cuddapah, S., Zhao, K., 2005. Active chromatin domains are defined by acetylation islands revealed by genome-wide mapping. *Genes Dev.* 19, 542–552.
- Rundlett, S.E., Carmen, A.A., Kobayashi, R., Bavykin, S., Turner, B.M., Grunstein, M., 1996. HDA1 and RPD3 are members of distinct yeast histone deacetylase complexes that regulate silencing and transcription. *Proc. Natl. Acad. Sci. USA* 93, 14503–14508.
- Ryan, M.P., Jones, R., Morse, R.H., 1998. SWI-SNF complex participation in transcriptional activation at a step subsequent to activator binding. *Mol. Cell Biol.* 18, 1774–1782.
- Sanchez-Elsner, T., Gou, D., Kremmer, E., Sauer, F., 2006. Noncoding RNAs of trithorax response elements recruit *Drosophila* Ash1 to Ultrabithorax. *Science* 311, 1118–1123.
- Sandman, K., Grayling, R.A., Dobrinski, B., Lurz, R., Reeve, J.A., 1994. Growth-phase-dependent synthesis of histones in the archaeon *Methanothermobacter fervidus*. *Proc. Natl. Acad. Sci. USA* 91, 12624–12628.
- Sandman, K., Pereira, S.L., Reeve, J.N., 1998. Diversity of prokaryotic chromosomal proteins and the origin of the nucleosome. *Cell. Mol. Life Sci.* 54, 1350–1364.
- Sandman, K., Reeve, J.N., 2000. Structure and functional relationships of archaeal and eukaryal histones and nucleosomes. *Arch. Microbiol.* 173, 165–169.
- Schübeler, D., Francastel, C., Cimbor, D.M., Reik, A., Martin, D.L., Groudine, M., 2000. Nuclear localization and histone acetylation: a pathway for chromatin opening and transcriptional activation of the human beta-globin locus. *Genes Dev.* 14, 940–950.
- Schubert, H.L., Blumenthal, R.M., Cheng, X., 2003. Many paths to methyltransferase: a chronicle of convergence. *Trends Biochem. Sci.* 28, 329–335.
- Schuettengruber, B., Chourrout, D., Vervoort, M., Leblanc, B., Cavalli, G., 2007. Genome regulation by polycomb and trithorax proteins. *Cell* 128, 735–745.
- Schwartz, Y.B., Pirrotta, V., 2008. Polycomb complexes and epigenetic states. *Curr. Opin. Cell Biol.* 20, 266–273.
- Sedighi, M., Sengupta, A.M., 2007. Epigenetic chromatin silencing: bistability and front propagation. *Phys. Biol.* 4, 246–255.
- Segal, E., Widom, J., 2009. What controls nucleosome positions? *Trends Genet.* 25, 225–243.
- Shi, Y., Berg, J.M., 1995. Specific DNA–RNA hybrid binding by zinc finger proteins. *Science* 268, 282–284.
- Shi, Y., Lan, F., Matson, C., Mulligan, P., Whetstone, J.R., Cole, P.A., Casero, R.A., Shi, Y., 2004. Histone demethylation mediated by the nuclear amine oxidase homolog LSD1. *Cell* 119, 941–953.
- Siegel, T.N., Hekstra, D.R., Kemp, L.E., Figueiredo, L.M., Lowell, J.E., Fenyo, D., Wang, X., Dewell, S., Cross, G.A., 2009. Four histone variants mark the boundaries of polycistronic transcription units in *Trypanosoma brucei*. *Genes Dev.* 23, 1063–1076.
- Singh, S.K., Choudhury, S.R., Roy, S., Sengupta, D.N., 2008. Sequential, structural, and phylogenetic study of BRCT module in plants. *J. Biomol. Struct. Dyn.* 26, 235–245.
- Slesarev, A.I., Belova, G.I., Kozyavkin, S.A., Lake, J.A., 1998. Evidence for an early prokaryotic origin of histones H2A and H4 prior to the emergence of eukaryotes. *Nucleic Acids Res.* 26, 427–430.
- Smith, J.S., Brachmann, C.B., Celic, I., Kenna, M.A., Muhammad, S., Starai, V.J., Avalos, J.L., Escalante-Semerena, J.C., Grubmeyer, C., Wolberger, C., Boeke, J.D., 2000. A phylogenetically conserved NAD<sup>+</sup>-dependent protein deacetylase activity in the Sir2 protein family. *Proc. Natl. Acad. Sci. USA* 97, 6658–6663.
- Starai, V.J., Escalante-Semerena, J.C., 2004. Identification of the protein acetyltransferase (Pat) enzyme that acetylates acetyl-CoA synthetase in *Salmonella enterica*. *J. Mol. Biol.* 340, 1005–1012.
- Sun, W., Xu, X., Pavlova, M., Edwards, A.M., Joachimiak, A., Savchenko, A., Christendat, D., 2005. The crystal structure of a novel SAM-dependent methyltransferase PH1915 from *Pyrococcus horikoshii*. *Protein Sci.* 14, 3121–3128.
- Talbert, P.B., Henikoff, S., 2009. Chromatin-based transcriptional punctuation. *Genes Dev.* 23, 1037–1041.
- Tan, E., Besant, P.G., Zu, X.L., Turck, C.W., Bogoyevitch, M.A., Lim, S.G., Attwood, P.V., Yeoh, G.C., 2004. Histone H4 histidine kinase displays the expression pattern of a liver oncodevelopmental marker. *Carcinogenesis* 25, 2083–2088.
- Tariq, M., Paszkowski, J., 2004. DNA and histone methylation in plants. *Trends Genet.* 20, 244–251.
- Taunton, J., Hassig, C.A., Schreiber, S.L., 1996. A mammalian histone deacetylase related to the yeast transcriptional regulator Rpd3p. *Science* 272, 408–411.
- Taverna, S.D., Li, H., Ruthenburg, A.J., Allis, C.D., Patel, D.J., 2007. How chromatin-binding modules interpret histone modifications: lessons from professional pocket pickers. *Nat. Struct. Mol. Biol.* 14, 1025–1040.
- Taylor, B.L., 2004. An alternative strategy for adaptation in bacterial behavior. *J. Bacteriol.* 186, 3671–3673.
- Taylor, W.R., 2006. Transcription and translation in an RNA world. *Philos. Trans. R. Soc. B* 361, 1751–1760.
- Thanbichler, M., Wang, S.C., Shapiro, L., 2005. The bacterial nucleoid: a highly organized and dynamic structure. *J. Cell. Biochem.* 96, 506–521.

- Thatcher, T.H., Gorovsky, M.A., 1994. Phylogenetic analysis of the core histones H2A, H2B, H3, and H4. *Nucleic Acids Res.* 22, 174–179.
- Torres-Padilla, M.E., Parfitt, D.E., Kouzarides, T., Zernicka-Goetz, M., 2007. Histone arginine methylation regulates pluripotency in the early mouse embryo. *Nature* 445, 214–218.
- Travers, A., Muskhelishvili, G., 2005. Bacterial chromatin. *Curr. Opin. Genet. Dev.* 15, 507–514.
- Treweek, S.C., McLaughlin, P.J., Allshire, R.C., 2005. Methylation: lost in hydroxylation? *EMBO Rep.* 6, 315–320.
- Triana, O., Galanti, N., Olea, N., Hellman, U., Wernstedt, C., Lujan, H., Medina, C., Toro, G.C., 2001. Chromatin and histones from *Giardia lamblia*: a new puzzle in primitive eukaryotes. *J. Cell. Biochem.* 82, 573–582.
- Tsukada, Y., Fang, J., Erdjument-Bromage, H., Warren, M.E., Borchers, C.H., Tempst, P., Zhang, Y., 2006. Histone demethylation by a family of JmjC domain-containing proteins. *Nature* 439, 811–816.
- Verdel, A., Vavasseur, A., Le Gorrec, M., Touat-Todeschini, L., 2009. Common themes in siRNA-mediated epigenetic silencing pathways. *Int. J. Dev. Biol.* 53, 245–257.
- Vetting, M.W., Magnet, S., Nieves, E., Roderick, S.L., Blanchard, J.S., 2004. A bacterial acetyltransferase capable of regioselective *N*-acetylation of antibiotics and histones. *Chem. Biol.* 11, 565–573.
- Vilkaitis, G., Suetake, I., Klimasauskas, S., Tajima, S., 2005. Processive methylation of hemimethylated CpG sites by mouse Dnmt1 DNA methyltransferase. *J. Biol. Chem.* 280, 64–72.
- Wang, H., Zhai, L., Xu, J., Joo, H.-Y., Jackson, S., Erdjument-Bromage, H., Tempst, P., Xiong, Y., Zhang, Y., 2006. Histone H3 and H4 ubiquitylation by the CUL4-DDB-ROC1 ubiquitin ligase facilitates cellular response to DNA damage. *Mol. Cell* 22, 383–394.
- Wang, Y., Wysocka, J., Sayegh, J., Lee, Y.H., Perlin, J.R., Leonelli, L., Sonbuchner, L.S., McDonald, C.H., Cook, R.G., Dou, Y., Roeder, R.G., Clarke, S., Stallcup, M.R., Allis, C.D., Coonrod, S.A., 2004. Human PAD4 regulates histone arginine methylation levels via demethylation. *Science* 306, 279–283.
- Wardleworth, B.N., Russell, R.J., Bell, S.D., Taylor, G.L., White, M.F., 2002. Structure of alba: an archaeal chromatin protein modulated by acetylation. *EMBO J.* 21, 4654–4662.
- Washietl, S., Machné, R., Goldman, N., 2008. Evolutionary footprints of nucleosome positions in yeast. *Trends Genet.* 24, 583–587.
- Weake, V.M., Workman, J.L., 2008. Histone ubiquitination: triggering gene activity. *Mol. Cell* 28, 653–663.
- Wenkert, D., Allis, C.D., 1984. Timing of the appearance of macronuclear-specific histone variant hv1 and gene expression in developing new macronuclei of *Tetrahymena thermophila*. *J. Cell. Biol.* 98, 2107–2117.
- White, M.F., Bell, S.D., 2002. Holding it together: chromatin in the archaea. *Trends Genet.* 18, 621–626.
- Wilson, D., Pethica, R., Zhou, Y., Talbot, C., Vogel, C., Madera, M., Chothia, C., Gough, J., 2009. SUPERFAMILY—comparative genomics, datamining and sophisticated visualisation. *Nucleic Acids Res.* 37, D380–D386.
- Wolf, Y.I., Koonin, E.V., 2007. On the origin of the translation system and the genetic code in the RNA world by means of natural selection, exaptation, and subfunctionalization. *Biol. Direct* 2, 14.
- Wong, J.T.Y., New, D.C., Wong, J.C.W., Hung, V.K.L., 2003. Histone-like proteins of the dinoflagellate *Cryptothecodinium cohnii* have homologies to bacterial DNA-binding proteins. *Eukaryotic Cell* 2, 646–650.
- Xu, C., Cui, G., Botuyan, M.V., Mer, G., 2008. Structural basis for the recognition of methylated histone H3K36 by the Eaf3 subunit of histone deacetylase complex Rpd3S. *Structure* 16, 1740–1750.
- Yang, J., Medvedev, S., Yu, J., Schultz, R.M., Hecht, N.B., 2006. Deletion of the DNA/RNA-binding protein MSY2 leads to post-meiotic arrest. *Mol. Cell. Endocrinol.* 250, 20–24.
- Yoon, H.S., Hackett, J.D., Pinto, G., Bhattacharya, D., 2002. The single, ancient origin of chromist plastids. *Proc. Natl. Acad. Sci. USA* 99, 15507–15512.
- Zhang, L., Eugeni, E.E., Parthun, M.R., Freitas, M.A., 2003. Identification of novel histone post-translational modifications by peptide mass fingerprinting. *Chromosoma* 112, 77–86.
- Zhang, L., Jones, K., Gong, F., 2009. The molecular basis of chromatin dynamics during nucleotide excision repair. *Biochem. Cell Biol.* 87, 265–272.
- Zhou, W., Wang, X., Rosenfeld, M.G., 2009. Histone H2A ubiquitination in transcriptional regulation and DNA damage repair. *Int. J. Biochem. Cell Biol.* 41, 12–15.