The evolution and diversity of Polycomb in the Metazoa

Author: Dana Frank - S3587185
Supervisors: Prof. Bregje Wertheim & Dr. Thomas Verschut
Date: 06-12-2021
University of Groningen, Faculty of Science and Engineering
Image source: Courtney Chromatini by David Sweatt

Table of Contents

Abstract

With their ability to mediate transcriptional memory, few genes have gained the same eminent status as the evolutionarily conserved members of the Polycomb group. Best known for silencing HOX genes along the anterior-posterior axis in *Drosophila*, these developmental regulators were thought to constitute a higher level of regulatory complexity exclusive to bilaterians. However, recent studies of Polycomb have extended our perspectives on its evolutionary history, and with the knowledge that Polycomb is present in nearly every animal phylum, we are beginning to create a more complete picture of its structural diversity. Furthermore, the Polycomb group has been implicated in numerous other biological processes such as TAD formation. Here, I examine the consequences of these new findings in the light of previous interpretations of developmental regulation in bilaterian and non-bilaterian animals. Specifically, I illustrate that most subunits from Polycomb Repressive Complex 1 and Polycomb Repressive Complex 2 underwent a multifaceted evolutionary history, rich with duplication and deletion events, causing them to differ considerably among animal lineages. I also suggest that Polycomb group (PcG) gene expansion and diversification occurred earlier in the metazoan lineages than previously assumed. I further stipulate that in order to unravel the evolutionary track of PcG, it is imperative to examine taxa in basal positions of the animal lineage, which for the most part have been heavily underrepresented in phylogenomic and functional studies in the context of developmental genes.

Introduction

For a long time, the relative complexity of organisms and the phenotypic differences between them were exclusively attributed to numbers of genes. In this sense, a larger and more intricate genetic repertoire would correlate with a higher functional sophistication of various biological systems. By extension, the position of animals on the phylogenetic tree was thought to represent their degree of genomic complexity. Earlier diverging phyla – which are, with some exceptions, morphologically more simple – were assumed genomically less complex in terms of protein-coding gene numbers than later diverging ones. This misconception was eventually revealed following the sequencing of the human genome [1,2,] and those of non-bilaterians [3]. Through comparison with other organisms, we discovered that differences among genome sequences were not sufficient to explain the vast diversity of life on earth; the human genome did not contain nearly as many genes as expected, with insufficient differences compared to other animals. When the vision of monogenetic explanations for biological questions was shattered, a sense of mystery swept through the scientific community. Understanding the mechanisms behind the collective complexity of animals, among which lies an enormous collection of behaviors, modular organizations, and forms, was pushed into the distant future.

It took several years before evolutionary developmental biology (evo-devo) brought the field forward. By inferring the relatedness of organisms based on their developmental processes rather than genetics alone, the notion of genetic regulation by epigenetic factors as the chief modulator of complexity was formed. In particular, discovering HOX genes, a family of transcription factors controlling the body plan along the anterior-posterior axis, which specifies segment identity of embryonic tissues, was paramount [4]. Using *Drosophila* as the central model system directed attention away from established dogmas about the drivers of gene evolution on the basis of major genomic changes and toward changes in regulatory networks as the primary source of biological diversity. The strong influence of epigenetics on phenotype, which consists of sequence-independent factors, was also emphasized. However, working nearly exclusively on *Drosophila, C. elegans,* and mice resulted in many key taxa being ignored and, indeed, false generalizations on the biology of genes in animals. It also set a precedent of hyper-focusing on a small subgroup of the animal kingdom as model organisms of genetics and development.

Not long after the discovery of HOX genes, the Polycomb gene was isolated as a repressor of HOX gene expression. When disrupted in *Drosophila*, Polycomb derepresses HOX genes in particular body segments. Most notably, a decline in Polycomb group (PcG) function often manifests as a homeotic transformation of posterior legs to anterior legs, which have a distinctive set of comb-like bristles [5,6]. Hence, the name Polycomb. Since then, it has been revealed that Polycomb regulates many more genes in a variety of elegant mechanisms. This functional diversity arises from the different members of the Polycomb family, acting in disparate developmental stages, often in a cell-specific manner, to modify chromatin for repression. Furthermore, Polycomb, and its regulatory associates, are highly conserved in almost every animal phylum extending beyond the bilaterian clan, with unique modifications scattered throughout the lineages [4,7–12]. The assumptions that the morphological simplicity of non-bilaterians corresponds to a lack of regulatory intricacy have clouded progress in uncovering the evolutionary history of Polycomb by setting the focus on more recently-diverging animals. We now know that numerous regulatory complexes, including Polycomb, appeared much earlier in metazoan phylogeny than previously assumed [7,9,12–14]. Still, the

under-representation of non-standard model organisms persists as an unfortunate trend in scientific inquiry.

Here, I will provide an overview of the molecular principles of Polycomb Repressive Complexes (PRCs) and explain the biochemical activities that enable their functioning based on foundational work in *Drosophila*. Next, I outline the structural and functional differences of PRCs among animal lineages, and finally, I will examine the direct evidence concerning the evolution of Polycomb based on the existing data surrounding PcG in early-diverging animals. I argue that studying early-diverging animals is vital for understanding the evolutionary history of metazoan Polycomb. Specifically, they are required as reference points to deduce the expansion and deletion events of PcG protein subunits, as well as uncovering their structural and functional diversification.

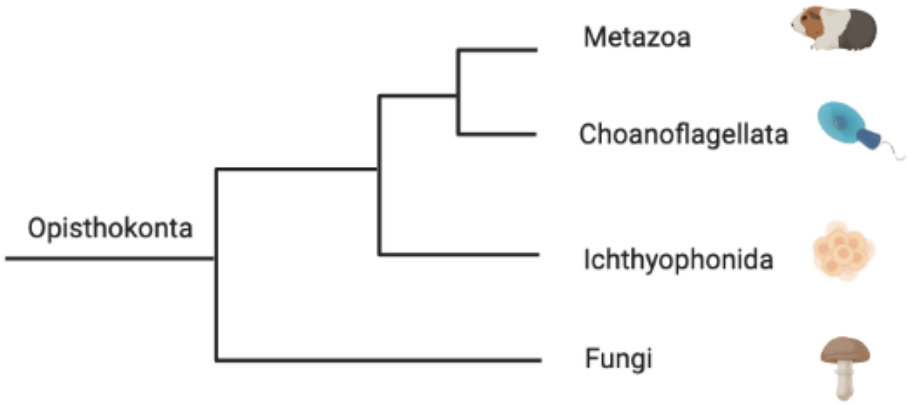

Figure 1. The Opisthokonta clade. The group constitutes a broad selection of eukaryotes, including the metazoan and fungus kingdoms. Phylogenetic topology was inferred from recent phylogenetic analysis. Created in Biorender.

The Polycomb system

The Polycomb group is now recognized as one of the most prominent families of chromatin-modifying proteins present in the Opisthokonta clade (Fig. 1.). It is implicated in regulating a broad spectrum of gene expression pathways, particularly gene repression. By working in opposition with its antagonistic counterpart, Trithorax, these chromatin-modifying proteins induce heritable states of repression and activation of gene expression, respectively [15–17]. The deceptively multifaceted interaction between these systems is maintained to allow for dynamic responses to varying conditions and thus ensures appropriate activation states of their target genes [18–20]. In this way, stable gene expression patterns throughout the entire span of an organism's development are achieved while also preserving cellular identity [18]. Here, I discuss PcG in *Drosophila* as representative of complexes in other bilaterian animals to be described below. Unfortunately, functional work on Polycomb complexes has not yet been performed in non-bilaterians and thus cannot be addressed.

The two major Polycomb repressive systems are Polycomb Repressive Complex 1(PRC1) and Polycomb Repressive Complex 2 (PRC2) (Fig. 2), which mainly repress transcription factors and signaling pathways of developmental genes. In the traditional model, PRC2 functions prior to PRC1 by epigenetically marking histones for repression, later recruiting

PRC1. PRC2 consists of four protein subunits: enhancer of zeste (E(Z)), extra sex combs (ESC), a zinc finger called suppressor of zeste 12 (SU(Z)12), and chromatin assembly factor 1 subunit (CAF1/NURF55) [15,21,22]. The E(Z) subunit constitutes the catalytic segment required for histone methyltransferase activity, specifically, the di- and trimethylation of histone 3 on lysine 27 (H3K27me2/3) [22]. This is facilitated by the SET domain of E(Z) and is enhanced by ESC through a boost in enzymatic rate [23]. SU(Z)12 and CAF1 are necessary for recruitment and binding to regulatory sites, such as Polycomb response elements (PREs) [21,24]. Jointly, these protein subunits constitute the first step in creating a repressive mark on their target histone, leading to chromatin compaction. This limits transcription factor binding and is followed by the further repressive action of PRC1.

Similarly, PRC1 also consists of four protein subunits: Polycomb (PC), Sex combs extra (SCE), Posterior sex combs (PSC), and polyhomeotic (PH) [25,26]. PC aids in recruiting PRC1 to H3K27 by recognizing the trimethylation mark deposited by PRC2 via its chromodomain [27]. The E3 ubiquitin ligase activity of SCE monoubiquitylates lysine 118 of histone H2A (H2AK118ub), constituting an additional chromatin silencing moiety [25]. PH is later involved in the inhibition of chromatin remodeling [28]. By binding to nucleosomes and depositing a second repressive moiety, PRC1 induces gene repression by physically blocking access of transcription factors to the target area, as well as generating another biochemical mark of repression.

The PRC1 complex can be further subdivided into canonical PRC1 (cPRC1) and noncanonical PRC1 (ncPRC1), distinguished by the absence of PC in the noncanonical form [29]. The majority of H2A ubiquitylation is catalyzed by this noncanonical system, although only the cPRC1 can be recruited by its PC subunit to the methylation site of histone 3, allowing it the flexibility to mediate long-range interactions as well as local ones [28,29]. This may be why HOX genes in *Drosophila* are exclusively regulated by cPRC1, even though the noncanonical version is also present in *Drosophila*. However, the reason behind the complex-specific regulation of HOX has not been empirically examined.

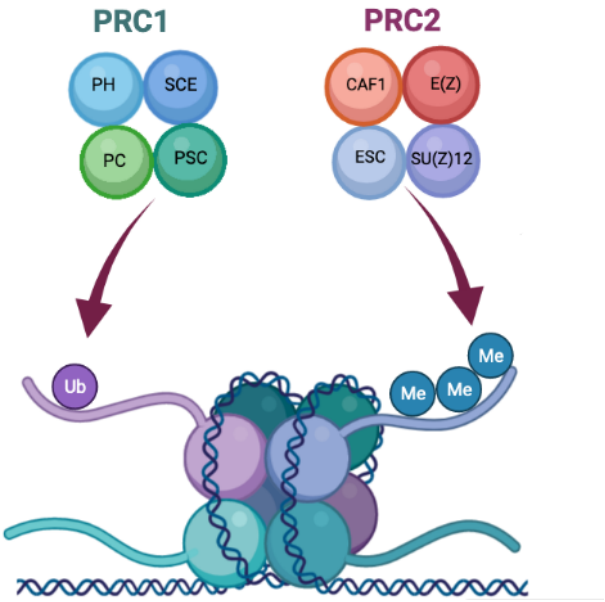

Figure 2. Classical model of Polycomb function in Drosophila. PRC2 is recruited to histone tail 3 and tri-methylates lysine 27. PRC1 is subsequently recruited by the PRC2 H3K27 trimethylated nucleosome and ubiquitinates lysine 118 of histone H2A. Together, these epigenetic marks generate a repressive state to the target region. Created in Biorender.

This classical model has been elaborated upon extensively in recent years, revealing a more complex interplay between PRC1 and PRC2 components beyond the idea that cPRC1 always recruits PRC2. The current view posits a less stepwise relationship between the complexes, but rather a mutual signal strengthening, where in many cases, PRC1 action precedes that of PRC2 (Fig. 2) [30]. In addition, the vast diversity of PcG genes means that many PRCs have specialized functions, yielding distinctly different phenotypes when mutated. This implies that the various targets of PcG are complex specific, whereby certain genes/cell types are regulated by particular PRCs exclusive to those targets. Moreover, PcG does not only function in chromatin modification but also in other mechanisms such as DNA methylation of CpG islands that regulate the recruitment and activity of both complexes. The subsequent transcriptional repression is believed to play a large part in the diversity of polycomb function and degree of specificity.

Interestingly, it has been shown that histone modifications, coordinated by the Polycomb system, are maintained through multiple cell divisions. This implies a robust mechanism that allows Polycomb to withstand the dramatic chromatin remodeling and biochemical modifications accompanying DNA replication and cell division. The details of this are not yet understood. However, the sequence-independent heredity posited by traditional epigenetic models does not seem to apply to PcG. While it has been demonstrated that PRC2-methylated histones can propagate their epigenetic state to first-generation daughter cells, this alone is not sufficient to further transmit the repressive mark [31]; excision of PcG responsive elements (PRE) loci cause cumulative dilution of trimethylation marks with newly incorporated nucleosomes during mitosis [32]. Thus, a coupled mechanism, whereby H3K27 trimethylated nucleosomes anchor PRC2 at the PRE binding site, in addition to the sequence-independent memory of repression via H3K27 trimethylated nucleosomes, is necessary for memory of repression [32,33]. This is an intriguing example of epigenetic inheritance acting in contradiction to its definition (sequence-independent inheritance). Perhaps it is time to revise this characterization.

Furthermore, the mechanisms by which Polycomb triggers gene repression are multifaceted. Polycomb-dependent long-range promoter–promoter and promoter–enhancer of numerous developmental genes, including HOX genes, can modify the 3D structure of the genome by creating and stabilizing the architecture of topologically associated domains (TADs) [19,20,26]. After the midblastula transition in embryogenesis, repressive chromatin loops are formed to mediate contact between repressed genes in distal regions of the genome, governing gene silencing during development. However, the mechanism behind the folding of Polycomb-associated domains inside nuclear structures is not yet understood [34]. Nevertheless, when disrupted, TAD boundary aberrations have been implicated in a myriad of developmental abnormalities, including abnormal gene-enhancer interactions and cancer, illustrating the importance of hierarchical organization in genome function [19]. Interestingly, TADs containing developmental genes have a higher degree of conservation than broadly expressed genes among *Drosophila* species [19]. In fact, most TADs are non-orthologous among *Drosophila* species, with boundary elements frequently re-organized by chromosomal rearrangements; TADs containing developmental genes seem to be more highly conserved than TADs containing other genes, for unknown reasons [19].

Structural and functional differences in bilaterians

The core components of the multimeric PcG complexes are well conserved among many animal lineages. However, vertebrates possess a significantly bigger repertoire of alternate subunits than other taxa. This is especially the case for PRC1, whose four core subunits have several homologous copies, with varying numbers, in vertebrate taxa. Specifically, they correspond to at least two and up to six variants per subunit each in vertebrates [35,36]. This larger pool allows for much more diversity of complex functions. As such, many papers have posited that the increase in PRC subunit combinations in vertebrates points to a lucrative developmental expansion of complexity. In this way, they suggest that a more sophisticated array of developmental tools exists in vertebrates compared to other animal lineages [35–44]. Of course, this was claimed far before any rigorous sampling of PcG was performed in non-bilaterians (Fig. 3). Due to the insufficient functional knowledge of PcG in early-diverging taxa, I will only be describing their structural and functional differences as they exist among bilaterians in this section.

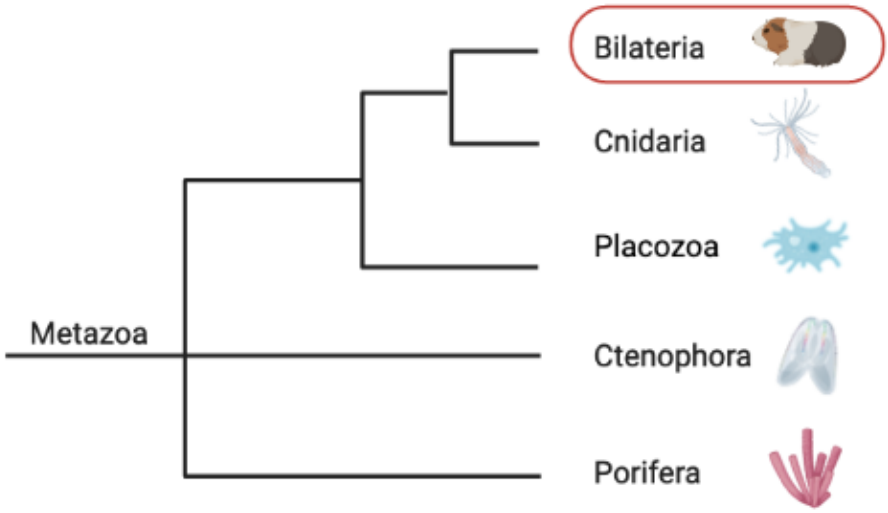

Figure 3. The extent of metazoan sampling for Polycomb performed in functional studies to date. The encircled bilaterian clade includes all organisms used as models to study Polycomb function thus far. Though Polycomb genes were identified in other animal lineages, the degree of homolog specificity, as well as the functional similarity, is ambiguous. Metazoan topology was inferred from recent phylogenetic analysis. Created in Biorender.

One such example is CBX7 (vertebrate ortholog form of *Drosophila* PC), which functions most abundantly in embryonic stem cells and is responsible for maintaining pluripotency. It also acts as a repressor of CBX2, CBX4, and CBX8 [45,46]. While the mechanism behind different CBX functions is unclear, there may be discrepancies between binding affinities of CBX to various chromodomains of target histones that could contribute to altered functioning [45]. In humans, however, CBX6, CBX7, CBX8, RING1, and RING2 show the ability to localize in overlapping sites; this redundancy seems to be unique to humans [42]. By contrast, the single PC subunit in insects performs all the roles of vertebrate CBX, from its embryonic to adult form, leaving the functional diversification of CBX variants in vertebrates unclear.

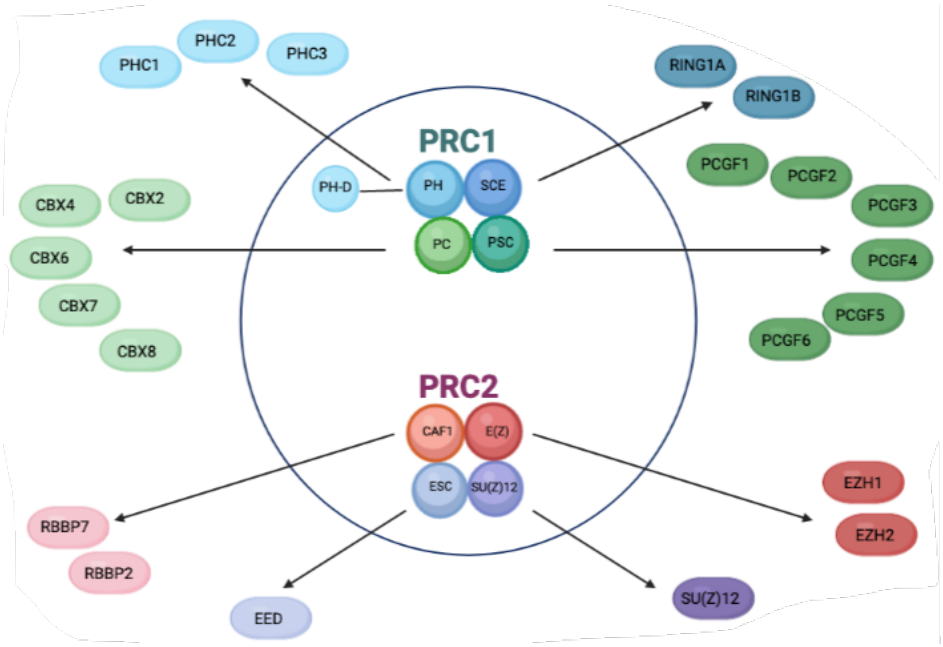

Fig. 4. Expansion of polycomb repressive complexes in vertebrates. *Drosophila* homologs are represented as circles, and vertebrate equivalents are represented as oblong. The insect PRC1 and PRC2 members are shown inside the large circle, and the corresponding vertebrate homologs are shown as branches outside the circle. Redrawn from Sowpati et al., 2015. Created in Biorender.

Similarly, the two Ph subunits in *Drosophila* (Ph-proximal and Ph-distal) correspond to three vertebrate homologs (PHC1, PHC2, and PHC3) [47]. Here, the zinc-finger, SAM, and HD1 domains are conserved between insects and vertebrates, with differences only present in the expansion of Serine-Threonine rich motif of vertebrate homologs surrounding their N-termini [48]. Glycosylation of this domain is essential for its function; thus, expansions could further enhance Ph function [49]. The vertebrate homologs also differ from each other with the presence of either a coil or helix-turn-helix motif between the SAM and zinc-finger motif. The unique structures pertaining to these motifs are not understood but also point to a specialization in mechanism compared to the functional generalization in insect PRC1.

In contrast, some PcG subunits have redundant functions with each other. For example, in the case of RING1A and RING1B (homologs of SCE in *Drosophila*) of PRC1, the enzymes both have the ability to ubiquitinate histone 2A [50], but this may be because the ubiquitination step is the most essential function of PRC1 so that an extra copy could be advantageous. This also seems the case for PRC2 subunits, although EZH1 and EZH2 also have a functional divergence, where EZH2 works similarly to its homolog in *Drosophila*, whereas EZH1 is mainly active in non-dividing cell types [51–53].

The evolutionary history of Polycomb

Expansion and contraction of PRC subunits

Gene duplication, either of a regulatory element or a target gene, allows for subsequent divergence in function and may result in gene family expansion [54,55]. The PcG is one such group that is likely to have undergone multiple gene duplications and losses at various stages of its evolution [7,44,56]. It has been suggested that a period of extensive gene duplication events, and even two whole-genome duplications, occurred during the evolution of

vertebrates from invertebrate ancestors. The PRC1 complex has been implicated in this expansion [57,58]. Extant invertebrates, such as *Drosophila* and *Echinodermata*, have single copies of PRC1 subunits with the exception of Psc [59,60] (Fig. 4). By contrast, vertebrate species have multiple homologs of all PRC1 members [36] (Fig. 4). Given the information on these taxa, the conclusion that PRC1 subunits underwent a significant expansion seems highly plausible.

In agreement with this line of reasoning, HOX genes also underwent multiple duplication events during vertebrate evolution, where invertebrates possess a single HOX gene cluster [12,61,62]. It, therefore, seems plausible that a conserved association between HOX and Polycomb genes, in a co-evolutionary scenario, resulted in the development of a higher degree of complexity and is functionally coupled to the diversification of bilaterian body plans. This appears even more likely when one considers the specialized functions of the PcG members; if Polycomb complexes are target specific, then an expansion of the number of targets implies an expansion of their regulators.

This massive diversity of PRC1 was most dramatic in the "expanded" collection of CBX and PCGF subunits, which have five and six homologs respectively in vertebrates, compared to the single copies in insects [7,36,56]. Indeed, the elegance of specialized Polycomb homologs in the vertebrate system coinciding with a substantially enriched complexity in gene regulation is a highly alluring idea. Bioinformatics studies too supported the notion that the enhanced variety of unique signature motifs could explain their divergent functions and act as a means of creating an unrivaled regulatory system [36].

Recently, however, these ideas were challenged by a comparative study [7]. Until then, the evidence supporting the expansion of PRC1 during vertebrate evolution had been deduced solely from members of the bilaterian clade (Fig. 3). By examining the genomes of a wide range of metazoans for cPRC1 and ncPRC1 subunits with the aid of phylogenetic methods, they found that the cnidarian anthozoan, *Nematostella vectensis*, possesses a full, vertebrate-like PCGF complement to all homologs (Fig. 5). A partial PCGF cluster was also identified in other cnidarians, and phylogenetic analysis reflected the evolutionary relationship between them: *PCGF5* and *PCGF3* have the highest sequence similarity and are located adjacently to each other. At the same time, *PCGF1* is the least similar and is also located further away. This cluster organization is retained in protostome genomes but not in deuterostomes. These data indicate that the expansion of PRC1 group proteins took place before the last common ancestor of cnidarians and bilaterians and not in the vertebrate lineage.

Whether the same evolutionary scenario of ancient duplications followed by gene loss in some animal lineages also applies to other subunits of PCR1 is yet unknown. Multiple copies of the CBX and PHC subunits were not found in invertebrates and could have indeed undergone duplication during vertebrate evolution. Further examination of metazoan taxa is required to determine their origin. Moreover, each vertebrate homolog group usually contains one member that highly resembles its *Drosophila* counterpart. It seems plausible that this vertebrate member is the ancestral subunit while the others evolved through vertebrate-specific duplications and functional drifting; however, this has not yet been addressed.

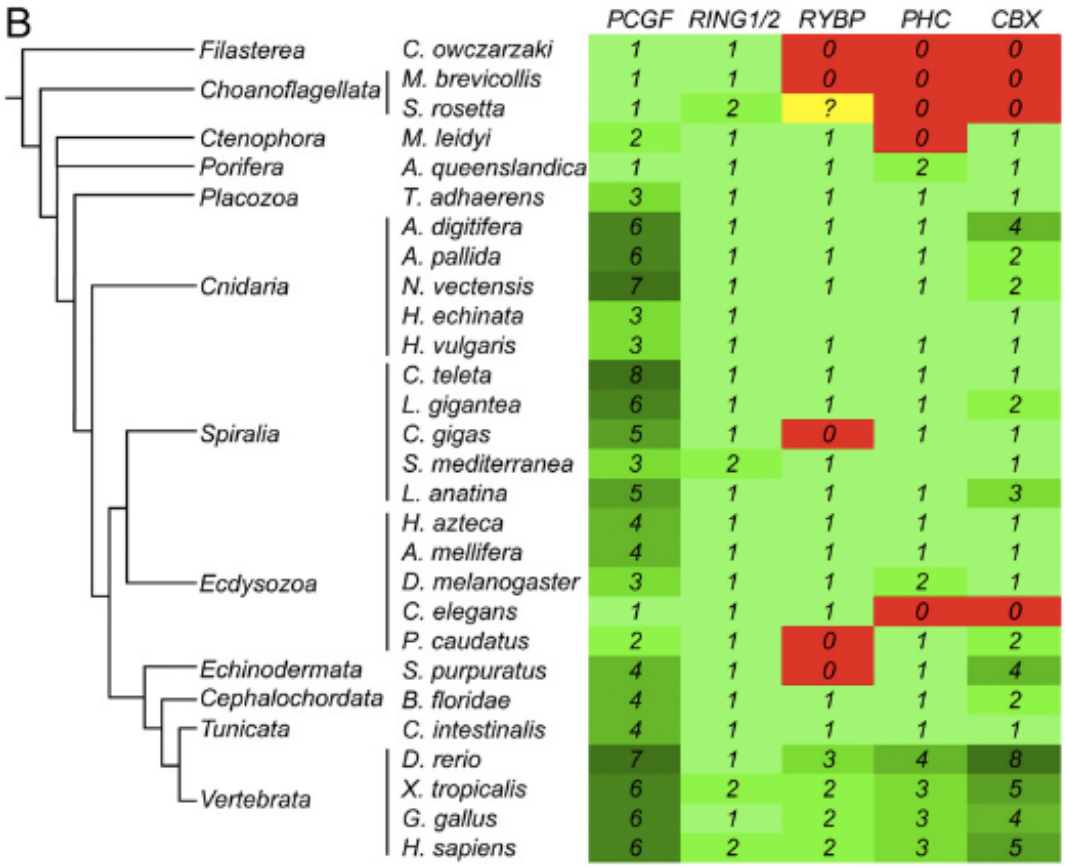

Figure 5. Phylogenetic distribution of PRC subunits in animals. A simplified phylogenetic tree of animals and their close relatives is shown on the left. The presence of PRC1 subunits is denoted by the numbers in the corresponding table, with red indicating their absence and green indicating their presence. The intensity of the green color also corresponds with the number of subunit homologs. Yellow boxes with question marks represent disagreement in the identity of predicted proteins orthologs as bona fide PRC subunits. Adapted from Gahan et al., 2020.

In a complementary approach, examining epigenetic marks, such as histone modifications, led to the discovery of PRC2 in the sponge, *Amphimedon queenslandica* [63]. Even though Gahan et al. did not identify PRC1 homologs in Poriferans, sponge homologs of bilaterian PRC2 components, including four copies of E(z), two copies of ESC, and one copy of SU(z)12, respectively, were found to be highly conserved between sponges and other, better-studied lineages. Moreover, binding motifs and homeodomain regulators were similar to those found in *Drosophila* [59,64–67]. This aligns with the previously identified presence of H3K27me3 silenced regions in the sponge genome [68,69]. Thus, the quantity of genomic regulatory features fails to reflect the degree of morphological complexity in animals, at least in this case. The fact that regulation seems to have the same foundational components from the Poriferan phylum to the Bilaterian lineage means that something else must be driving the morphological differences in animals. Possibly, the number of regulatory components possessed by animals, rather than qualitative changes in gene expression regulation itself, determine the intricacy of the body plans of different metazoans.

The functional reasons for why the expansion of gene families occurred at a given time are unknown. Such expansions or contractions of gene families can result from natural selection or may merely be the result of chance events, namely mutation and drift. To distinguish between these two explanations is often challenging in practice. Recent work employs a

sequence of statistical models and algorithmic methods to detect if gene families evolved under the influence of selection [70–72]. Given the diversity in function of CBX and PCGF in vertebrates, retention by chance alone would seem unlikely. It will be interesting to explore if anthozoan PRC1 diversification also reflects new functions.

PCGF and its implications

As previously discussed, it has become evident that the last common ancestor of the cnidarian and bilaterian lineages probably had a minimum of five PCGF subunits. Moreover, there was likely only a single gene duplication event of the PCGF during vertebrate evolution. The fact that anthozoan cnidarians have one of the largest collections of PCGF genes sets the stage for countless contraction events in the animal kingdom [7] (Fig. 5.). Interestingly though, phylogenies reveal that non-canonical PCGF (ncPCGF) genes (PCGF1, PCGF3, PCGF5, & PCGF6) are more closely related to each other than to canonical PCGF genes (PCGF 2 &PCGF4), implying that the ncPCGF family descended from one ancestral gene that underwent multiple duplications. This is further supported by their sequential presence in a single genomic cluster, although some clades lack PCGF5, likely through deletion [29]. Interestingly, its loss in both the protostome and non-vertebrate deuterostomes lineages occurred independently. Further losses have ensued in specific protostomes, for instance, deletion of *PCGF6* and the loss of *PCGF3* in hydrozoans.

It is unclear whether this cluster organization is required for its function, as documented in HOX clusters. Crucially, however, the absence of a PCGF1 homolog in *Drosophila*, coupled with the presence of noncanonical PCGF homologs in its canonical PRC1 systems, is puzzling [73]. This phenomenon implies that natural selection may favor a particular PCGF component over another under given conditions, thus "switching" the composition of PRC1. Whether this is an isolated case in *Drosophila* alone or a more common occurrence is unknown.

Choanoflagellates, the sister group of animals (Fig. 1), possess PRC1 components such as RYBP but not the central cPRC1 components, such as CBX. As such, it was previously proposed that ncPRC1 evolved prior to the cPRC1. However, if the switching phenomenon observed in *Drosophila* occurred in other lineages, it may indicate that PRC1 and its noncanonical form evolved in the same evolutionary time frame. Therefore, broader and more basal sampling in the metazoan lineage would be the most effective in answering the question of the emergence timing of different Polycomb subunits.

Box 1. Origin of Polycomb

PRC2 is believed to have appeared during early eukaryotic evolution, likely in an ancestral unicellular organism [81]. The apparent absence of PRC2 in unicellular fungi also suggests its emergence coinciding with the evolution of multicellularity. It is suggested that it may have been involved in gene silencing through H3K27 methylation, particularly in defense against transposable elements or insertion of new genes. Given that this function is partially redundant, this mode of gene silencing gained a more specific function as multicellular organisms attained cell-specific lineages.

By contrast, PRC1 exists in a narrower phylogenetic distribution in extant eukaryotes, and thus could have evolved more recently in specific lineages. Notably though, PRC1 subunits exist in the last common ancestor of seed plants, where they play a similar regulatory role. Further insights about PRC1 origin are not known.

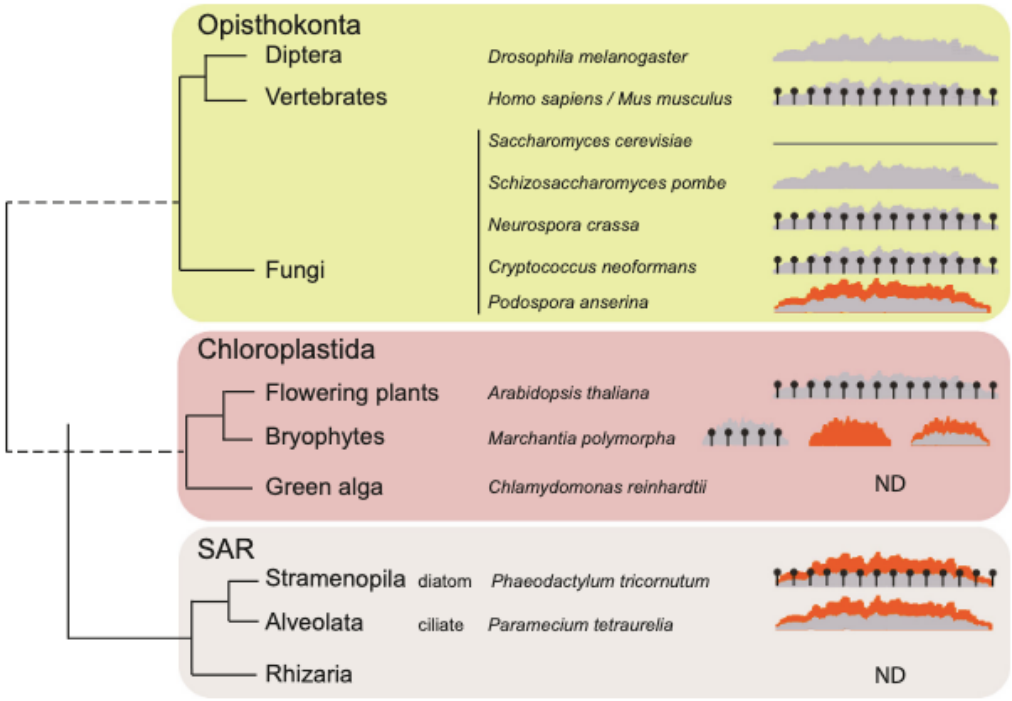

Figure 6. Function of PCR2-mediated H3K27 methylation in silencing transposable elements in eukaryotes. The phylogenetic tree represents eukaryotic clades possessing PRC2. The presence of orange topology represents instances where H3K27 is methylated as a silencing mechanism for transposable elements. Grey represents other histone silencing marks. DNA methylation is denoted by lollipops. ND (not determined) represents cases where main silencing mechanisms for transposons are not yet determined. Adapted from Deleris et al., 2021.

Discussion

Polycomb function has been revealed to extend far beyond its HOX-specific function. From constructing *de novo* 3D genomic architecture during development to controlling cell proliferation and cell identity, the scope of biological functions mediated by this single gene

family is immense [18,19,26]. By understanding the molecular mechanisms behind PcG, we open the door to understanding a myriad of cellular processes.

However, the misconceptions surrounding the evolution of complexity in early-diverging animals, as well as a heavy bias toward bilaterian model organisms, have meant that our progress has been severely constrained. Specifically, the view that bilaterian attributes represent a higher level of functioning and therefore must harness the most interesting and complex regulatory mechanisms has been exceptionally difficult to eliminate. The decades-long focus on vertebrate and insect models shows this, particularly in the unsubstantiated interpretation of vertebrate PRC1 homologs as expanded subunits without knowledge of ancestral equivalents.

The finding that corals and sea anemones contain identical *PCGF* homologs as vertebrates disproves the hypothesis that vertebrate-specific traits evolved through PRC1 expansion. It also offers an opportunity to attain new knowledge concerning PRC1 evolution, which is still poorly understood. For example, one of the specific properties of PCGF2/4 in *Drosophila* is to compact chromatin with the aid of its unique C-terminal region [7]. This function seems to be conserved in PCGF2/4 among many species of invertebrates [74]. Conversely, in vertebrates, the two disparate PRC1 components, EMF1 and CBX2, serve the same function, which intriguingly is also the case in plants [74]. As such, the ancestral function of PRC1 subunits remains unclear.

An interesting question is why PRC1 is less conserved than PRC2. In particular, the marked difference in the number of subunits between PRC1 and PRC2 among animal lineages is stunning and yet the topic remains undiscussed. One possible explanation for the difference in diversification between PRC1 and PRC2 is that the ancestral function of PRC2 may have been silencing transposable elements. This would have put a constraint on functional divergence as mutations in PRC2 are detrimental to its functioning [75]. In modern metazoans, however, H3K9me3 and DNA methylation at the 5th position of cytosine (5mC) are the prevailing silencing mechanisms for transposable elements. Early duplication of PRC1 to generate canonical PRC1 (cPRC1) and noncanonical PRC1 (ncPRC1) allowed further diversification of ncPRC1. A successful duplication of PRC2 never occurred, leaving a single PRC2 till today.

The lack of available genome sequences for early-diverging animals and further lack of functional understanding of PcG function mean our existing knowledge is incomplete and highly biased. It is almost trivial how self-evident the solution appears since the challenge lies in encouraging more curiosity and research in non-standard model organisms to achieve a broad sampling of developmental genes among the branches of the animal family tree. There is, in fact, very little to revolutionize here; much more can be accomplished, and much faster, by performing baseline characterization of PcG complexes, particularly in early-diverging lines, than scrutinizing the molecular details of better-studied animals in order to find out the same information.

It should also be more widely recognized that morphological simplicity does not correspond to regulatory or evolutionary inferiority. For instance, the basic anatomical features of poriferans (sponges) have earned them the position of basal metazoan for a long time. However, recent evidence for the existence of an intricate array of regulatory systems, familiar to us from bilaterian studies, have now been shown to exist in sponges, despite their lack of muscles, nerves, and fully differentiated tissues [74–78]. Natural selection of regulatory

agents determines the level of diversity we see among the organisms in our kingdom, but only insofar as they are different, not inferiorly built. The morphological characteristics that make sponges successful in surviving since their divergence appear simple, but this does not imply anything about their regulatory complexity. As such, Polycomb and other genetic regulatory agents should be studied across the entire animal phylogeny, not just for the most derived clades. The more basal lineages include numerous promising model organisms and studying these will elucidate how regulatory mechanism differ among metazoan taxa.

The diversity of gene regulatory mechanisms in animals reflects their complexity. Understanding the intricacies of these processes is ever compelling, but, as of yet, our attempts have been inadequate in encapsulating the true complexity of developmental processes. The static and often isolated experimental approaches in functional studies of genes are not fully representative of dynamic cellular processes. Determining the persistently changing activities of Polycomb necessitates live, single-cell molecular measurements. This is because Polycomb function changes drastically throughout different points in development, performing different roles during different biological stages. In particular, uncovering the mechanisms where PcG marks are retained through cell divisions and generations constitutes a challenge for future studies. This is important because faithful copying of epigenetic marks is essential for maintaining structural integrity in organisms and preventing malignancy.

These approaches will also be of relevance in comparing the functions of Polycomb to the yet undescribed early-diverging lineages. Especially in the case of deducing evolutionary history, the absence of reference points in early animal phylogeny makes creating a holistic and accurate picture impossible. In a sense, this situation is analogous to trying to solve an algebraic equation with multiple missing variables. As the interest in Polycomb and other epigenetic regulators grows, the coming years will inevitably be filled with rapidly emerging breakthroughs and exciting discoveries illustrating the potency of developmental modulation. With the help of a more complete archive of metazoan PRCs, we may gain crucial insights into one means by which gene regulation creates the diverse phenotypic forms of the animal kingdom.

List of acronyms and abbreviations

CAF1 – Chromatin assembly factor 1 p55; subunit of PRC2
cPRC1 – Canonical polycomb repressive complex 1
CBX2/4/6/7/8 – Chromobox 2/4/6/7/8; subunit of PRC1
E(z) – Enhancer of zeste; subunit of PRC2
E(z)H1/2 – Enhancer of zeste homolog 1/2; subunit of PRC2
EED – Embryonic ectoderm development protein; subunit of PRC2
ESC – Extra sex combs; subunit of PRC2
EZH1/2 – EZ homolog 2; subunit of PRC2
H3K27 – Histone H3 lysine 27
H3K27me – Lysine residue at N-terminal position 27 of the histone H3
HD1 – homology domain 1
HOX – Homeobox; class of homeobox genes
ncPCGF – Noncanonical PCGF
ncPRC1 – Noncanonical polycomb repressive complex 1
PC – Polycomb; subunit of PRC1
PCGF – Polycomb group ring finger proteins; subunit of PRC1

PH(-D) – Plant homeodomain; subunit of PRC1
PHC1/2/3 – Polyhomeotic homolog 1/2/3; subunit of PRC1
PRC1 – Polycomb repressive complex 1
PRC2 – Polycomb repressive complex 2
PRE – Polycomb response element
PSC – Posterior sex combs; subunit of PRC1
PcG – Polycomb group
Ph – Polyhomeotic homolog; subunit of PRC1
RBP2/7 – RNA binding protein; subunit of PRC2
RING1A/B – Really interesting new gene 1A/B; subunit of PRC1
RYBP – Ring1 and YY1 binding protein; subunit of PRC2
SAM – S. adenosyl methionine
SCE – Sex combs extra; subunit of PRC1
SET domain – Su(var)3-9, Enhancer-of-zeste and Trithorax
SU(z)12 – SUZ12; subunit of PRC2

References

1. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
2. Venter, J. C. *et al.* The Sequence of the Human Genome. *Science* **291**, 1304–1351 (2001).
3. Kortschak, R. D., Samuel, G., Saint, R. & Miller, D. J. EST Analysis of the Cnidarian Acropora millepora Reveals Extensive Gene Loss and Rapid Sequence Divergence in the Model Invertebrates. *Current Biology* **13**, 2190–2195 (2003).
4. Finnerty, J. R. & Martindale, M. Q. Ancient origins of axial patterning genes: Hox genes and ParaHox genes in the Cnidaria. *Evolution & development* **1**, 16–23 (1999).
5. Bienz, M. & Müller, J. Transcriptional silencing of homeotic genes in drosophila. *BioEssays* **17**, 775–784 (1995).
6. Biggin, M. D. & Tjian, R. Transcription factors that activate the Ultrabithorax promoter in developmentally staged extracts. *Cell* **53**, 699–711 (1988).
7. Gahan, J. M., Rentzsch, F. & Schnitzler, C. E. The genetic basis for PRC1 complex diversity emerged early in animal evolution. *Proceedings of the National Academy of Sciences* **117**, 22880–22889 (2020).
8. Hauenschild, A., Ringrose, L., Altmutter, C., Paro, R. & Rehmsmeier, M. Evolutionary Plasticity of Polycomb/Trithorax Response Elements in Drosophila Species. *PLoS Biology* **6**, e261 (2008).
9. Yanze, N., Spring, J., Schmidli, C. & Schmid, V. Conservation of Hox/ParaHox-related genes in the early development of a cnidarian. *Developmental biology* **236**, 89–98 (2001).
10. Fortunato, S. A. V *et al.* Calcisponges have a ParaHox gene and dynamic expression of dispersed NK homeobox genes. *Nature* **514**, 620–623 (2014).
11. Schierwater, B. *et al.* The early ANTP gene repertoire: insights from the placozoan genome. *PloS one* **3**, e2457 (2008).
12. Chourrout, D. *et al.* Minimal ProtoHox cluster inferred from bilaterian and cnidarian Hox complements. *Nature* **442**, 684–687 (2006).
13. Jakob, W. *et al.* The Trox-2 Hox/ParaHox gene of Trichoplax (Placozoa) marks an epithelial boundary. *Development genes and evolution* **214**, 170–175 (2004).
14. Moroz, L. L. *et al.* The ctenophore genome and the evolutionary origins of neural systems. *Nature* **510**, 109–114 (2014).

15. Schuettengruber, B. *et al.* Functional Anatomy of Polycomb and Trithorax Chromatin Landscapes in Drosophila Embryos. *PLoS Biology* **7**, e1000013 (2009).
16. Schuettengruber, B., Chourrout, D., Vervoort, M., Leblanc, B. & Cavalli, G. Genome Regulation by Polycomb and Trithorax Proteins. *Cell* **128**, 735–745 (2007).
17. Alvarez-Venegas, R. Regulation by Polycomb and Trithorax Group Proteins in Arabidopsis. *The Arabidopsis Book* **8**, e0128 (2010).
18. Schuettengruber, B. & Cavalli, G. Recruitment of Polycomb group complexes and their role in the dynamic regulation of cell fate choice. *Development* **136**, 3531–3542 (2009).
19. Torosin, N. S., Anand, A., Golla, T. R., Cao, W. & Ellison, C. E. 3D genome evolution and reorganization in the Drosophila melanogaster species group. *PLOS Genetics* **16**, e1009229 (2020).
20. Noordermeer, D. & Duboule, D. Chromatin looping and organization at developmentally regulated gene loci. *WIREs Developmental Biology* **2**, 615–630 (2013).
21. Cao, R. & Zhang, Y. SUZ12 Is Required for Both the Histone Methyltransferase Activity and the Silencing Function of the EED-EZH2 Complex. *Molecular Cell* **15**, 57–67 (2004).
22. Kurzhals, R. L., Tie, F., Stratton, C. A. & Harte, P. J. Drosophila ESC-like can substitute for ESC and becomes required for Polycomb silencing if ESC is absent. *Developmental Biology* **313**, 293–306 (2008).
23. Cao, R. *et al.* Role of Histone H3 Lysine 27 Methylation in Polycomb-Group Silencing. *Science* **298**, 1039–1043 (2002).
24. Song, Y. *et al.* CAF-1 is essential for Drosophila development and involved in the maintenance of epigenetic memory. *Developmental Biology* **311**, 213–222 (2007).
25. Saurin, A. J., Shao, Z., Erdjument-Bromage, H., Tempst, P. & Kingston, R. E. A Drosophila Polycomb group complex includes Zeste and dTAFII proteins. *Nature* **412**, 655–660 (2001).
26. Pachano, T., Crispatzu, G. & Rada-Iglesias, A. Polycomb proteins as organizers of 3D genome architecture in embryonic stem cells. *Briefings in Functional Genomics* **18**, 358–366 (2019).
27. Francis, N. J., Kingston, R. E. & Woodcock, C. L. Chromatin Compaction by a Polycomb Group Protein Complex. *Science* **306**, 1574–1577 (2004).
28. Lo, S. M. & Francis, N. J. Inhibition of Chromatin Remodeling by Polycomb Group Protein Posterior Sex Combs Is Mechanistically Distinct from Nucleosome Binding. *Biochemistry* **49**, 9438–9448 (2010).
29. Conway, E. M. & Bracken, A. P. Unraveling the Roles of Canonical and Noncanonical PRC1 Complexes in *Polycomb Group Proteins* (ed. Pirrotta, V. B. T.-P. G. P.) 57–80 (Elsevier, 2017).
30. van Mierlo, G., Veenstra, G. J. C., Vermeulen, M. & Marks, H. The Complexity of PRC2 Subcomplexes. *Trends in Cell Biology* **29**, 660–671 (2019).
31. Coleman, R. & Struhl, G. Causal role for inheritance of H3K27me3 in maintaining the OFF state of a Drosophila HOX gene. *Science* **356**, eaai8236 (2017).
32. Laprell, F., Finkl, K. & Müller, J. Propagation of Polycomb-repressed chromatin requires sequence-specific recruitment to DNA. *Science* **356**, 85–88 (2017).
33. Gaydos, L. J., Wang, W. & Strome, S. H3K27me and PRC2 transmit a memory of repression across generations and during development. *Science* **345**, 1515–1518 (2014).
34. Abed, J. A. *et al.* De novo recruitment of Polycomb-group proteins in Drosophila embryos. *Development* **145**, (2018).

35. Sowpati, D. T., Ramamoorthy, S. & Mishra, R. K. Expansion of the polycomb system and evolution of complexity. *Mechanisms of Development* **138**, 97–112 (2015).
36. Senthilkumar, R. & Mishra, R. K. Novel motifs distinguish multiple homologues of Polycomb in vertebrates: expansion and diversification of the epigenetic toolkit. *BMC Genomics* **10**, 549 (2009).
37. Bauer, M., Trupke, J. & Ringrose, L. The quest for mammalian Polycomb response elements: are we there yet? *Chromosoma* **125**, 471–496 (2016).
38. Vasanthi, D., Nagabhushan, A., Matharu, N. K. & Mishra, R. K. A functionally conserved Polycomb response element from mouse HoxD complex responds to heterochromatin factors. *Scientific Reports* **3**, 3011 (2013).
39. Connelly, K. E. & Dykhuizen, E. C. Compositional and functional diversity of canonical PRC1 complexes in mammals. *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms* **1860**, 233–245 (2017).
40. Geng, Z. & Gao, Z. Mammalian PRC1 Complexes: Compositional Complexity and Diverse Molecular Mechanisms. *International Journal of Molecular Sciences* **21**, 8594 (2020).
41. Aranda, S., Mas, G. & Di Croce, L. Regulation of gene transcription by Polycomb proteins. *Science Advances* **1**, e1500737 (2015).
42. Pemberton, H. *et al.* Genome-wide co-localization of Polycomb orthologs and their effects on gene expression in human fibroblasts. *Genome Biology* **15**, R23 (2014).
43. Majewski, I. J. *et al.* Opposing roles of polycomb repressive complexes in hematopoietic stem and progenitor cells. *Blood* **116**, 731–739 (2010).
44. Whitcomb, S. J., Basu, A., Allis, C. D. & Bernstein, E. Polycomb Group proteins: an evolutionary perspective. *Trends in Genetics* **23**, 494–502 (2007).
45. Morey, L. *et al.* Nonoverlapping Functions of the Polycomb Group Cbx Family of Proteins in Embryonic Stem Cells. *Cell Stem Cell* **10**, 47–62 (2012).
46. O'Loghlen, A. *et al.* MicroRNA Regulation of Cbx7 Mediates a Switch of Polycomb Orthologs during ESC Differentiation. *Cell Stem Cell* **10**, 33–46 (2012).
47. Hodgson, J. W. *et al.* The polyhomeotic locus of Drosophila melanogaster is transcriptionally and post-transcriptionally regulated during embryogenesis. *Mechanisms of Development* **66**, 69–81 (1997).
48. Dura, J.-M. *et al.* A complex genetic locus, polyhomeotic, is required for segmental specification and epidermal development in D. melanogaster. *Cell* **51**, 829–839 (1987).
49. Gambetta, M. C., Oktaba, K. & Müller, J. Essential Role of the Glycosyltransferase Sxc/Ogt in Polycomb Repression. *Science* **325**, 93–96 (2009).
50. de Napoles, M. *et al.* Polycomb Group Proteins Ring1A/B Link Ubiquitylation of Histone H2A to Heritable Gene Silencing and X Inactivation. *Developmental Cell* **7**, 663–676 (2004).
51. Kuzmichev, A., Nishioka, K., Erdjument-Bromage, H., Tempst, P. & Reinberg, D. Histone methyltransferase activity associated with a human multiprotein complex containing the Enhancer of Zeste protein. *Genes & Development* **16**, 2893–2905 (2002).
52. Cao, R. *et al.* Role of Histone H3 Lysine 27 Methylation in Polycomb-Group Silencing. *Science* **298**, 1039–1043 (2002).
53. Stojic, L. *et al.* Chromatin regulated interchange between polycomb repressive complex 2 (PRC2)-Ezh2 and PRC2-Ezh1 complexes controls myogenin activation in skeletal muscle cells. *Epigenetics & Chromatin* **4**, 16 (2011).
54. Kondrashov, F. A., Rogozin, I. B., Wolf, Y. I. & Koonin, E. V. Selection in the evolution of gene duplications. *Genome biology* **3**, (2002).

55. NEI, M. Gene Duplication and Nucleotide Substitution in Evolution. *Nature* **221**, 40–42 (1969).
56. Sowpati, D. T., Ramamoorthy, S. & Mishra, R. K. Expansion of the polycomb system and evolution of complexity. *Mechanisms of Development* **138**, 97–112 (2015).
57. Blomme, T. *et al.* The gain and loss of genes during 600 million years of vertebrate evolution. *Genome Biology* **7**, R43 (2006).
58. Holland, P. W., Garcia-Fernàndez, J., Williams, N. A. & Sidow, A. Gene duplications and the origins of vertebrate development. *Development (Cambridge, England). Supplement* 125–133 (1994).
59. Kassis, J. A. & Brown, J. L. Polycomb Group Response Elements in Drosophila and Vertebrates. in *Advances in genetics* vol. 81 83–118 (2013).
60. Gustafson, E. A. & Wessel, G. M. Polycomb group gene expression in the sea urchin. *Developmental Dynamics* **237**, 1851–1861 (2008).
61. Brooke, N. M., Garcia-Fernàndez, J. & Holland, P. W. The ParaHox gene cluster is an evolutionary sister of the Hox gene cluster. *Nature* **392**, 920–922 (1998).
62. Larroux, C. *et al.* The NK homeobox gene cluster predates the origin of Hox genes. *Current biology : CB* **17**, 706–710 (2007).
63. Gaiti, F. *et al.* Landscape of histone modifications in a sponge reveals the origin of animal cis-regulatory complexity. *eLife* **6**, e22194 (2017).
64. Simon, J. A. & Kingston, R. E. Mechanisms of Polycomb gene silencing: knowns and unknowns. *Nature Reviews Molecular Cell Biology* **10**, 697–708 (2009).
65. Müller, J. & Kassis, J. A. Polycomb response elements and targeting of Polycomb group proteins in Drosophila. *Current Opinion in Genetics & Development* **16**, 476–484 (2006).
66. Strutt, H. Co-localization of Polycomb protein and GAGA factor on regulatory elements responsible for the maintenance of homeotic gene expression. *The EMBO Journal* **16**, 3621–3632 (1997).
67. Brown, J. L. An Sp1/KLF binding site is important for the activity of a Polycomb group response element from the Drosophila engrailed gene. *Nucleic Acids Research* **33**, 5181–5189 (2005).
68. Fernandez-Valverde, S. L., Calcino, A. D. & Degnan, B. M. Deep developmental transcriptome sequencing uncovers numerous new genes and enhances gene annotation in the sponge Amphimedon queenslandica. *BMC Genomics* **16**, 387 (2015).
69. Fernandez-Valverde, S. L. & Degnan, B. M. Bilaterian-like promoters in the highly compact Amphimedon queenslandica genome. *Scientific Reports* **6**, 22496 (2016).
70. Demuth, J. P., De Bie, T., Stajich, J. E., Cristianini, N. & Hahn, M. W. The Evolution of Mammalian Gene Families. *PLoS ONE* **1**, e85 (2006).
71. Hahn, M. W., De Bie, T., Stajich, J. E., Nguyen, C. & Cristianini, N. Estimating the tempo and mode of gene family evolution from comparative genomic data. *Genome research* **15**, 1153–1160 (2005).
72. Boussau, B., Karlberg, E. O., Frank, A. C., Legault, B.-A. & Andersson, S. G. E. Computational inference of scenarios for -proteobacterial genome evolution. *Proceedings of the National Academy of Sciences* **101**, 9722–9727 (2004).
73. Lagarou, A. *et al.* dKDM2 couples histone H2A ubiquitylation to histone H3 demethylation during Polycomb group silencing. *Genes & Development* **22**, 2799–2810 (2008).
74. Beh, L. Y., Colwell, L. J. & Francis, N. J. A core subunit of Polycomb repressive complex 1 is broadly conserved in function but not primary sequence. *Proceedings of the National Academy of Sciences* **109**, E1063–E1071 (2012).

75. Déléris, A., Berger, F. & Duharcourt, S. Role of Polycomb in the control of transposable elements. *Trends in Genetics* **37**, 882–889 (2021).
76. Larroux, C. *et al.* Developmental expression of transcription factor genes in a demosponge: insights into the origin of metazoan multicellularity. *Evolution & Development* **8**, 150–173 (2006).
77. Conaco, C. *et al.* Transcriptome profiling of the demosponge Amphimedon queenslandica reveals genome-wide events that accompany major life cycle transitions. *BMC Genomics* **13**, 209 (2012).
78. Srivastava, M. *et al.* The Amphimedon queenslandica genome and the evolution of animal complexity. *Nature* **466**, 720–726 (2010).
79. Nakanishi, N., Sogabe, S. & Degnan, B. M. Evolutionary origin of gastrulation: insights from sponge development. *BMC Biology* **12**, 26 (2014).
80. Peterson, K. J. & Sperling, E. A. Poriferan ANTP genes: primitively simple or secondarily reduced? *Evolution & development* **9**, 405–408 (2007).
81. Shaver, S., Casas-Mollano, J. A., Cerny, R. L. & Cerutti, H. Origin of the polycomb repressive complex 2 and gene silencing by an E(z) homolog in the unicellular alga Chlamydomonas. *Epigenetics* **5**, 301–312 (2010).