Old cogs, new tricks: the evolution of gene expression in a chromatin context

Paul B. Talbert, Michael P. Meers* and Steven Henikoff*

Abstract | Sophisticated gene-regulatory mechanisms probably evolved in prokaryotes billions of years before the emergence of modern eukaryotes, which inherited the same basic enzymatic machineries. However, the epigenomic landscapes of eukaryotes are dominated by nucleosomes, which have acquired roles in genome packaging, mitotic condensation and silencing parasitic genomic elements. Although the molecular mechanisms by which nucleosomes are displaced and modified have been described, just how transcription factors, histone variants and modifications and chromatin regulators act on nucleosomes to regulate transcription is the subject of considerable ongoing study. We explore the extent to which these transcriptional regulatory components function in the context of the evolutionarily ancient role of chromatin as a barrier to processes acting on DNA and how chromatin proteins have diversified to carry out evolutionarily recent functions that accompanied the emergence of differentiation and development in multicellular eukaryotes.
The origin of the eukaryotic cell (eukaryogenesis) is controversial, particularly the origin of the nucleus (karyogenesis). The chimeric nature of eukaryotes became clear when the endosymbiotic origin of mitochondria and chloroplasts from an α-proteobacterium and cyanobacterium, respectively, became generally accepted.\(^ \text{158,159} \) Around the same time, Woese and colleagues\(^ \text{160,161} \) discovered that archaea were a distinct domain of life separate from bacteria and eukaryotes, with particular signatures in tRNA and ribosomal RNA (rRNA) molecules and lipid membranes distinct from those of bacteria and eukaryotes. Woese argued that these three domains were anciently diverged and referred to the ancestral lineage of modern eukaryotes as urkaryotes (three-domains hypothesis; see the figure, part a). Directional models of evolution applied to protein domain data sets of all three cellular domains support the primary divergence of the eukaryotic lineage.\(^ \text{162} \)

The close relationship between informational molecules of archaea and eukaryotes has led to a popular class of models in which the ancestral host cell for the mitochondrion was not an urkaryote but an archaean cell (two-domains hypothesis; see the figure, part b), most recently proposed to be among the Asgardarchaeota\(^ \text{163} \), although this remains disputed.\(^ \text{164,165} \) This scenario requires an unprecedented replacement of the cell membrane and virome of the archaean host.\(^ \text{11} \) Alternative models propose that the mitochondrial host was a chimeric cell resulting from an endosymbiotic archaean in a bacterial cell\(^ \text{166,167} \) (see the figure, part c). Regardless of the nature of the host cell, the eukaryotic genome encodes as many protein fold superfamilies specifically shared with viral genomes as with archaea.\(^ \text{11} \) Models of viral eukaryogenesis have attributed the origin of the nucleus to virus factories of nucleocytoplasmic large DNA viruses (NCLDV’s) (see the figure, part d) or proposed its evolution as a protection from viral infections.\(^ \text{11} \) Large viruses have contributed numerous genes to eukaryotes,\(^ \text{163} \) possibly including the mRNAs-capping enzyme, certain DNA and RNA polymerases, topoisomerase IIA,\(^ \text{11,120} \), and even core histones.\(^ \text{11} \)

**Metamonads**

Anaerobic cells typically with two pairs of basal bodies with one posterior and three anterior flagella. Metamonads include the diplomonad Giardia intestinalis and the parabasalid Trichomonas vaginalis, among others, and may represent one of the earliest branches of the eukaryotic phylogenetic tree.

**Kinetoplastids**

Flagellates that have a dense mass of DNA called a kinetoplast, which contains many copies of the mitochondrial genome. Kinetoplastids include Bodonids and trypanosomes and are thought to represent an early branch of the eukaryotic phylogenetic tree.

---

**Box 1 | Models of eukaryogenesis**

<table>
<thead>
<tr>
<th>a Endokaryogenesis</th>
<th>b Endokaryogenesis</th>
<th>c Archaeal karyogenesis</th>
<th>d Viral karyogenesis</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Urkaryotic host</strong></td>
<td><strong>Archaean host</strong></td>
<td><strong>Bacterial host</strong></td>
<td><strong>Urkaryotic host</strong></td>
</tr>
<tr>
<td><strong>LECA</strong></td>
<td><strong>Proto-eukaryote</strong></td>
<td><strong>Prokaryote</strong></td>
<td></td>
</tr>
</tbody>
</table>

---

The complexity of LECA suggests dramatic evolution in a proto-eukaryote stage between FECA and LECA, which already possessed most features of modern chromatin regulation.\(^ \text{11} \) Thus, there are few clues to the intermediate steps in the remarkable transition to nucleosome-based gene regulation. The chimeric genome of LECA had already assembled the component genes of chromatin regulation from homologues in bacteria, archaea and viruses and invented many new genes.\(^ \text{31} \) Despite the dearth of information on this transition, we find it of interest to place the available data in an evolutionary context.

A fundamental role of nucleosomes in the evolution of eukaryotic genomes has been as a nonspecific, passive barrier to DNA-templated processes. We propose that the evolution of chromatin proteins from prokaryotic homologues has been shaped by their increasing participation in the dynamic processes whereby nucleosome assembly competes with TF binding and transcription by RNA polymerase II (RNAPII) or other RNAPs, both of which require transient unwrapping of nucleosomes. It is likely that the acquisition of features that enable or reduce nucleosome mobilization by some histone modifiers facilitated the evolution of genome complexity in multicellular eukaryotes and established a generalized modular framework for complex cell and tissue differentiation regimes that continues to evolve today.

We focus here on histones and non-histone chromatin proteins that mobilize nucleosomes for transcription initiation by RNAPII, including topoisomerases, TFS, remodelers, histone acetyltransferases (HATs), histone deacetylases (HDACs), the histone variant H2A.Z, and PTMs associated with initiation. We also discuss Polycomb group negative regulators of transcription. We do not discuss in detail PTMs of transcriptional elongation or constitutive heterochromatin, as excellent reviews exist on the descriptive and mechanistic aspects of these topics.\(^ \text{14–16} \) Also outside of our scope are the regulatory roles of long non-coding RNAs, as our limited understanding of the evolutionary history of these RNA is discussed elsewhere.\(^ \text{17} \)

**Transcriptional machinery**

In the bacterium *Escherichia coli*, regulation of ~4,500 genes is achieved by seven different sigma factors that recruit RNAP to different sets of genes. Transcription is further modulated by 300 TFs that bind between
one and hundreds of genes each, together with RNAP components, and stabilize or occlude RNAP binding (reviewed elsewhere 18). By contrast, eukaryotes are often considered to be the sisters or descendants of archaea, because eukaryotic RNAPII and its associated general transcription factors TATA-binding protein (TBP), TFIIB and TFIIE — as well as RNAPI, RNAPIII and their own paralogous general transcription factors — all have homologues in archaea. Both archaea and eukaryotes assemble a pre-initiation complex (PIC) in which TBP binds to a TATA box, bending DNA, and is stabilized by TFIIB or its homologues, which together recruit the RNAP19. TFIIE and its homologues may open or stabilize the transcription bubble20. TBP is highly conserved in eukaryotes, but archaea may have several paralogues of TBP and the TFIIB homologue TFB, although to what extent these bind to different sequences or TFs is not clear19. Archaea lack homologues of the eukaryotic factors TFIIA, TFII F, TFIIH, the TBP-associated factor (TAF) proteins that together with TBP form TFIID, and the mediator complex, as well as polymerase-specific factors for RNAPI and RNAPIII19,21. Some TAF proteins make additional DNA contacts that facilitate TBP binding to promoters that lack a TATA box19, and TAF1 inhibits the TBP-specific remodeler MOT1 from removing TBP from these lower-affinity sites22. A recent analysis proposes that eukaryotic RNAPIII is the direct homologue of archaeal RNAP and was transferred from an urkaryote or proto-urkaryote to the ancestor of nucleocytoplasmic large DNA viruses (NCLDVs), where it diversified along with them. Subsequently, RNAPII and the large subunit of RNAPI were acquired from different NCLDV lineages by a proto-eukaryote before LECA23.

**Origin of eukaryotic nucleosomes**

LECA packaged its genome in nucleosomes similar to those of modern eukaryotes, with PTMs of histone tails and variant histones such as cenH3 in centromeres1. HFD proteins are found in all three cellular domains, although the functions and DNA-binding abilities of some HFD families in bacteria and archaea are unknown24. Many archaea have one or more histones of the HMfB family that bind to DNA and are more similar to eukaryotic histones, although none is specifically related in sequence to eukaryotic histones, depending on the model of eukaryogenesis (BOX 1). Archaeal histones mostly lack the unstructured tails of eukaryotic histones, but recently some have been found to have tails containing lysine, raising the possibility that they may be modified25. Archaeal histones fold together in dimers, usually homomeric, which can be further polymerized through four-helix bundles to form tetrameric HFD particles that wrap 60 bp of DNA or more extended DNA-wrapping polymers26 termed hypernucleosomes.

**SET domain**

A protein domain generally associated with protein lysine methyltransferase activity.

**Last eukaryotic common ancestor (LECA)**: The most recent cell that was ancestral to all extant eukaryotes, inferred to be a complex nucleated cell with an endomembrane system, a cytoskeleton, a mitochondrion, and linear chromosomes that were packaged in nucleosomes and underwent mitosis and meiosis.

**First eukaryotic common ancestor (FECA)**: A hypothetical first cell that was ancestral to all eukaryotes and distinct from bacteria and archaea, variously conceived as an urkaryote or a chimeric organism composed of host and endosymbiont.
Urkaryote
A hypothetical cell ancestral to the eukaryotic cytoplasmin with distinct features of the translational apparatus that distinguish it from archaea and bacteria and that arose independently from the last universal cellular ancestor.

Endosymbiont
A cell or organism living inside another cell or organism.

Proto-eukaryote
A cell in any stage of eukaryotic evolution between the first eukaryotic common ancestor to the last eukaryotic common ancestor.

General transcription factors
The factors that together with RNA polymerase II make up the pre-initiation complex for transcription.

Nucleocytoplasmic large DNA viruses (NCLDV)
A diverse monophyletic group of fairly large, complex DNA viruses, including giant viruses such as the Marseilleviridae as well as other more modestly sized viruses.

CenH3
Histone H3 variant specific to centromeric nucleosomes, examples of which include mammalian CENP-A and budding yeast Cse4.

Eukaryogenesis
The process of evolving from the first eukaryotic common ancestor to the last eukaryotic common ancestor.

Hypernucleosomes
Archaeal histone polymers of variable length that wrap DNA.

Monophyletic group
A group of genes, proteins or organisms that includes all the descendants of a single common ancestor and excludes others.

Capsid
A proteinaceous structure that encloses a viral genome for protection and dispersion.

Winged helix domain
A protein domain with combined specific and nonspecific DNA-binding affinity characterized by a helix–turn–helix motif flanked by β-sheets on one or both sides.

[FIG. 1a]. In vitro, archaeal histones can repress transcription initiation and slow elongation, but promoter occupancy of histones can be outcompeted by a TF. A few archaeal histones have tandemly linked HFD doublets that are constrained to fold together, enabling divergence between the two HFDs in the same protein (FIG. 1b). In Haloferax volcanii, such doublets can then dimerize to form a structure with four HFDs that shows regular spacing on the chromosome, including depletion of such particles at promoters, similar to eukaryotic nucleosomes. Such histone doublets were proposed to be intermediates in the evolution of nucleosomes, facilitating the diversification of the four families of core histones in specific pairs: H2A with H2B and H3 with H4.

The giant viruses of the family Marseilleviridae within the NCLDV's encode HFD doublets that form obligate heterodimers, and these HFDs have unstructured tails and are orthologous to the eukaryotic core histone pairs, with an H2B–H2A-like doublet and a H4–H3-like doublet (FIG. 1c). Each class of Marseilleviridae HFD orthologues of the four core histone classes forms a monophyletic group that is a sister to all eukaryotic histones of the corresponding class, branching at the base of the eukaryotic tree before variants within a class, such as H2A and its variant H2A.Z or H3 and its variant CenH3. The doublet structures make unlikely the scenario that all four core histone classes were separately acquired from modern eukaryotes (FIG. 1d) and then rapidly diverged and reorganized into doublets in Marseilleviridae. It is more parsimonious to suggest that a common ancestor of viral and eukaryotic histones existed before the diversification of histone variants. Marseilleviridae histones may have been acquired from a proto-eukaryote to protect the viral genome from host nucleases and may preserve the predicted doublet intermediate stage of nucleosome evolution. Alternatively, doublet histones may have been acquired from an archaeon and subsequently specialized to package large viral genomes into the constrained space of a capsid. The viral histones could then have been acquired by a proto-eukaryote. Acquisition of a major chromatin-packaging protein family from NCLDVs has precedent in dinoflagellates, which do not package the bulk of their large condensed chromosomes in nucleosomes but nevertheless encode many divergent histones that may be present at transcribed genes. DNA-packaging proteins donated by the NCLDV family Phycodnaviridae to the dinoflagellate ancestor* possibly drove histone depletion.

All eukaryotes encode homologues of the core histone families H2A, H2B, H3 and H4; however, the linker histone H1, which does not have an HFD, has an independent origin and may not have been present in LECA. Homologues of the carboxy-terminal portion of animal H1 are found in bacteria, but not archaea, and are present in most protists except the early diverging metamonads. The winged helix domain of H1 found in animals and plants, however, is absent in several protist groups, suggesting that this domain may have been independently acquired in the animal, fungal and plant lineages* (TABLE 1).

Whether eukaryotic nucleosomes derive from an urkaryote, archaean symbiont, or a giant virus, their packaging function is required for mitotic chromosome condensation* and may have facilitated the expanded size of eukaryotic genomes. However, they may have been initially selected to restrict access to the genome by viruses and transposons. Nucleosomes are barriers to access of genomic DNA by DNA transposons, which prefer to insert into exposed DNA. Although DNA methylation was absent in LECA and is not found in early-diverging eukaryotes, later-diverging eukaryotes repeatedly adapted bacterial DNA methylation enzymes (TABLE 1), used in prokaryotes to discriminate host DNA from foreign DNA, to silence transposons. However, because nucleosomes are also barriers to DNA methylation, chromatin remodelers are necessary to methylate DNA in nucleosomes, and histone modifications help regulate where methylation occurs (BOX 2).

Supercoiling and nucleosomes
Perhaps the most basic consequence of wrapping DNA around nucleosomes is that it constrains negative DNA supercoils throughout the eukaryotic genome, altering the roles of DNA superhelicity or torsion in gene regulation and genome organization. In E. coli, TFs are unknown for the majority of genes, and, instead, superhelicity has a major role in transcriptional regulation. Topoisomerases regulate superhelicity by relaxing or adding supercoils that can form when DNA is constrained in a circle or by anchoring proteins or cellular structures (reviewed elsewhere). Type I topoisomerases make transient single-stranded breaks in DNA, and type II topoisomerases make transient double-stranded breaks, passing DNA through the break before resealing it with an altered superhelicity. Supercoils are generated by polymerases as they unwind helical DNA for transcription or replication, and topoisomerases relax the (overwound) positive supercoils generated ahead of RNAP during transcription and the (underwound) negative supercoils formed in its wake. Although the positive and negative supercoils would be expected to cancel out any net change to the superhelicity of the DNA, nucleoid-associated DNA-binding proteins can constrain the negative supercoils and direct their energy towards promoters, whereas no abundant DNA-binding proteins constrain positive supercoils, resulting in a net negative superhelicity that helps to unwind DNA for transcription or replication. In highly transcribed operons loaded with multiple polymerases, the negative supercoils behind a leading polymerase can cancel the positive supercoils ahead of a following polymerase, facilitating multiple rounds of transcription. Transcription-coupled supercoils can quickly propagate to nearby genes to create local supercoiling domains that coordinate expression from neighbouring genes in temporal expression patterns. In Caulobacter crescentus swarmer cells, interactions between nearby loci form chromosomal interaction domains that are dependent on transcriptional elongation and on DNA gyrase (topoisomerase type IIA family), which can introduce negative supercoils. The interaction domains probably arise from supercoiled plectonemes (twisted loops) separated by more open, highly expressed genes.

Whereas fluid superhelicity is a major factor in bacterial and archaean transcription regulation, eukaryotic
nucleosomes restrict the free propagation of supercoils. Changes in superhelicity are largely accommodated by topoisomerases acting at the linker regions between nucleosomes and at nucleosome-free promoters. The linker histone H1 can bind to positively or negatively supercoiled DNA and may help to control the torsional state of the chromatin. Positive torsional stress induced by RNAPII progression promotes H2A–H2B dimer loss and can displace nucleosomes downstream of elongating RNAPII in Drosophila melanogaster S2 cells, releasing the constrained negative supercoiling energy that was stored in the evicted nucleosomes to counteract positive torsion while eliminating the nucleosome barrier that otherwise induces polymerase backtracking and arrest.

Table 1 | Phylogenomic distribution of chromatin proteins

<table>
<thead>
<tr>
<th>Protein families or subfamilies</th>
<th>Bacteria</th>
<th>Archaea</th>
<th>NCLDVs</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>(+)</td>
<td>(+)</td>
</tr>
<tr>
<td>Histones H2A, H2B, H3, H4</td>
<td>–</td>
<td>(+)</td>
<td>(+)</td>
</tr>
<tr>
<td>H2A.Z</td>
<td>–</td>
<td>–</td>
<td>(+)</td>
</tr>
<tr>
<td>H2A.Z-specific H2B</td>
<td>–</td>
<td>–</td>
<td>H2B,V</td>
</tr>
<tr>
<td>H2A,W</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Short H2As</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>H1</td>
<td>H1-cterm</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>RNA pol</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>TBP</td>
<td>–</td>
<td>+</td>
<td>Some</td>
</tr>
<tr>
<td>TFIIA</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Mediator</td>
<td>–</td>
<td>–</td>
<td>+</td>
</tr>
<tr>
<td>Capping enzyme</td>
<td>–</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Topo type IA</td>
<td>Topo IA,</td>
<td>Topo III,</td>
<td>Topo III</td>
</tr>
<tr>
<td>Topo III, reverse gyrase</td>
<td>–</td>
<td>reverse gyrase</td>
<td>–</td>
</tr>
<tr>
<td>Topo III</td>
<td>Topo III</td>
<td>Topo III</td>
<td></td>
</tr>
<tr>
<td>Topo type IB</td>
<td>+</td>
<td>Thaum</td>
<td>Pox</td>
</tr>
<tr>
<td>Topo type IIA</td>
<td>Gyrase 6</td>
<td>Topo II</td>
<td></td>
</tr>
<tr>
<td>Topo IV</td>
<td>Gyrase (some)</td>
<td>Topo II</td>
<td></td>
</tr>
<tr>
<td>Topo type IIB</td>
<td>Some</td>
<td>Topo VI</td>
<td></td>
</tr>
<tr>
<td>SWI/SNF ATPases</td>
<td>(+)</td>
<td>(+)</td>
<td>6</td>
</tr>
<tr>
<td>HAT families</td>
<td>(+)</td>
<td>(+)</td>
<td>4</td>
</tr>
<tr>
<td>HDAC families</td>
<td>(+)</td>
<td>(+)</td>
<td>2</td>
</tr>
<tr>
<td>SET domain methylases</td>
<td>(+)</td>
<td>(+)</td>
<td>5</td>
</tr>
<tr>
<td>Histone demethylases</td>
<td>(+)</td>
<td>–</td>
<td>8</td>
</tr>
<tr>
<td>DNA methylases</td>
<td>(+)</td>
<td>(+)</td>
<td>–</td>
</tr>
</tbody>
</table>

Presence of proteins is indicated with + or a number representing the minimum number of protein families or subfamilies. (+) indicates homologous proteins in archaea and bacteria that do not form or act on octameric nucleosomes. H1 proteins may be homologous only to the carboxyl terminus of human H1 proteins (H1-cterm) or additionally have a winged helix (WH) domain. RNAPII* indicates that, in metamonads, RNAPII lacks the characteristic heptad repeat at the carboxyl terminus. Spo11 is the recombination double-strand break enzyme of eukaryotes that lacks topoisomerase activity. CMT, chromomethylase; Dnmt, DNA methyltransferase; HAT, histone acetyltransferase; HDAC, histone deacetylase; NCLDVs, nucleocytoplasmic large DNA viruses; Pox, poxviruses; RNA Pol, RNA polymerase; TBP, TATA-binding protein; Thaum, Thaumarchaeota; Topo, topoisomerase. Eukaryotic tree simplified from REF. 12. Data compiled from REFs 12, 21, 31, 35, 36, 58, 59, 115, 168, 185, 196, 200.

Superhelicity
The degree of torsion or supercoiling in a DNA molecule.

Plectonemes
Writhed loops resulting from supercoiling.
In bacteria, cytosine methylation protects DNA from host-encoded restriction endonucleases. In eukaryotes, methylation has multiple functions, and there are several families of DNA methyltransferases independently acquired from bacteria, which may be preferentially specialized for de novo (Dnmt3/DRM and CMT/Dim-2 families) or for maintenance methylation during replication (Dnmt1/MET1 and Dnmt5 families) \(^{162,163,164}\).

Nucleosomes are barriers to methylation, and in three kingdoms of marine algae, CG methylation by Dnmt5 forms a periodic pattern in linker regions that disfavours nucleosomes, suggesting that mutual antagonisms between methylation and nucleosomes may reinforce nucleosome positions \(^{144}\). Dnmt1/MET1 methylates linkers around well-positioned nucleosomes in the absence of linker histone H1 and of the chromatin remodelers DDM1 in Arabidopsis thaliana and LSH in mouse, which enable methylation of nucleosomes \(^{170}\). Loss of DDM1 or methyltransferases results in activation of transposons \(^{173}\). Transcriptional silencing of transposons is a key function of DNA methylation in plants, fungi and vertebrates \(^{174}\), and methylation also silences gene promoters during mammalian development and differentiation and establishes genomic imprinting (see recent reviews \(^{169,172}\)).

In bacteria, the generation of negative supercoils upstream of eukaryotic promoters can have regulatory consequences but with new ‘twists’. Underwinding DNA can favour the formation of non-B-form DNA structures that exclude nucleosomes \(^{141}\) and enable binding of specific regulatory factors \(^{11}\). Sequence motifs associated with such non-B structures are largely absent from prokaryotic genomes \(^{184}\). In yeast, the spread of transcription-coupled supercoils appears to co-regulate clusters of seven or more contiguous genes \(^{144}\). In human cells, genome-wide mapping of supercoils revealed variably sized supercoiling domains of median size 100 kb \(^{19}\), reminiscent of the chromatin interaction domains of C. crescentus. Underwound domains are associated with transcription, DNase I hypersensitive sites, and topoisomerase I and are depleted for topoisomerase II. They depend on both transcription and topoisomerase activity, have a more decondensed structure than overwound domains, and become compacted if transcription is inhibited \(^{165}\). Transcription of short initiating RNAs is sufficient to decompact the domains when elongation is prevented \(^{166}\). Supercoiling domains may bring enhancers and promoters together in proximity. Non-coding RNAs transcribed from enhancers, known as enhancer RNAs (eRNAs), have been proposed to promote a superhelicity that is favourable for transcription of the promoters on which they act \(^{167}\).

The different superhelical environment imposed by nucleosomes possibly favoured different topoisomerases than those in bacteria and archaea. Eukaryotic type II topoisomerases (type IIA) are not related to the type II topoisomerases (type IIB) used by most archaea \(^{168}\) but instead branch as sisters to the topoisomerases of the Marseilleviridae \(^{169}\). Indeed, topoisomerase IIA was present in the ancestor of all NCLDV s that pre-dated LECA \(^{170,171}\). The relationship of giant viruses to eukaryotes is controversial \(^{172}\), but most topoisomerases may have originated in viruses and been transferred to the three cellular domains after their divergence \(^{173}\) (TABLE 1). Interestingly, topoisomerase II is found along the scaffold of mitotic chromatids \(^{174}\) and is required for decatenation of sister chromatids. This activity is driven by positive supercoiling that is introduced by the combined action of condensin with topoisomerase II in the presence of mitotic spindles \(^{175}\). Topoisomerase II is found with cohesin and CTCF sites at the boundaries of topologically associating domains (TADs), where it may help regulate the superhelicity of TADs \(^{176}\). A model of chromatin condensation proposes that as transcription ceases in metaphase, topoisomerase II and condensin compact DNA into condensed, positively supercoiled domains that form the structure of mitotic chromosomes \(^{177}\). It seems possible that the role of nucleosomes working with condensin and topoisomerase II to compact and decatenate mitotic chromosomes \(^{178}\) may have been a key factor that facilitated the expansion of eukaryotic genomes.

**TFs and remodellers in nucleosome depletion**

Bacterial and archaean TFs share a common origin, with ~53% of archaean TFs having at least one bacterial homologue \(^{179}\). By contrast, only ~2% of archaean TFs have homologues uniquely in eukaryotes, with another 6% having homologues in both eukaryotes and bacteria. Among the latter are 6–11 families with helix–turn–helix (HTH) domains, which are abundant in prokaryotes \(^{180}\). HTH families in eukaryotes, such as the MYB and homeodomain families, are only distantly related to those in prokaryotes, and some eukaryotic HTH proteins may derive from transposons. In addition to HTH proteins, a profusion of ~55 distinct eukaryotic-specific DNA-binding domain families were present in LECA, such as HMGI, AT-hook, C_H, zinc-fingers, and MADS-box proteins \(^{181}\). This diversification might be due, in part, to the expansion of genomes and gene regulatory targets made possible by nucleosomal packaging and perhaps also due to a reduced regulatory role of superhelicity in genomes largely constrained by nucleosomes. Lineage-specific amplifications of TF families are widespread in eukaryotes \(^{182}\) and may underlie eukaryotic morphological diversity.

Diversification of TFs was accompanied by specialization and diversification of the Snf2 family of proteins that use the energy of ATP to slide or evict nucleosomes or other proteins to make TF-binding sites available. The Snf2 family of ATP-dependent remodellers is divided into four major subfamilies, SWI/SNF, ISWI/SNF2L,
CHD/Mi-2, and INO80 (REF. 67), and was represented in LECA by at least six members12. Distant homologues of remodelers comprising two prokaryotic Snf2 subfamilies exist in bacteria and some archaea, where they are probably nonessential and show no relationship to the presence or absence of archaeal histones68, indicating that the eukaryotic remodelling functions evolved de novo to accommodate the acquisition of nucleosomes. An Snf2 homologue has a conserved role in transcription initiation69 in several NCLDVs59, suggesting that the ancestral NCLDV Snf2 may also have been ancestral to eukaryotic remodelers (TABLE 1). The subfamilies of remodelers are distinguished by additional chromatin- binding domains and by their mechanisms of action (Box 3) (reviewed elsewhere67). Members in each subfamily form multiple lineage- specific complexes with other subunits that can interact with DNA, TFs, and histone modifications and thereby determine their specificity67,70,71.

A nucleosome- depleted region (NDR) is found at budding yeast promoters upstream of active transcription start sites (TSSs) and is flanked by two well- positioned nucleosomes72,73, termed the −1 nucleosome (upstream) and +1 nucleosome (downstream). NDRs are maintained by poly(dA:dT) sequence elements that have evolved to resist nucleosome formation74, by the action of remodelers, and by the binding of TFs, PIC components and RNAPII75. NDR length in yeast is correlated with levels of transcription, binding of TBP, and trimethylation of H3 on lysine 4 (H3K4me3) of adjacent nucleosomes and is anticorrelated with H1 occupancy and H2A.Z occupancy76. The NDR is widely conserved in eukaryotes, and TF binding and nucleosome depletion are highly concordant throughout plant and animal development77,78. In zebrafish, depleting nuclear histone concentration while retaining genome- wide nucleosome density causes early activation of zygotic transcription by enabling TFs to quickly outcompete nucleosomes bound at their target sites79. In D. melanogaster, DNA replication removes both nucleosomes and TFs, with recolonization behind the replication fork by nucleosomes that are subsequently outcompeted by TFs in a process that is likely to be both concentration- dependent and remodeler- dependent80. These observations suggest a simple mass- action model of DNA occupation by nucleosomes and TFs, with adjustments by remodelers, which could apply in diverse developmental contexts.

+1 Nucleosome
A highly conserved nucleosome positioned downstream of the nucleosome- depleted region, which acts as a barrier to transcription elongation by RNA polymerase.

Fig. 2 | Supercoiling in prokaryotes and eukaryotes. Supercoiling can manifest as a twist of the DNA double helix about its axis or as a writhe of the DNA looping on itself. Melting and unwinding by polymerases progressing along the DNA produce positive supercoiling ahead of the polymerase and negative supercoiling behind the polymerase. Nucleoid- associated proteins in prokaryotes and nucleosomes in eukaryotes constrain supercoils. In eukaryotes, nucleosomes force supercoiling into the linker regions between nucleosomes. Positive supercoils ahead of a polymerase can evict nucleosomes, whereas negative supercoils behind can favour transcription factor (TF) binding or nucleosome assembly. Topoisomerases (Topos) break one or both strands of DNA to relax supercoils, relieving superhelical torsion. RNAP, RNA polymerase.
**Box 3 | Mechanisms of ATP-dependent remodellers**

The Snf2 family catalytic subunits of ATP-dependent remodellers belong to the superfamily 2 DNA and RNA helicases but have lost helicase activity while retaining the ability to translocate on DNA43. The translocase domains of remodellers bind to DNA on the nucleosome surface and utilize ATP hydrolysis to introduce superhelical torsion or twist, generating the power stroke for remodelling that breaks histone–DNA contacts44,45. Other than INO80, which binds to DNA close to superhelical location 6 (SHL6) at the edge of nucleosomes46,47, the translocase domains of all other remodellers bind at SHL2 close to the centre of nucleosomes to initiate DNA translocation48,49. The translocase tracks the nucleosomal DNA unidirectionally by pulling DNA towards itself and pumping DNA past the site of translocation while remaining tethered by another DNA-anchor or histone-anchor. Nicks or gaps in the DNA backbone of the tracking DNA strand greatly impede translocation100.

Remodellers tune this basic DNA-tracking and DNA-propelling action to bring about specific nucleosomal changes such as nucleosome spacing (ISWI, CHD and INO80), octamer eviction (SWI/SNF), or histone dimer exchange (INO80). Remodeller-specific regulatory domains within the catalytic subunit or associated subunits in the remodeller complex positively or negatively regulate the ATPase and DNA-translocation activities83,84 (reviewed elsewhere79), which are often modulated by histone interactions.

For example, the H4 amino-terminal tail and the acidic patch of H2a or H2a.Z70 for another DNA-anchor or histone-anchor. Nicks or gaps in the DNA backbone of the tracking DNA strand greatly impede translocation180.

Besides translocating DNA on the histone octamer surface, remodellers can also deform the histone octamer core85. Preventing deformation near SHL2 inhibits sliding by the ISWI subfamily SNF2h ATPase, but not by INO80, and increases nucleosome eviction by the SWI/SNF subfamily RSC complex. In yeast, RSC-bound +1 nucleosomes are asymmetrically disrupted in their DNA contacts107, consistent with a cryo-electron-microscopy reconstruction of an RSC–nucleosome complex in which DNA and an H2A–H2B dimer108 are displaced from the nucleosome surface.

**Yeast TFs differ widely in their intrinsic ability to deplete nucleosomes.** Six of 104 TFs studied in vivo bind to DNA tightly at a single motif, are highly expressed, and have strong nucleosome depletion activity, including the ubiquitous so-called general regulatory factors Abf1, Rap1 and Reb1 (REF10). Other TFs, such as Rsc3, a subunit of the RSC remodelling complex, generally require multiple binding sites to deplete nucleosomes, and two-thirds of the 104 TFs are unable to deplete nucleosomes, presumably relying on other TFs, remodellers, or transient unwrapping to access nucleosomal DNA. General regulatory factors and RSC can act independently to deplete nucleosomes at NDRs62. RSC slides the +1 nucleosome towards the gene body, oriented by GC-rich and poly(A) motifs in the NDR, to better expose the TATA promoter element to enable binding of TBP and PIC formation.

Similar to yeast general regulatory factors, pioneer factors in animals are thought to be able to displace nucleosomes at enhancers and promoters to initiate tissue-specific developmental programmes when other TFs cannot (reviewed elsewhere15). The pioneer factors FOXA1 (also known as HNF3A) and FOXA2 (also known as HNF3B) have domains that are structurally similar to histone H1, which binds to entry and exit DNA. They can displace H1, which may enable access by other TFs84. By contrast, in mouse 3134 cells, the glucocorticoid receptor, which binds to DNA upon activation by hormone, binds mostly to nucleosomes, but also in NDRs, and recruits the SWI/SNF remodeller BRG1 to establish an NDR at the receptor binding site85. The pioneer factors Zelda and Gaf act during zygotic transcriptional activation in *D. melanogaster* embryogenesis to deplete nucleosomes and facilitate acetylation and binding by other TFs86.

**Acetylation and nucleosome disruption**

PTMs of histone amino-terminal tails of the four core histones constitute a major mechanism for directing chromatin proteins that bind to specific PTMs to their histone targets. Well-studied PTMs are predominantly found on the highly conserved tails of H3 and H4 (REF11), perhaps because H3–H4 tetramers have lower turnover than H2A–H2B dimers42. The smaller number of PTMs on H2A and H2B tails might be the reason they are less strictly conserved43 than H3 and H4 tails.

Acetylation was proposed decades ago to structurally alter nucleosomes and promote their disruption through neutralizing the charge on the lysines of the tails, reducing DNA–histone binding and limiting their ability to form hydrogen bonds12. Histone acetylation improves the efficiency with which RNAPII traverses chromatin in vitro, similar to the transcriptional effect of removing the tails entirely83. The discovery of acetylation turnover by HATs and HDACs on histones revealed histone acetylation to be an active regulatory mechanism12.

*N*-lysine acetyltransferase homologues of HATs are found in archaea, bacteria and viruses12,45, where they acetylate non-histone substrates. Indeed, HATs and HDACs are more properly known as lysine acetyltransferases (KATs) and lysine deacetylases (KDACs) because they acetylate and deacetylate thousands of other proteins besides histones (reviewed elsewhere12). In addition, they can add and remove longer-chain acyl groups to and from lysine, although in HeLa cells, except for propionylation, which is nearly as abundant as acetylation, longer-chain acylations are more 200-fold less abundant than acetylations11, meaning that any possibilities of metabolic functions are only speculative (reviewed elsewhere12). Acetylation of the nucleoid-associated protein MthU in *Mycobacterium tuberculosis* reduces DNA interaction and decompacts the genome13, attempts to acetylate archaeal histones were unsuccessful14, although archaea acetylate the chromatin protein Alba, which represses in vitro transcription after deacetylation by a sirtuin15. Eukaryotic histones have increased affinity for DNA and less flexibility than archaeal histones, suggesting these features may have necessitated their acetylation for mobilization. LECA had at least four families of HATs in the GNAT/MYST domain superfamily, as well as the unrelated TAF1 HAT family associated with TBP46. The RPD3 and sirtuin HDAC superfamilies were present in LECA, and additional HDACs have been acquired from bacteria at multiple different points in eukaryotic evolution (TABLE 1).

In human cells, both HATs and HDACs associate with active genes and correlate with H3 and H4 acetylation levels, RNAPII levels and gene expression levels, and inhibition of either HDACs or p300/CBP HAT activity inhibits gene expression and reduces RNAPII occupancy, suggesting that acetylation turnover is more important than static acetylation41, perhaps to first loosen chromatin for transcription and then to reset the chromatin landscape to prevent inappropriate transcription initiation. Acetylation of promoters is not dependent on transcription in a yeast RNAPII mutant42, and transcription-coupled acetylation, despite...
being important for RNAPII transit\(^9\), appears to constitute only a small fraction of steady-state histone acetylation\(^9\).

HATs are frequently found in multisubunit complexes and can be recruited to promoters by the interaction of these subunits with the acidic activation domains of TFs\(^59\). Tethering of the p300 acetyltransferase core to promoters or enhancers of inactive genes is sufficient to drive their robust expression\(^100\), suggesting that a major role of TFs is to target acetylation to the genes they activate. However, p300 may also have a role in recruiting RNAPII. The \textit{D. melanogaster} homologue of p300, Cbp, is required to maintain paused RNAPII and to overcome the barrier of the +1 nucleosome, which it acetylates. At highly paused promoters, Cbp may recruit RNAPII through an interaction with TFIIB\(^101\).

In addition to a structural role in loosening DNA contacts, acetyl-lysine provides binding sites for three different protein domains: bromodomodomains, double plant homeodomain fingers (DPFs), and YEATS domains. These acetyl-binding domains are often found in multifunctional proteins or protein complexes with other chromatin-binding and enzymatic domains\(^102\), consistent with the notion that acetyl-lysine plays a critical role in localizing and/or regulating these proteins at promoters. For HATs with acetyl-binding domains, this localization is self-reinforcing. At least four bromodomodomains were present in LECA\(^11\). The bromodomodomains of the HAT Gcn5 and the remodeller Swi2/Snf2 are required in yeast for the stable occupancy at promoter nucleosomes of the SAGA transcription complex and the SWI/SNF remodelling complex\(^31\). Similarly, DPFs are found in the HATs MOZ and MORF where they bind to H3K9ac or H3K14ac and facilitate association with chromatin\(^30\). The YEATS domain protein Yaf9 preferentially binds to H3K27ac\(^38\) and is a component of NuA4 and of the SWR1 complex that deposits H2A.Z. Yaf9 is required for acetylation of H2A.Z by NuA4 and for H2A.Z incorporation at a third of yeast promoters\(^39\). Similarly, in mouse embryonic stem cells (ESCs), the YEATS domain of GAS41, a component of the TIP60 and SRCAP complexes that deposit H2A.Z, binds to H3K27ac and is necessary for H2A.Z deposition at bivalent promoters\(^40\). These findings suggest that acetylation was already an important promoter landmark in LECA.

SET domain proteins serve as histone methyltransferases in all eukaryotes. At least five were present in LECA, and SET domain proteins of unknown function are widespread in bacteria\(^41,108\). Trimethylation of H3K4 by SET1 family histone methyltransferases, which had a homologue in LECA\(^12\), appears to be important for directing acetylation to promoters\(^42\). A minor fraction of HATs and HDACs are found at inactive human genes, and those inactive genes marked with H3K4me1, H3K4me2, or H3K4me3 at their promoters are more likely to become acetylated and occupied by RNAPII upon inhibition of HDACs\(^35\), consistent with the fact that in animals, p300/CBP is necessary for dynamic acetylation on H3 tails marked with K4me3 (REF.\(^11\)). Several other HATs and HDACs contain plant homeodomain (PHD) fingers or tandem tudor domains that bind to H3K4me3 (REF.\(^15\)). In yeast, the NuA3 acetyltransferase has PHD-finger and PWWP domains that independently recruit NuA3 to H3K4me and H3K36me, respectively, although recruitment of NuA3 did not necessarily result in acetylation, suggesting an additional layer of regulation\(^107\). Mammalian BRWD2/PHIP colocalizes extensively with H3K4me through its cryptotudor domain, and depletion of the \textit{D. melanogaster} homologue disrupts the pattern of H3K27ac\(^108\). H3K4 methyltransferases, in turn, can be directed to promoters by the Ser5-initiating form of the RNAPII CTD co-transcriptionally or by TFs, especially at mammalian CpG islands\(^15\). The mammalian MLL H3K4 methyltransferase contains a bromodomain\(^102\), suggesting that HATs and MLL may recruit each other. H3K4me might be largely redundant with TFs in directing HATs, which could explain why loss of H3K4me has few phenotypic consequences other than a reduced proliferative rate in \textit{D. melanogaster} cells\(^11\), despite H3K4me being conserved throughout eukaryotes\(^12–14\).

### Chromatin diversification after LECA

While many features of chromatin regulation of promoters are highly conserved across plants, animals, and fungi, variations occur in early-diverging eukaryotes.

**Promoter architecture.** The root of the eukaryotic tree remains controversial, but metamonads branch near the likely root\(^1\), and the kinetoplastid \textit{Trypanosoma brucei} probably also represents an early branch. The metamonads \textit{Giardia intestinalis} and \textit{Trichomonas vaginalis} have simple promoter regions of generally <100 bp. They lack the conserved heptapeptide repeat of the RNAPII CTD, the general TF TFIIB, H2A.Z, H1, p300/CBP acetyltransferases, histone demethylases, and DNA methyltransferases\(^12,35,115\), which could reflect either loss or a primitive absence of these proteins (TABLE 1). In \textit{T. vaginalis}, H3K27ac and H3K4me3 are associated with active genes\(^32\). In \textit{G. intestinalis}, HDACs and the silencing mark H3K9me are important for encystation and antigenic variation\(^34\). Kinetoplastids such as \textit{T. brucei} have unique polycistronic transcription units, with dispersed initiation of transcripts over approximately the first 2 kb of an ~10 kb region of more accessible nucleosomes containing H4ac, H2A.Z, and H2B.V, which is an H2B variant specifically paired with H2A.Z\(^37,116\). The mRNAs for individual genes are processed with the addition of a 39-nucleotide capped \textit{trans}-spliced leader, and NDRs are found at the splice site upstream of each gene start codon rather than in the region where transcripts are initiated. These observations suggest that the most common eukaryotic promoter architecture was not yet fixed in early-diverging eukaryotes, but the basic activating and silencing PTMs were already present in LECA.

**H2A.Z and nucleosome stability.** Histone variants replace their cognate core histones in nucleosomes and change nucleosome properties, including how much DNA a nucleosome wraps (BOX 4). We have previously reviewed histone variants and their dynamics\(^5,35\), and here we focus on recent advances in understanding
Box 4 | H2A wrapping variants

*Arabidopsis thaliana* has four types of H2A variant: conventional H2A, H2A.X, H2A.Z and plant-specific H2A.W, which all form homotypic nucleosomes122. H2A.Z is the least stable to temperature, and H2A.W is the most stable, with stability largely dependent on their differing Loop1 regions and docking domains. The longer carboxyl terminus of H2A.W protects an additional 10–15 bp of linker DNA beyond the 147 bp of most nucleosomes. H2A.W is found in heterochromatin132, where it may serve a silencing function similar to HP1 proteins found in animals. It can be phosphorylated in the DNA damage response, similar to H2A.X in euchromatin188.

Like H2A.W, macroH2A in animals and their holozoan relatives185 protects 10 bp of extranucleosomal DNA with its basic protein linker region that connects the histone fold domain (HFD) to the macrodomain that distinguishes macroH2A from other H2As188. The linker region facilitates condensation189, and Loop1 stabilizes DNA binding to the histone octamer190. MacroH2A nucleosomes are often repressive, as they are less preferred by remodelers, inhibit acetylation by p300 [REFS191,192], and are found on the inactive X chromosome190, but they stabilize both active and inactive cell-specific gene expression patterns, presenting a barrier to reprogramming cells193,194.

Placental mammals have four families of short H2As (H2A.B, H2A.L, H2A.P, and H2A.Q) that have shortened docking domains and reduced acidic patches, and they wrap only 110–130 bp of DNA195. All are encoded on the X chromosome, have stage-specific expression in tests, and are evolving rapidly, suggesting a role in sperm and sperm nuclei198. Mammals also have H2A.J, differing from H2A mostly at its carboxyl terminus198. Through an unknown mechanism, H2A.J promotes senescence-associated inflammatory gene expression in cells with persistent DNA damage.

H2A.Z, which is present in nearly all eukaryotes and is enriched at the +1 nucleosome of genes130, where it appears to poise genes for transcription. H2A.Z is absent in metamonads131, raising the possibility that this variant was absent in LECA and originated after the divergence of metamonads, but its strong conservation in nearly all other eukaryotes attests to its key importance in gene regulation (TABLE 1). H2A.Z has an extended acidic patch that stimulates ATP-dependent remodelers194, and, in yeast, a complex of RSC bound to H2A.Z-containing nucleosomes with asymmetric DNA contacts occurs at ~5% of +1 nucleosomes, which may represent remodelling interventional complexes that facilitate NDR formation and/or RNAPII transit182. The yeast SWR1 complex, which is a member of the INO80 subfamily of remodelers, binds to the NDR and replaces H2A–H2B dimers with H2A.Z–H2B dimers on the +1 and −1 nucleosomes183. Deposition is facilitated by nucleosome acetylation196,197,198,210. The DNA-binding Swc4 subunit can also direct SWR1 to specific genes in *Arabidopsis thaliana*211. The INO80 remodeller preferentially carries out the reverse exchange of H2A.Z–H2B for H2A–H2B in vitro122,124,125, although this has been disputed212. Deletion of Ino80 in yeast results in global mislocalization of H2A.Z213 and, more specifically, increased H2A.Z levels around unresolved DNA double-strand breaks127.

In yeast H2A.Z sits over the TSS, where it promotes RNAPII recruitment122, and is displaced by PIC components194. At the mating-type locus and telomeres, it resists the spread of silent chromatin199. In vitro, H2A.Z nucleosomes are more mobile than H2A nucleosomes, with a lower breaking force131 and lower thermal stability125, although an earlier in vitro study found that they have greater stability towards salt than H2A nucleosomes do, even when acetylated132. In vivo, in *D. melanogaster* cells, where the +1 nucleosome is downstream of the TSS, H2A.Z lowers the barrier to transcription presented by the +1 nucleosome134. Progression of RNAPII through the +1 nucleosome can result in loss of an H2A.Z–H2B dimer and/or its DNA contacts on the promoter-proximal side in association with stalled RNAPII or on the distal side in association with elongation and positive torsion134. Such loss of dimer–DNA contact may underlie the eviction of H2A.Z without loss of H3 in *A. thaliana* temperature-responsive gene promoters135 and the enrichment of subnucleosomal DNA fragments and increased accessibility found around sites occupied by H2A.Z and p300 in enhancers in mouse ESCs136.

It has long been puzzling that H2A.Z can be either activating or inhibiting for transcription, but if H2A.Z is more easily disrupted at the +1 nucleosome and stimulates remodelers, these different responses may reflect the different contexts in which H2A.Z nucleosomes reside. In the *A. thaliana* thermal response, H2A.Z is inhibitory until temperature-stimulated heat shock factors bind to promoters and stimulate transcription137. In mouse pituitary gonadotropes, the position of H2A.Z relative to the TSS regulates high or low expression of luteinizing hormone subunits138. In mouse ESCs, H2A.Z colocalizes with H3K4me3 at both promoters and enhancers and facilitates targeting of the histone methyltransferases for the active mark H3K4me3 and the repressive mark H3K27me3, especially at enhancers139. In turn, H3K4me3 promotes H2A.Z incorporation at enhancers in an apparent positive feedback loop. We speculate that such a feedback loop could explain why both H2A.Z and H3K4me2 and/or H3K4me3 are anticorrelated with DNA methylation in *A. thaliana*137,138, as H3K4me3 inhibits DNA methylation, and DNA methylation exuudes H2A.Z (BOX 2).

**Interaction domains and complexes.** The domain architectures of HATs, HDACs, and especially histone methyltransferases in metamonads are the simplest in eukaryotes, indicating that the interaction domains of chromatin proteins have increased with organismal complexity in later-diverging eukaryotes12, both through the addition of PTM recognition domains to proteins and through the addition of subunits to complexes. For example, in *Candida albicans*, the NuA4 acetyltransferase and the SWR1 remodeler are separate complexes in the hyphal state but are combined in the yeast state, and domains from both complexes have been combined into one protein independently in several eukaryotes, including in human p400 of the TIP60 complex139.

A more complex example comes from Polycomb repressive complex 1 (PRC1) and PRC2, which have essential roles in controlling cell-type-specific developmental gene expression in multicellular eukaryotes. Diversification of these complexes may have facilitated the advent of cell differentiation in multicellular organisms by serving as a flexible, modular silencing apparatus that selectively inactivates a range of *cis* elements in response to developmental cues.
In animals, developmental silencing is dependent on methylation of H3K27 by PRC2 and H3K27me3-directed gene silencing by PRC1 (REFS 140,141). A homologue of enhancer of zeste (E(z)) — the methyltransferase component of PRC2 that catalyses methylation of H3K27 — was present in LECA12 and might have originally served merely to block H3K27 acetylation, perhaps to prevent transcription of transposons and repetitive sequences, as in Chlamydomonas reinhardtii 142, or of telomeric repeats, as in fungi 143. Classical PRC2 in D. melanogaster (Fig. 3a) contains four subunits, which are present in most eukaryotes, although individual proteins have been lost in some lineages 142. Flowering plants and vertebrates have multiple E(z) paralogues, which form multiple PRC2-like complexes that differ in cell-specific expression (Fig. 3b) and can dynamically exchange subunits to target particular genes 144,145. Specific subunits can target PRC2 to CpG islands in mouse cells 146 or even

---

**Fig. 3 | PRC1 and PRC2 in animals and plants.** Although both plants and animals share many components of the Polycomb repressive complexes PRC1 and PRC2, they are differently deployed. **a, b |** PRC2 is widely conserved in eukaryotes and has four core components in Drosophila melanogaster, E(s), Su(z)12, p55, and the histone methyltransferase E(z), most of which are present in multiple copies in Arabidopsis thaliana, forming at least three distinct complexes that regulate distinct developmental programmes. The A. thaliana chromodomain protein LHP1 binds to H3K27me3 and together with enhancer of zeste (E(z)) homologue curly leaf (CLF) acts to spread H3K27me3. **c |** PRC1 in D. melanogaster contains the chromodomain protein Polycomb (Pc) that binds to H3K27me3 and ring finger E3 ubiquitin ligases Psc and Sce that ubiquitylate H2AK119; however, silencing depends on chromatin compaction by polymerization of Polyhomeotic (Ph). The complex can be localized by transcription factors (TFs) such as Pleiohomeotic (Pho). **d |** In A. thaliana, PRC1 complexes are not well characterized, but two complexes have been proposed containing BMI and RING1, homologues of Psc and Sce, respectively, along with plant-specific components with PHD fingers that can bind to H3K27me3 (SHL and EBS) or H3K4me3 (AL). The latter complex is proposed to shut off active genes to transition to repressed chromatin marked with H3K27me3 and H2AKub. Shapes coloured identically represent homologous proteins.
direct methylation to H1K26 instead of H3K27 [REF.14]. Human PRC2 member EED (Esc homologue) binds to H3K27me3 and positions the E(z) homologue EZH2 for methylation of an adjacent nucleosome, facilitating H3K27me3 spreading to form H3K27me3 domains148.

Classical PRC1 in D. melanogaster also contains four core subunits, including the chromodomain protein Polycomb, which binds to H3K27me3, and two RING-type zinc-finger E3 ubiquitin ligases (Sce and Psc) that together ubiquitylate H2A119, as well as variable substoichiometric subunits including HDACs and the DNA-binding protein Pleiohomeotic that targets PRC1 to specific genes144 (FIG. 5c). H2A119ub is not essential for PRC1-mediated silencing in animals149,150. H3K27me3 silencing is thought to occur through inhibition of transcription initiation and elongation by PRC1-mediated chromatin compaction151, which changes during differentiation and depends on the PRC1 component Polyhomeotic150, which can polymerize152, potentially bringing disparate PRC1 complexes together (FIG. 5c). Vertebrates have 3–5 Polycomb homologues and several partially redundant paralogues of PRC1 Sce (RING1) and Psc (BMI) proteins146. Although LECA had several chromodomain proteins and RING E3 ligases, a homologue of the RING1 proteins of PRC1 was not among them144, but RING1 homologues are present in plants, animals and other eukaryotes. However, fungi including Cryptococcus neoformans and Neurospora crassa lack PRC1 homologues, indicating that PRC1 is dispensable for general PRC-mediated silencing144,151.

Plants have no polyhomeotic homologue, and other PRC1 components are deployed differently in plants144 (FIG. 5d). The plant chromodomain protein LHP1 binds to H3K27me3 [REF.14] similarly to Polycomb but has a chromoshadow domain like animal H3K9me3-binding HP1, to which it is more closely related144. It copurifies with one of the PRC2-like complexes containing E(z) homologue curly leaf (CLF), rather than with PRC1 components, and aids in spreading H3K27me3, repressing floral genes in seedlings148. By contrast, RING1 and BMI in combination with CLF and the E(z) homologue SWN (also known as EZA1) repress embryonic genes. These two CLF complexes are associated with different sets of TF-binding motifs, suggesting they may be targeted by specific TFs. Two PRC1 complexes containing RING1, BMI, and plant-specific proteins have been proposed. To ubiquitylate H2A, RING1 and BMI require EMF1, which can bind to plant-specific SHL and EBS, which each bind to H3K27me3 through their PHD fingers150. RING1 and BMI can also bind to AL paralogues that bind to H3K4me3 with their PHD fingers and are proposed to switch off active H3K4-containing genes150.

Conclusions and perspectives

Proto-eukaryotes assembled the enzymes and domains of chromatin proteins from precursors in bacteria, archaea, viruses and perhaps urkaryotes and invented new uses for them in response to the acquisition or invention of nucleosomes, which may have first served a genome defence role. Nucleosomes altered the role of DNA superhelicity, required mobilization by remodelers and HATs, and probably enabled the expansion of genomes through more efficient packaging and condensation. Such opportunities for genome expansion are likely to have facilitated dramatic evolution in the form of gene duplication, regulatory element proliferation, and deployment of an expanding and progressively interactive set of chromatin domains and proteins both before and since LECA, including the diversification of PRC complexes to control developmental programmes in multicellular eukaryotes. Chromatin evolution is ongoing in the rapid evolution of short H2A histone variants in mammals [BOX 4].

Continuing investigation of cell-type-specific chromatin changes in both model and early-diverging eukaryotes will better illuminate chromatin regulation and evolution. The interaction of HATs, HDACs, and remodelers with TFs, PIC components, PTMs, and each other is a very active area of investigation that is likely to flesh out details of chromatin regulation in the immediate future. Similarly, the association of different PRC complexes with specific developmental programmes in animals and plants is proceeding apace. Further investigation of viral genomes, which are less intensively studied, using existing technologies may provide insight into the ‘missing’ proto-eukaryote stages in the evolution of chromatin proteins. Similarly, investigation of transcriptional regulation in metamonads, particularly free-living metamonads, may illuminate details of a more primitive state of chromatin regulation of transcription, and knowledge of the chromatin proteins in other early-branching eukaryotes, such as species of Malawimonas, Collodictyon, Ancyromonas, and others, would help to illuminate the early stages of eukaryotic chromatin evolution. Our understanding of supercoiling domains in both gene regulation and chromosomal compaction is in its infancy. Psoralen-based methods for mapping negative supercoils153,154 and recent technologies for detecting single-strand and double-strand breaks155 and non-B-form DNA156 on a genome-wide scale show promise for better illuminating the role of DNA torsion, although development of additional new methods at both genomic and gene-specific scales is likely to speed progress in this area.

Published online: 18 March 2019

methytransferase CLF and aids spreading of methylation to repress floral development, while PRC1 components BMI1 and RING1 associate with histone methytransferases CLF and SWR and regulate gene expression beyond gene boundaries.


Super-resolution microscopy of mammalian cells identifies PRC1 compacted chromatin domains that depend on Polycomb but not on PRC1-dependent and that are lost as PRC1 binding is lost during differentiation.


Acknowledgements The authors declare no competing interests.

Author contributions P.B.T. and M.P.M. researched content for the article. All authors contributed to the discussion of content, writing the article and reviewing and editing the manuscript before submission.

Competing interests The authors declare no competing interests.

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Reviewer information Nature Reviews Genetics thanks K. Struhl and the other, anonymous reviewer(s) for their contribution to the peer review process of this work.