nature communications

Article

https://doi.org/10.1038/s41467-025-64570-0

Chromatin profiling identifies putative dual roles for H3K27me3 in regulating cell type-specific genes and transposable elements in choanoflagellates

Received: 7 June 2024

Accepted: 19 September 2025

Published online: 29 October 2025

Check for updates

James M. Gahan[1,2,9] ✉, Lily W. Helfrich[3,10], Laura A. Wetzel[3,11], Natarajan V. Bhanu[4], Zuo-Fei Yuan[5], Benjamin A. Garcia[4], Robert J. Klose[2], Alex de Mendoza[6,7] & David S. Booth[1,8] ✉

Chromatin-based mechanisms contribute to the exquisite regulation of gene expression during animal development. But how those mechanisms evolved remains elusive. Here we investigate chromatin regulatory features in the closest relatives of animals, choanoflagellates. In a model choanoflagellate *Salpingoeca rosetta*, we compare chromatin accessibility and histone modifications to gene expression. Accessible genomic regions in *S. rosetta* primarily correspond to gene promoters, and we find no evidence of distal gene regulatory elements that resemble enhancers deployed to regulate developmental genes in animals. Remarkably, the histone modification H3K27me3 decorates genes with cell type-specific expression, revealing a functional similarity in *S. rosetta* and animals. Additionally, H3K27me3 marks LTR retrotransposons, retaining a potential ancestral role in regulating these elements. We further uncover a putative bivalent chromatin state at cell type-specific genes that consists of H3K27me3 and H3K4me1. Together, these data support the emergence of gene-associated histone modification states that underpin development before the evolution of animal multicellularity.

Animal development depends on the ability to precisely regulate gene expression in time and space to specify cell types with different functions as well as the ability to maintain those cell identities after differentiation. Although the mechanisms underlying these processes have been extensively studied within animals[1,2], their evolution is less clear due to the lack of information in the closest relatives of animals, the unicellular holozoans. A mechanistic account of gene regulation from diverse representatives within these groups will help clarify how animal developmental gene regulation evolved, which aspects of this process predate animals, and which emerged concomitant with animal evolution.

The regulation of genes during animal development occurs at all stages of gene expression with transcriptional regulation primarily

[1]Department of Biochemistry and Biophysics, University of California, San Francisco, San Francisco, CA, USA. [2]Department of Biochemistry, University of Oxford, Oxford, UK. [3]Department of Molecular and Cell Biology, Howard Hughes Medical Institute, University of California, Berkeley, Berkeley, CA, USA. [4]Department of Biochemistry and Molecular Biophysics, Washington University School of Medicine, St Louis, MO, USA. [5]Center for Proteomics and Metabolomics, St. Jude Children's Research Hospital, Memphis, TN, USA. [6]School of Biological and Behavioural Sciences, Queen Mary University of London, London, UK. [7]Centre for Epigenetics, Queen Mary University of London, London, UK. [8]Chan Zuckerberg Biohub, San Francisco, CA, USA. [9]Present address: Centre for Chromosome Biology, School of Biological and Chemical Sciences, University of Galway, Galway, Ireland. [10]Present address: Benchling, San Francisco, CA, USA. [11]Present address: BioMarin Pharmaceutical Inc, San Francisco, CA, USA. ✉e-mail: james.gahan@universityofgalway.ie; David.Booth@ucsf.edu

driving differences between cell types. At the level of DNA sequence, animals not only rely on promoter proximal cis-regulatory sequences to regulate transcription but also utilize distal elements called enhancers[3,4]. Distal enhancers are cis-acting regulatory elements that can act over long distances (at least in bilaterians) to regulate target gene expression[4]. Distal enhancers have been identified in most major animal groups and are likely an ancestral component of the animal gene regulation toolkit[5–9]. Studies on the filasterean *Capsaspora owczarzaki*, however, did not identify any major contribution of distal enhancer-like gene regulation in this group, leading to the hypothesis that distal regulation emerged in the animal lineage[10,11] work which was supported by the absence of long-range chromatin loops in close animal relatives[8]. Whether such elements truly emerged only in animals, however, is still not clear as more sampling within unicellular holozoans is required. In addition, a clear definition of enhancers and what would constitute long-range activity in early branching metazoans and unicellular holozoans is lacking as is clear functional evidence for enhancer activity (reviewed in ref. 12).

The regulation of chromatin and in particular histone post-translational modifications (hPTMs) is also key to animal development[13–17]. hPTMs are used in a variety of ways to regulate the expression of genes[17,18]. hPTMs have generally conserved genomic distributions across eukaryotes, particularly at active genes, while repressive states seem to vary more between species[19]. Histone H3 lysine 4 tri-methylation and lysine 27 acetylation (H3K4me3 and H3K27ac), for example, are found at active genomic regions like promoters or active enhancers in diverse organisms[20–25]. Other modifications like H3K4me1 have varying distributions. In animals H3K4me1 is associated with enhancer activity and the relative levels of H3K4me3 and H3K4me1 have been used to distinguish promoters versus enhancers[26,27]. In *Arabidopsis thaliana*, H3K4me1 is found on the bodies of active genes[20] while in the green algae, *Chlamydomonas reinhardtii*, H3K4me1 is broadly distributed throughout the genome but excluded from active regions[28] and may play a role in silencing[29].

While the capacity to turn genes on when they are required is central to cell-type specification, it is equally important to ensure that genes which should be inactive are maintained in this state and the ability to stably repress genes is essential for cell differentiation. Two prominent hPTMs associated with repression are methylation of H3K9 and H3K27[30]. H3K27me3 is deposited by Polycomb Repressive Complex 2 (PRC2)[31,32] and is a key regulator of developmental genes in animals[32–35]. Interestingly, Polycomb complexes play analogous roles in plants but this may have evolved independently as there are several mechanistic differences and many of the key players in animals are not found in plants[36–39]. Additionally, recent work found a role for PRC2/H3K27me3 in transposable element repression in diverse eukaryotes[40–54] leading to the hypothesis that this is an ancestral role for PRC2[42]. Despite their importance, repressive PTMs have thus far not been studied in unicellular holozoans beyond describing their presence/absence[10,19,55]. In the filasterean *Capsaspora owczarzaki and i*chthyosporean *Creolimax fragrantissima*, where active PTMs are well characterized, both the H3K9 and H3K27 methylation machinery have been lost[10,19,55].

The choanoflagellates are the sister group to animals and therefore a key group for understanding early animal evolution[56–58] (Fig. 1a). Choanoflagellates are a genetically diverse clade of unicellular heterotrophs which are found in virtually all aquatic environments[59]. The species *Salpingoeca rosetta* has emerged in recent years as a leading research organism due to several factors[56,60]. Firstly, *S. rosetta* has a dynamic life history with numerous cell types[61–63] (Fig. 1b). These include two swimming cell types: Slow swimmers which are the default state in high nutrient conditions and fast swimmers which are a dispersal stage induced in low nutrients. Their life history also includes a facultative multicellular state known as a rosette colony which is induced via specific environmental cues[64,65]. Recent work has shown that the formation of rosettes does not involve major transcriptional changes[66]. Instead, however, another cell type called thecate cells are very transcriptionally distinct from all swimming cell types providing a model for cell type-specific gene regulation. Thecate cells are physically attached to the substrate and are likely the diploid progeny of mating between haploid swimming cells[67,68]. The *S. rosetta* genome has been sequenced[62,69] and the availability of functional tools including transfection[70] and CRISPR-Cas9 genome editing[71,72] make it a powerful system to dissect pre-metazoan gene-regulatory capabilities.

Despite their important phylogenetic position, there is currently little knowledge on gene regulation or chromatin in any choanoflagellate species. The role of one transcription factor, cRFX, in ciliogeneis has been reported[73]. To overcome this, we have mapped chromatin accessibility in *S. rosetta* using ATAC-seq and show that the majority of regulatory elements are in close proximity to transcriptional start sites (TSS). We have also cataloged the complement of hPTMs and described the genome-wide localization of several key hPTMs. This revealed a conserved pattern of hPTMs surrounding active genes while inactive cell type-specific genes are occupied by H3K27me3-marked nucleosomes. H3K27me3 is also found on a subset of long terminal repeat (LTR) retrotransposons indicating a dual role for this modification in both repression of transposable elements and cell type-specific gene regulation. Finally, we show a putative bivalent state demarcating cell type specific genes when they are repressed.

Results

Gene regulation predominantly relies on promoter-proximal elements in *S. rosetta*

In all eukaryotes, the core promoter of genes drives gene expression and can mediate complex environmental sensing and differentiation[2,3]. Distal elements, like enhancers that impact transcriptional regulation of a gene add another layer of transcriptional control in animals. To understand what features of the cis-regulatory landscape in *S. rosetta* are shared with animals, their closest relatives, or other holozoans that lack distal enhancers, we employed ATAC-seq, a method which utilizes differential accessibility to Tn5 transposase to map accessible chromatin regions genome-wide across cell types[74] (Supplementary Fig. 1a, b). Visualization of mapped reads showed clear peaks throughout the genome and high enrichment at transcription start sites (TSSs) (Fig. 2a, b and Supplementary Fig. 1c). ATAC-seq signal showed a clear correlation between TSS enrichment and transcript levels (Fig. 2c). The majority of called peaks were in promoter regions or overlapped with predicted TSSs with a very small number of peaks distal to the TSS (Supplementary Fig. 1d/e) which was also evident from manual inspection of the reads (Fig. 2a). A more detailed analysis revealed that ~75% of peaks directly overlapped a predicted TSS while more than 80% had their midpoint with a distance of −500 to +100 bps from a predicted TSS (Fig. 2d). In addition, manual inspection of peaks that

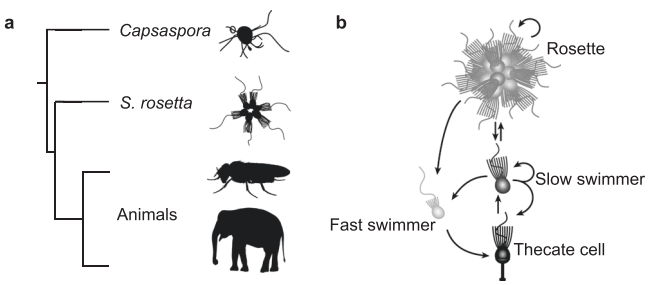

Fig. 1 | *Salpingoeca rosetta* as a model for pre-animal cell differentiation. **a** Phylogenetic tree highlighting the position of choanoflagellates (e.g., *S. rosetta*) as the sister group to animals. The filastereans (e.g., *Capsaspora owczarzaki*) are the sister group to animals and choanoflagellates. Silhouettes are from http://phylopic.org/ under a CC0 1.0 license **b** The life history of *S. rosetta* consists of several distinct cell types representing different life history stages including solitary swimming cells (both slow swimmers and fast swimmers), a facultative multicellular stage called a rosette colony and a substrate-attached cell type called a thecate cell.

https://doi.org/10.1038/s41467-025-64570-0

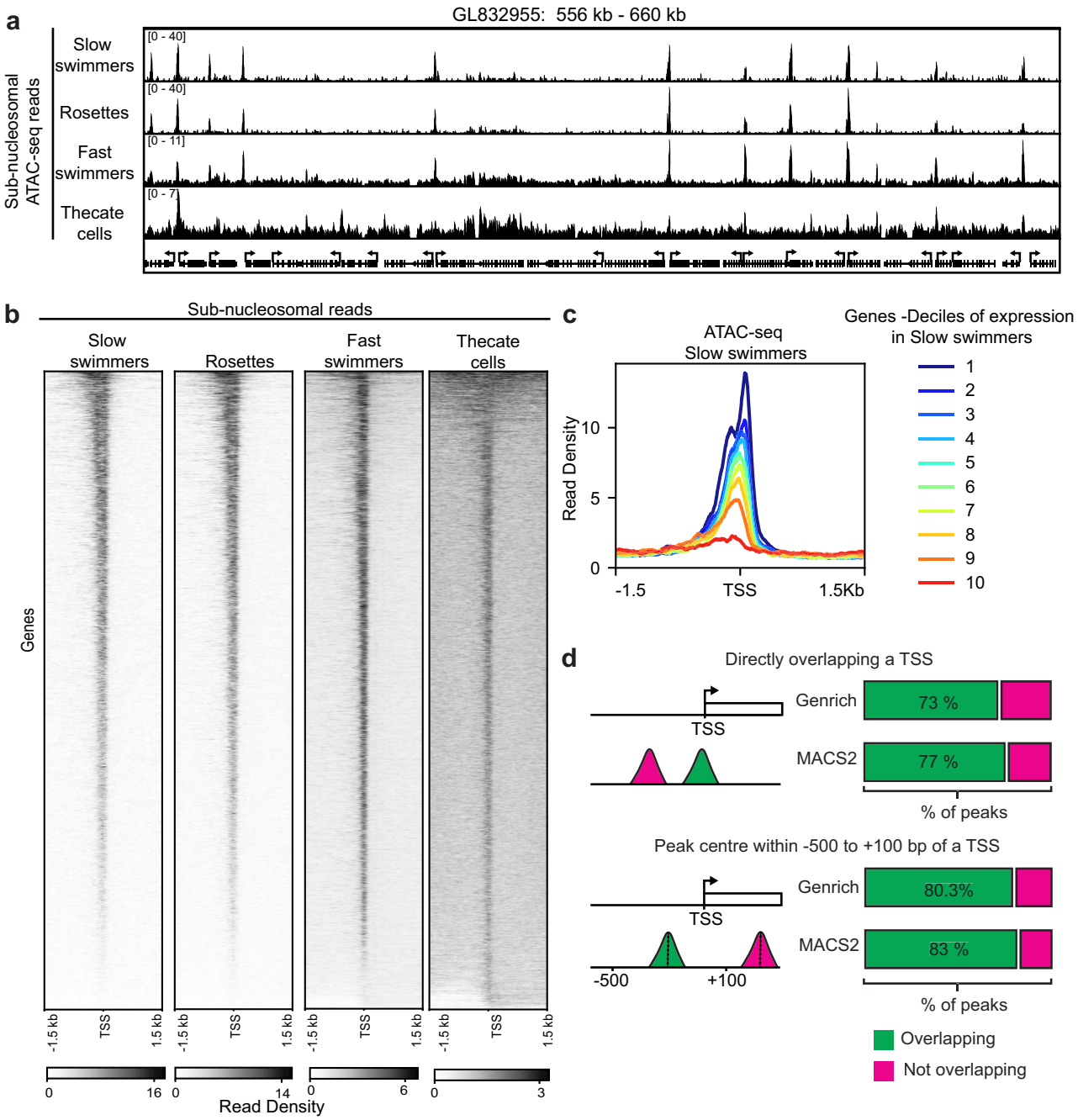

Fig. 2 | ATAC-seq on different cell types in *S. rosetta*. **a** Genome browser snapshot of ATAC-seq in the different *S. rosetta* cell types. Only sub-nucleosomal reads are shown. The cell type is shown on the left and the contig and region are shown on top. Annotated genes are shown along the bottom. **b** Heatmap of sub-nucleosomal ATAC-seq read density surrounding the TSS of all annotated genes ($n = 11{,}731$). Cell type is shown on top. **c** Metaplot of ATAC-seq reads from slow swimmers surrounding the TSS for genes divided into ten deciles based on RNA expression (i.e., RPKM values) in slow swimmers with 1 being the highest expressed and 10 the lowest. **d** Quantification of peaks which either overlap with predicted TSSs or with a midpoint located in a −500 to +100 bp window surrounding predicted TSSs. Peaks were called with either MACS2 or Genrich.

were predicted to be outside of these ranges, i.e., putative distal peaks, showed that a large fraction correspond to un-annotated TSSs rather than true distal peaks (Supplementary Fig. 2). Taken together, these data show that most cis-regulatory elements in *S. rosetta* are TSS-proximal with few distal regulatory elements.

Genome-wide profiling of hPTMs reveals signatures for active and repressed genes

Given the prominent role of hPTMs in animal developmental gene regulation we next sought to understand the hPTM landscape in *S. rosetta*. Having defined the histones present (Supplementary Fig. 3a, b and Supplementary Table 1) we performed quantitative histone mass-spectrometry to identify hPTMs in samples extracted from both slow swimmers and thecate cells. We identified methylated and acetylated lysine residues in SrH3.1 and SrH4 (Supplementary Fig. 3c). *S. rosetta* has all the common histone lysine methylation and acetylation sites seen in other eukaryotes, e.g., methylation of H3K4 and H3K36 and acetylation of H3K27 and H4K16. Importantly, we also identified both H3K9 and H3K27 methylation indicating *S. rosetta* likely has both classical facultative and constitutive heterochromatin (Supplementary

https://doi.org/10.1038/s41467-025-64570-0

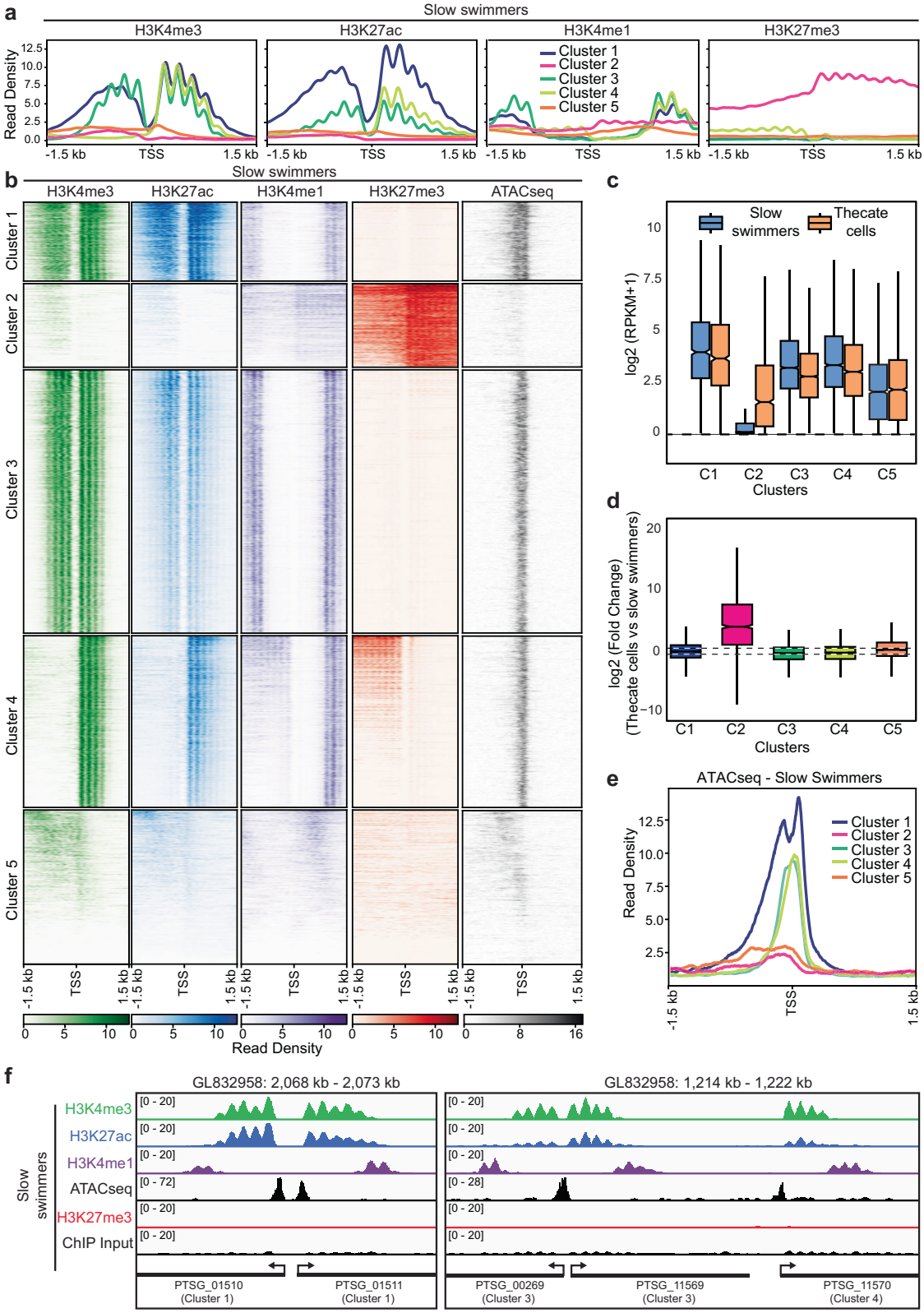

Fig. 3c). We next mapped the genome-wide localization of 4 modifications: H3K4me3 and H3K27ac, two hPTMs present at active regulatory elements genome-wide across eukaryotes; H3K4me1 which has variable distributions in different organisms but is generally present at both promoters and enhancers in animals; and the PRC2-dependent modification H3K27me3. We optimized a native chromatin immunoprecipitation with sequencing (ChIP-seq) protocol using micrococcal nuclease (MNase) digested chromatin and performed ChIP-seq for the 4 hPTMs in duplicate in slow swimmers (Supplementary Fig. 4). K-means clustering of the distribution of hPTMs around the TSS of all genes identified 5 broad groups of genes (Fig. 3a, b). Three clusters (Clusters 1,3 and 4) are associated with high levels of H3K4me1/3 and H3K27ac. A fourth cluster (Cluster 2) is depleted of these PTMs but was characterized by high levels of H3K27me3. The final cluster consists of

Fig. 3 | ChIP-seq of hPTMs reveals chromatin states of active and repressed genes. a Metaplot analysis of ChIP-seq on slow swimmers with the indicated antibodies for each cluster of genes. Clustering was performed by k-means clustering. **b** Heatmap of data shown in A along with ATAC-seq data from slow swimmers. **c** Box plot showing the $\log_2$ transformed RPKM values (+1) for genes in each cluster in slow swimmers and thecate cells. **d** Box plot showing the $\log_2$ fold change between slow swimmers and thecate cells for genes in each cluster. The dashed lines are at +1 and −1. The boxes show interquartile range, center line represents median, whiskers extend by 1.5× interquartile range (IQR) or the most extreme point (whichever is closer to the median), while notches extend by 1.58× IQR/sqrt(n), giving a roughly 95% confidence interval for comparing medians. ($n = 3$ independent biological replicates) **e** Metaplot analysis showing ATAC-seq signal from slow swimmers surrounding the TSS for genes from each cluster. **f** Genome browser snapshots of genes representative of Clusters 1,3 and 4 showing ChIP-seq with the indicated antibodies along with input and ATAC-seq in slow swimmers. The genes are shown on the bottom as well as which cluster they belong to. The genomic scaffolds and positions are shown on top. Annotated genes are shown along the bottom. Cluster 1 ($n = 1232$), Cluster 2 ($n = 1305$), Cluster 3 ($n = 4112$), Cluster 4 ($n = 2681$), Cluster 5 ($n = 2401$).

genes with no obvious pattern (Cluster 5) which we excluded from further analysis. We then looked at the expression of genes from each cluster in both slow swimmers and thecate cells. We focus on thecate cell because they are very transcriptionally different from slow swimmers, while fast swimmers and rosettes are not[66]. Genes from Clusters 1, 3 and 4 are expressed in slow swimmers while Cluster 2 genes show lower levels of expression (Fig. 3c). Interestingly, Cluster 2 contains genes that are upregulated in thecate cells while genes from the other clusters generally do not change in expression between cell types (Fig. 3c, d). Finally, analysis of ATAC-seq reads shows that genes in Clusters 1, 3, and 4 have an open, nucleosome-free region surrounding their TSSs while Cluster 2 genes do not (Fig. 3b, e). Inspection of genes from Clusters 1, 3 and 4 revealed the major defining difference to be orientation within the genome (Fig. 3f). All genes in Clusters 1 and 3 are oriented with another divergent active TSS located directly upstream. In each case, overlapping domains of H3K4me1/3 and H3K27ac are found flanking a nucleosome-free region at the TSS. In the case of Cluster 4 genes these modifications are only seen downstream of the TSS but this may reflect the compact nature of the genome as an actively transcribed or H3K27me3-marked gene are always directly upstream (Fig. 3a, b, f). Taken together this shows that hPTMs differentially demarcate active and inactive genes in *S. rosetta*.

H3K27me3 and H3K4me1 co-occur on cell type-specific genes when they are repressed

Given the lack of knowledge on Polycomb-mediated repression in unicellular holozoans and its prominent role in animal developmental gene regulation, we decided to focus more closely on the Cluster 2 genes which are strongly associated with the PRC2-mark H3K27me3. Looking at the expression of Cluster 2 genes we see that ~75% are upregulated in thecate cells (Fig. 4a) but they only represent a small proportion of all thecate-upregulated genes (Supplementary Fig. 5a). To understand what distinguished the Cluster2/upregulated genes from the other upregulated genes we first looked at the level of upregulation and see that the Cluster2 genes are generally very highly upregulated in thecate cells compared to others (Supplementary Fig. 5b, c) and represent 86% of the top 500 upregulated genes (Fig. 4b). Further, in slow swimmers, when these genes are H3K27me3 marked, we see that they are very lowly expressed or not expressed at all (Fig. 4c). Taken together, this shows that H3K27me3 marks a population of genes that are regulated in a cell-type-specific manner. We then looked closer at the pattern of hPTMs on these genes in slow swimmers and looked at all genes as well as specific examples, e.g., PTSG_09715, a gene recently validated as being thecate-specific[66] (Fig. 4d, e). As predicted, they were marked by H3K27me3, which was the defining feature of Cluster 2. There were, however, two unexpected aspects of their hPTM signature. Firstly, H3K27me3 is very specific to the gene body of these genes rather than being localized on their promoters. In some cases, we saw H3K27me3 upstream of the TSS, but clustering of these genes revealed the presence of another H3K27me3-marked gene directly upstream (Supplementary Fig. 5a, d). Secondly, H3K4me1 co-occurred with H3K27me3 on these repressed genes. Together, these hPTM patterns show that cell type-specific genes in *S. rosetta* are marked both by H3K27me3 and by H3K4me1 when repressed, a situation potentially analogous to bivalent chromatin in animals.

H3K27me3 labels a subset of LTR retrotransposons

Given the emerging roles of H3K27me3 in transposable element repression in diverse eukaryotes[42], we wondered whether such a role may also be present in *S. rosetta*. A visual inspection revealed high levels of H3K27me3 at the end of many of the super-contigs in the genome. Given that some of the super-contigs are in fact full chromosomes[69] we investigated if H3K27me3 deposition occurred at sub-telomeric regions that were discernable by telomeric repeats within super-contigs[62]. Indeed, H3K27me3 levels were high in regions adjacent to the telomeric repeats (Supplementary Fig. 6). Additionally, we noticed that some H3K27me3-marked regions overlapped areas resembling transposable elements. To further examine this, we utilized RepeatMasker[75] to annotate the genome using a previously published transposable element annotation[76]. We then looked at hPTMs and ATAC-seq on LTR retrotransposons and DNA transposons (Fig. 5 and Supplementary Fig. 7). This analysis showed that many LTR retrotransposons exhibited high levels of H3K27me3 (Fig. 5). As TEs are repetitive sequences that show a high degree of sequence conservation we also filtered ChIP-seq reads to include only those with a high mapping score, i.e., to control for multi-mapping on highly similar copies (Supplementary Fig. 7). Although less signal was present on the TEs themselves, we still saw high levels surrounding the same families of LTR/retrotransposons as would be expected for large families with multiple similar copies (Supplementary Fig. 7). We then looked at individual copies and compared ChIP-seq data to RNA-seq data from multiple cell types (Supplementary Fig. 8). In general, the LTR retrotransposons are not expressed but, in some cases, we do see high expression across cell types (e.g., a subset of copia-like elements (Srosspv6)). In this case, H3K27me3 levels are lower than over less-expressed families (e.g., Srosspv2/3) (Supplementary Fig. 8). Together, this suggests that H3K27me3 may be acting to repress a subset of TEs in *S.rosetta*. In contrast, DNA transposons generally have higher levels of expression and low levels of H3K27me3 (Fig. 5 and Supplementary Fig. 7/9). Some families of DNA transposons like SrorTig1 and SrosTm showed consistent ATAC-seq signal on both the 5′ and 3′ ends in most insertions as well as enrichment of active histone PTMs, implying that these DNA transposons evade direct epigenetic silencing in *S. rosetta*. Taken together, these data show that H3K27me3 decorates a subset of transposable elements in the *S. rosetta* genome with an inverse correlation with RNA expression.

Discussion

Here we present the an analysis of chromatin accessibility and hPTMs in choanoflagellates and present findings that inform the origins of animal gene regulation. We observe no clear evidence for distal enhancers like those found in animals (Fig. 2). Together with previous work[8,10], these data suggest that animals elaborated on core chromatin and transcription machinery from their holozoan ancestor to regulate gene expression with distal cis-regulatory elements. More mechanistic work will uncover the evolutionary changes that enabled distal enhancers to become a widespread feature of

https://doi.org/10.1038/s41467-025-64570-0

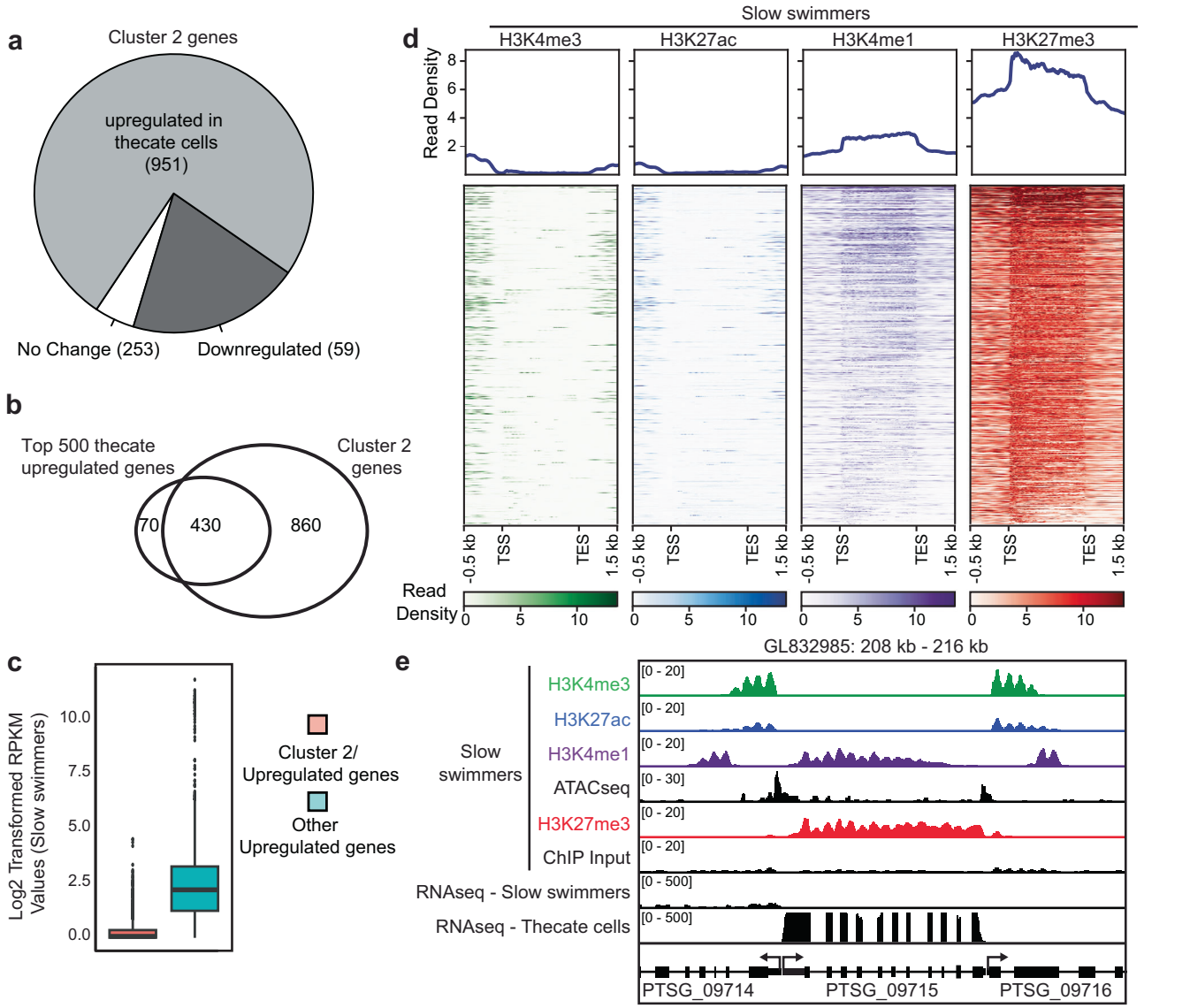

Fig. 4 | Thecate-specific genes are marked by H3K27me3 and H3K4me1 when silent. a Pie chart showing Cluster 2 genes and whether they are upregulated, downregulated on unchanged between thecate cells and slow swimmers. Log_2 fold change of ±1, as determined using DESeq2, was considered as a significant change. **b** Venn diagram showing the overlap between genes from Cluster 2 and the top 500 genes upregulated in thecate cells vs slow swimmers. **c** Box plot showing log_2 transformed RPKM values (+1) from RNA-seq data from slow swimmers ($n = 3$ independent biological replicates). Gene upregulated in thecate cells are shown and separated based on whether they overlap with Cluster 2 genes or not. The boxes show interquartile range, center line represents median, whiskers extend by 1.5× IQR or the most extreme point (whichever is closer to the median) and dots show outliers. **d** Heatmap and metaplot showing ChIP-seq in slow swimmers with the indicated antibodies over Cluster 2/thecate upregulated genes ($n = 951$). **e** Genome browser snapshot of a gene representative of Cluster 2 showing ChIP-seq with the indicated antibodies along with input, ATAC-seq in slow swimmers and RNA-seq in both slow swimmers and thecate cells. The genomic scaffolds and positions are shown on top. Annotated genes are shown along the bottom.

developmental regulation in animals. We cannot exclude the possibility that *S. rosetta* may have some features of enhancers that are difficult to disentangle in a condensed genome with a mean intergenic distance of 885 bp[12,77]. For example, the core promoter from one gene may influence the expression of neighboring genes, similar to the original discovery of enhancers whereby an SV40 promoter enhanced the transcription of a distant beta-globin gene[78]. It could, therefore, be the case that *S. rosetta* promoters not only act on their proximal genes but also as enhancers of more distal genes. One alternative cause may be the small size of the genome (~55 Mb)[62]. The similarly compact genome of *C. owczarzaki*[8,10,79] may also explain its lack of distal-regulatory regions. Animals that have undergone massive genome size reductions sometimes have limited intergenic enhancers but often retain distal enhancer-type elements within introns[80–82]. We do not see any evidence for such intronic regulatory sequences, arguing against a loss of enhancers during a reduction of the genome size. Studying other unicellular holozoan lineages with larger genome sizes and/or larger intergenic distances, such as *Amoebidium appalachense* (~200 Mb)[83], will help resolve these issues but work in *Sphaeroforma arctica* (~145 Mb)[84] showed no evidence for chromatin looping or distal regulatory elements. Notably, transcriptomes from diverse choanoflagellate species suggest that some have a higher degree of gene conservation than the two model species, *S. rosetta* and *Monosiga brevicollis*, that have sequenced genomes[59]. Some of these species have already revealed surprises such as the presence of POU and Sox transcription factors, which were originally thought to be absent in choanoflagellates[85]. Additional genomes may therefore reveal other features of choanoflagellate chromatin biology and help unravel the origin and evolution of animal chromatin-based gene regulation.

https://doi.org/10.1038/s41467-025-64570-0

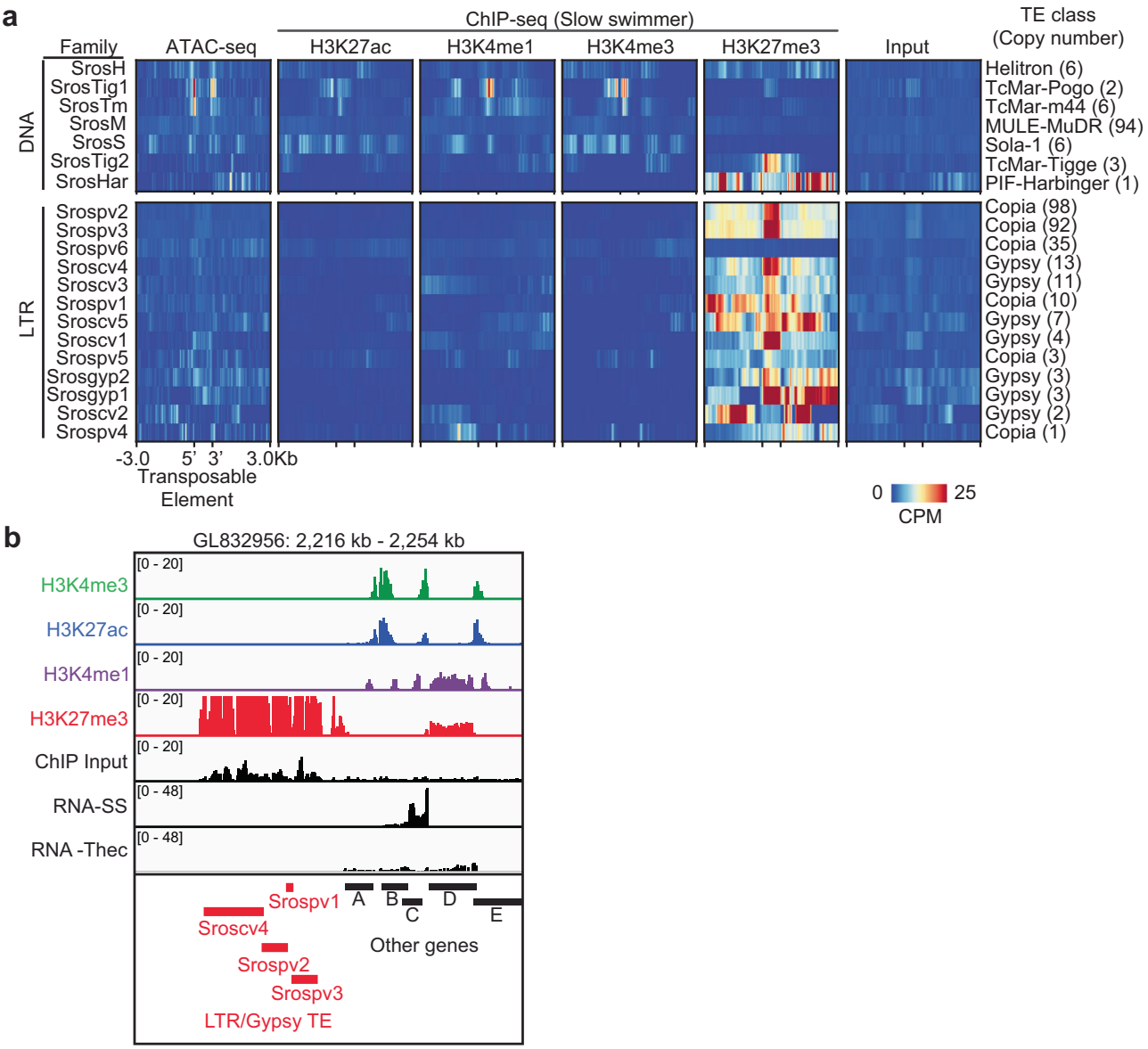

Fig. 5 | H3K27me3 is enriched over LTR retrotransposons. a Heatmap showing average ATAC-seq and ChIP-seq in slow swimmers over annotated *Salpingoeca rosetta* transposable element families[76], divided by Class. Antibodies used for ChIP-seq are indicated on top. The TE family is shown on the left and the Class and number of copies of each are shown on the right, heatmap scale bottom right. Only inserts spanning at least 50% of the consensus sequence were used. **b** Genome browser snapshot showing an example of some LTR/Gypsy retrotransposons. ChIP-seq data is shown for the indicated antibodies as well as input and RNA-seq in both slow swimmers (SS) and thecate cells (Thec). The genomic scaffolds and positions are shown on top. Annotated transposable elements and the family they belong to as well as other genes are shown along the bottom. A = PTSG_01036, B = PTSG_01037, C = PTSG_01038, D = PTSG_11676, E = PTSG_11677.

The ability to stably repress genes in a cell type-specific manner is a crucial component of animal developmental gene regulation. This is achieved partly by the Polycomb-repressive system. Here, we present evidence for a putative dual role for PRC2 and its modification H3K27me3 in the regulation of transposable elements and cell type-specific genes in *S. rosetta*.

Recent work has highlighted the previously underappreciated role of PRC2 in regulating transposable elements across diverse eukaryotes leading to a hypothesis that this role may have descended from the last eukaryotic common ancestor[42,48]. Our data suggest this function was retained in *S. rosetta* for silencing retrotransposons (Fig. 5). Yet, we also show that H3K27me3 is largely absent on DNA transposons. In the case of some DNA transposons, we see both ATAC-seq signal and active hPTMs present along with RNA-seq reads, potentially indicating they are active. It is possible that H3K9me3 may silence these TEs in *S. rosetta* as in diverse eukaryotes[42], but we were unable to profile H3K9me3 in *S. rosetta*. Commercial antibodies for H3K9me3 failed in our hands–perhaps due to a difference of the histone H3 of *S. rosetta* in which the tenth position is a threonine rather than the serine that is present in the vast majority of species. This experimental hurdle and the importance of H3K9me3 for TE silencing in diverse eukaryotes compels future characterization of histone modifications in *S. rosetta*. We do find enrichment for H3K27me3 in the sub-telomeric regions on the few contigs where telomeres are annotated, but in these cases, we do not see any transposable elements in these regions. Unfortunately, most super-contigs in the current genome sequence do not extend all the way to the telomere yet those telomeres do share a common sub-telomeric sequence[62]. It may therefore be that they have an accumulation of transposable elements. Future improvements in the genome assembly will reveal if all chromosomes have sub-telomeric H3K27me3, as is the case in other species like *Cryptococcus neoformans*[86], and its possible functions.

The potential role of H3K27me3 in regulating cell type-specific genes is more complex. We show that H3K27me3 marks genes in a pattern consistent with a similar role in animals, i.e., repressing cell-type specific gene (Fig. 4). This function of H3K27me3 may have been independently co-opted in the *S. rosetta* lineage, or both animals and *S. rosetta* retained this function from their last common ancestor. The mechanisms of PRC2 recruitment to chromatin are currently unknown in *S. rosetta*; therefore, further studies of those molecular mechanisms will better facilitate comparisons with animals to potentially resolve these evolutionary scenarios. Furthermore, overcoming technical challenges to characterize the distribution of H3K27me3 in thecate cells will aid in establishing the mechanism of H3K27me3 recruitment and function.

The presence of H3K27me3 over gene bodies could suggest an antagonistic role between transcription (or transcriptional coupled hPTMs like H3K36 methylation) as has been shown in other scenarios[87,88]. Gene-body H3K27me3 also anticorrelates with levels of expression in the unicellular algae *Cyanidioschizon merolae* and the diatom *Phaeodactylum tricornutum*[41,89]. In addition, the role of the other major Polycomb-repressive complex, PRC1, is not known in *S. rosetta* and is, in fact, not well understood in any system outside of plants and bilaterian animals. Choanoflagellates do possess a variant PRC1 complex, which is the evolutionary precursor of animal PRC1 complexes[39,90]. PRC1 is responsible for H2A monoubiquitylation but we were unable to determine if this modification is present in *S. rosetta* from our mass-spectrometry data. Future work on the role of this complex will be necessary to determine whether it co-operates with PRC2 in the regulation of transposable elements and/or cell type-specific genes in *S. rosetta*.

In addition to being marked by H3K27me3, cell-type-specific genes are also marked by H3K4me1 when repressed (Figs. 3 and 4). H3K4me1 may therefore some roles at repressed genes in addition to its canonical roles at active genes. This situation resembles "bivalent" chromatin domains in animals which are regions marked by both H3K4me3 and H3K27me3[91,92]. There are, however, several key differences. The bivalent state was first identified in embryonic stem cells[92,93] and later in other animal species[94,95] and is mostly associated with the co-occurrence of H3K27me3 and H3K4me3 on promoters[92,96–98], not H3K4me1. Some regions are, however, marked by H3K27me3 and H3K4me1 such as poised enhancers[91,98–103]. In *S. rosetta*, H3K27me3 and H3K4me1 co-occur over gene bodies rather than at regulatory elements (Fig. 4). Although bivalent chromatin was originally believed to prime genes for rapid activation upon differentiation, more recent studies suggest that H3K4me3 at bivalent regions prevents DNA methylation to prevent irreversible silencing[104–107]. It is possible that H3K4me1 in *S. rosetta* also prevents further repression of cell type-specific genes and bookmarks them for later activation. It is highly unlikely that this occurs through DNA methylation because that machinery is absent in *S. rosetta*[83], so H3K4me1 may prevent accumulation of other chromatin modifications. Notably, a similar type of bivalent chromatin was recently described in the plant *Brassica napus* where it is associated with tissue-specific genes[108]. Alternatively, H3K4me1 may in fact be required for repression of cell-type specific genes in *S. rosetta*. Indeed, there are several instances where H3K4me1 has been implicated in silencing (e.g., in *Chlamydomonas reinhardtii)*[29], but there is no direct functional evidence for this. Finally, in *C. owczarzaki* repressed genes have atypical hPTM signatures[10]. In that organism, a subset of genes display gene-body H3K27ac and a peak of H3K4me1 downstream of the TSS. Although this markedly differs from *S. rosetta*, it may indicate a shared silencing role for H3K4me1 at some genes. However, *Capsaspora* lacks H3K27me3 and H3K9me3 and may therefore have evolved its own specific mechanisms for repression. Further studies will be needed to elucidate the functional relevance of this putative bivalent state and how it may relate to bivalent promoters and/or poised enhancers in animals.

In conclusion, our data show a putative dual role for H3K27me3 in repressing cell-type specific genes and a subset of transposable elements. Future work will unravel the mechanisms of action of this modification at both these locations and further refine whether these represent ancestral features shared with animals or whether independent co-option may have led to the observed similarities. We further uncover a putative bivalent state at cell-type specific genes which suggests a form of priming of gene expression. Functional work will be needed to test this hypothesis and, if true, understand the mechanisms underlying the phenomenon.

Methods

Choanoflagellate cultures and media preparation

All experiments were performed using *Salpingoeca rosetta* in co-culture with a single bacterial food source, *Echinicola pacifica* (ATCC PRA-390, strain designation: SrEpac)[67,109].

Artificial sea water – Keller formula (AKSW), high nutrient media (HN) and cereal grass media (CG) were prepared as previously described[67,70,109].

ATSM D 1141 Artificial Seawater (hereafter ASW) was sourced commercially (ASTM D 1141, Ricca Cat No. 8363-5). RA media was prepared from *Porphyra umbilicalis* as previously described[65]. Briefly 10 g of dried algae was added to 1 L of ASW and incubated mixing for 1 h at room temperature. This was then sterile filtered and RA media was prepared by diluting this 1:4 in ASW along and 1:1000 dilutions of 1000× (Potassium Iodide, Sodium Nitrate, Sodium Phosphate), 1000× L1 vitamins[110], and 1,000x L1 trace metals[110].

ATAC-seq

Nuclei were isolated from different cell types: (1) slow swimmers, (2) rosettes, (3) fast swimmers, and (4) thecate cells with two independent replicates. Cultures containing single cell types were produced as described previously[66,73]. Slow swimmers were generated by maintaining cells in HN media. Rosettes were induced with outer membrane vesicles from *Algoriphagus machipongonensis* as previously described[111]. Fast swimmers were grown 3 days in HN and then heat shocked at 30 °C for 2.75 h. Slow swimmers, rosettes, and fast swimmers were harvested by pelleting at 2400 × *g* for 5 min, washed with 50 mL AKSW, re-pelleted at 2400 × *g* for 5 min, counted with the Luna cell counter (Logos Biosystems), diluted to 50 million cells/mL, and 10 million cells were pelleted at 2700 × *g*. Thecate cells were derived from an isolate of SrEpac, called HD1[67], and maintained in 10% CG in AKSW (vol/vol) in petri dishes. To harvest thecate cells, plates were washed with 16.7 mL of AKSW, cells lifted from the plate with a cell scraper and filtered onto a 3 µm polycarbonate membrane filter to concentrate. Filtered cells were pelleted at 2700 × *g* for 5 min, washed with 50 mL AKSW two times, re-pelleted at 2700 × *g* for 5 min, counted with Luna cell counter, diluted to 50 million cells/mL, and 10 million cells were pelleted at 2700 × *g*. All cell types were resuspended in 200 µL freshly prepared pretreatment buffer (10 mM citric acid, 100 mM Lithium Acetate, 10% (w/v) PEG 8000 pH 8.5 with Tris, 100 nM papain, and 10 mM thioglycolic acid) and incubated at room temperature for 22 min. Nuclei were isolated in four steps: wash, strip, lyse, and purify. To wash, cells were pelleted at 1200 × *g* for 5 min, the supernatant discarded, and resuspended in 200 µL of 0.7 M sorbitol in 1x PBS and 1% (w/v) BSA, and pelleted at 1200 × *g* for 5 min. Pellets were resuspended in 250 µL cold buffer L (10 mM HEPES-KOH pH 7.9, 0.2 mM $MgCl_2$, 10 mM KCl, 0.1 mM EDTA-KOH pH 8.0, 0.5 mM EGTA-KOH pH 8.0, 0.5 mM DTT, 0.5 mM Pefabloc-SC, and 1 × Roche Complete EDTA-free protease inhibitor cocktail) and incubated for 10 min on ice. To lyse cells, 0.05% IGEPAL CA-630 was added, cells were incubated on ice for 10 min, and then samples were passed through a 30 G needle ten times. Lysed cells were pelleted at 1000 × *g* for 5 min at 4 °C and the supernatant was removed. Pellets were resuspended in 250 µL buffer L with sucrose (Buffer L, 250 mM sucrose, 0.5 mM DTT, 0.5 mM

pefabloc, and 1 × Roche protease inhibitor solution), spun at 1000 × *g* for 5 min at 4 °C, and both steps repeated.

For transposition, nuclei were pelleted at 1000 × *g* for 5 min at 4 °C, resuspended in 25 µL of 2× Tagmentation DNA buffer and 2.5 µl of Tn5 transposase from the Nextera DNA Library Prep kit (Illumina, San Diego, CA), and incubated at 37 °C for 30 min. DNA was purified using the MinElute kit (Qiagen, 28004) per PCR purification protocol provided by the manufacturer. Transposed DNA was originally amplified and barcoded in a PCR reaction using NEBnext PCR master mix (NEB, M0544) and 1.25 µM forward and reverse primers originally described in ref. 112, using the following PCR conditions: 72 °C for 5 min; 98 °C for 30 s; and thermocycling at 98 °C for 10 s, 63 °C for 30 s and 72 °C for 1 min. To reduce GC and size bias in the PCR, we monitored the PCR reaction using qPCR in order to stop amplification before saturation. To do this, we amplified the full libraries for five cycles, after which we took an aliquot of the PCR reaction and added 10 µl of PCR cocktail with Sybr Green at a final concentration of 0.6x. We ran this reaction for 20 cycles to determine the additional number of cycles needed for the remaining 45 µL reaction (the reaction less the aliquot removed for qPCR). The libraries were purified using a Qiagen PCR cleanup kit. Libraries were amplified for a total of 10–12 cycles. An additional 0.9X SPRI bead cleanup (Beckman Coulter) was performed according to the manufacturers protocol to eliminate a contaminating 50 bp peak. Samples were pooled, quantified using qPCR and sequenced on an Illumina HiSeq 2500 to generate 50 bp PE reads.

ChIP-seq

For ChIP-seq, ~500 million cells were used per experiment. Cells were seeded at ~10,000 cells per ml in RA and grown for 24 h at 27 °C. Cells were centrifuged at 2400 × *g* for 5 min at 4 °C (all further centrifugation steps and washed used these parameters unless otherwise specified). They were then washed once with ASW and resuspended in ASW, this time combining them into one 50 ml tube and cell number was quantified. Following centrifugation, the pellet was resuspended in 5 ml ice-cold lysis buffer (50 mM HEPES pH 7.6, 100 mM NaCl, 0.5 mM $MgCl_2$, 10% (v/v) glycerol, 1% (v/v) Triton x-100, 10 mM sodium butyrate, 2 mM DTT, 2 mM Pefabloc-SC, 2x Complete EDTA-free Protease inhibitor cocktail) and incubated on ice for 30 s. Released nuclei were collected by centrifugation and washed once in nuclease digestion buffer (10 mM Tris-HCl pH 8.0, 10 mM NaCl, 3 mM MgCl2, 0.1% (v/v) IGEPAL CA-630, 0.25 M sucrose, 3 mM $CaCl_2$, 2 × Complete EDTA-free Protease inhibitor cocktail). The nuclei were then resuspended in 1 ml of nuclease digestion buffer per 500 million cells and placed in a dry bath at 37 °C for 5 min. Micrococcal nuclease (MNase) (ThermoFisher Scientific, EN0181) was added to the warmed nuclei at a final concentration of 450 U/ml and cells were incubated for exactly 5 min at 37 °C with gentle mixing by inversion every minute. Digestion was halted by addition of 8 µL 0.5 M EDTA pH 8.0 per ml on ice. The samples were then centrifuged, and the supernatant was kept and placed on ice (S1). The remaining pellet was resuspended in nucleosome release buffer (10 mM Tris-HCl pH 7.5, 10 mM NaCl, 0.2 mM EDTA, 2 × Complete EDTA-free Protease inhibitor cocktail), placed rotating end over end at 4 °C for 1 h and passed 5 times through a 27 G needle. Following centrifugation, the supernatant (S2) was retained and then added to the S1 (nucleosome solution). Digestion to majority mononucleosomes was confirmed by extracting DNA from 50 µL followed by agarose gel electrophoresis.

For each ChIP experiment 100 µL of mononucleosomes per antibody (+100 µL for input) was diluted 10-fold with ChIP incubation buffer (70 mM NaCl, 10 mM Tris-HCl pH 7.5, 2 mM $MgCl_2$, 2 mM EDTA, 0.1% Triton, 1 × Roche Complete EDTA-free Protease inhibitor cocktail). 1 ml aliquots were then made and the antibodies (Listed in Supplementary Table 2) added and incubated rotating at 4 °C overnight. The next day, CaptivA® Protein A Affinity Resin (Repligen, CA-PRI-0025) (40 µl slurry per ChIP, hereafter called "beads") was pre-blocked in ChIP incubation buffer supplemented with 1 mg/ml BSA and 1 mg/ml yeast tRNA, for 1 h at 4 °C. The beads were collected by centrifugation at 1000 × *g* for 1 min at 4 °C (all centrifugations and washes from this point were performed using these parameters) and washed three times in ChIP incubation buffer. After being aliquoted into 1 ml tubes, the beads were collected by centrifugation and the antibody-nucleosome solutions added. This was incubated rotating at 4 °C for 1 h before the beads were collected by centrifugation and washed 5 times in ChIP wash buffer (20 mM Tris-HCl pH 7.5, 2 mM EDTA, 125 mM NaCl, 0.1% (v/v) Triton X-100) with 5-min incubations rotating at 4 °C between each wash. Following a final wash with TE buffer, the beads were resuspended in 100 µL fresh elution buffer (1% (w/v) SDS, 0.1 M $NaHCO_3$) and incubated shaking at room temperature for 30 min. The beads were then collected by centrifugation and the supernatant retained. ChIP DNA was purified using the ChIP DNA Clean and Concentrator kit (Zymo, D5206) using the manufacturer's instructions and eluted in 10 µl.

ChIP libraries were generated by the QB3-Berkeley Functional Genomics Laboratory (FGL) and sequenced at Vincent J. Coates Genomics Sequencing Laboratory (GSL) at UC Berkeley. Samples were checked for concentration and size using the Qubit dsDNA High Sensitivity assay (ThermoFisher Scientific, Q32851) and Agilent Fragment Analyzer with the DNA High Sensitivity NGS assay (Aligent, 5067-4626). Subsequently, the input material was size-selected using a double-sided bead cleanup at a 0.55x/1.8x bead ratio using Kapa Pure Beads. Library preparation was carried out using the KAPA HyperPrep kit for DNA (KK8504). Truncated universal stub adapters were used for ligation, and indexed primers were used during PCR amplification to complete the adapters and to enrich the libraries for adapter-ligated fragments. After PCR, the libraries were cleaned at 0.9x bead ratio to remove smaller insert fragments and dimers. Samples were then checked for quality on an AATI (now Agilent) Fragment Analyzer. Illumina sequencing library molarity was measured with quantitative PCR with the Kapa Biosystems Illumina Quant qPCR Kits on a BioRad CFX Connect thermal cycler. Libraries were then pooled evenly by molarity and sequenced on a shared Illumina NovaSeq6000 150PE S4 flowcell. Raw sequencing data was converted into fastq format sample specific files using the Illumina BCL Convert V4 software on the sequencing center's local Linux server system.

Histone extraction and mass spectrometry

For histone extraction, thecate cells or slow swimmers were seeded at ~10,000 cells per ml in RA and grown for 24 h. Approximately 200 million cells were used per replicate. Cells were centrifuged at 2400 × *g* for 5 min at 4 °C (all further centrifugation steps and washed used these parameters unless otherwise specified) in 50 ml tubes. They were then washed once with ASW and following centrifugation each tube was resuspended in 1 ml ASW. This was then carefully layered on top of 2 ml percoll solution (160 µl percoll brought to 2 ml with ASW) in a 15 ml tube and centrifuged at 1000 × *g* for 10 min at 4 °C with the brakes off. The supernatant was carefully removed, the pellets resuspended in ASW and combined into one 50 ml tube and cells were counted. Following centrifugation, cells were resuspended in 1 ml lysis buffer (see above) per 50 million cells and incubated on ice for 30 s. The nuclei were collected by centrifugation and directly resuspended in 800 µL ice-cold 0.4 N H_2SO_4 and left at 4 °C for 2 h. They were then centrifuged at 16,000 × *g* for 10 min at 4 °C and the pellet discarded. The histones were precipitated by drop-wise addition of 264 µL ice-cold 100% TCA followed by overnight incubation at 4 °C. Histones were pelleted by centrifugation at 16,000 × *g* for 10 min at 4 °C followed by three washes with ice-cold acetone. The histone pellet was allowed to air dry for ~15 min at room temperature and resuspended in 50–100 µL of water. Histone concentrations were determined using Bradford assay with BSA as standard. Three independent biological replicates were used for each cell type.

Sample preparation and mass spectrometry were performed as previously described[113]. Propionic acid derivatization was carried out in 20 µg of purified histones. About 4-19 amino acid-long peptides were generated using propionic anhydride derivatization, followed by digestion with 1 µg trypsin and desalting for bottom-up mass spectrometry. The peptides were then analyzed using a Thermo Scientific Acclaim PepMap 100 C18 HPLC Column (250 mm length, 0.075 mm I.D., Reversed Phase, 3 um particle size) fitted on an Vanquish™ Neo UHPLC System (Thermo Scientific, San Jose, Ca, USA) using the HPLC gradient: 2% to 45% solvent B (*A* = 0.1% formic acid; *B* = 95% MeCN, 0.1% formic acid) over 50 min, to 95% solvent B in 10 min, all at a flow-rate of 300 nL/min. The sample (5 µl of 1 µg/µl) was analyzed in a QExactive-Orbitrap mass spectrometer (Thermo Scientific) using data-independent acquisition (DIA). It consisted of full scan MS (m/z 295–1100) acquired in Orbitrap with a resolution of 70,000 and an AGC target of 1 × 106. Tandem MS was acquired in centroid mode in the ion trap using sequential isolation windows of 24 m/z, AGC target of 2 × 105, CID collision energy of 30 and maximum injection time of 50 ms.

Data analysis
All analysis software used are listed in Supplementary Table 3.

ATAC-seq data analysis
ATAC reads were initially processed using Trimmomatic[114] to remove adapters. They were then mapped to the genome[62] using Bowtie2[115] (with the "-very-sensitive" option) and converted to BAM files and sorted using SAMtools[116]. PCR duplicates were removed using Picard and low-quality reads were removed using SAMtools view ("-q 30" option). SAMtools was used to extract sub-nucleosome sizes reads, i.e., reads with insert sizes less than 100 bps. DeepTools[117] was used to generate PCA and Pearson correlation plots. Replicates were merged together for visualization and BigWig files were generated using bamCoverage (--binsize 10 –effectiveGenomesize 55000000 –normalizeUsing RPGC) and visualized using the Integrative Genomics Viewer (IGV)[118]. Heatmaps and profile plots were generated using deepTools (v3.4.3). Peak calling was performed either with Genrich (-j -y -r -v options) which takes both replicates as input or with MACS 2[119] for individual samples (-f -BAMPE option --keep-dup all). Sub-nucleosomal reads were used as input for Genrich while all reads were used for MACS2 due to poor performance of MACS2 with only sub-nucleosome reads. In the case of MACS2 consensus peaks sets for each cell-type were derived by taking only peaks which were present in both replicates using BEDtools[120] intersect (-wa option). For both softwares, consensus peak sets were derived taking all peaks present in one or more cell type using BEDtools intersect (-wa option). Annotation of peaks was performed using ChIPseeker[121] in R studio. GenomicRanges[122] was used to measure overlaps and distances between TSSs and peaks.

ChIP-seq data analysis
Reads were first processed using Trimmomatic to remove adapters and they were trimmed to 100 bps. They were then processed as above for ATAC-seq.

RNA-seq analysis
RNA-seq data for the different cell types was published previously[66]. We re-analyzed the data by first mapping the reads to the genome using STAR[123] (--quantMode GeneCounts options). BigWig files were generated using bamCoverage (--binsize 10–effectiveGenomesize 55000000 –normalizeUsing RPKM) and visualized using IGV. Differential expression analysis was performed using DESeq2[124] using default parameters. Normalized RPKM values from DESeq2 were used to calculate the log (RPKM + 1) values. Visualizations were generated using ggplot2[125].

Analysis of transposable elements
The fasta file with the annotated transposable elements from Southworth et al. was downloaded and used as input for a RepeatMasker[75]) search using default parameters. The resulting repeat annotation was then filtered to select insertions that spanned at least 50% of the length of the consensus sequences, to avoid spurious short annotations that might represent highly derived insertions. The resulting annotation was then converted for compatibility with TElocal part of the TEtranscripts pipeline[126]. RNA-seq was mapped against the genome using HISAT2[127], and then each replicate quantified with TElocal, with counts then normalized in DESeq2, using both protein coding genes and repeat count data. For epigenomic visualization of hPTMs and ATAC-seq, each TE family, as defined by Southworth et al., was then used as input for deepTools (3.5.0), using a bin size of 50 bp, including input to control for potential mappability artifacts on transposable elements (requiring MAPQ > 30). The order of inserts produced by deepTools clustering was then used to generate a heatmap of transcriptional values with matched order using pheatmap. Summary TE family heatmaps were generated using plotProfile function in deepTools.

Mass-spectrometry data analysis
Histone Mass spectrometry data was processed using EpiProfile 2[128]. Briefly, *S. rosetta* histone sequences underwent in silico digestion into peptides, with cleavage occurring after Arginine. Each peptide was then assessed for common potential post-translational modifications (PTMs), such as H3 3-8 unmodified, K4me1, K4me2, K4me3, and K4ac. The area under the curve (AUC) for all peptides is extracted from the raw data. To normalize and facilitate group comparisons, the percentage of each peptide within the same sequence is calculated by dividing its AUC by the summed AUC. The program is available for download at GitHub (https://github.com/zfyuan/EpiProfile2.0_Family/blob/master/EpiProfile2.1_S.rosetta.zip). A summary of the data analysis is available as Supplementary data file 1 which includes AUC, calculated ratios of modified peptides and retention times. This method was not capable of determining the exact stoichiometry of individual modifications due to the lack of spike-in controls.

Reporting summary
Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability
We used the *S. rosetta* genome[62] (GenBank assembly accession: GCA_000188695.1) available from Ensembl protists (https://protists.ensembl.org/Salpingoeca_rosetta_gca_000188695/Info/Index?db=core) for all analyses. RNA-seq data was published previously[66] and have been deposited in NCBI's Gene Expression Omnibus and are accessible through GEO Series accession number GSE267344. ATAC-seq and ChIP-seq data generated in this study have been deposited to the NCBI Short Read Archive. ATAC-seq is bioproject PRJNA1107385 (ID 1107385 - BioProject - NCBI) and ChIP-seq data is bioproject PRJNA1112805 (ID 1112805 - BioProject - NCBI). Mass spectrometry data has been deposited in the MassIVE database under dataset ID: MSV000094416 (MassIVE Dataset Summary). Source data are provided with this paper.

References
1. Zeitlinger, J. & Stark, A. Developmental gene regulation in the era of genomics. *Dev. Biol.* **339**, 230–239 (2010).
2. Levine, M., Cattoglio, C. & Tjian, R. Looping back to leap forward: transcription enters a new era. *Cell* **157**, 13–25 (2014).
3. Levine, M. Transcriptional enhancers in animal development and evolution. *Curr. Biol.* **20**, R754–R763 (2010).

4. Long, H. K., Prescott, S. L. & Wysocka, J. Ever-changing landscapes: transcriptional enhancers in development and evolution. *Cell* **167**, 1170–1187 (2016).
5. Sebe-Pedros, A. et al. Early metazoan cell type diversity and the evolution of multicellular gene regulation. *Nat. Ecol. Evol.* **2**, 1176–1188 (2018).
6. Schwaiger, M. et al. Evolutionary conservation of the eumetazoan gene regulatory landscape. *Genome Res* **24**, 639–650 (2014).
7. Gaiti, F. et al. Landscape of histone modifications in a sponge reveals the origin of animal cis-regulatory complexity. *Elife* **6**, e22194 (2017).
8. Kim, I. V. et al. Chromatin loops are an ancestral hallmark of the animal regulatory genome. *Nature* **642**, 1097–1105 (2025).
9. Wong, E. S. et al. Deep conservation of the enhancer regulatory code in animals. *Science* **370**, eaax8137 (2020).
10. Sebé-Pedrós, A. et al. The dynamic regulatory genome of Capsaspora and the origin of animal multicellularity. *Cell* **165**, 1224–1237 (2016).
11. Ruiz-Trillo, I. & de Mendoza, A. Towards understanding the origin of animal development. *Development* **147**, dev192575 (2020).
12. Coyle, M. C. & King, N. The evolutionary foundations of transcriptional regulation in animals. *Nat. Rev. Genet.* https://doi.org/10.1038/s41576-025-00864-9 (2025).
13. Yadav, T., Quivy, J.-P. & Almouzni, G. Chromatin plasticity: a versatile landscape that underlies cell fate and identity. *Science* **361**, 1332–1336 (2018).
14. Dambacher, S., Hahn, M. & Schotta, G. Epigenetic regulation of development by histone lysine methylation. *Heredity* **105**, 24–37 (2010).
15. Chen, T. & Dent, S. Y. Chromatin modifiers and remodellers: regulators of cellular differentiation. *Nat. Rev. Genet* **15**, 93–106 (2014).
16. Perino, M. & Veenstra, G. J. Chromatin control of developmental dynamics and plasticity. *Dev. Cell* **38**, 610–620 (2016).
17. Zhou, V. W., Goren, A. & Bernstein, B. E. Charting histone modifications and the functional organization of mammalian genomes. *Nat. Rev. Genet* **12**, 7–18 (2011).
18. Millan-Zambrano, G., Burton, A., Bannister, A. J. & Schneider, R. Histone post-translational modifications—cause and consequence of genome function. *Nat. Rev. Genet* **23**, 563–580 (2022).
19. Navarrete, C., Montgomery, S. A., Mendieta, J., Lara-Astiaso, D. & Sebé-Pedrós, A. Diversity and evolution of chromatin regulatory states across eukaryotes. *bioRxiv* https://doi.org/10.1101/2025.03.17.643675 (2025).
20. Zhang, X., Bernatavichute, Y. V., Cokus, S., Pellegrini, M. & Jacobsen, S. E. Genome-wide analysis of mono-, di- and trimethylation of histone H3 lysine 4 in Arabidopsis thaliana. *Genome Biol.* **10**, R62 (2009).
21. Barski, A. et al. High-resolution profiling of histone methylations in the human genome. *Cell* **129**, 823–837 (2007).
22. Yan, W. et al. Dynamic control of enhancer activity drives stage-specific gene expression during flower morphogenesis. *Nat. Commun.* **10**, 1705 (2019).
23. Wu, Y. & Tirichine, L. Chromosome-wide distribution and characterization of H3K36me3 and H3K27Ac in the marine model diatom Phaeodactylum tricornutum. *Plants* **12**, 2852 (2023).
24. Wu, Y., Chaumier, T., Manirakiza, E., Veluchamy, A. & Tirichine, L. PhaeoEpiView: an epigenome browser of the newly assembled genome of the model diatom Phaeodactylum tricornutum. *Sci. Rep.* **13**, 8320 (2023).
25. Bourdareau, S. et al. Histone modifications during the life cycle of the brown alga Ectocarpus. *Genome Biol.* **22**, 12 (2021).
26. Heintzman, N. D. et al. Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature* **459**, 108–112 (2009).
27. Herz, H.-M. et al. Enhancer-associated H3K4 monomethylation by Trithorax-related, the Drosophila homolog of mammalian Mll3/Mll4. *Genes Dev.* **26**, 2604–2620 (2012).
28. Strenkert, D. et al. The landscape of Chlamydomonas histone H3 lysine 4 methylation reveals both constant features and dynamic changes during the diurnal cycle. *Plant J.* **112**, 352–368 (2022).
29. van Dijk, K. et al. Monomethyl histone H3 lysine 4 as an epigenetic mark for silenced euchromatin in Chlamydomonas. *Plant Cell* **17**, 2439–2453 (2005).
30. Allshire, R. C. & Madhani, H. D. Ten principles of heterochromatin formation and function. *Nat. Rev. Mol. Cell Biol.* **19**, 229–244 (2018).
31. Margueron, R. & Reinberg, D. The Polycomb complex PRC2 and its mark in life. *Nature* **469**, 343–349 (2011).
32. Blackledge, N. P. & Klose, R. J. The molecular principles of gene regulation by Polycomb repressive complexes. *Nat. Rev. Mol. Cell Biol.* **22**, 815–833 (2021).
33. Piunti, A. & Shilatifard, A. The roles of Polycomb repressive complexes in mammalian development and cancer. *Nat. Rev. Mol. Cell Biol.* **22**, 326–345 (2021).
34. Deevy, O. & Bracken, A. P. PRC2 functions in development and congenital disorders. *Development* **146**, dev181354 (2019).
35. Schuettengruber, B., Bourbon, H. M., Di Croce, L. & Cavalli, G. Genome regulation by polycomb and trithorax: 70 years and counting. *Cell* **171**, 34–57 (2017).
36. Jiao, H., Xie, Y. & Li, Z. Current understanding of plant Polycomb group proteins and the repressive histone H3 Lysine 27 trimethylation. *Biochem. Soc. Trans.* **48**, 1697–1706 (2020).
37. Calonje, M. PRC1 marks the difference in plant PcG repression. *Mol. Plant* **7**, 459–471 (2014).
38. Berke, L. & Snel, B. The plant Polycomb repressive complex 1 (PRC1) existed in the ancestor of seed plants and has a complex duplication history. *BMC Evol. Biol.* **15**, 44 (2015).
39. de Potter, B., Raas, M. W., Seidl, M. F., Verrijzer, C. P. & Snel, B. Uncoupled evolution of the Polycomb system and deep origin of non-canonical PRC1. *Commun. Biol.* **6**, 1144 (2023).
40. Carlier, F. et al. Loss of EZH2-like or SU (VAR) 3–9-like proteins causes simultaneous perturbations in H3K27 and H3K9 trimethylation and associated developmental defects in the fungus Podospora anserina. *Epigenetics Chromatin* **14**, 22 (2021).
41. Veluchamy, A. et al. An integrative analysis of post-translational histone modifications in the marine diatom Phaeodactylum tricornutum. *Genome Biol.* **16**, 1–18 (2015).
42. Déléris, A., Berger, F. & Duharcourt, S. Role of Polycomb in the control of transposable elements. *Trends Genet.* **37**, 882–889 (2021).
43. Frapporti, A. et al. The Polycomb protein Ezl1 mediates H3K9 and H3K27 methylation to repress transposable elements in Paramecium. *Nat. Commun.* **10**, 2710 (2019).
44. Montgomery, S. A. et al. Chromatin organization in early land plants reveals an ancestral association between H3K27me3, transposons, and constitutive heterochromatin. *Curr. Biol.* **30**, 573–588. e7 (2020).
45. Xu, J. et al. A Polycomb repressive complex is required for RNAi-mediated heterochromatin formation and dynamic distribution of nuclear bodies. *Nucleic Acids Res.* **49**, 5407–5425 (2021).
46. Zhao, X. et al. RNAi-dependent Polycomb repression controls transposable elements in Tetrahymena. *Genes Dev.* **33**, 348–364 (2019).
47. 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).
48. Hisanaga, T. et al. The Polycomb repressive complex 2 deposits H3K27me3 and represses transposable elements in a broad range of eukaryotes. *Curr. Biol.* **33**, 4367–4380.e9 (2023).
49. Miro-Pina, C. et al. Paramecium Polycomb repressive complex 2 physically interacts with the small RNA-binding PIWI protein to repress transposable elements. *Dev. Cell* **57**, 1037–1052.e8 (2022).
50. Kramer, H. M., Seidl, M. F., Thomma, B. P. & Cook, D. E. Local rather than global H3K27me3 dynamics are associated with differential gene expression in Verticillium dahliae. *MBio* **13**, e03566–21 (2022).
51. Zhao, X. et al. Genome wide natural variation of H3K27me3 selectively marks genes predicted to be important for cell differentiation in Phaeodactylum tricornutum. *New Phytol.* **229**, 3208–3220 (2021).
52. Huang, T.-C. et al. Sex-specific chromatin remodelling safeguards transcription in germ cells. *Nature* **600**, 737–742 (2021).
53. Akkouche, A. et al. A dual histone code specifies the binding of heterochromatin protein Rhino to a subset of piRNA source loci. *bioRxiv* https://doi.org/10.1101/2024.01.11.575256 (2024).
54. Walter, M., Teissandier, A., Pérez-Palacios, R. & Bourc'His, D. An epigenetic switch ensures transposon repression upon dynamic loss of DNA methylation in embryonic stem cells. *Elife* **5**, e11418 (2016).
55. Grau-Bove, X. et al. A phylogenetic and proteomic reconstruction of eukaryotic chromatin evolution. *Nat. Ecol. Evol.* **6**, 1007–1023 (2022).
56. Booth, D. & King, N. The history of Salpingoeca rosetta as a model for reconstructing animal origins. *Curr. Top. Dev. Biol.* **147**, 73–91 (2021).
57. Brunet, T. & King, N. The origin of animal multicellularity and cell differentiation. *Dev. Cell* **43**, 124–140 (2017).
58. Brunet, T. & King, N. The single-celled ancestors of animals: a history of hypotheses. in *The Evolution of Multicellularity* 251–278 (CRC Press, 2022).
59. Richter, D. J., Fozouni, P., Eisen, M. B. & King, N. Gene family innovation, conservation and loss on the animal stem lineage. *eLife* **7**, e34226 (2018).
60. Hoffmeyer, T. T. & Burkhardt, P. Choanoflagellate models—Monosiga brevicollis and Salpingoeca rosetta. *Curr. Opin. Genet Dev.* **39**, 42–47 (2016).
61. Dayel, M. J. et al. Cell differentiation and morphogenesis in the colony-forming choanoflagellate Salpingoeca rosetta. *Dev. Biol.* **357**, 73–82 (2011).
62. Fairclough, S. R. et al. Premetazoan genome evolution and the regulation of cell differentiation in the choanoflagellate Salpingoeca rosetta. *Genome Biol.* **14**, R15 (2013).
63. Fairclough, S. R., Dayel, M. J. & King, N. Multicellular development in a choanoflagellate. *Curr. Biol.* **20**, R875–R876 (2010).
64. Alegado, R. A. et al. A bacterial sulfonolipid triggers multicellular development in the closest living relatives of animals. *Elife* **1**, e00013 (2012).
65. Perotti, O., Viramontes Esparza, G. & Booth, D. S. A red algal polysaccharide influences the multicellular development of the choanoflagellate Salpingoeca rosetta. *bioRxiv* https://doi.org/10.1101/2024.05.14.594265 (2024).
66. Leon, F. et al. Cell differentiation controls iron assimilation in the choanoflagellate Salpingoeca rosetta. *mSphere* **10**, e00917-24 (2025).
67. Levin, T. eraC. & King, N. Evidence for sex and recombination in the choanoflagellate Salpingoeca rosetta. *Curr. Biol.* **23**, 2176–2180 (2013).
68. Woznica, A., Gerdt, J. P., Hulett, R. E., Clardy, J. & King, N. Mating in the closest living relatives of animals is induced by a bacterial chondroitinase. *Cell* **170**, 1175–1183.e11 (2017).
69. Schultz, D. T. et al. Ancient gene linkages support ctenophores as sister to other animals. *Nature* **618**, 110–117 (2023).
70. Booth, D. S., Szmidt-Middleton, H. & King, N. Transfection of choanoflagellates illuminates their cell biology and the ancestry of animal septins. *Mol. Biol. Cell* **29**, 3026–3038 (2018).
71. Booth, D. S. & King, N. Genome editing enables reverse genetics of multicellular development in the choanoflagellate Salpingoeca rosetta. *Elife* **9**, e56193 (2020).
72. Combredet, C., Ansel, M. & Brunet, T. A selection-based knockout approach for a choanoflagellate reveals regulation of multicellular development by Hippo signaling. *Cell Rep.* **44**, 116345 (2025).
73. Coyle, M. C. et al. An RFX transcription factor regulates ciliogenesis in the closest living relatives of animals. *Curr. Biol.* **33**, 3747–3758. e9 (2023).
74. Buenrostro, J. D., Wu, B., Chang, H. Y. & Greenleaf, W. J. ATAC-seq: a method for assaying chromatin accessibility genome-wide. *Curr. Protoc. Mol. Biol.* **109**, 21–29 (2015).
75. Smit, A., Hubley, R. & Green, P. RepeatMasker at http://repeatmasker.org. (2013–2015).
76. Southworth, J., Grace, C. A., Marron, A. O., Fatima, N. & Carr, M. A genomic survey of transposable elements in the choanoflagellate Salpingoeca rosetta reveals selection on codon usage. *Mob. DNA* **10**, 1–19 (2019).
77. Sebé-Pedrós, A., Degnan, B. M. & Ruiz-Trillo, I. The origin of Metazoa: a unicellular perspective. *Nat. Rev. Genet.* **18**, 498–512 (2017).
78. Banerji, J., Rusconi, S. & Schaffner, W. Expression of a β-globin gene is enhanced by remote SV40 DNA sequences. *Cell* **27**, 299–308 (1981).
79. Suga, H. et al. The Capsaspora genome reveals a complex unicellular prehistory of animals. *Nat. Commun.* **4**, 2325 (2013).
80. Martín-Durán, J. M. et al. Conservative route to genome compaction in a miniature annelid. *Nat. Ecol. Evol.* **5**, 231–242 (2021).
81. Navratilova, P. et al. Sex-specific chromatin landscapes in an ultra-compact chordate genome. *Epigenetics Chromatin* **10**, 3 (2017).
82. Denoeud, F. et al. Plasticity of animal genome architecture unmasked by rapid evolution of a pelagic tunicate. *Science* **330**, 1381–1385 (2010).
83. Sarre, L. A. et al. DNA methylation enables recurrent endogenization of giant viruses in an animal relative. *Sci. Adv.* **10**, eado6406 (2024).
84. Dudin, O. et al. A unicellular relative of animals generates a layer of polarized cells by actomyosin-dependent cellularization. *Elife* **8**, e49801 (2019).
85. Gao, Y. et al. The emergence of Sox and POU transcription factors predates the origins of animal stem cells. *Nat. Commun.* **15**, 9868 (2024).
86. Dumesic, P. A. et al. Product binding enforces the genomic specificity of a yeast polycomb repressive complex. *Cell* **160**, 204–218 (2015).
87. Riising, E. M. et al. Gene silencing triggers polycomb repressive complex 2 recruitment to CpG islands genome wide. *Mol. Cell* **55**, 347–360 (2014).
88. Streubel, G. et al. The H3K36me2 methyltransferase Nsd1 demarcates PRC2-mediated H3K27me2 and H3K27me3 domains in embryonic stem cells. *Mol. Cell* **70**, 371–379 e5 (2018).
89. Mikulski, P., Komarynets, O., Fachinelli, F., Weber, A. P. & Schubert, D. Characterization of the polycomb-group mark H3K27me3 in unicellular algae. *Front. Plant Sci.* **8**, 607 (2017).

90. Gahan, J. M., Rentzsch, F. & Schnitzler, C. E. The genetic basis for PRC1 complex diversity emerged early in animal evolution. *Proc. Natl. Acad. Sci. USA* **117**, 22880–22889 (2020).
91. Blanco, E., Gonzalez-Ramirez, M., Alcaine-Colet, A., Aranda, S. & Di Croce, L. The bivalent genome: characterization, structure, and regulation. *Trends Genet.* **36**, 118–131 (2020).
92. Bernstein, B. E. et al. A bivalent chromatin structure marks key developmental genes in embryonic stem cells. *Cell* **125**, 315–326 (2006).
93. Azuara, V. et al. Chromatin signatures of pluripotent cell lines. *Nat. Cell Biol.* **8**, 532–538 (2006).
94. Dattani, A. et al. Epigenetic analyses of planarian stem cells demonstrate conservation of bivalent histone modifications in animal stem cells. *Genome Res.* **28**, 1543–1554 (2018).
95. Cheng, Q. & Xie, H. Genome-wide analysis of bivalent histone modifications during Drosophila embryogenesis. *Genesis* **60**, e23502 (2022).
96. Pan, G. et al. Whole-genome analysis of histone H3 lysine 4 and lysine 27 methylation in human embryonic stem cells. *Cell Stem Cell* **1**, 299–312 (2007).
97. Voigt, P. et al. Asymmetrically modified nucleosomes. *Cell* **151**, 181–193 (2012).
98. Weiner, A. et al. Co-ChIP enables genome-wide mapping of histone mark co-occurrence at single-molecule resolution. *Nat. Biotechnol.* **34**, 953–961 (2016).
99. van der Velde, A. et al. Annotation of chromatin states in 66 complete mouse epigenomes during development. *Commun. Biol.* **4**, 239 (2021).
100. Voigt, P., Tee, W. W. & Reinberg, D. A double take on bivalent promoters. *Genes Dev.* **27**, 1318–1338 (2013).
101. Mas, G. et al. Promoter bivalency favors an open chromatin architecture in embryonic stem cells. *Nat. Genet.* **50**, 1452–1462 (2018).
102. Rada-Iglesias, A. et al. A unique chromatin signature uncovers early developmental enhancers in humans. *Nature* **470**, 279–283 (2011).
103. Zentner, G. E., Tesar, P. J. & Scacheri, P. C. Epigenetic signatures distinguish multiple classes of enhancers with distinct cellular functions. *Genome Res.* **21**, 1273–1283 (2011).
104. Kumar, D., Cinghu, S., Oldfield, A. J., Yang, P. & Jothi, R. Decoding the function of bivalent chromatin in development and cancer. *Genome Res* **31**, 2170–2184 (2021).
105. Eckersley-Maslin, M. A. et al. Epigenetic priming by Dppa2 and 4 in pluripotency facilitates multi-lineage commitment. *Nat. Struct. Mol. Biol.* **27**, 696–705 (2020).
106. Gretarsson, K. H. & Hackett, J. A. Dppa2 and Dppa4 counteract de novo methylation to establish a permissive epigenome for development. *Nat. Struct. Mol. Biol.* **27**, 706–716 (2020).
107. Hu, D. et al. The Mll2 branch of the COMPASS family regulates bivalent promoters in mouse embryonic stem cells. *Nat. Struct. Mol. Biol.* **20**, 1093–1097 (2013).
108. Zhang, Q. et al. Asymmetric epigenome maps of subgenomes reveal imbalanced transcription and distinct evolutionary trends in Brassica napus. *Mol. Plant* **14**, 604–619 (2021).
109. Levin, T. C., Greaney, A. J., Wetzel, L. & King, N. The rosetteless gene controls development in the choanoflagellate S. rosetta. *Elife* **3**, e04070 (2014).
110. Hallegraeff, G. M., Anderson, D. M., Cembella, A. D. & Enevoldsen, H. O. *Manual on Harmful Marine Microalgae* (UNESCO, 2004).
111. Woznica, A. et al. Bacterial lipids activate, synergize, and inhibit a developmental switch in choanoflagellates. *Proc. Natl. Acad. Sci. USA* **113**, 7894–7899 (2016).
112. Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. methods* **10**, 1213 (2013).
113. Sidoli, S., Bhanu, N. V., Karch, K. R., Wang, X. & Garcia, B. A. Complete workflow for analysis of histone post-translational modifications using bottom-up mass spectrometry: from histone extraction to data analysis. *J. Vis. Exp.* https://doi.org/10.3791/54112 (2016).
114. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
115. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
116. Li, H. et al. The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
117. Ramírez, F., Dündar, F., Diehl, S., Grüning, B. A. & Manke, T. deepTools: a flexible platform for exploring deep-sequencing data. *Nucleic Acids Res.* **42**, W187–W191 (2014).
118. Robinson, J. T. et al. Integrative genomics viewer. *Nat. Biotechnol.* **29**, 24–26 (2011).
119. Zhang, Y. et al. Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, 1 (2008).
120. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
121. Yu, G., Wang, L.-G. & He, Q.-Y. ChIPseeker: an R/Bioconductor package for ChIP peak annotation, comparison and visualization. *Bioinformatics* **31**, 2382–2383 (2015).
122. Lawrence, M. et al. Software for computing and annotating genomic ranges. *PLoS Comput. Biol.* **9**, e1003118 (2013).
123. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
124. Love, M., Anders, S. & Huber, W. Differential analysis of count data–the DESeq2 package. *Genome Biol.* **15**, 10–1186 (2014).
125. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis*. ISBN 978-3-319-24277-4 (Springer-Verlag, 2016).
126. Jin, Y., Tam, O. H., Paniagua, E. & Hammell, M. TEtranscripts: a package for including transposable elements in differential expression analysis of RNA-seq datasets. *Bioinformatics* **31**, 3593–3599 (2015).
127. Kim, D., Paggi, J. M., Park, C., Bennett, C. & Salzberg, S. L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.* **37**, 907–915 (2019).
128. Yuan, Z. F. et al. EpiProfile 2.0: a computational platform for processing epi-proteomics mass spectrometry data. *J. Proteome Res.* **17**, 2533–2541 (2018).

Acknowledgements

We would like to thank Nicole King for generously supporting ATAC-seq experiments at an early phase of this project. We would also like to thank Neil Blackledge and the rest of the Klose lab for help and advice with ChIP-seq and analysis. We thank Uri Frank for constructive comments on the manuscript. This work was funded by a Wellcome Trust, Sir Henry Wellcome Postdoctoral Fellowship (222767/Z/21/Z) to J.M.G. and an NIH award to D.S.B. (R35GM147404). The Klose lab is supported by the Wellcome Trust (209400/Z/17/Z). A.d.M. is supported by a European Research Council Starting Grant (950230).

Author contributions

Conceptualization, J.M.G. and D.S.B.; methodology, J.M.G., L.W.H., L.A.W., N.V.B., D.S.B.; software, Z.F.Y.; formal analysis, Z.F.Y., A.d.M., J.M.G.; investigation, J.M.G., L.W.H., L.A.W., N.V.B.; writing, J.M.G. and D.S.B.; supervision, R.J.K., B.A.G., and D.S.B.; funding acquisition, J.M.G., R.J.K., and D.S.B. All authors edited and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-025-64570-0.

Correspondence and requests for materials should be addressed to James M. Gahan or David S. Booth.

Peer review information *Nature Communications* thanks the anonymous reviewers for their contribution to the peer review of this work. A peer review file is available.