Title: Evolution of the *D. melanogaster* chromatin landscape and its associated proteins

Authors and affiliations: Elise Parey[1, 2] and Anton Crombach*[1,3]

(1) Center for Interdisciplinary Research in Biology (CIRB), Collège de France, CNRS, INSERM, PSL Université Paris, 75005 Paris, France

(2) (current address) Institut de Biologie de l'Ecole Normale Supérieure (IBENS), Ecole Normale Supérieure, CNRS, INSERM, PSL Université Paris, 75005 Paris, France

(3) (current address) Inria, Antenne Lyon La Doua, 69603 Villeurbanne, France

Author for Correspondence (*): Anton Crombach, Center for Interdisciplinary Research in Biology (CIRB), Collège de France, CNRS, INSERM, PSL Université Paris, 75005 Paris, France, anton.crombach@college-de-france.fr

Keywords:

phylogenomics, chromatin-associated proteins, chromatin types, histone modifications, centromere drive, *D. melanogaster*.

Abstract (221w, max 250w)

In the nucleus of eukaryotic cells, genomic DNA associates with numerous protein complexes and RNAs, forming the chromatin landscape. Through a genome-wide study of chromatin-associated proteins in *Drosophila* cells, five major chromatin types were identified as a refinement of the traditional binary division into hetero- and euchromatin. These five types are defined by distinct but overlapping combinations of proteins and differ in biological and biochemical properties, including transcriptional activity, replication timing and histone modifications. In this work, we assess the evolutionary relationships of chromatin-associated proteins and present an integrated view of the evolution and conservation of the fruit fly *D. melanogaster* chromatin landscape. We combine homology prediction across a wide range of species with gene age inference methods to determine the origin of each chromatin-associated protein. This provides insight into the emergence of the different chromatin types. Our results indicate that the two euchromatic types, YELLOW and RED, were one single activating type that split early in eukaryotic history. Next, we provide evidence that GREEN-associated proteins are involved in a centromere drive and expanded in a lineage-specific way in *D. melanogaster*. Our results on BLUE chromatin support the hypothesis that the emergence of Polycomb Group proteins is linked to eukaryotic multicellularity. In light of these results, we discuss how the regulatory complexification of chromatin links to the origins of eukaryotic multicellularity.

Introduction

The chromatin landscape consists of DNA, histones and other associated proteins and RNAs, and plays a fundamental role in development, cellular memory, and integration of external signals. As a unique feature of the eukaryotic cell, it is closely tied to the evolution of eukaryotes, both regarding their origin and the major transition(s) to multicellularity (Newman 2005; Aravind et al. 2014; Gombar et al. 2014; Penny et al. 2014; Miyamoto et al. 2015; Sebé-Pedrós et al. 2017). At a basic level, chromatin is responsible for maintenance, organization, and correct use of the genome. Histone proteins package and condense DNA in the nucleus, and act as a docking platform for hundreds of structural and regulatory proteins. A variety of reversible post-translational modifications of histones, known as epigenetic marks, promote the recruitment of specific proteins. This creates a local context for nuclear processes such as transcriptional activity, replication, as well as DNA-repair. These and other epigenetic mechanisms involved in chromatin modification have been extensively characterized in variety of eukaryotic species, which led to the observation that the chromatin landscape is effectively subdivided into a small set of distinct chromatin states (Filion et al. 2010; Ernst et al. 2011; Roudier et al. 2011). A largely open question, however, is how these chromatin states have (co-)evolved. In this work, we assess the evolutionary relationships of chromatin-associated proteins (CAPs) and present an integrated view of the evolution and conservation of the fruit fly *D. melanogaster* chromatin landscape.

Classically, chromatin is divided into two states, namely heterochromatin and euchromatin, the former a compacted DNA state in which transcription is mostly repressed and the latter an open, transcriptionally active configuration. This classification has been refined into multiple types of chromatin. In particular, a breakthrough result was presented by Filion *et al.,* who established five major chromatin types in *D. melanogaster,* named with the colors YELLOW, RED, GREEN, BLUE, and BLACK. To do so, they used genome-wide binding profiles of CAPs obtained via DamID (Vogel et al. 2007; Filion et al. 2010; van Bemmel et al. 2013). This approach is complementary to more commonly used genome-wide histone mark profiling techniques, such as ChiP-seq. Nevertheless, both are consistent with each other and serve as independent validation. Indeed, the five types can be mapped to an alternative classification into nine chromatin states, that is derived from histone modifications (Kharchenko et al. 2011).

The five chromatin types have different biological and biochemical properties. YELLOW and RED are two types of euchromatin. YELLOW mainly marks ubiquitously expressed housekeeping genes. In contrast, the genes harbored in RED show more restricted expression patterns and are linked to specific tissues and developmental processes. Both euchromatin types are replicated in early S phase, and of the two, RED tends to be replicated first (Filion et al. 2010). GREEN, BLUE, and BLACK are three types of heterochromatin. GREEN is considered a type of constitutive heterochromatin. It is identified by HP1-related proteins and is especially prevalent in pericentric regions as well as on chromosome 4. BLUE is facultative heterochromatin and concerns mostly genes specifically repressed during

development. It is notably composed of the Polycomb Group (PcG) proteins, which were originally discovered in *D. melanogaster* to repress Hox genes, and were later found to have a general role in development. (Lewis 1978; Duncan 1982; Boyer et al. 2006; Lee et al. 2006; Nègre et al. 2006). Finally, BLACK is a major repressive type, covering 65% of silent genes, whose underlying repressive molecular mechanisms remain poorly characterized (Filion et al. 2010).

From an evolutionary point of view, although prokaryotes have specialized proteins associated with their DNA, they do not share homology with eukaryotic CAPs (Luijsterburg et al. 2008). In general, evolution of chromatin and diversification of epigenetic mechanisms are suggested to be tightly linked with eukaryotic evolution, from its origin to the transition to multicellularity (Newman 2005; Aravind et al. 2014; Gombar et al. 2014; Penny et al. 2014; Miyamoto et al. 2015; Sebé-Pedrós et al. 2017). Indeed, the Last Eukaryotic Common Ancestor (LECA) is considered to possess the key components of eukaryotic epigenetics, including most histone modification enzymes and some histone mark readers (Aravind et al. 2014). In addition, a current hypothesis on the transition to multicellularity is that complexification of the regulatory genome, via the emergence of repressive chromatin contexts and distal regulatory elements, permitted to generate the cell-type-specific transcriptional programs required for multicellularity (Larroux et al. 2006; Mendoza et al. 2013; Sebé-Pedrós et al. 2016, 2017; Arenas-Mena 2017; Hinman & Cary 2017). Recently, a system-level view of the evolution of the chromatin modification machinery was provided by (On et al. 2010). They demonstrated the high conservation of a core of chromatin proteins

across four model organisms (human, yeast, fruit fly, and worm), accompanied with diverse species-specific innovations.

Here, we investigate the evolutionary relationships of the CAPs studied by (Filion et al. 2010; van Bemmel et al. 2013), using homology prediction, gene age inference methods, functional annotations, and protein domain annotations. Taken together, the work provides an integrated view of the conservation of a chromatin landscape across eukaryotes. Our phylogenomic analysis leads us to propose that the chromatin types YELLOW and RED derive from a single ancestral euchromatin-like type. With respect to GREEN chromatin, we provide evidence that some of its associated proteins are undergoing an evolutionary Red Queen process called centromere drive, while others expanded in a lineage specific manner in *D. melanogaster*. Finally, our results support the association between the emergence of BLUE chromatin with its Polycomb proteins, and animal and plant multicellularity.

Material and Methods

Data set

Our data set contains all CAPs whose chromatin types have been assigned by (Filion et al. 2010; van Bemmel et al. 2013). As a convention throughout the work, a CAP is assigned the color of the chromatin type(s) it binds over more than 10% (fraction of 0.1). The set contains 107 *D. melanogaster* proteins, which include 65 well-characterized CAPs selected to cover a wide range of known chromatin complexes plus 42 previously unknown proteins putatively linked with chromatin. All have also been selected on expressibility in Kc167 cell-lines (derived from *D. melanogaster* embryonic hemocytes). This set was used to search for homologs in 53 species, covering 15 prokaryotes, 15 non-metazoan eukaryotes, and 23 metazoa (Supplementary Table 1, Supplementary Figure 1). The selection of species was guided by the quality of their PhylomeDB entry.

Homology prediction

All homolog predictions for the set of 107 *D. melanogaster* proteins were extracted using MetaPhOrs (http://orthology.phylomedb.org/) (Pryszcz et al. 2011), a repository of phylogeny-based orthology and paralogy predictions computed through popular homology prediction services: PhylomeDB, Ensembl, EggNOG, OrthoMCL, COG, Fungal Orthogroups, and TreeFam (Tatusov et al. 1997; Chen et al. 2006; Wapinski et al. 2007; Flicek et al. 2008; Ruan et al. 2008; Muller et al. 2010; Huerta-Cepas et al. 2014).

In a first round, we extracted all *D. melanogaster* homology predictions for the 107 CAPs in the other species of interest. We used the common assumption that protein function tends to be conserved in homologs across species, between orthologs and less systematically between paralogs (Koonin & Galperin 2003). We retained only homology hits (i.e orthology and/or paralogy) that had sufficient sequence similarity with the corresponding *D. melanogaster* protein. In all cases, a sequence similarity criterion of 25% and a maximum gap proportion of 60% (i.e. minimum 40% overlap) were applied after Needleman-Wunsch global pairwise alignment with the *D. melanogaster* protein. The maximum gap proportion avoids hits that share very conserved domains in otherwise unconserved sequences. The similarity threshold for homology was chosen to be consistent with knowledge for well-studied proteins, including Polycomb, HP1, SU(VAR)3-9, Sir2, RNA pol, TBP, CTCF, PCNA, SU(HW), BEAF-32 (Klenk et al. 1992; Lanzendörfer et al. 1993; Marsh et al. 1994; Rowlands et al. 1994; Krauss et al. 2006; Lomberk et al. 2006; Whitcomb et al. 2007; Greiss & Gartner 2009; Chia et al. 2010; Schoborg & Labrador 2010; Heger et al. 2013).

The homology prediction of MetaPhOrs is based on searching over half a million pre-computed gene trees. These trees usually focus on subsets of species, for instance, a tree can be restricted to vertebrates only. This may generate false negatives in our first round of homology search, since some species are less likely to appear in trees with *D. melanogaster.* Therefore, a second round of homology search was conducted to cover also the less-studied species as follows. For each protein of a particular organism lacking a hit in the first round, the predicted homologs of the two closest species to that particular organism were used to seed a

second search for an ortholog in this organism. For instance, during the first round a homolog of the *D. melanogaster* protein HP6 (HP6_Dme) was found in *D. simulans* as HP6_Dsi, but not in the ant *A. cephalotes*. In the second round, the homology search in *A. cephalotes* was seeded with HP6_Dsi. Then finding an ortholog in *A. cephalotes* points to a candidate homolog of *D. melanogaster* HP6_Dme. We encountered 190 cases of a successful second round of homology search.

Despite the two rounds of homology search, strictly speaking we cannot prove the absence of homologs observed in certain species, as we cannot rule out that it is related to biological and/or technical challenges, such as rapid sequence divergence, limited sequencing depth and/or genome coverage, or the sensitivity of the homology search.

Different amino acid substitution matrices were used to account for different evolutionary distances: Blossum45 to compare with prokaryotes, Blossum62 with eukaryotes, and Blossum80 with metazoa. Finally, we note that instead of *D. melanogaster* Su(var)3-9, the well-characterized human homolog SUV39H2 was used as a seed for homolog search, since this gene and the eukaryotic translation initiation factors eiF2 are fused in *D. melanogaster* (Krauss et al. 2006) and attract false positive hits.

Gene age inference

The binary vectors of homolog absence/presence of the 107 CAPs for each species were clustered using partitioning around medoids (PAM) (Kaufman & Rousseeuw 1990), with

simple matching distance (SMD) as dissimilarity measure, and followed by silhouette optimization. The resulting clustering and age groups are robust, as confirmed by re-runs of PAM and by using the Jaccard distance measure.

Similar to (Arcas et al. 2014), we verify our clustering by independently applying the Dollo parsimony method, which associates gene age to the most recent common ancestor (MRCA). We relate each gene to the age of the most distant hit, defining 5 age groups: Pre-Eukaryotes, Eukaryotes, Opisthokonta, Metazoa, and Arthropods. For instance, since the most distant homolog of *Deformed Wings* (DWG) is in the spreading earthmoss *P. patens,* we assign it to Eukaryotes. We confirm that the trends in Figure 3 and Figure 5B remain unaffected (Supplementary Table 2, Supplementary Figure 2).

Finally, to determine if *D. melanogaster* CAPs are enriched at certain ages, we used ProteinHistorian (Capra et al. 2012) (http://lighthouse.ucsf.edu/proteinhistorian/). ProteinHistorian regroups databases of *D. melanogaster* proteomes with protein age assigned by different methods. We calculated enrichment using five different sets of protein family prediction of the Princeton Protein Orthology Database (Heinicke et al. 2007) (DROME_PPODv4 clustered with OrthoMCL, Multiparanoid, Lens, Jaccard and Panther7) and two different methods (Wagner and Dollo parsimony) to account for the expected differences according to the different phylogenies and data sets (Supplementary Table 3).

Reader/Writer/Eraser of histone marks analysis

From the literature known *D. melanogaster* histone modifiers and histone marks readers were extracted in addition to the ones present in the initial set (Bannister et al. 2001; Cao et al. 2002; Schotta et al. 2002; Byrd & Shearn 2003; Smith et al. 2004; Stabell et al. 2006; Steward et al. 2006; Wysocka et al. 2006; Eissenberg et al. 2007; Larschan et al. 2007; Rudolph et al. 2007; Seum et al. 2007; Srinivasan et al. 2008; Smith et al. 2008; Moore et al. 2010; Rechtsteiner et al. 2010; Wagner & Carpenter 2012). Homologs of these proteins among our species set were searched applying the same method as described in the above section 'Homology Prediction'.

Coding sequences extraction for dN/dS calculation and positive selection tests

For all 107 *D. melanogaster* CAPs, MetaPhOrs was used to retrieve orthologs within ten other *Drosophila* species (*D. yakuba, D. sechellia, D. pseudoobscura, D. willistoni, D. virilis, D. simulans, D. persimilis, D. erecta, D. ananassae*, and *D. mojavensis*). Using Flybase (http://flybase.org/, version *FB2017_01*, released February 14, 2017), we extracted all corresponding coding sequences (CDS). To avoid different isoforms and different within-species paralogs, only the protein with the highest alignment score to its corresponding *D. melanogaster* protein was retained for each species. Next, with these *Drosophila* species we inferred phylogenetic tree topologies, we estimated dN/dS, and we performed positive selection tests. We elaborate on each of these steps below.

Sequence alignment and tree topology inference for dN/dS calculation and positive selection tests

To prepare the homology sets for dN/dS calculation and positive selection tests with PAML (Yang 2007), CDSs of each set were multiple-aligned and a tree topology inferred. First, CDSs were translated and multiple aligned with Clustal Omega 2.1 (Chenna et al. 2003) Translation, alignment, cleaning and translation reversion is done with TranslatorX local version (Abascal et al. 2010) (available at http://translatorx.co.uk/), with the following parameters for Gblocks cleaning: '-b1=6 -b2=6 -b3=12 -b4=6 -b5=H' (Castresana 2000). In short, the Gblocks parameters b1 to b4 tune which amino acid (sub)sequences are considered conserved and/or non-conserved. They were chosen to relax cleaning on variable regions and retain diversity. The parameter -b5=H permits to clean sites with gaps in more than half of the sequences, following the recommendation from the PAML documentation to remove such sites. We refer to Gblocks documentation for details.

To account for possible differences between gene trees and species tree, positive selection tests were run on maximum likelihood trees computed from CDS alignments with phyml (Guindon et al. 2010) and also on *Drosophila* species trees extracted from TimeTree (Kumar et al. 2017) (http://www.timetree.org/). Phyml was run with default parameters to return the topology maximizing the likelihood function.

dN/dS estimation

From multiple CDS alignments and inferred tree topology (see previous section), PAML fits codon substitution models and estimates both branch length and dN/dS by maximum likelihood. For each of these alignments, a single dN/dS was estimated using Model 0 of codeml included in PAML (Yang 2007). We verified that dN/dS values are similar with the two tree topology inference methods (Supplementary Figure 3).

Positive selection tests

In order to detect positive selection among amino acid sites and along branches of the *Drosophila* tree, tests were carried out on gene and species trees with codeml from PAML using branch-site codon substitution models (Yang 2007). Since PAML fits models by maximum likelihood, it allows to put constraints on the dN/dS parameter and compare models via their likelihood. Following the approach of "Test 2" (see PAML documentation), we predicted positive selection by comparing Model A to the Null Model. In these models, different constraints can be put on a candidate branch, the so-called foreground branch, and all other branches in the tree, i.e. background branches. Model A allows dN/dS to vary among sites and lineages on the specified foreground branch, thus allowing for positive selection. The Null Model fixes dN/dS to 1 on both foreground and background branches, thus allowing only for neutral selection. This process was automated for all branches in the trees. Finally, for every (Model A, Model Null) pair, likelihood ratio tests (LRT) with Bonferroni correction for multiple testing were applied. The Null model was rejected where the adjusted p-value was <

0.01. Finally, Bayes empirical Bayes (BEB) calculates the posterior probabilities for sites to be under positive selection when the LRT is significant.

Protein domain annotation

To search for over-represented domains among the proteins in each of the inferred age clusters, domain annotations for the 107 *D. melanogaster* CAPs were extracted from InterPro database v63 (Finn et al. 2017). DNA-binding domains and their location in D1 proteins from 10 *Drosophila* species were inferred from protein sequence by searching Pfam or Prosite domains using InterProScan v5 (Jones et al. 2014).

Gene Ontology Annotation

PANTHER is a multifaceted database, classifying proteins via their evolutionary history and function. Functional annotations are provided both by downloading them directly from the GO Consortium and by inferring them from the phylogeny. We conducted a functional classification analysis per cluster of CAPs (see 'Gene age inference' section) with PANTHER. We combined clusters I and II (Figure 1) into a single pre-eukaryotic cluster. From this analysis, we extracted two types of GO terms. We used GO slim terms, which are high-level GO terms that serve as an overview of ontology content. Moreover, we used specific fine-grained terms by taking the deepest children of a corresponding GO slim category.

Results

The *Drosophila* chromatin landscape is biased towards eukaryotic age

Taking the dataset from (Filion et al. 2010; van Bemmel et al. 2013), we searched for homologs across 53 species and clustered the resulting phylogenetic profile in order to gain insight into the conservation and evolution of the *D. melanogaster* chromatin landscape. The clustering reveals six clusters (Figure 1, left side, I – VI). We associated these clusters to five major age groups: pre-eukaryotic genes (I & II), eukaryotic genes (III), multicellular plant and metazoan genes (IV), metazoan genes (V) and arthropod genes (VI), with ages assigned on the basis of stable blocks of conserved CAPs in multiple species. For phylogenetic positioning and dates of our five age groups, see Supplementary Figure 2.

We made several major observations on the inferred clusters. We find two dominant clusters, one referring to eukaryotes in general (III) and one specific to metazoans (V), and a third large cluster indicating lineage specific diversification (VI). Next, we observe a regular lack of CAPs across evolutionary times, in particular in fungal and parasitic species (for instance *S. pombe* and *S. japonicus*, respectively *Spo* and *Sja* in Figure 1). For fungal species the lack of CAPs may be due to lineage specific divergence, such that we do not detect any homologs, though we cannot rule out lineage specific loss. With respect to parasitic species, loss of CAPs is more likely.

In order to understand what biological functions are found in each of the clusters, we used PANTHER GO Slim annotations from the domains ‘Biological Process’ and ‘Molecular

Function', as well as corresponding specific terms that are at a lower level in the GO hierarchy (Mi et al. 2017). The oldest age groups (I, II, III) contain a more diverse set of functional annotation terms than the youngest groups (V, VI) (Figure 2A and B). Analysing the occurrence of different annotations and terms, we find that the pre-eukaryotic clusters (I, II) contain CAPs with roles in basic nuclear processes: translation, transcription, replication, and splicing. The eukaryotic cluster (III) is the richest in annotation terms, containing proteins involved in transcription regulation, mitosis, cellular transport, post-translational modifications, and cell-cycle regulation. The three youngest clusters (IV, V, VI) are dominated by transcription factors and co-factors, some of which are annotated with chromatin remodeling activity.

These annotations suggest that most chromatin-related processes are ancient and were present in the last common ancestor of eukaryotes. We strengthened this hypothesis by independent age enrichment tests against the *D. melanogaster* proteome, with age assigned to each protein by means of Dollo and Wagner parsimony (Csurös 2010). Indeed, we find that CAPs are significantly enriched in genes that date back to the origin of eukaryotes (Supplementary Table 2). Moreover, our analysis suggests that evolution towards more complex eukaryotic organisms was accompanied by the acquisition of new regulatory interactions. This is consistent with the paradigm that the evolution of increasingly complex transcriptional regulation is one of the key features in (animal) multicellularity, enabling the establishment of precise spatio-temporal patterns of gene expression and regulation (Larroux

et al. 2006; Mendoza et al. 2013; Sebé-Pedrós et al. 2016, 2017; Arenas-Mena 2017; Hinman & Cary 2017).

In summary, chromatin-associated proteins appear to have been established early in eukaryotic evolution, after which they continuously diversified and specialized. In the next sections, we assess the conservation of the *D. melanogaster* chromatin landscape in eukaryotes and we highlight three major dynamics in chromatin evolution.

YELLOW and RED emerged from an ancient single euchromatin type

Of the five chromatin types, YELLOW and RED are the two euchromatic types, associated to transcriptionally active regions in the genome. The key biological differences between them are gene expression patterns, broad in YELLOW and specific in RED, and replication timing, which is early in YELLOW and very early in RED (Filion et al. 2010).

We hypothesized that YELLOW and RED are derived from one ancestral active chromatin type (Figure 1). To shed light on the idea, we examined the phylogenetic profile of chromatin-associated proteins. The distribution of CAPs across clusters I—VI supports the idea of a single ancestral euchromatic type in two ways. First of all, proteins binding either YELLOW or RED are most abundant amongst pre-eukarotic and eukaryotic ones (Figure 3A, cluster I—III). This suggests a rather conserved (i.e. ancient) composition of both euchromatin types. Second, CAPs in the older clusters I—III more often associate with both YELLOW and RED, while younger CAPS appear to be more specialized (Figure 3B, “pre-euk” and “euk”).

To strengthen the above observations, we explored complementary lines of evidence. First, we investigated the origin of the histone marks specific to YELLOW and RED (H3K4me3) and specific to YELLOW (H3K36me3). The starting point was evidence that the last eukaryotic common ancestor (LECA) had a lysine (K) at the amino acid positions indicated by H3K4 and H3K36 (Aravind et al. 2014). To understand if these lysines were indeed part of an ancient "epigenetic code", we summarized the rich literature of histone modifiers in a phylogenetic profile across the 53 species, similar to the profile that we made for CAPs (Figure 4, see Methods for used literature). We focused on three classes of proteins: *writers* that do the histone modification (i.e. methylation, acetylation, etc.), *readers* that interpret the mark, and *erasers* that remove the mark. We identified the first writer for both H3 lysine marks in one basal eukaryote (*Phaeodactylum tricornutum*) and three Viridiplantae (*Physcomitrella patens*, *Oryza sativa* and *Arabidopsis thaliana*). And we found one H3K4me3 reader and one H3K36me3 eraser in four basal eukaryotes (*Guillardia theta*, *Emiliania huxleyi*, *Bigelowiella natans*, and *Phaeodactylum tricornutum*). Moreover, genome-wide histone modification studies in yeasts, plants, as well as *Capsaspora owczarzaki,* which is a close unicellular relative of metazoa, reveal abundant use of both H3K4me3 and H3K36me3 (Bernstein et al. 2002; Suzuki et al. 2016; Roudier et al. 2011; Sebé-Pedrós et al. 2016). Finally, basal unicellular eukaryotes such as *Tetrahymena*, *Euglena*, *Stylonychia*, and *Trichomonas* make use of H3K4me3, but not H3K36me3 (Garcia et al. 2007; Postberg et al. 2010), suggesting H3K4me3 to be older than H3K36me3. In summary, H3K4 and H3K36

methylation appear indeed ancient, functional epigenetic marks, which supports our hypothesis of an early euchromatin split.

Second, a substantial decrease in proteins that associate to both YELLOW and RED takes place from eukaryotes to metazoans (Figure 3B). The decrease coincides with the major evolutionary transition to (animal) multicellular life. One hypothesis on the origin of this transition is that a unicellular ancestor with a complex life cycle transitioned from temporally regulated differentiation to a spatiotemporal one (Sebé-Pedrós et al. 2017). The complex life cycle of such a unicellular organism is based on two main features controlled by environmental stimuli, namely cell-cycle control and directional cell type transitions. In support of this hypothesis, we find that proteins involved in replication and cell-cycle control are in the eukaryotic cluster III (CAF-1, PCAF, ASF1, RAD21, and TRIP1) and that they are amongst the oldest RED-associated proteins. At first sight, four proteins in the arthropod cluster (MNT, PROD, SUUR and SSP) invalidate this "rule". However, these may be considered exceptions, as they are linked to a specialized process of proliferation control through endoreplication, a replication without cell division in *D. melanogaster* salivary glands. All in all, it suggests the first RED proteins to be involved in cell-cycle control. Developmental RED proteins are then observed in the multicellular and metazoan clusters (see Figure 1; JRA, DSP1,GRO, PHO, LOLAL, ECR, TRL, PHOL, ERR, BCD).

Summarizing, we have shown several lines of evidence for the hypothesis that YELLOW and RED were once a single euchromatin type. If we take the hypothesis to hold, it allows for three different scenarios: RED could derive from an ancestral type functionally

closest to current YELLOW, YELLOW could derive from an ancestral type functionally closest to current RED, or both types could derive from a distinct ancestral type. As RED is more complex and more specialized (Filion et al. 2010), we favour the scenario that it derived from an ancestral general euchromatin type, that was similar to *Drosophila's* current YELLOW. On the basis of Figure 3B, we suggest that the split was initiated before the acquisition of multicellularity. Indeed, the overlap between RED and YELLOW has its most substantial decrease between the eukaryotic cluster (III) and the multicellular cluster (IV).

GREEN emerged in metazoa and expanded in a lineage-specific way in *Drosophila*

GREEN chromatin is best characterized as constitutive, classic heterochromatin, and encompasses regions with high content in repetitive DNA and transposable elements (Sun et al. 1997; Filion et al. 2010). It is marked by HP1, a protein family that is involved in chromatin packaging and that binds di- and trimethylated histone H3 (H3K9me2/3) (Bannister et al. 2001). Classic proteins linked with HP1 heterochromatin are conserved (Saksouk et al. 2015) and indeed we find HP1, HP1c, and SU(VAR)3-9 across metazoa (cluster IV). Yet, eleven GREEN proteins, from a total of 25 in the whole dataset, are assigned to the arthropod cluster (the youngest gene cluster VI). Thus, as opposed to YELLOW and RED, the fraction of proteins bound in GREEN increases through evolutionary times (Figure 5A). At first view, this observation is paradoxical, since GREEN proteins are involved in genome integrity, in particular centromere maintenance. One expects to find them conserved across metazoa.

In a first step, we explored the conservation of GREEN chromatin proteins in the context of a previously established *Drosophila* chromatin protein network. In a pioneering effort van Bemmel et al. (van Bemmel et al. 2013) applied Bayesian network inference on binding profiles of CAPs to model interactions among chromatin components. In this model, GREEN is the only chromatin type to be divided into multiple regions of the network (Figure 5A), which has lead to the suggestion that GREEN chromatin is decomposable into three distinct subtypes (van Bemmel et al. 2013).

We propose this fragmentation to be linked to gene age. In the network, Region 1 contains 3 proteins, RAD21, MRG15, and CC35, that bind both GREEN and YELLOW chromatin. They belong to the oldest group of GREEN proteins. RAD21 and MRG15 are found across eukaryotes (cluster III), while CC35 is predicted to be of metazoan origins (cluster IV). Region 2 consists of proteins of all age clusters, from eukaryotes to arthropods, marking the 3 heterochromatin types (GREEN, BLUE, and BLACK). The region is organized around SUUR, a key player in chromatin silencing on polytene chromosomes (Makunin et al. 2002). Finally, region 3 contains mostly young GREEN proteins from the arthropod age group, organized around two metazoan proteins, HP1 and SU(VAR)3-9. Matching the three regions to the protein age clusters, we find that regions 2 and 3 are most strongly involved in the specific expansion of GREEN in *Drosophila*. Moreover, their peripheral location in the chromatin network compared to region 1 is consistent with this explanation (Zhang et al. 2015).

***D1 chromosomal protein* evolves under the centromere drive model**

We asked if poor conservation of many GREEN proteins may be due to the fact that they are fast evolving, which would lead to the rapid divergence of homologs. We estimated dN/dS, the ratio of non-synonymous nucleotide substitutions versus synonymous substitutions among different *Drosophila* species for all CAPs (Figure 5B). Under neutral evolution, non-synonymous substitutions and synonymous substitutions occur with the same probabilities and dN/dS ~ 1. If positively selected, amino acids change rapidly and dN/dS > 1. On the other hand, under purifying selection amino acid variation is reduced and results in dN/dS < 1. The ratio averaged over all sites and all lineages is however almost never > 1, since positive selection is unlikely to affect all sites over long periods of time. Our analysis revealed that Green CAPs from the arthropod cluster (Green Arthropod Cluster, GAC) show significantly more elevated dN/dS than other CAPs (8 GACs among a total of 16 CAPs with elevated dN/dS, p-value = $7.48\ 10^{-5}$) (Figure 5B).

Next, we asked if those 8 GAC candidates (green labeled proteins in Figure 5B) evolve under relaxed selective constraint or under positive selection. In particular, we wondered if they fit in the centromere-drive model proposed in (Henikoff & Malik 2002; Brown & O'Neill 2014). In this model, some heterochromatin proteins evolve under positive selection to suppress the deleterious effect of genetic drive in meiosis. This genetic drive is the consequence of a selfish behavior of chromosomes, which compete for preferential transmission in female meiosis by increasing affinity for microtubule attachment. Chromosomes with more satellite DNA sequences gain an advantage, if heterochromatin

proteins involved in recruitment of microtubules do not correct the bias by changing binding specificity. If a centromere drive is left unchecked, it breaks meiotic parity and has a deleterious effect on fitness both at the organism level and at the species level. Chromatin proteins repressing the drive must therefore contain both a role in binding satellite DNA and a role in recruitment of other heterochromatic or centromere proteins.

Of the 8 GACs candidates, HP6 and LHR have been proposed to be evolving under this model (Brideau et al. 2006; Ross et al. 2013). We carried out a positive selection test under a branch-site model and found recurrent positive selection for D1. D1 presents the features of heterochromatin proteins evolving through centromere-drive: it is capable of binding satellite DNA and is involved in heterochromatin propagation (Levinger & Varshavsky 1982). To the best of our knowledge, it has not been previously reported as a centromere drive protein. We also propose CC29 as a potential candidate. Although we have not been able to detect positive selection using the branch site model, CC29 has DNA binding domains, shows elevated dN/dS, and is part of a centromeric complex with HMR and LHR (Thomae et al. 2013).

For a better characterization of positive selection affecting D1 and to corroborate the hypothesis that it is involved in the centromere drive, we investigated more closely at which amino acids positive selection took place. We detected that positively selected sites (Figure 6A) are within or close to AT-HOOK domains. AT-HOOK domains enable D1 to bind to DNA: the domain is organized around a so-called GRP core, which is able to insert itself into the minor groove of DNA (Aravind & Landsman 1998). Many negatively charged amino

acids around this core are then involved in DNA-protein interactions. *Drosophila* species have nine to eleven copies of AT-HOOK in D1 (Figure 6B). Moreover, their locations in sequences vary between species (Figure 6B), highlighting domain-level differences in D1 proteins amongst *Drosophila,* possibly related to DNA binding specificity. As an example of a positively selected amino acid in an AT-HOOK motif, Leucine 83 is replaced by an Alanine directly before the GRP core (Figure 6C). We verified that positively selected sites are equivalent between the two tree topology inference methods, i.e species tree and gene tree (Supplementary Figure 4, Supplementary Table 5). In summary, D1 shows strong signs of evolving under positive selection in *Drosophila* and we propose that it tunes the specificity of its DNA-binding motifs to counterbalance fast-evolving satellite DNA.

Recent GREEN proteins associate with the expansion of the BESS protein domain in *Drosophila*

After we established four recent GREEN proteins are involved in the centromere drive model, we studied the evolution of the GREEN proteins that lacked signs of positive selection. Notably, in the *Drosophila* genus, the HP1 family has been demonstrated to present little evidence of positive selection. Nevertheless, this protein family is numerous with about 25 members, of which only four are conserved across a large number of drosophilids, and others are evolutionarily restricted to particular *Drosophila* species (Levine et al. 2012). This diversification of the HP1 family is thought to be a lineage-specific expansion driven by karyotype evolution, where events of chromosome rearrangements (fusion/fission) correlate

with losses and gains of HP1 proteins (Levine et al. 2012). We explored if other GREEN-associated proteins showed signs of lineage-specific expansions in *Drosophila*.

By studying protein domains, we found evidence that a subset of young GREEN proteins are part of the family of proteins with BESS domains that is expanding in the *Drosophila* lineage. BESS domains direct protein-protein interactions, including with itself. Among all known proteins (not just the ones in our data set) with an inferred BESS domain (InterPro database), more than 80% are restricted to insects and more than 50% are restricted to diptera. A comparison among *Drosophilids* has shown that the BESS domain family expanded through duplications in a lineage-specific way approximately 40 million years ago (Shukla et al. 2014). In our dataset, five of 107 proteins have a BESS domain (SU(VAR)3-7, LHR, BEAF-32, CC20, and CC25). They are all found in the arthropod cluster (VI), and with the exception of CC20, they are GREEN-associated. Therefore, we propose that these GREEN CAPs evolve rapidly through lineage-specific expansion. And we suggest that BESS domains are involved in directing protein-protein interactions in GREEN chromatin in *Drosophila*.

BLUE is related to the origin of multicellularity

Central in BLUE chromatin are the Polycomb group (PcG) proteins, which are recruited to Polycomb Response Elements (PREs) to silence specific target genes during development, such as Hox genes. PcG proteins form two multiprotein complexes, PRC1 and PRC2. Their catalytic signatures are well-characterized; PRC2 trimethylates histone H3K27 into

H3K27me3; this modified histone is bound by PRC1, which in turn ubiquitylates histone H2A. Extensive study on the evolution and conservation of PRC1 and PRC2 has suggested that expansion and diversification of PcG proteins contributed to the complexity of multicellular organisms (Trojer & Reinberg 2006; Whitcomb et al. 2007; Köhler & Villar 2008; Gombar et al. 2014).

In this study, the PcG proteins are represented by the main components of PRC2, namely E(Z) and PCL, and PRC1, with SCE and PC, in addition to three PRE-binders, respectively PHO, LOLAL, and PHOL. PRE-binders are found in RED chromatin, though, as they trigger the transition from active developmentally controlled chromatin to the PcG repressed state. Of the PcG proteins, the oldest ones that lay down key heterochromatin histone marks, are found in the multicellular cluster (IV). They are the writers E(Z) and SCE, which, respectively, tri-methylate H3K27 and ubiquitinate H3K118. Another key BLUE protein, PC, which reads H3K27me3 marks, is metazoan (Cluster V). This is in support of the hypothesis that PRC1, which contains PC, is younger than PRC2. Summarizing, both complexes are conserved across metazoans, suggesting the repression mediated by the PcG proteins as described above, was established at the origins of animal multicellularity (Whitcomb et al. 2007).

Several BLUE proteins are found in cluster II and III, and thus are older than PcG proteins. We mention the three most prominent ones: EFF, IAL, and LAM. All three are conserved in all eukaryotes, with functions unrelated to Polycomb-controlled repression. EFF is involved in protein ubiquitination and degradation, and is suggested to have a general role in

chromatin organization (Cipressa & Cenci 2013). IAL is mainly involved in mitosis (Adams et al. 2001) and LAM recruites chromatin to the nuclear envelope (Gruenbaum et al. 1988). We argue that these are not BLUE specialized proteins but rather general heterochomatic proteins recruited by GREEN, BLUE, and BLACK chromatin to form a repressed state.

Discussion

We have presented an integrated view of the evolution and conservation of a chromatin-associated proteome across eukaryotes. The creation and analysis of a phylogenetic profile of protein presence/absence resulted in three major findings. First, we presented evidence that YELLOW and RED chromatin originate from a single euchromatic type. Second, GREEN-associated proteins were found to be relatively specific to arthopods (or even restricted to dipterans). We connected two processes to this observation, namely a Red Queen type of evolution due to centromere drive, and lineage-specific expansion of proteins with BESS domains. Finally, our analysis of BLUE chromatin confirmed existing hypotheses on the importance of Polycomb repressive proteins for the evolutionary success of multicellular life forms. BLACK has not been addressed in this work. It is hard to interpret because it is mechanistically poorly understood and overlaps strongly with BLUE chromatin.

To place these results in context, we mention some critical points of our study. The evolutionary view on an epigenetic landscape that we have provided here is, of course, restricted in the sense that it is defined explicitly from a *D. melanogaster* angle. Notably, the *Drosophila* genome is particular, as it appears to lack DNA methylation and is known for an original mechanism of telomere maintenance by specialized non-LTR retrotransposons (Pardue & DeBaryshe 1999). Also, the homologs of *D. melanogaster* CAPs in other species do not necessarily share the same interactions and global assembly to form similar chromatin types. Indeed, in distant species that are separated by more evolutionary time, they are more

likely to be functionally different. To counter such false positives, we used a strict similarity cut-off for all protein-protein comparisons. The cut-off indeed helped us to reject functional homology prediction. For instance, it did not accept the *A. thaliana* HP1 homolog, LHP1, which appears to function both in a "classical" HP1-fashion and as a PcG protein (Zhang et al. 2007). Nevertheless, we cannot exclude that even if sequences and domains are very similar, the exact role in chromatin organization may be different.

Histone modifications, gene regulation, and the origins of multicellularity

The evolution of (animal) multicellularity is one of the major transitions in evolution. Within the area of (epi)genomics, it has been hypothesized that complexification of chromatin states and in particular the emergence of distinct heterochromatin states lay at the origin of multicellular life (Sebé-Pedrós et al. 2016; Hinman & Cary 2017). For instance, general heterochomatic proteins are already present in unicellular eukaryotes such as *S. cerevisiae* and *T. thermophila*, while more specific ones are found in mammals, which indeed have more complex repressive chromatin states (Garcia et al. 2007). Similar observations are made in studies focused on the large repertoire of histone modifiers in mammals and in work on PcG proteins. In summary, these studies propose that an elaboration of chromatin states is based on (unique) combinations of histone modifications.

Our phylogenomic profile supports the above idea of regulatory complexification. Indeed, we find that older proteins are more general than recent ones, in the sense that the older proteins tend to be found in multiple types of chromatin. Moreover, both multicellular

and metazoan clusters (IV and V) highlight complexification of histone modifications throughout eukaryotic evolution. In the eukaryotic cluster (III), proteins linked with histone modification are acetylation/deacetylation proteins (RPD3, DMAP1, SIN3A, PCAF), H3K36me3 reader (MRG15) and H3K4me3 writer (CC10). New repressive histone marks appeared in the multicellular and metazoan clusters, respectively H3K9me3 (SU(VAR)3-9) and H3K27me3 (E(Z)). We confirmed these results through an additional analysis of the conservation of *Drosophila* histone modifiers (Figure 4). It is interesting to note that in well-studied unicellular organisms (*T. thermophila*, S. cerevisiae, C. *owczarzaki*), repressive methylated histones H3K9 and H3K27 are often absent or present only at a very low level, while they are abundant in the multicellular fungi *N. crassa* (Garcia et al. 2007; Roudier et al. 2011; Ernst et al. 2011; Jamieson et al. 2013; Sebé-Pedrós et al. 2016). Thus we find diversification of histone marks and the acccompanying proteins, which as mentioned above, allow for a more fine-grained regulatory control over the genome.

Connected to the modulation of accessibility through histone modifications, our work also supports new regulatory elements to be linked with the transition to multicellularity. We find that the multicellular and metazoan clusters (IV and V) display the first insulator (DWG, CTCF, CC27) and enhancer binding proteins (JRA). Indeed, enhancers and insulators are mechanistically linked: enhancers being distal regulatory regions, they rely on looping with help of insulators to influence the expression of their targets (Krivega & Dean 2012; Phillips-Cremins & Corces 2013).

Taken together, we affirm the importance of regulatory complexification in the success of multicellular life. Like other studies, our work suggests this regulatory complexification to be linked with the need to control chromatin states and their propagation in an increasingly complex landscape of active and repressive genomic regions.

Outlook

We have enhanced our understanding of the evolution of the chromatin landscape through the epigenomic proteome in *Drosophila*. This is a good starting point, and we need additional studies that focus on other species to deepen and broaden that knowledge. Tackling other model organisms is a straightforward extension, such as the worm *C. elegans* and the plant *A. thaliana*. One future breakthrough we hope for, is that such studies could provide insight into new BLACK-associated proteins and perhaps lead to a better molecular and evolutionary characterization of this type. Moreover, some classes of proteins are better studied in species other than *Drosophila*. For instance, in our dataset, five proteins are responsible of histone acetylation/deacetylation, but substrate specificity and links with previously inferred chromatin states are not well-investigated in fly species. Contrastingly, acetylases (HAT) and deacetylases (HDAC) specificity are well-characterized in human (Seto & Yoshida 2014) and thus *H. sapiens* could be a better subject for questions in this area. Furthermore, non-coding RNAs are tightly associated to both active and inactive chromatin in eukaryotes, including in *S. pombe* (Martienssen et al. 2005), in various mammals (Saksouk et al. 2015), and in *D. melanogaster* (Fagegaltier et al. 2009). Thus we advocate for an inclusion of ncRNA

functionality within the analyses on different chromatin states across species. Clearly our current study is but an introduction that shows the potential exists for new insights into the evolution of the chromatin landscape.

Acknowledgments

We thank Joke van Bemmel for help interpretating the genome-wide binding data and the network model of chromatin organization. AC kindly acknowledges Fondation Bettencourt Schueller.

Abascal F, Zardoya R, Telford MJ. 2010. TranslatorX: multiple alignment of nucleotide sequences guided by amino acid translations. Nucleic Acids Res. 38:W7–W13. doi: 10.1093/nar/gkq291.

Adams RR, Maiato H, Earnshaw WC, Carmena M. 2001. Essential roles of Drosophila inner centromere protein (INCENP) and aurora B in histone H3 phosphorylation, metaphase chromosome alignment, kinetochore disjunction, and chromosome segregation. J. Cell Biol. 153:865–880.

Aravind L, Burroughs AM, Zhang D, Iyer LM. 2014. Protein and DNA modifications: evolutionary imprints of bacterial biochemical diversification and geochemistry on the provenance of eukaryotic epigenetics. Cold Spring Harb. Perspect. Biol. 6:a016063. doi: 10.1101/cshperspect.a016063.

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

Arcas A, Fernández-Capetillo O, Cases I, Rojas AM. 2014. Emergence and evolutionary analysis of the human DDR network: implications in comparative genomics and downstream analyses. Mol. Biol. Evol. 31:940–961. doi: 10.1093/molbev/msu046.

Arenas-Mena C. 2017. The origins of developmental gene regulation. Evol. Dev. 19:96–107. doi: 10.1111/ede.12217.

Bannister AJ et al. 2001. Selective recognition of methylated lysine 9 on histone H3 by the HP1 chromo domain. Nature. 410:120–124. doi: 10.1038/35065138.

van Bemmel JG et al. 2013. A network model of the molecular organization of chromatin in Drosophila. Mol. Cell. 49:759–771. doi: 10.1016/j.molcel.2013.01.040.

van Bemmel JG et al. 2013b. A network model of the molecular organization of chromatin in Drosophila. Mol. Cell. 49:759–771. doi: 10.1016/j.molcel.2013.01.040.

Bernstein BE et al. 2002. Methylation of histone H3 Lys 4 in coding regions of active genes. Proc. Natl. Acad. Sci. 99:8695–8700. doi: 10.1073/pnas.082249499.

Boyer LA et al. 2006. Polycomb complexes repress developmental regulators in murine embryonic stem cells. Nature. 441:349–353. doi: 10.1038/nature04733.

Brideau NJ et al. 2006. Two Dobzhansky-Muller Genes Interact to Cause Hybrid Lethality in *Drosophila*. Science. 314:1292–1295. doi: 10.1126/science.1133953.

Brown JD, O'Neill RJ. 2014. The Evolution of Centromeric DNA Sequences. In: eLS. John Wiley & Sons Ltd, editor. John Wiley & Sons, Ltd: Chichester, UK. doi: 10.1002/9780470015902.a0020827.pub2.

Byrd KN, Shearn A. 2003. ASH1, a Drosophila trithorax group protein, is required for methylation of lysine 4 residues on histone H3. Proc. Natl. Acad. Sci. U. S. A. 100:11535–11540. doi: 10.1073/pnas.1933593100.

Cao R et al. 2002. Role of histone H3 lysine 27 methylation in Polycomb-group silencing. Science. 298:1039–1043. doi: 10.1126/science.1076997.

Capra JA, Williams AG, Pollard KS. 2012. ProteinHistorian: Tools for the Comparative Analysis of Eukaryote Protein Origin. PLOS Comput. Biol. 8:e1002567. doi: 10.1371/journal.pcbi.1002567.

Castresana J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol. Biol. Evol. 17:540–552.

Chen F, Mackey AJ, Stoeckert CJ, Roos DS. 2006. OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. Nucleic Acids Res. 34:D363–D368. doi: 10.1093/nar/gkj123.

Chenna R et al. 2003. Multiple sequence alignment with the Clustal series of programs. Nucleic Acids Res. 31:3497–3500.

Chia N, Cann I, Olsen GJ. 2010. Evolution of DNA Replication Protein Complexes in Eukaryotes and Archaea. PLoS ONE. 5. doi: 10.1371/journal.pone.0010866.

Cipressa F, Cenci G. 2013. Effete, an E2 ubiquitin-conjugating enzyme with multiple roles in Drosophila development and chromatin organization. Fly (Austin). 7:256–262. doi: 10.4161/fly.26567.

Csurös M. 2010. Count: evolutionary analysis of phylogenetic profiles with parsimony and likelihood. Bioinforma. Oxf. Engl. 26:1910–1912. doi: 10.1093/bioinformatics/btq315.

Duncan IM. 1982. Polycomblike: A Gene That Appears to Be Required for the Normal Expression of the Bithorax and Antennapedia Gene Complexes of DROSOPHILA MELANOGASTER. Genetics. 102:49–70.

Eissenberg JC et al. 2007. The trithorax-group gene in Drosophila little imaginal discs encodes a trimethylated histone H3 Lys4 demethylase. Nat. Struct. Mol. Biol. 14:344–346. doi: 10.1038/nsmb1217.

Ernst J et al. 2011. Systematic analysis of chromatin state dynamics in nine human cell types. Nature. 473:43–49. doi: 10.1038/nature09906.

Fagegaltier D et al. 2009. The endogenous siRNA pathway is involved in heterochromatin formation in Drosophila. Proc. Natl. Acad. Sci. U. S. A. 106:21258–21263. doi: 10.1073/pnas.0809208105.

Filion GJ et al. 2010. Systematic protein location mapping reveals five principal chromatin types in Drosophila cells. Cell. 143:212–224. doi: 10.1016/j.cell.2010.09.009.

Finn RD et al. 2017. InterPro in 2017—beyond protein family and domain annotations. Nucleic Acids Res. 45:D190–D199. doi: 10.1093/nar/gkw1107.

Flicek P et al. 2008. Ensembl 2008. Nucleic Acids Res. 36:D707–D714. doi: 10.1093/nar/gkm988.

Garcia BA et al. 2007. Organismal differences in post-translational modifications in histones H3 and H4. J. Biol. Chem. 282:7641–7655. doi: 10.1074/jbc.M607900200.

Gombar S, MacCarthy T, Bergman A. 2014. Epigenetics decouples mutational from environmental robustness. Did it also facilitate multicellularity? PLoS Comput. Biol. 10:e1003450. doi: 10.1371/journal.pcbi.1003450.

Greiss S, Gartner A. 2009. Sirtuin/Sir2 Phylogeny, Evolutionary Considerations and Structural Conservation. Mol. Cells. 28:407–415. doi: 10.1007/s10059-009-0169-x.

Gruenbaum Y et al. 1988. Drosophila nuclear lamin precursor Dm0 is translated from either of two developmentally regulated mRNA species apparently encoded by a single gene. J. Cell Biol. 106:585–596.

Guindon S et al. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst. Biol. 59:307–321. doi: 10.1093/sysbio/syq010.

Heger P, George R, Wiehe T. 2013. Successive Gain of Insulator Proteins in Arthropod Evolution. Evol. Int. J. Org. Evol. 67:2945–2956. doi: 10.1111/evo.12155.

Heinicke S et al. 2007. The Princeton Protein Orthology Database (P-POD): a comparative genomics analysis tool for biologists. PloS One. 2:e766. doi: 10.1371/journal.pone.0000766.

Henikoff S, Malik HS. 2002. Centromeres: Selfish drivers. Nature. 417:227–227. doi: 10.1038/417227a.

Hinman V, Cary G. 2017. Multicellularity: The evolution of gene regulation. eLife. 6:e27291. doi: 10.7554/eLife.27291.

Huerta-Cepas J, Capella-Gutiérrez S, Pryszcz LP, Marcet-Houben M, Gabaldón T. 2014. PhylomeDB v4: zooming into the plurality of evolutionary histories of a genome. Nucleic Acids Res. 42:D897–D902. doi: 10.1093/nar/gkt1177.

Jamieson K, Rountree MR, Lewis ZA, Stajich JE, Selker EU. 2013. Regional control of histone H3 lysine 27 methylation in Neurospora. Proc. Natl. Acad. Sci. 110:6027–6032. doi: 10.1073/pnas.1303750110.

Jones P et al. 2014. InterProScan 5: genome-scale protein function classification. Bioinforma. Oxf. Engl. 30:1236–1240. doi: 10.1093/bioinformatics/btu031.

Kaufman L, Rousseeuw PJ. 1990. Partitioning Around Medoids (Program PAM). In: Finding Groups in Data. John Wiley & Sons, Inc. pp. 68–125. doi: 10.1002/9780470316801.ch2.

Kharchenko PV et al. 2011. Comprehensive analysis of the chromatin landscape in Drosophila melanogaster. Nature. 471:480–485. doi: 10.1038/nature09725.

Klenk H-P, Palm P, Lottspeich F, Zillig W. 1992. Component H of the DNA-Dependent RNA Polymerases of Archaea is Homologous to a Subunit Shared by the Three Eucaryal Nuclear RNA Polymerases. Proc. Natl. Acad. Sci. U. S. A. 89:407–410.

Köhler C, Villar CBR. 2008. Programming of gene expression by Polycomb group proteins. Trends Cell Biol. 18:236–243. doi: 10.1016/j.tcb.2008.02.005.

Koonin EV, Galperin MY. 2003. *Evolutionary Concept in Genetics and Genomics*. Kluwer Academic https://www.ncbi.nlm.nih.gov/books/NBK20255/.

Krauss V, Fassl A, Fiebig P, Patties I, Sass H. 2006. The evolution of the histone methyltransferase gene Su(var)3-9 in metazoans includes a fusion with and a re-fission from a functionally unrelated gene. BMC Evol. Biol. 6:18. doi: 10.1186/1471-2148-6-18.

Krivega I, Dean A. 2012. Enhancer and promoter interactions — long distance calls. Curr. Opin. Genet. Dev. 22:79–85. doi: 10.1016/j.gde.2011.11.001.

Kumar S, Stecher G, Suleski M, Hedges SB. 2017. TimeTree: A Resource for Timelines, Timetrees, and Divergence Times. Mol. Biol. Evol. 34:1812–1819. doi: 10.1093/molbev/msx116.

Lanzendörfer M et al. 1993. Structure and Function of the DNA-Dependent RNA Polymerase of Sulfolobus. Syst. Appl. Microbiol. 16:656–664. doi: 10.1016/S0723-2020(11)80337-1.

Larroux C et al. 2006. Developmental expression of transcription factor genes in a demosponge: insights into the origin of metazoan multicellularity. Evol. Dev. 8:150–173. doi: 10.1111/j.1525-142X.2006.00086.x.

Larschan E et al. 2007. MSL complex is attracted to genes marked by H3K36 trimethylation using a sequence-independent mechanism. Mol. Cell. 28:121–133. doi: 10.1016/j.molcel.2007.08.011.

Lee TI et al. 2006. Control of developmental regulators by Polycomb in human embryonic stem cells. Cell. 125:301–313. doi: 10.1016/j.cell.2006.02.043.

Levine MT et al. 2012. Phylogenomic Analysis Reveals Dynamic Evolutionary History of the Drosophila Heterochromatin Protein 1 (HP1) Gene Family. PLOS Genet. 8:e1002729. doi: 10.1371/journal.pgen.1002729.

Levinger L, Varshavsky A. 1982. Protein D1 preferentially binds A + T-rich DNA in vitro and is a component of Drosophila melanogaster nucleosomes containing A + T-rich satellite DNA. Proc. Natl. Acad. Sci. U. S. A. 79:7152–7156.

Lewis EB. 1978. A gene complex controlling segmentation in Drosophila. Nature. 276:565–570. doi: 10.1038/276565a0.

Lomberk G, Wallrath L, Urrutia R. 2006. The Heterochromatin Protein 1 family. Genome Biol. 7:228. doi: 10.1186/gb-2006-7-7-228.

Luijsterburg MS, White MF, Driel R van, Dame RT. 2008. The Major Architects of Chromatin: Architectural Proteins in Bacteria, Archaea and Eukaryotes. Crit. Rev. Biochem. Mol. Biol. 43:393–418. doi: 10.1080/10409230802528488.

Makunin IV et al. 2002. The Drosophila suppressor of underreplication protein binds to late-replicating regions of polytene chromosomes. Genetics. 160:1023–1034.

Marsh TL, Reich CI, Whitelock RB, Olsen GJ. 1994. Transcription factor IID in the Archaea: sequences in the Thermococcus celer genome would encode a product closely related to the TATA-binding protein of eukaryotes. Proc. Natl. Acad. Sci. U. S. A. 91:4180–4184.

Martienssen RA, Zaratiegui M, Goto DB. 2005. RNA interference and heterochromatin in the fission yeast Schizosaccharomyces pombe. Trends Genet. TIG. 21:450–456. doi: 10.1016/j.tig.2005.06.005.

Mendoza A de et al. 2013. Transcription factor evolution in eukaryotes and the assembly of the regulatory toolkit in multicellular lineages. Proc. Natl. Acad. Sci. 110:E4858–E4866. doi: 10.1073/pnas.1311818110.

Mi H et al. 2017. PANTHER version 11: expanded annotation data from Gene Ontology and Reactome pathways, and data analysis tool enhancements. Nucleic Acids Res. 45:D183–D189. doi: 10.1093/nar/gkw1138.

Miyamoto T, Furusawa C, Kaneko K. 2015. Pluripotency, Differentiation, and Reprogramming: A Gene Expression Dynamics Model with Epigenetic Feedback Regulation. PLoS Comput. Biol. 11:e1004476. doi: 10.1371/journal.pcbi.1004476.

Moore SA, Ferhatoglu Y, Jia Y, Al-Jiab RA, Scott MJ. 2010. Structural and biochemical studies on the chromo-barrel domain of male specific lethal 3 (MSL3) reveal a binding preference for mono- or dimethyllysine 20 on histone H4. J. Biol. Chem. 285:40879–40890. doi: 10.1074/jbc.M110.134312.

Muller J et al. 2010. eggNOG v2.0: extending the evolutionary genealogy of genes with enhanced non-supervised orthologous groups, species and functional annotations. Nucleic Acids Res. 38:D190–D195. doi: 10.1093/nar/gkp951.

Nègre N et al. 2006. Chromosomal distribution of PcG proteins during Drosophila development. PLoS Biol. 4:e170. doi: 10.1371/journal.pbio.0040170.

Newman SA. 2005. The pre-Mendelian, pre-Darwinian world: shifting relations between genetic and epigenetic mechanisms in early multicellular evolution. J. Biosci. 30:75–85.

On T et al. 2010. The evolutionary landscape of the chromatin modification machinery reveals lineage specific gains, expansions, and losses. Proteins. 78:2075–2089. doi: 10.1002/prot.22723.

Pardue M-L, DeBaryshe PG. 1999. Drosophila telomeres: two transposable elements with important roles in chromosomes. Genetica. 107:189–196. doi: 10.1023/A:1003905210770.

Penny D, Collins LJ, Daly TK, Cox SJ. 2014. The Relative Ages of Eukaryotes and Akaryotes. J. Mol. Evol. 79:228–239. doi: 10.1007/s00239-014-9643-y.

Phillips-Cremins JE, Corces VG. 2013. Chromatin Insulators: Linking genome organization to cellular function. Mol. Cell. 50:461–474. doi: 10.1016/j.molcel.2013.04.018.

Postberg J, Forcob S, Chang W-J, Lipps HJ. 2010. The evolutionary history of histone H3 suggests a deep eukaryotic root of chromatin modifying mechanisms. BMC Evol. Biol. 10:259. doi: 10.1186/1471-2148-10-259.

Pryszcz LP, Huerta-Cepas J, Gabaldón T. 2011. MetaPhOrs: orthology and paralogy predictions from multiple phylogenetic evidence using a consistency-based confidence score. Nucleic Acids Res. 39:e32. doi: 10.1093/nar/gkq953.

Rechtsteiner A et al. 2010. The Histone H3K36 Methyltransferase MES-4 Acts Epigenetically to Transmit the Memory of Germline Gene Expression to Progeny. PLOS Genet. 6:e1001091. doi: 10.1371/journal.pgen.1001091.

Ross BD et al. 2013. Stepwise evolution of essential centromere function in a Drosophila neogene. Science. 340:1211–1214. doi: 10.1126/science.1234393.

Roudier F et al. 2011. Integrative epigenomic mapping defines four main chromatin states in Arabidopsis. EMBO J. 30:1928–1938. doi: 10.1038/emboj.2011.103.

Rowlands T, Baumann P, Jackson SP. 1994. The TATA-binding protein: a general transcription factor in eukaryotes and archaebacteria. Science. 264:1326–1329.

Ruan J et al. 2008. TreeFam: 2008 Update. Nucleic Acids Res. 36:D735-740. doi: 10.1093/nar/gkm1005.

Rudolph T et al. 2007. Heterochromatin formation in Drosophila is initiated through active removal of H3K4 methylation by the LSD1 homolog SU(VAR)3-3. Mol. Cell. 26:103–115. doi: 10.1016/j.molcel.2007.02.025.

Saksouk N, Simboeck E, Déjardin J. 2015. Constitutive heterochromatin formation and transcription in mammals. Epigenetics Chromatin. 8:3. doi: 10.1186/1756-8935-8-3.

Schoborg TA, Labrador M. 2010. The phylogenetic distribution of non-CTCF insulator proteins is limited to insects and reveals that BEAF-32 is Drosophila lineage specific. J. Mol. Evol. 70:74–84. doi: 10.1007/s00239-009-9310-x.

Schotta G et al. 2002. Central role of Drosophila SU(VAR)3-9 in histone H3-K9 methylation and heterochromatic gene silencing. EMBO J. 21:1121–1131. doi: 10.1093/emboj/21.5.1121.

Sebé-Pedrós A et al. 2016. The Dynamic Regulatory Genome of Capsaspora and the Origin of Animal Multicellularity. Cell. 165:1224–1237. doi: 10.1016/j.cell.2016.03.034.

Sebé-Pedrós A, Degnan BM, Ruiz-Trillo I. 2017. The origin of Metazoa: a unicellular perspective. Nat. Rev. Genet. 18:498–512. doi: 10.1038/nrg.2017.21.

Seto E, Yoshida M. 2014. Erasers of Histone Acetylation: The Histone Deacetylase Enzymes. Cold Spring Harb. Perspect. Biol. 6. doi: 10.1101/cshperspect.a018713.

Seum C et al. 2007. Drosophila SETDB1 is required for chromosome 4 silencing. PLoS Genet. 3:e76. doi: 10.1371/journal.pgen.0030076.

Shukla V, Habib F, Kulkarni A, Ratnaparkhi GS. 2014. Gene duplication, lineage-specific expansion, and subfunctionalization in the MADF-BESS family patterns the Drosophila wing hinge. Genetics. 196:481–496. doi: 10.1534/genetics.113.160531.

Smith ER et al. 2008. Drosophila UTX Is a Histone H3 Lys27 Demethylase That Colocalizes with the Elongating Form of RNA Polymerase II. Mol. Cell. Biol. 28:1041–1046. doi: 10.1128/MCB.01504-07.

Smith ST et al. 2004. Modulation of heat shock gene expression by the TAC1 chromatin-modifying complex. Nat. Cell Biol. 6:162–167. doi: 10.1038/ncb1088.

Srinivasan S, Dorighi KM, Tamkun JW. 2008. Drosophila Kismet Regulates Histone H3 Lysine 27 Methylation and Early Elongation by RNA Polymerase II. PLOS Genet. 4:e1000217. doi: 10.1371/journal.pgen.1000217.

Stabell M et al. 2006. The Drosophila G9a gene encodes a multi-catalytic histone methyltransferase required for normal development. Nucleic Acids Res. 34:4609–4621. doi: 10.1093/nar/gkl640.

Steward MM et al. 2006. Molecular regulation of H3K4 trimethylation by ASH2L, a shared subunit of MLL complexes. Nat. Struct. Mol. Biol. 13:852–854. doi: 10.1038/nsmb1131.

Sun X, Wahlstrom J, Karpen G. 1997. Molecular Structure of a Functional Drosophila Centromere. Cell. 91:1007–1019.

Suzuki S et al. 2016. Histone H3K36 trimethylation is essential for multiple silencing mechanisms in fission yeast. Nucleic Acids Res. 44:4147–4162. doi: 10.1093/nar/gkw008.

Tatusov RL, Koonin EV, Lipman DJ. 1997. A genomic perspective on protein families. Science. 278:631–637.

Thomae AW et al. 2013. A Pair of Centromeric Proteins Mediates Reproductive Isolation in Drosophila Species. Dev. Cell. 27:412–424. doi: 10.1016/j.devcel.2013.10.001.

Trojer P, Reinberg D. 2006. Histone Lysine Demethylases and Their Impact on Epigenetics. Cell. 125:213–217. doi: 10.1016/j.cell.2006.04.003.

Vogel MJ, Peric-Hupkes D, van Steensel B. 2007. Detection of in vivo protein–DNA interactions using DamID in mammalian cells. Nat. Protoc. 2:1467–1478. doi: 10.1038/nprot.2007.148.

Wagner EJ, Carpenter PB. 2012. Understanding the language of Lys36 methylation at histone H3. Nat. Rev. Mol. Cell Biol. 13:115–126. doi: 10.1038/nrm3274.

Wapinski I, Pfeffer A, Friedman N, Regev A. 2007. Automatic genome-wide reconstruction of phylogenetic gene trees. Bioinforma. Oxf. Engl. 23:i549-558. doi: 10.1093/bioinformatics/btm193.

Whitcomb SJ, Basu A, Allis CD, Bernstein E. 2007. Polycomb Group proteins: an evolutionary perspective. Trends Genet. TIG. 23:494–502. doi: 10.1016/j.tig.2007.08.006.

Wysocka J et al. 2006. A PHD finger of NURF couples histone H3 lysine 4 trimethylation with chromatin remodelling. Nature. 442:86–90. doi: 10.1038/nature04815.

Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. 24:1586–1591. doi: 10.1093/molbev/msm088.

Zhang W, Landback P, Gschwend AR, Shen B, Long M. 2015. New genes drive the evolution of gene interaction networks in the human and mouse genomes. Genome Biol. 16:202. doi: 10.1186/s13059-015-0772-4.

Zhang X et al. 2007. The Arabidopsis LHP1 protein colocalizes with histone H3 Lys27 trimethylation. Nat. Struct. Mol. Biol. 14:869–871. doi: 10.1038/nsmb1283.

Figures

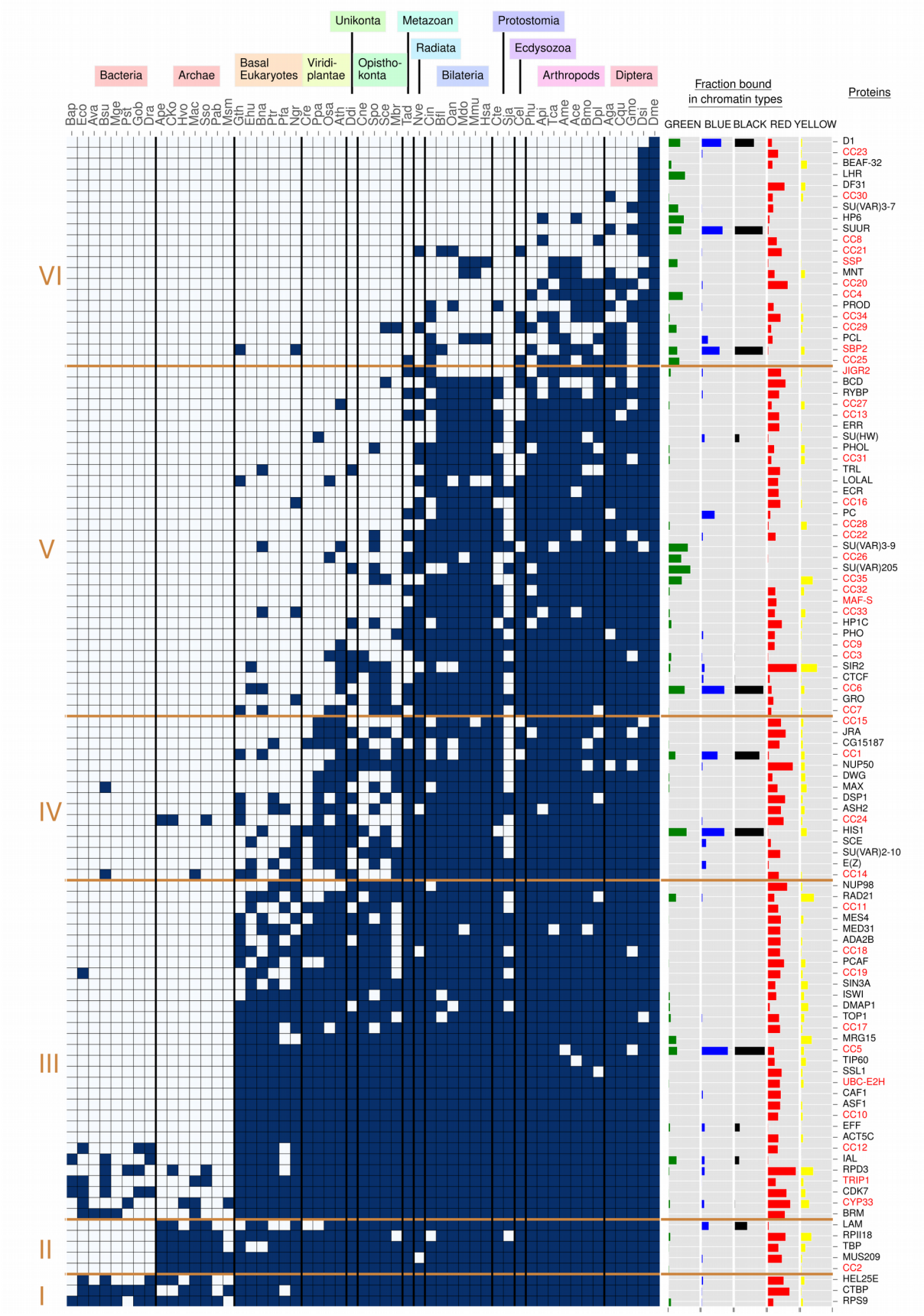
Unikonta
Metazoan
Protostomia
Radiata
Ecdysozoa
Bacteria
Archae
Basal Eukaryotes
Viridi-plantae
Opistho-konta
Bilateria
Arthropods
Diptera
Fraction bound in chromatin types
Proteins
GREEN BLUE BLACK RED YELLOW
Bap Eco Ava Bsu Mge Pst Gob Dra Ape Cko Hvo Mac Ssp Pab Nsm Gth Ehu Bna Ptr Pfa Ngr Cre Ppa Osa Ath Ddi Cne Spo Sce Mbr Tad Nve Cin Bfl Oan Mdo Mmu Hsa Cte Sja Cel Phu Api Tca Ame Ace Bmo Dpl Aga Cqu Gmo Dsi Dme
VI
V
IV
III
II
I
D1
CC23
BEAF-32
LHR
DF31
CC30
SU(VAR)3-7
HP6
SUUR
CC8
CC21
SSP
MNT
CC20
CC4
PROD
CC34
CC29
PCL
SBP2
CC25
JIGR2
BCD
RYBP
CC27
CC13
ERR
SU(HW)
PHOL
CC31
TRL
LOLAL
ECR
CC16
PC
CC28
CC22
SU(VAR)3-9
CC26
SU(VAR)205
CC35
CC32
MAF-S
CC33
HP1C
PHO
CC9
CC3
SIR2
CTCF
CC6
GRO
CC7
CC15
JRA
CG15187
CC1
NUP50
DWG
MAX
DSP1
ASH2
CC24
HIS1
SCE
SU(VAR)2-10
E(Z)
CC14
NUP98
RAD21
CC11
MES4
MED31
ADA2B
CC18
PCAF
CC19
SIN3A
ISWI
DMAP1
TOP1
CC17
MRG15
CC5
TIP60
SSL1
UBC-E2H
CAF1
ASF1
CC10
EFF
ACT5C
CC12
IAL
RPD3
TRIP1
CDK7
CYP33
BRM
LAM
RPII18
TBP
MUS209
CC2
HEL25E
CTBP
RPS9

Figure 1. Phylogenetic profile of chromatin-associated proteins. To the left, six protein age clusters are indicated with Roman numerals (I–VI). In the matrix, dark blue rectangles represent the presence of a homolog, grey rectangles its absence. On top, 13 species groups are defined to aid the reader, three letter codes refer to species names as given in Supplementary Table 1. The five columns "Fraction bound in chromatin types" display the fraction of chromatin type (GREEN, BLUE, BLACK, RED, YELLOW) bound by each CAP. To the right, the column "Proteins" contains protein names, with unknown proteins in a red font.

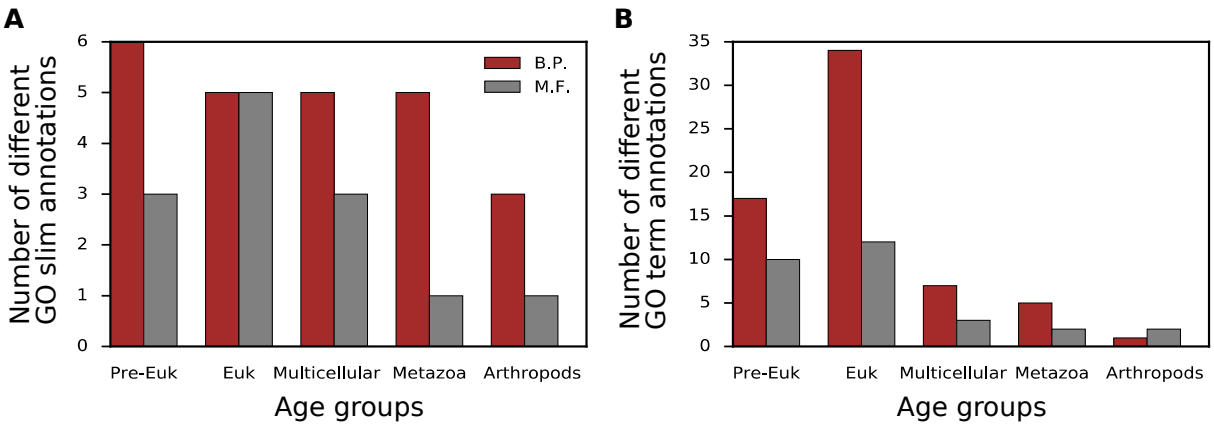

Figure 2. Diversity in functional annotations at evolutionary age groups. (A) Number of different high-level GO slim annotations per age group (see Methods for details). **(B)** Per age group, the number of different specific GO term annotations is displayed. Specific GO terms are defined as the deepest children of a GO slim category. In both panels, B.P. stands for Biological Process, M.F. for Molecular Function, *Pre-Euk* means pre-eukaryotic gene age (cluster I and II), *Euk* is eukaryotic age (cluster III), *Multicellular* is multicellular plant and

metazoan age (cluster IV), *Metazoa* is metazoan age (cluster V), and *Arthropods* is arthropod age (cluster VI).

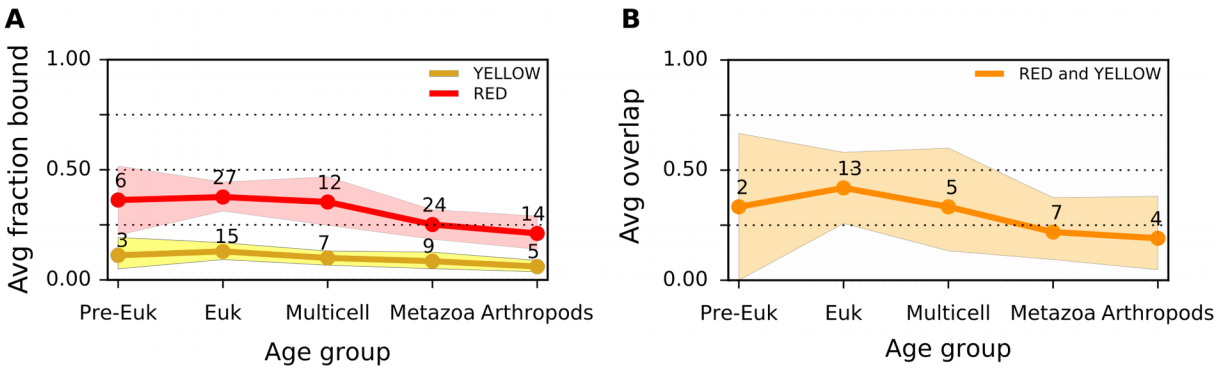

Figure 3. Average fraction of genome bound by proteins over evolutionary age groups. **(A)** Per age group, the average fraction of YELLOW and RED chromatin to which CAPs bind, with 95% confidence intervals obtained by bootstrap analysis. Points annotated with numbers, indicate the number of proteins per age group classified as bound to a given chromatin type (fraction for a chromatin type > 0.1). **(B)** Average proportion of CAPs of each evolutionary age group that bind both YELLOW and RED (fraction for both chromatin types > 0.1). Again, 95% confidence intervals were obtained by bootstrap analysis, and annotated points are number of proteins bound per age group. In both panels, the bootstrap procedure was as follows: for each age group, say of size *n*, we resample *n* proteins (with replacement) 1000 times. For panel A, we compute each time the average fraction of YELLOW and RED bound by the resampled proteins, while for panel B, we count the CAPs that bind both chromatin types > 0.1. See Fig. 1 "Fraction bound in chromatin types" for the fraction of chromatin bound by individual proteins.

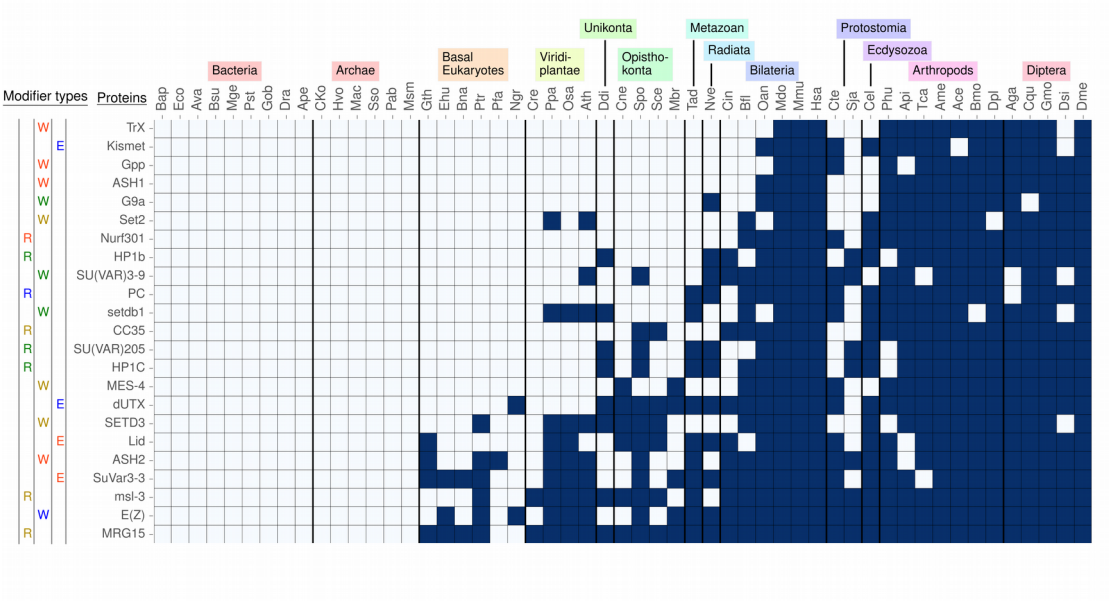

Figure 4. Phylogenetic profile of *D. melanogaster* histone mark modifiers. On the left, reader (R), writer (W), and eraser (E) proteins are specified with a color code corresponding to the chromatin types and their typical histone modifications. YELLOW refers to H3K36me3, YELLOW and RED are shown in orange and indicate H3K4me3, BLUE is H3K27me3, GREEN is H3K9me2/3. In the matrix, dark blue rectangles represent the presence of a homolog, grey rectangles its absence. On top, 13 species groups are defined to aid the reader and the three letter codes refer to species names as given in Supplementary Table 1.

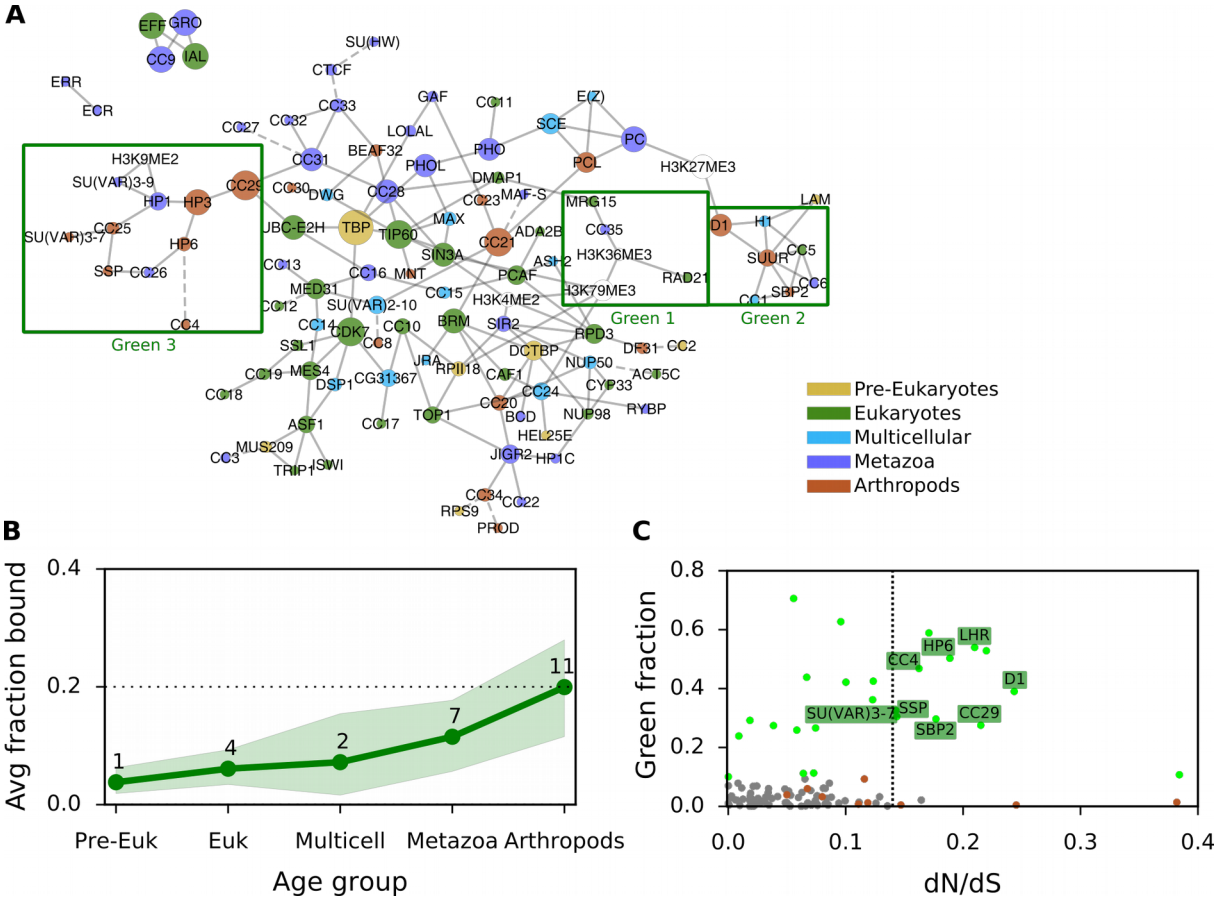

Figure 5: GREEN-associated proteins over evolutionary age and their dN/dS ratio. **(A)** Bayesian network of CAPs adapted from (van Bemmel et al. 2013). Proteins and histone marks are connected by a solid line if the predicted interaction has a confidence score of at least 70%. Dashed lines indicate the highest scoring interaction for proteins with all confidence scores below 70%. Three GREEN chromatin subnetworks are highlighted with a green box. **(B)** Average fraction of GREEN chromatin bound by CAPs of each evolutionary age group, with 95% confidence intervals obtained through bootstrap analysis. See Fig. 1 "Fraction bound in chromatin types" for the individual proteins and their fraction of chromatin bound and Fig. 3 for details on the bootstrap method. **(C)** The ratio of non-synonymous over synonymous aminoacid mutations (dN/dS) against the GREEN fraction bound for each of the

107 CAPs. The dotted vertical line at 0.135 divides CAPs with dN/dS > 0.135 (15% higher of distribution) from the rest. Chromatin proteins are grey dots, with GREEN ones as green dots. GREEN Arthropod Cluster proteins (GAPs) are labeled with their names, while other proteins of the Arthropod Cluster are brown dots.

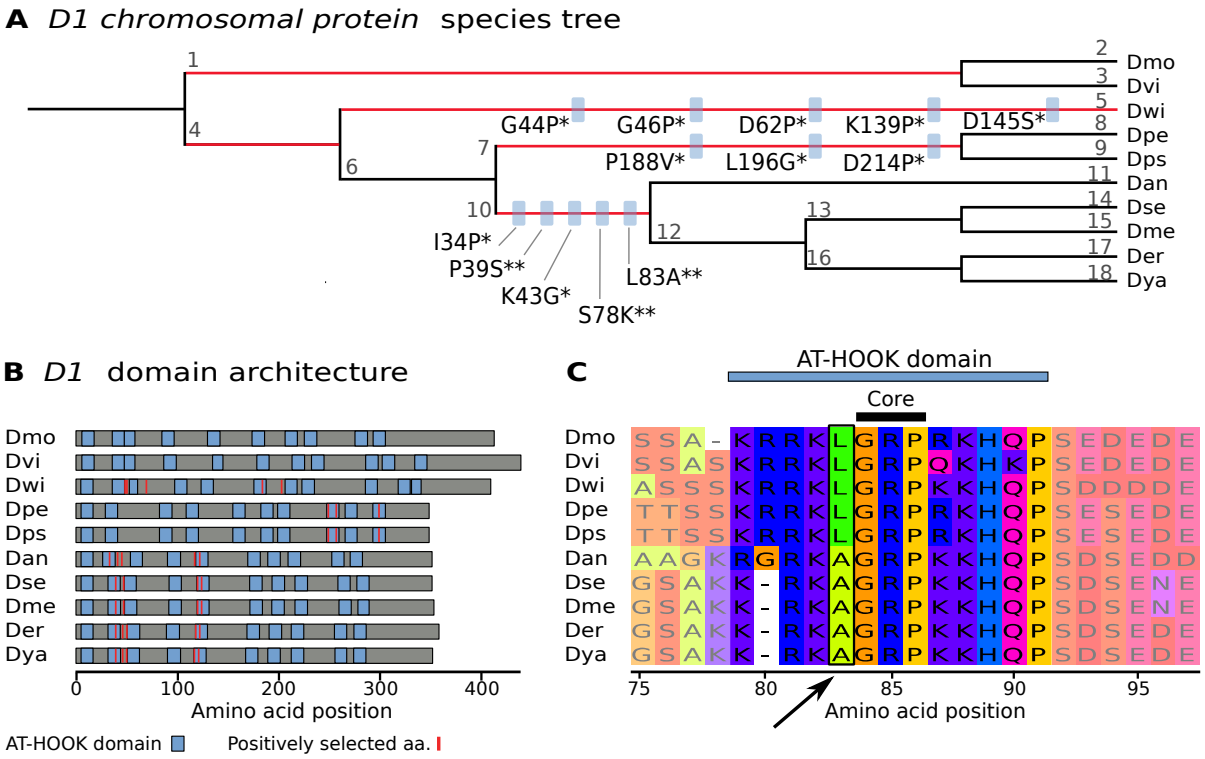

Figure 6. *D1 Chromosomal protein* is undergoing positive selection at AT-HOOK domains. (A) Drosophilid species tree (with arbitrary branch lengths). The five branches with positive selection events are highlighted in red (p< 0.01, Bonferroni correction). On branches with more than one positively selected site, blue boxes indicate the amino acid substitution under positive selection, with the significance given as posterior probability of dN/dS > 1 (* for Pr > 0.95, ** for Pr > 0.99). For instance, L83A indicates the substitution of lysine with

alanine at position 83. The *Drosophila* species are indicated by 3 letter abbreviations: Dmo is *D. mojavensis*, Dvi is *D. virilis*, Dwi is *D. willistoni*, Dpe is *D. persimilis*, Dps is *D. pseudoobscura*, Dan is *D. ananassae,* Dse is *D. sechellia*, Dme is *D. melanogaster*, Der is *D. erecta*, and Dya is *D. yakuba.* Each branch is identified with a number that links to details on positive selection tests (Supplementary Table 4). **(B)** Domain architecture of D1 proteins. Per species, protein length is given by grey horizontal bars. AT-HOOK domains are represented by blue boxes and positively selected amino acids (aa) are displayed as red vertical lines. **(C)** An example of positive selection in an AT-HOOK domain. The multiple alignment of D1 is zoomed in on the region 75–97aa, showing a change of L83 directly in front of the core motif. Visualisation is done with MSAViewer (http://msa.biojs.net/app/).