

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

Corresponding Author: Dr James Gahan

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)

The manuscript by Gahan et al. examines chromatin accessibility and 4 histone modifications genome-wide in the choanoflagellate *Salpingoeca rosetta*, a sister group to animals. The data are of outstanding quality and the analyses provide a thorough description of histone marks with respect to gene expression levels and accessibility. This is an interesting, well-done, yet entirely descriptive study.

The work is sound but would improve with the following comments and suggestions:

Major comments

H3K9me3 mapping would be a clear plus to complete the study, especially considering this repressive modification is generally associated with repeats and given the observed enrichment of H3K27me3 over transposable elements in *Salpingoeca rosetta*.

It seems that the genome is rather compact and that the intergenic (promoters) regions are very small. The window -1.5 kb to +1.5 kb around the TSS often covers in fact the promoters of two divergent genes (example on figure 3). Segmentation of the genes according to their relative orientation (divergent or not) would clarify the observed patterns, and especially that of cluster 4. If the RNA-seq data is strand-specific these can be incorporated into the heatmaps to show directionality.

Minor comments

1. The ATAC-seq signal is found mostly near gene promoter regions and reveals a clear correlation between TSS enrichment and transcript levels. While “80% had their midpoint with a distance of -500 to +100 bps from a predicted TSS” (Lanes 133-134), I wonder what the 20% remaining correspond to. Could the authors explicit what “the few distal regulatory elements (Lane 135-136)” are? It would be appreciated to see these data. Maybe an example screenshot. Additionally, an expanded discussion on the possibility of the existence of enhancers would be appreciated (as opposed to simply a lack of accessibility outside promoter regions). Is there simply no room for enhancers, considering average intergenic length? What about average gene length (intronic enhancers)? Is the *C. owczarzaki* genome also very compact with little intergenic DNA?

2. Histone modifications were identified by mass spectrometry. Can the data be used to evaluate the abundance of each of modification detected? Or at least those that are the focus of the manuscript? Were modifications on H2A found, such as H2AK119ub?

3. Please explain why the ChIP analyses are performed on slow swimmers (Lane 152). Is it because slow swimmers are the default state? This important point could be reiterated.

4. Cluster 2 genes are slightly enriched for H3K27me3 but what is striking is the levels of H3K27me3 upstream the TSS on the heatmap Figures 3A-B. What is the explanation? If the upstream H3K27me3 persists after segmenting the analysis into + and – strand genes, can you explain why?

5. In Figure 4D, the upregulation of the gene in thecate cells marked with H3K27me3 in swimming cells is not obvious (due to scaling with highly expressed adjacent upstream gene), and intron annotations are missing. Perhaps a log scale for RNAseq data?

The snapshot on Figure 5A is not very clear: 1. The scale is not shown. 2. it would be nice to display H3K27me3 enrichment at all subtelomeric regions mentioned and 3. transposon annotations are missing (are these LTRs?). Please add the gene annotations to the snapshot as well.

What does “the remaining contig had a large gap in this region” (Lane 198) mean? Please clarify. Please show this entire chromosome in a supplement with the gap shown.

6. The section on H3K27me3 at repeats is incomplete. Apparently, only a subset of transposable elements (LTR/gypsy-like retrotransposons and mutator-like element (MULE) DNA transposons) is included in the analysis. I do not understand why any annotated transposable elements should not be included in this first genomics analysis. You could show the relative enrichment of annotated TEs as a pie chart. This would allow a better description of which transposons are marked by H3K27me3. Again, the comparison to H3K9me3 would be of great interest here.

Please include the RNAseq data analysis as one would like to see the transcriptional status of these transposons in the different cell types.

Lanes 206-7: these sentences belong to the figure legend. “The example in Fig. 5C is the same genomic region as in Fig. 5E but zoomed out. The cluster 2 gene shown in Fig. 4E is annotated with a green arrow.”

Cite and discuss comparison with Dumesic et al. Cell 2015 (Cryptococcus subtelomeres).

7. Figure 3C/D: What is the rationale for comparing slow swimmers vs thecate cells and not other cell types? E.g. fast swimmers or rosette.

8. In general, the ATAC profiles of slow swimmers would be welcome in the heatmaps. Figure S3E nicely shows lack of accessibility of Cluster 2 for example.

9. The presence of H3K27me3 in gene bodies could be interpreted by a lack of transcription-coupled H3K36me3, i.e. promiscuous PRC2 activity.

10. The presence of H3K4me1 in gene bodies could be interpreted as residual transcription-coupled H3K4 histone methyltransferase activity. Perhaps there are no dedicated H3K4me1 histone methyltransferases as there are in animals, and all H3K4me1 is simply “on its way” to becoming H3K4me3. i.e. H3K4me1 over a transcriptionally silent gene body could be a consequence of being surrounded by two strongly transcribed promoters. Are there examples of very long and silent genes where H3K4me1 cover the entire gene body?

11. Materials and Methods: some editing is needed there.

Lane 303: “Luna cell counter (manufacturer ...it’s in my genome editing paper)”

Lane 320: What is the TD buffer?

Lane 331: it is unclear what “the remaining 45 µL reaction” is. No volume mentioned before. Please clarify.

Lane 332: what is “SPRI bead cleanup”?

Lane 431: what is the reason “to extract sub-nucleosome sizes reads, i.e. reads with insert sizes less than 100 bps” for ATAC-seq data analysis?

Lane 438: Please explain why “Sub-nucleosomal reads were used as input for Genrich while all reads were used for MACS2”.

Why do the authors state they do not do peak calling in the Reporting Summary file?

Lane 459: any normalisation used in DESeq2? Also, which version? Default behaviour changed at some update.

12. Data availability: Please add the raw MS data repository.

13. Figures:

- In all heatmaps: The unit for the scale is missing.

- Please define “IQR” in Figure 3.

- Figure S4 typo “DNA(MULE) transposons”

- What is the difference between Figure 5B and Figure S4C? The figure legends are confusing.

Lane 645. Reference 69 is incomplete.

Lanes 201-202: “sitting on top of”...

Lane 206: no 5E – should be 4E?

Reviewer #2

(Remarks to the Author)

“I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.”

(Remarks to the Author)

Review of the manuscript entitled "Chromatin profiling identifies putative dual roles for H3K27me3 in regulating transposons and cell type-specific genes in choanoflagellates", authored by James Gahan et al.

The present manuscript is about the characterization of histone post translation modifications (hPTMs) in *S. rosetta*, a choanoflagellate species. Because Choanoflagellata is the sister clade to Metazoa, this research directly impacts the understanding of the evolution of genome regulation and organization. Moreover, the direct implication of the Polycomb Repressive Complex (PRC) in gene regulation in developmental processes in animals, adds the interest of understanding if the transition of the different life cycle morphologies of *S. rosetta* could also be governed by the PRC, more exactly (PRC2). The manuscript presents interesting data which is technically appropriate, nevertheless, these reviewers think insufficient to sustain the claims proposed by the authors.

Claims of the Manuscript

1) *S. rosetta* appears to be devoid of the long-distance regulatory regions, such as called the enhancers present in animals

2) H3K27me3 decorates those genes with cell-type specific expression

3) A subset of genes with the presence of H3K27me3, also contain H3K4me1, suggesting a bivalent state in these cases

4) Altogether, the above evidences suggest that the histone code that signals for genes involved in development, was already present before animal multicellularity appeared

Following we specify our comments concerning each of these claims, in major and minor, and also suggest experiments and literature additions to improve the work.

Major comments:

About experimental data:

1- The authors perform Mass spect in slow swimmers and thecate cells. Chip-Seq for slow swimmers. ATAC-Seq for all three types of cells, and they use RNA-seq data from a different study.

One of the main claims of the study is that a subset of the genes upregulated in thecate cells correspond to genes in Cluster 2. Cluster 2 genes are the ones marked by H3K27me3 in slow swimmers. This is a nice correlation that suggest this claim, but because it is an important claim, should be well demonstrated. In this study the authors are not really demonstrating that in thecate cells those upregulated genes show a lower level of H3K27me3 and therefore can not conclude that this might be the regulatory mechanism.

Following the authors make a second claim about a putative bivalent state working for regulating these genes, which involved the H3K4me1 mark. Again, this is very interesting but the authors do not have any direct indication that this mark is involved in activating transcription or avoiding completed repression. Actually, in other organisms this mark has not been involved in activation but in repression.

Moreover, these two marks are unexpectedly over the gene body instead of at the promoter regions. It is possible that chromatin regulation in *S. rosetta* responds to a different distribution, but I think they should try to experimentally demonstrate both claims. How would these marks help to open the chromatin at the TSS from the inside the gene body? It would be important to compare the ATAC-seq data with CHIP-Seq (or cut&tag data) for thecate cells from the same culture.

Interestingly, in this publication from a unicellular red algae red alga *Cyanidioschyzon merolae* (Mikulski P, et al. (2017)

Characterization of the Polycomb-Group Mark H3K27me3 in Unicellular Algae. *Front. Plant Sci.* 8:607. doi:

10.3389/fpls.2017.00607) they demonstrate that genes with H3K27me3 inside the gene body are the ones with higher repression.

The authors do not give any explanation in the text about why they use different substrates (cell types) for each experiment. If they have experimental difficulties for doing CHIP-seq in thecate cells, they could explain the reasons. Although this would be the ideal solution, alternatively, they could target specific genes expressed in thecate cells and CHIP for H3K27me3 and H3K4me1 with primers for the promoter region and for body gene. Taking some examples for cluster 2 genes, two genes upregulated in thecate, two genes not upregulated in thecate and two from another cluster (1,3 or 4) as control would be enough.

Moreover, it would be important that Chip-seq data and RNA-seq data is done with the same ongoing culture, under the same laboratory conditions and with the same number of passages for the cells. Alternatively, they could perform q-RT-PCR for the same genes chosen for the suggested experiment above and do it in parallel with the same culture.

Related with the above: There is poor correlation in the level of explanation along the text, the plots in the figures and the corresponding figure legends. This together with the fact that not all experiments have been done with the same cell types, makes the interpretation of the data quite confusing. For example, from line 162 to line 168, the results from ATAC-seq (that are done in all cell types) are related to chromatin marks (H3K27ac, H3K4 me1/2, and later H3K27me3) which are done only in slow swimmers, and then conclusions are taken. Even looking at the figures, one has to concentrate in which type of cells each assay is performed and therefore try to understand from where the general conclusion comes from. We think the data shows interesting indications but does not provide proofs.

2) The authors also claim that the decoration of H3K27me3 over Transposable Elements (TEs) correlates with the specificity of this mark seen in in TEs other organisms. The authors localize retrotransposons at the subtelomeric regions and after that they further look to identify and annotate other TEs.

2a) It is not clear where the other two main superfamilies identified, LTR/gypsy-like and MULE are also located at the subtelomeric regions or elsewhere in the genome.

2b) The authors use another study (Southworth, J et al.) to better annotate the TEs. In (Southworth, J et al.) it was

demonstrated that several families of TEs in *S. rosetta* are active and are being expressed. Nevertheless, the authors do not mention or discuss this piece of data. It would be important to know if Gaham et al. are referring to only the ones located at the subtelomeric regions or also elsewhere in the genome.

2c) The authors claim that TEs in *S. rosetta* are likely silenced by H3K27me3 as has been demonstrated in other eukaryotes. Nevertheless, it has also been demonstrated in several organisms that subtelomeric regions are often decorated by H3K27me3 and PRC2 has been shown to be important for repressing expression in this telomeric compartment. This presents the opportunity to clarify if H3K27me3 is marking and silencing TEs or subtelomeric regions, or both. Again, as in point 1, it would be desirable to have individual Chip-seq in thecate cells for some TEs known to be in the subtelomeric regions or otherwise

2d) Lines 204 to 209 are really confusing. It is hard to navigate the results explained in this paragraph with the different sections of the different figures cited. I think some reorganization and better explanation correlating results and figures could help.

These referees do not have any comment of the results of the ATAC-seq experiments and the conclusions that mostly cis-regulatory elements in *S. rosetta* seem to be close to TSS and the actual data suggests that, as in *C. owczarzaki*, enhancers or long-distance regulatory elements are not present in this genome.

About the text and bibliography

-The manuscript is missing some key references that are directly related to the topic. One of them is Grau-Bové et al. 2022. In this publication there is information regarding PRC2 and PRC1 in other unicellular Holozoans and also in two Choanoflagellate species. The authors cite the paper in the Supplementary section (the old bioRxiv version) regarding the methods used, but it is not cited in the main text, and it should be there. Also Mikulski et al 2017 is missing, where they characterise H3K27me3 in unicellular algae.

- In the introduction (lines 86-90), the authors claim that “despite their importance, repressive PTMs have thus far not been studied in unicellular holozoans”. This is not true, they are studied in the forementioned paper (Grau-Bové et al. 2022) where there is information on *Capsaspora owczarzaki* (Filasterea), *Creolimax fragrantissima* (Ichthyosporea), *Corallochytrium limacisporum* (Corallochytrrea), and two Choanoflagellates (*S. rosetta* and *Monosiga brevicollis*).

-Related with the above. There is no discussion on the other unicellular Holozoans at the end of the discussion (line 276), and this is needed. The discussion, in general, could be much more elaborated, having into account previous works. Again in (lines 243-246), reference to the other unicellular organisms studied in other works, such as Grau-Bové et al 2022 is missing.

-Also in the discussion when they comment on the results of the different clusters where they only observe modifications downstream of the TSS, they must mention that this was also observed in *Capsaspora* in Sebé-Pedrós et al. 2016.

-There are many references to future work to be done, but some of it could have been done for this paper. For instance, screening the other Choanoflagellates taking advantage of the data already available in the search for PRC1 or the variant PRC1 complex that is found in *S. rosetta*.

Comments on methodology

-What was the criteria to decide the number of cells used for each experiment? 500 million cells for ChIP-seq experiments seems too much. Was there a particular reason for that?

-The protocol used to treat the cells before extracting the nuclei seems too aggressive. Why didn't the authors use a less stressful technique, such as digitonin permeabilization?

-Was any filtering applied for contamination in the raw RNAseq data? Please, cite it.

Minor comments

-One reference is missing in line 303.

-There is a typo in line 134, where Fig 1D is cited instead of Fig 2D.

-Again and in general, it is not clear what experiments are the authors referring to in the text. While explaining the results, the text jumps from slow swimmers to fast swimmers or thecates. There are no results shown or discussed about the rosettes. It is important to be clearer on which experiments are being described in the results section, and that they are also further discussed and put in context with the literature.

-In line 158 the authors describe results about slow swimmers and thecate cells, but they don't mention fast swimmers, why?

-Clusters 1, 2, 3, 4, 5 are cited in different ways (cluster 2, cluster2, Cluster 2...)

Reviewer #4

(Remarks to the Author)

I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

Reviewer #5

(Remarks to the Author)

Gahan and colleagues have produced an excellent manuscript on DNA modifications in the genome of the choanoflagellate

Salpingoeca rosetta. The work appears to have been performed to a high standard and is an important study, as information on gene regulation in the closest relatives of animals remains limited. The authors show a second holozoan protist that lacks distal gene enhancers, however promoter sequences are characterized by histone modifications. The conclusions drawn by the authors are appropriate and appear justified on the presented results. I should state at this point I have no past experience in working with ATAC-seq.

I believe the manuscript will have broad appeal to readers from a variety of research fields. There would be a clear interest from the choanoflagellate community, but the research will also be of considerable interest to workers studying gene regulation, both in animals and other model organisms.

One issue is the emphasis on transposons in the title, which did not appear justified in the manuscript. A simple solution to this would be for the authors to remove transposon (which is not used in the technically correct manner) from the title, so that the emphasis is on host gene regulation. A more satisfying approach would be to expand their analyses on DNA modifications in transposable element sequences in the *S. rosetta* genome.

Major Comments

Transposons are mentioned before host genes in the title of the manuscript, so it is a little surprising that they only appear in one paragraph at the end of the Results section. As full-length annotations of 20 *S. rosetta* transposable element families are available (Southworth et al. 2019, *Mobile DNA*, 10:44), it would improve the manuscript if the authors could perform a more detailed analysis of the chromatin modification and the different TEs. The Southworth study analyses showed that individual families have quite different population dynamics and it would improve the submitted manuscript if the authors could show whether these differences are reflected in the DNA modifications present in TE DNA.

The authors have analysed LTR/gypsy-like retrotransposons and report on them as a group, however there are a minimum of seven distinct gypsy-like families within the *S. rosetta* genome, with both chromoviral and non-chromoviral gypsy-like families present. Published RNA-Seq data indicate considerable variation in the expression levels between families. It would improve the paper if they could report on analysis of individual gypsy-like families.

Although not commented upon here in the current manuscript, the Southworth study provided multiple lines of evidence (e.g. identical paralogous copy number, RNA-Seq read number and strong selection on codon usage) to show that copia-like families, specifically *Srospv2* and *Srospv3*, rather than gypsy-like elements or transposons, are the most active TE families in the *S. rosetta* genome. It feels an omission on behalf of the authors to not report on DNA modification within those two highly active families and the copia-like families in general. Furthermore, the analyses of transposons would be improved if the authors looked beyond the MULE-like family and also analysed the other transposon families known to be present in the genome.

Minor Comments

1. General: The term transposon should really only be used for Class II, DNA-based, transposable elements and should not be used as a generic term for all transposable elements. Ideally, the authors should replace the term transposon with transposable element, or TE, to avoid confusion, unless they are specifically discussing Class II transposable elements (transposons).
2. General: A brief explanation of ATAC-sequencing would assist the generalist reader.
3. Introduction, Page 1, lines 61-63: The absence of enhancer sequences in *Capsaspora* is thin evidence for their metazoan origin. The absence is also consistent with an earlier holozoan origin and their subsequent loss in the *Capsaspora* lineage. The authors should tone down, or reword, this statement.

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)

Most comments have been fully addressed and we congratulate the authors on the much-improved revised manuscript.

There are still some points listed below that should be addressed to improve clarity:

1. The authors argue that intergenic distal ATAC-seq peaks correspond to unannotated TSSs, and this is nicely illustrated in Supplementary Fig. 2, yet these correspond to intra and not inter-genic peaks. This should be corrected.
2. The fact that the mass spectrometry data cannot be used to determine stoichiometry of individual histone modifications is worth mentioning somewhere (perhaps in the Materials and Methods section). Same comment for the lack of detection of H2AK119ub.
3. Regarding H3K27me3, there seem to be three “flavors” of H3K27me3 domains that would be worth displaying in the main Figure 6 as screenshots. In addition to the snapshot of the cluster of LTR retrotransposons, the authors could display an inactive gene body and a gap-less sub-telomeric region (one shown in Supplementary Fig. 6).

4. The expanded ATAC-seq analyses are very welcome. One major concern about these new heatmaps based on histone PTM enrichment over individual TEs is whether individual TEs are unique in sequence in rosetta. i.e. are they mappable? Are the MAPQ scores for aligned reads over individual TEs above 10 for example? If not, perhaps a “small family”-based analysis is more appropriate (breaking down TEs into SrosH, SrosM, etc).

Assuming TEs in rosetta are mappable, I still have a minor concern about the legibility of the revised heatmaps. For both Fig. 5 and 7, I suggest you shuttle the families with a few individual elements to the supplemental, resulting in simplified and more legible main figures. Metaplots (tops of heatmaps) are uninterpretable with these many classes. You could apply a cutoff of 20 elements for example. I realize this may more or less recreate the TE heatmap from the initial version with the addition of copia elements. However, the full heatmaps, as currently presented in the updated Fig5 and 7, would still of interest as supplemental figures.

Alternatively, and this is a non-issue, only a suggestion, instead of heatmaps, 2D scatterplots showing individual element expression (x-axis) versus H3K27me3 enrichment over the TSS (y-axis) would be clearer. You could have one scatterplot for each TE family (small or large), or one big plot with data points colored by family. Expression vs K27me3 is the main comparison anyway, the other PTMs and accessibility could still be kept as supplementary information.

Reviewer #2

(Remarks to the Author)

I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

Reviewer #3

(Remarks to the Author)

Second revision of the manuscript entitled: Chromatin profile identifies putative dual roles H3K27me3 in regulating cell type-specific genes and Transposable Elements in choanoflagellates. Authored by James Gahan et al.

In this new version the text has substantially gained clarity and key references for the conceptual framework have been added improving the overall quality of the manuscript.

We agree that the authors demonstrate the role of the H3K27me3 chromatin modification as likely the on and off switch for some developmental genes in *S. rosetta*. We think this is a sound and important claim which the scientific community will receive as important knowledge concerning evolution of genome regulation and developmental control. Nevertheless, we have some important comments to different aspects of this work.

Major

- The authors have made an effort to expand their work on TEs and have included three figures (5, 6 and 7) where they combine their data with previous RNA-seq data (Southworth et al 2019 Mobile DNA). The insight from these figures is just reflected in a total of four lines in the results (216 to 220). These three figures contain quite a substantial amount of data that is poorly reflected, in our opinion, in the results and in the discussion sections.
- Moreover, figure legends of these figures are not sufficiently explanatory. For example, for Figures 5 and 7, what does the code color for the different subfamilies of TEs (right above corner) mean? It is unclear whether these colors represent the different TEs in the different profiles (ATAC-seq, ChIP-seq)? If that is so, at the scale that the profiles are done is impossible to see anything.
- Also for Figures 5 and 7, very important, are the horizontal lines inside the box of each TE subfamily, corresponding to different copies of each specific subfamily, so that one can follow the amount of H3K27me3 and expression in the different cell types for each copy?
- If, in both Figures 5 and 7, the horizontal lines from the RNA seq data and histone modifications correspond to individual copies of a specific TE there are several occasions where a specific copy is both highly expressed in slow swimmers and strongly decorated with H3K27me3. If this interpretation is correct, they should be consistently mentioned as “copies” rather than “each transposon” throughout the figure legends to avoid confusion.
- Checking the data from Southworth et al., Mobile DNA (2019), Srospv2 has 17 paralogous copies, and in the Figure 5 of Gahan et al., there are many more horizontal lines. Same for Figure 7. These are important figures for understanding the extent to which H3K27me3 and TE transcription support one of the main claims of the manuscript. Right now, these numbers are unclear, and the figure legends are minimal.
- While the Figure legends mention the presence or absence of H3K27me3 on different classes of transposable elements (TEs), they do not explain why this is biologically meaningful. Consider briefly describing somewhere in the text the functional significance—how does H3K27me3 relate to the transcriptional regulation or silencing of LTR retrotransposons versus DNA transposons?
- Ensure the description of ATAC-seq, ChIP-seq, and RNA-seq data is uniform across both figures. In Figure 5, RNA-seq is described as “DESeq2 normalized counts,” while the corresponding description in Figure 7 is slightly different. Consistent phrasing will enhance clarity. Please, clarify how TE subfamilies are defined (e.g., based on prior annotations) and whether copy numbers were filtered or normalized in the presentation of the heatmaps.

- Related to the above point, Line 217-218: The authors state that H3K27me3 levels drop in LTRs that are substantially expressed. Since Chip-seq data is only available for slow swimmers, it is unclear whether this statement refers specifically to expression during this stage or to highly expressed LTRs across all stages. Clarifying whether these observations reflect increased expression in slow swimmers or in other stages would strengthen the argument. We trust that further examination of this relationship could provide stronger support for the claim in Line 222 regarding H3K27me3 regulating TE expression. Maybe this result is already in Figure 5, but the current figure legend is not self-explanatory to make this interpretation clear.
- Also Figure 7 shows that H3K27me3 is not significantly present on DNA transposons, but there are signals of H3K4me1 and H3K4me3 that the authors could discuss in the discussion. Exploring whether these histone marks suggest active regulatory regions or other functional implications would provide a more comprehensive interpretation of the chromatin landscape surrounding DNA transposons.

- Also, it is necessary to have the reference from Southworth et al Mobile DNA (2019) correctly cited in the figure legends where the data is used and again in the M&M section, where it is only cited as Southworth et al.
- IMPORTANT: Also, in Figures 5 and 7 legends: TEs are classified in families and Superfamilies, not “bigger” families.
- Figure 6 legend should contain the name or code for the “other genes” shown in the figure.
- Last but not least, and very importantly, the final message is not clear. Is the main conclusion that H3K27me3 marks developmental genes, LTR transposable elements, and sub-telomeric regions? If so, which additional histone or chromatin marks would enable developmental genes to be turned on and off when needed? In other words, which chromatin factors keep LTRs and sub telomeric regions repressed? Since the authors have detected H3K9me3 in *S. rosetta*, I think it is likely and worth speculating on a possible role of H3K9me3 as this possible additional mark for repressing sub telomeric and TEs sequences. This speculation is particularly relevant since H3K9me3 is well-known to mediate transposon and heterochromatin (including subtelomeric sequences) silencing in other organisms.
- Related to the above, it would be important to mention in the discussion the lack of experiments regarding H3K9me3 exposing the particular experimental difficulties. Any reader with knowledge in genome regulation would wonder why this is not inspected, especially after mentioning that the authors have actually identified its presence in *S. rosetta*. Moreover, the particular difficulty with the antibody for this modification will be of interest for the scientific community working on non-model organisms which might present the same threonine substitution.

Minor comments

- Finally, the authors devote quite a substantial portion of the manuscript discussing enhancers, which *S. rosetta* does not seem to have, at least with the current dataset (Intro; line 57 to 67, Results; lines 122 to 141 Discussion; 225-243). While we understand that it is important to determine the existence of distal regulatory elements in unicellular holozoans, since choanoflagellates are the closest relatives of animals, and only in *C. owczarzaki* the presence of enhancers has been inspected, we feel that the discussion of enhancers may be disproportionate to the available evidence. This contrasts with the emphasis placed on TEs, which is highlighted in the title. It would be helpful to achieve a better balance between these two aspects, aligning the title's focus on TEs with an in-depth discussion of their role in *S. rosetta*.
- Line 64 “...enhancer are truly animal-specific”. We understand the intention of this definition might be to distinguish metazoans from unicellular holozoans since the authors are working inside the Holozoa branch, but enhancers are also present in plants, and therefore we think this sentence could be misleading.
- References not cited in the text: 44 and 47, which is a pity since they look interesting.
- Results from line 148 to 151 miss the figure citation
- Line 873-874 seems to contain an error. The term “grouped” should be replaced with “groups” to maintain grammatical consistency.

Reviewer #4

(Remarks to the Author)

I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

Reviewer #5

(Remarks to the Author)

Gahan and co-workers have undertaken many of my original suggestions, which I feel has improved the resulting manuscript. Where they have disagreed with my comments, I view their rebuttals to be fair and reasonable. I consider the resubmitted manuscript an excellent piece of work, which should be of considerable interest to researchers in a broad range of fields. As a result, I recommend it is accepted for publication in Nature Communications.

My only minor suggestion is that the authors tighten up their word formatting a little. Slow swimmers are consistently written with an upper case “S”, whilst fast swimmers are always written with a lower case “f”. Neither term appears to be a formal noun, so lower case letters would be more appropriate. This also applies to thecate cells, which is written as Thecate suggesting it is a formal, proper noun rather than a descriptive term. In addition, binomial species/genus names should be written consistently in italics, whereas at present there are a number of examples where names are presented in regular font (for example, on page 13 of the main manuscript, *Capsaspora owczarzaki* is written with the genus name in regular font, but species name in italics).

Version 2:

Reviewer comments:

Reviewer #1

(Remarks to the Author)

The authors have satisfactorily addressed all the comments and concerns raised in the previous review. I recommend that the manuscript be accepted for publication.

Reviewer #2

(Remarks to the Author)

I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

Reviewer #3

(Remarks to the Author)

We are pleased with all the changes that the authors have applied to this latest version of the manuscript. The manuscript reads much better, the figures are more correct and understandable and the bibliography and the discussion are much more complete.

We do not have any further comment and consider that this is an important investigation for the community working on developmental evolution and consider that is suitable for publication in Nature communications.

Reviewer #4

(Remarks to the Author)

I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

Open Access This Peer Review File is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

In cases where reviewers are anonymous, credit should be given to 'Anonymous Referee' and the source.

The images or other third party material in this Peer Review File are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons

license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

To view a copy of this license, visit <https://creativecommons.org/licenses/by/4.0/>

Response to Reviewers

We thank the reviewers for taking the time to evaluate our manuscript. We have included many of their suggestions in the revised version and we believe their input has helped to improve the manuscript overall. In particular, we now include a vastly expanded analysis of transposable elements which adds significantly to the findings we presented in the original version of the manuscript. Although the conclusions are the same, we believe this expanded analysis will make the manuscript even more interesting to a wider audience interested in TE biology. We include below a point-by-point response to the reviewers' comments.

REVIEWER COMMENTS

Reviewer #1 (Remarks to the Author):

The manuscript by Gahan et al. examines chromatin accessibility and 4 histone modifications genome-wide in the choanoflagellate *Salpingoeca rosetta*, a sister group to animals. The data are of outstanding quality and the analyses provide a thorough description of histone marks with respect to gene expression levels and accessibility. This is an interesting, well-done, yet entirely descriptive study.

We thank the reviewer for their comments and for highlighting the quality and interesting nature of our data.

The work is sound but would improve with the following comments and suggestions:

Major comments:

H3K9me3 mapping would be a clear plus to complete the study, especially considering this repressive modification is generally associated with repeats and given the observed enrichment of H3K27me3 over transposable elements in *Salpingoeca rosetta*.

We agree fully with the reviewer that understanding the distribution of H3K9me3 would be very interesting. We have tried using commercial antibodies against H3K9me3 for ChIP but this was unsuccessful. We believe this may be because in *S. rosetta* there is a threonine at position 10 in histone H3 as opposed to the serine present in most other species (Reviewer Figure 1). In the future, we plan to try to generate custom-made antibodies to recognize H3K9 methylation in *S. rosetta* but this is outside the scope of what is possible in this study.

SrH3.1	MARTKQTARKTTGGKAPRKQLATKAARKSAPSTGGVKKPHRFPGTVALREIRRYQKSTE	60
HsH3.1	MARTKQTARKSTGGKAPRKQLATKAARKSAPATGGVKKPHRYRPGTVALREIRRYQKSTE	60
	*****.*****.*****.*****	
SrH3.1	LLIRKLPFQRLVREIAQDFKTDLRFQSTAVSALQEAAEAYLVNLFEDTNLCAIHAKRVTI	120
HsH3.1	LLIRKLPFQRLVREIAQDFKTDLRFQSSAVMALQEACEAYLVGLFEDTNLCAIHAKRVTI	120
	*****.***.*****.*****	
SrH3.1	MPKDIQLARRIGERS	136
HsH3.1	MPKDIQLARRIGERA	136
	*****.	

Reviewer Figure 1. Alignment of *S. rosetta* and human histone H3 with S10 highlighted.

It seems that the genome is rather compact and that the intergenic (promoters) regions are very small. The window -1.5 kb to +1.5 kb around the TSS often covers in fact the promoters of two divergent genes (example on figure 3). Segmentation of the genes according to their relative orientation (divergent or not) would clarify the observed patterns, and especially that of cluster 4. If the RNA-seq data is strand-specific these can be incorporated into the heatmaps to show directionality.

The average intergenic distance in *S. rosetta* is 885bp so indeed many of the selected windows contain the TSSs of two genes. This actually fully explains the difference between cluster 1,3 and 4. We apologise that this was not clear in the original version of manuscript, but we have now stated this more explicitly (Lines 170-176). Almost all genes in Clusters 1 and 3 are oriented with another divergent active promoter directly upstream while cluster 4 genes are not. We could segment genes based on this, but it would produce the same result as what we already show.

The RNAseq is not strand specific and therefore we do not think would help to highlight this point.

Minor comments:

1. The ATAC-seq signal is found mostly near gene promoter regions and reveals a clear correlation between TSS enrichment and transcript levels. While “80% had their midpoint with a distance of -500 to +100 bps from a predicted TSS” (Lanes 133-134), I wonder what the 20% remaining correspond to. Could the authors explicit what “the few distal regulatory elements (Lane 135-136)” are? It would be appreciated to see these data. Maybe an example screenshot. Additionally, an expanded discussion on the possibility of the existence of enhancers would be appreciated (as opposed to simply a lack of accessibility outside promoter regions). Is there simply no room for enhancers, considering average intergenic length? What about average gene length (intronic enhancers)? Is the *C. owczarzaki* genome also very compact with little intergenic DNA?

We thank the reviewer for bringing up this point. The majority of the “other” peaks are, we believe, TSSs that are incorrectly annotated in the genome. We have now included a new supplementary figure (Supplementary Fig. 2) which shows two examples of intragenic “distal” peaks that are in fact clearly unannotated TSSs as well as commenting on this in the results (lines 137-139). We do not entirely exclude the possibility of additional peaks being genuine distal-peaks and in the future, we would be interested to investigate any potential role of those peaks. We are currently in the process of re-sequencing the *S. rosetta* genome and plan to then generate a better genome annotation which would aid in this.

We have added some sentences to the discussion to expand upon what we already had in relation to enhancers and to comment specifically on genome size and the lack of intronic enhancers (lines 230-247). We do not feel that more than this is necessary as anything beyond what we have already concluded would be speculation.

2. Histone modifications were identified by mass spectrometry. Can the data be used to evaluate the abundance of each of modification detected? Or at least those that are the focus of the manuscript? Were modifications on H2A found, such as H2AK119ub?

Unfortunately, the data we have cannot be used to determine stoichiometry of individual modifications. It can only be used to calculate relative changes between samples. This is because we are not correcting for ionization efficiency differences of individual peptides which can be very high (see (Lin et al., 2014)). In terms of ubiquitinylation of H2A, the preparation and analysis method we used would not have been able to detect ubiquitinated peptides but even if we did a preparation for ubiquitin detection, the peptide which would be generated from SrH2A containing the sites homologous to H2AK119ub (NDDELSKLLSGVTIAQGGVLP~~HI~~HSNLIPKGKGKKKAASQSQEY) would be too long to be detected.

3. Please explain why the ChIP analyses are performed on slow swimmers (Lane 152). Is it because slow swimmers are the default state? This important point could be reiterated.

Slow swimmers are the state in which we normally grow *S. rosetta* cells in the lab. The choice to perform ChIPseq only on slow swimmers was, however, technical. As discussed in detail below, we were unable to optimise a ChIP protocol that worked in Thecate cells. In the future we hope to overcome this technical difficulty but to date we have not been successful.

4. Cluster 2 genes are slightly enriched for H3K27me3 but what is striking is the levels of H3K27me3 upstream the TSS on the heatmap Figures 3A-B. What is the explanation? If the upstream H3K27me3 persists after segmenting the analysis into + and – strand genes, can you explain why?

Cluster 2 genes are characterized by high levels of gene-body H3K27me. This can be seen in Figure 3A/B where the signal downstream of the TSS is stronger than that upstream. It is also evident in Figure 4D/E. Dealing with this point from all cluster 2 genes is tricky as it contains some TEs, so we have now added a section later in the manuscript where we deal with this point directly for those genes in Cluster 2 which overlap with genes upregulated in thecate cells (Supplementary Fig. 5). We have sub-clustered these genes and show they fall into two groups: Cluster 2A are genes with H3K27me3 located both upstream and downstream of the

TSS and Cluster 2B are genes where H3K27me3 is located exclusively downstream. Similarly to the case of active genes this is related to gene order. Cluster 2A are genes which are labelled by H3K27me3 over their gene body where the upstream gene is also labelled by H3K27me3. Cluster 2B genes represent genes labelled by H3K27me3 where the upstream gene is not labelled by H3K27me3. We have provided examples of these two types of genes (Supplementary Fig. 5a, d) and discuss this in the text (lines 195-198).

5. In Figure 4D, the upregulation of the gene in thecate cells marked with H3K27me3 in swimming cells is not obvious (due to scaling with highly expressed adjacent upstream gene), and intron annotations are missing. Perhaps a log scale for RNAseq data?

We have now replaced this with a different gene where the upregulation in thecate cells is more obvious and have also included the annotation of introns.

The snapshot on Figure 5A is not very clear: 1. The scale is not shown. 2. it would be nice to display H3K27me3 enrichment at all subtelomeric regions mentioned and 3. transposon annotations are missing (are these LTRs?). Please add the gene annotations to the snapshot as well.

We have now moved this to the supplementary material and show all 6 super-contigs in the new Supplementary Fig. 6. We have also added gene annotations but there are no TEs in these regions.

What does “the remaining contig had a large gap in this region” (Lane 198) mean? Please clarify. Please show this entire chromosome in a supplement with the gap shown.

In fact, two out of the 6 have a large gap adjacent to the sub-telomeric sequence which we annotated in Fig. S6. This is a gap in the sequence which happens when mate-pairs are used during the sequencing, i.e. two sequences can be confidently placed on the same contig with a defined distance but the intervening sequence is not known. Showing the whole chromosome is not helpful because the gap is so small that it becomes invisible at that scale.

6. The section on H3K27me3 at repeats is incomplete. Apparently, only a subset of transposable elements (LTR/gypsy-like retrotransposons and mutator-like element (MULE) DNA transposons) is included in the analysis. I do not understand why any annotated transposable elements should not be included in this first genomics analysis. You could show the relative enrichment of annotated TEs as a pie chart. This would allow a better description of which transposons are marked by H3K27me3. Again, the comparison to H3K9me3 would be of great interest here. Please include the RNAseq data analysis as one would like to see the transcriptional status of these transposons in the different cell types.

We have now included a much more comprehensive analysis of TEs. In the new figures we include ChIP-Seq data as well as both ATAC-seq from slow swimmers and RNAseq data from all cell types. This new analysis reveals very similar trends as in the original scaled down analysis, i.e., H3K27me3 is enriched on LTR retrotransposons and not on DNA transposons. This also seems to correlate with expression, i.e. we see more expression of the DNA transposons and also within the LTR retrotransposons we see that those that are expressed

have less enrichment of H3K27me3 (e.g., *Srospv6*). There are now 3 figures (Fig. 5-7) covering this and the corresponding text has also changed (lines 203-223).

Regarding H3K9me3 I guide you to or response above. We do, however hope to do this in the future as we agree that analysing this modification would be very interesting.

Lanes 206-7: these sentences belong to the figure legend. "The example in Fig. 5C is the same genomic region as in Fig. 5E but zoomed out. The cluster 2 gene shown in Fig. 4E is annotated with a green arrow."

We have now removed this entirely as it was confusing.

Cite and discuss comparison with Dumesic et al. Cell 2015 (*Cryptococcus subtelomeres*).

We have added a line to the discussion to compare to *Cryptococcus*, mainly as a future plan! (Line 265-267)

7. Figure 3C/D: What is the rationale for comparing slow swimmers vs thecate cells and not other cell types? E.g. fast swimmers or rosette.

We choose to only compare these two as there are no significant transcriptional differences between slow swimmers and rosettes or fast swimmers (Leon et al., 2024). We also mention this directly in the text (lines 163-165)

8. In general, the ATAC profiles of slow swimmers would be welcome in the heatmaps. Figure S3E nicely shows lack of accessibility of Cluster 2 for example.

We have moved the ATAC heatmap to the main figure (Fig. 3b) and have also included them in our new TE analysis (Fig. 5/7).

9. The presence of H3K27me3 in gene bodies could be interpreted by a lack of transcription-coupled H3K36me3, i.e. promiscuous PRC2 activity.

We agree totally with the reviewer, and we also think that PRC2 may be recruited to genes by virtue of their inactivity whether this be lack of H3K36me2/3, other modifications or simply due to lack of transcription itself. We added a line on this to the discussion (lines 278-281).

10. The presence of H3K4me1 in gene bodies could be interpreted as residual transcription-coupled H3K4 histone methyltransferase activity. Perhaps there are no dedicated H3K4me1 histone methyltransferases as there are in animals, and all H3K4me1 is simply "on its way" to becoming H3K4me3. i.e. H3K4me1 over a transcriptionally silent gene body could be a consequence of being surrounded by two strongly transcribed promoters. Are there examples of very long and silent genes where H3K4me1 cover the entire gene body?

There do not appear to be any extraordinarily long genes in this list (although "long" is hard to define). Instead, we show an example of a gene where both adjacent genes are oriented with the TSS away from the H3K27me3/K4me1-labelled gene (Supplementary Fig. 5d). Here, there are no such active promoters directly adjacent to the gene but there is still H3K4me1 indicating it is unlikely to have spread from an adjacent, expressed gene. We therefore favour the

hypothesis that there is some mechanism to specifically deposit H3K4me1 on these genes. Functional work on KMT2 homologs in *S. rosetta* will be needed to test this.

11. Materials and Methods: some editing is needed there.
Lane 303: "Luna cell counter (manufacturer ...it's in my genome editing paper)"

We have removed this as we mention the manufacturer earlier in the protocol.

Lane 320: What is the TD buffer?

We have changed this to "Tagmentation DNA Buffer" which is supplied as part of the Illumina kit cited at the end of that same line (Line 361).

Lane 331: it is unclear what "the remaining 45 µL reaction" is. No volume mentioned before. Please clarify.

The remaining 45 µL refer to the volume left over after an aliquot is removed for qPCR. We have clarified this now (Line 371).

Lane 332: what is "SPRI bead cleanup"?

We have edited this to add the manufacturer and to indicate that the manufacturer's protocol was used for this (line 373).

Lane 431: what is the reason "to extract sub-nucleosome sizes reads, i.e. reads with insert sizes less than 100 bps" for ATAC-seq data analysis?

This is a standard approach during processing of ATAC-seq data and is based on the principle that reads less than this size are likely derived from nucleosome-free regions.

Lane 438: Please explain why "Sub-nucleosomal reads were used as input for Genrich while all reads were used for MACS2".

We tested both sets of reads using both programs and saw that they performed better with these different datasets. We have included a line in the methods to explain this (line 478-480).

Why do the authors state they do not do peak calling in the Reporting Summary file?

The reporting summary refers to ChIP-seq analysis and we have not performed any peak calling on our ChIP-seq data.

Lane 459: any normalisation used in DESeq2? Also, which version? Default behaviour changed at some update.

We have added the DESeq2 version (v1.38.3) to supplementary table 3 and have added to the methods that we used default parameters, i.e. size factor scaling for normalization (line 494).

12. Data availability: Please add the raw MS data repository.

We have added this to the data availability section.

13. Figures:

- In all heatmaps: The unit for the scale is missing.

The units in the heatmaps generated from deepTools are “read density” which is the normalized read density and is dependent on how the bigwig files were generated, this is described in the methods and we have added “Read density” to all the figures as necessary.

- Please define “IQR” in Figure 3.

We have added this to the figure legend.

- Figure S4 typo “DNA(MULE) transposons”

We no longer have this figure legend in the revised manuscript.

- What is the difference between Figure 5B and Figure S4C? The figure legends are confusing.

This is no longer part of the paper as we included a new analysis of TEs.

Lane 645. Reference 69 is incomplete.

We have fixed this reference

Lanes 201-202: “sitting on top of”...

This line is no longer present in the manuscript.

Lane 206: no 5E – should be 4E?

This is no longer present as we have radically changed this section.

Reviewer #2 (Remarks to the Author):

"I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts."

We thank this reviewer for taking the time to co-review our manuscript.

Reviewer #3 (Remarks to the Author):

Review of the manuscript entitled "Chromatin profiling identifies putative dual roles for H3K27me3 in regulating transposons and cell type-specific genes in choanoflagellates ", authored by James Gahan et al.

The present manuscript is about the characterization of histone post translation modifications (hPTMs) in *S. rosetta*, a choanoflagellate species. Because Choanoflagellata is the sister clade to Metazoa, this research directly impacts the understanding of the evolution of genome regulation and organization. Moreover, the direct implication of the Polycomb Repressive Complex (PRC) in gene regulation in developmental processes in animals, adds the interest of understanding if the transition of the different life cycle morphologies of *S. rosetta* could also be governed by the PRC, more exactly (PRC2).

The manuscript presents interesting data which is technically appropriate, nevertheless, these reviewers think insufficient to sustain the claims proposed by the authors. Claims of the Manuscript

1) *S. rosetta* appears to be devoid of the long-distance regulatory regions, such as called the called enhancers present in animals.

2) H3K27me3 decorates those genes with cell-type specific expression

3) A subset of genes with the presence of H3K27me3, also contain H3K4me1, suggesting a bivalent state in these cases

4) Altogether, the above evidences suggest that the histone code that signals for genes involved in development, was already present before animal multicellularity appeared. Following we specify our comments concerning each of these claims, in major and minor, and also suggest experiments and literature additions to improve the work.

We thank the reviewer for their reading of the manuscript and their comments. We do not agree, however that our claims are not substantiated by the evidence we have presented. We have outlined this in our detailed response to many of the comments below.

Major comments:

About experimental data:

1- The authors perform Mass spect in slow swimmers and thecate cells. Chip-Seq for slow swimmers. ATAC-Seq for all three types of cells, and they use RNA-seq data from a different study.

One of the main claims of the study is that a subset of the genes upregulated in thecate cells

correspond to genes in Cluster 2. Cluster 2 genes are the ones marked by H3K27me3 in slow swimmers. This is a nice correlation that suggest this claim, but because it is an important claim, should be well demonstrated. In this study the authors are not really demonstrating that in thecate cells those upregulated genes show a lower level of H3K27me3 and therefore can not conclude that this might be the regulatory mechanism.

We agree with the reviewer that more data will be needed in the future to establish a causative link between H3K27me3 and transcriptional regulation. We do not, however, state that in thecate cells these genes have a lower level of H3K27me3 as we do not have the data to show this, although this is what our proposed model suggests. What we do show is a strong correlation between genes upregulated in thecate cells and H3K27me3 in slow swimmers which we think is robust based on the data we present.

Following the authors make a second claim about a putative bivalent state working for regulating these genes, which involved the H3K4me1 mark. Again, this is very interesting but the authors do not have any direct indication that this mark is involved in activating transcription or avoiding completed repression. Actually, in other organisms this mark has not been involved in activation but in repression.

As above we believe we have been very careful in describing the patterns of modifications without overstating our results. We put forward a potential mechanism in our discussion which is plausible. We do, however, agree that we did not give enough attention to the possibility that K4me1 is involved in repression. We have added to discussion to more directly state this (lines 308-314). We are not, however, aware of any papers which show a direct role of H3K4me1 in repression.

Moreover, these two marks are unexpectedly over the gene body instead of at the promoter regions. It is possible that chromatin regulation in *S. rosetta* responds to a different distribution, but I think they should try to experimentally demonstrate both claims.

Unfortunately, we do not understand what the reviewer means here.

How would these marks help to open the chromatin at the TSS from the inside the gene body? It would be important to compare the ATAC-seq data with CHIP-Seq (or cut&tag data) for thecate cells from the same culture.

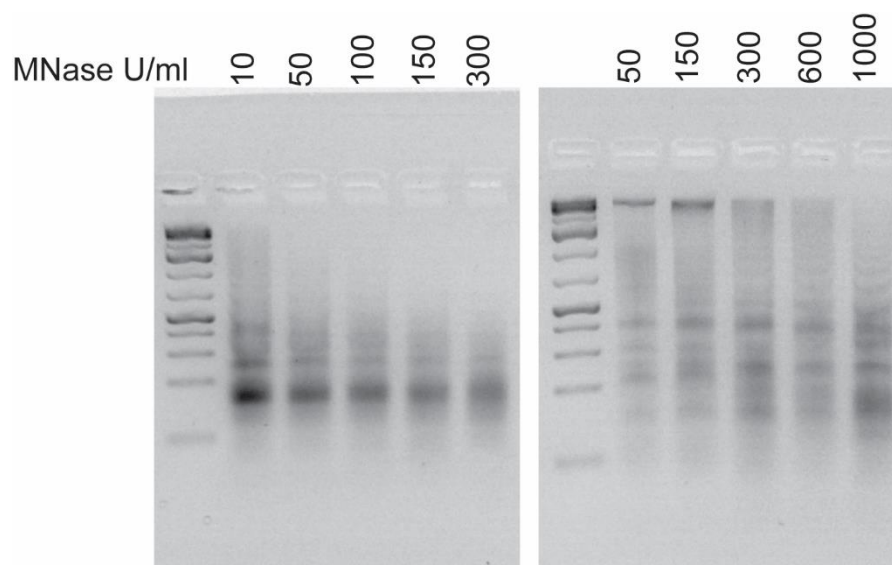
We have not claimed that these marks would “open the chromatin at the TSS from the inside the gene body”. In fact, we are agnostic to any potential mechanisms. We do not know how repeating the experiment in the way the reviewer mentions would be helpful at all in this context.

Interestingly, in this publication from a unicellular red algae red alga *Cyanidioschyzon merolae* (Mikulski P, et al. (2017) Characterization of the Polycomb-Group Mark H3K27me3 in Unicellular Algae. *Front. Plant Sci.* 8:607. doi: 10.3389/fpls.2017.00607) they demonstrate that genes with H3K27me3 inside the gene body are the ones with higher repression.

We thank the reviewer for highlighting this paper. We have included this in our discussion as well as data from *Phaeodactylum tricornutum* (Lines 281/282).

The authors do not give any explanation in the text about why they use different substrates (cell types) for each experiment. If they have experimental difficulties for doing CHIP-seq in thecate cells, they could explain the reasons. Although this would be the ideal solution, alternatively, they could target specific genes expressed in thecate cells and CHIP for H3K27me3 and H3K4me1 with primers for the promoter region and for body gene. Taking some examples for cluster 2 genes, two genes upregulated in thecate, two genes not upregulated in thecate and two from another cluster (1,3 or 4) as control would be enough.

We agree with the reviewer that having ChIP data from Thecate cells would be very useful and the fact we do not have any is not for lack of trying. For reasons we do not yet understand thecate cells are highly resistant to MNase digestion. Even at very high MNase concentrations we are not able to digest the chromatin to anything close to what we could use for ChIP. We have tried multiple ways to improve this, but it is still something we cannot resolve. Given this, it is not possible to perform these experiments. It would also not be possible to do a proper ChIP qPCR with these samples for the same reason. Given the massive difference in digestion, it would be impossible to compare samples. We hope to overcome these difficulties in the future. We have now added a line to the discussion about this (line 277-278).



Reviewer figure 2. MNase digestion of chromatin from swimming cells (left) and thecate cells (right) shows major differences in the ability of MNase to digest chromatin from the different cell types.

Moreover, it would be important that Chip-seq data and RNA-seq data is done with the same ongoing culture, under the same laboratory conditions and with the same number of passages for the cells. Alternatively, they could perform q-RT-PCR for the same genes chosen for the suggested experiment above and do it in parallel with the same culture.

We do not agree that it would be necessary to repeat the experiment in this way. We have no indication that different cultures of the same cell type would have any radical difference in their

gene expression. We also have no indication that passage number would be an issue with *S. rosetta* like it would be for cell culture, i.e., there is no reason to believe there would be any type of replicative senescence in *S. rosetta* cultures. We have, however, changed the example of a Cluster 2 gene to PTSG_09715 (Figure 4e). This gene has recently been experimentally validated as a true thecate-specific gene (Leon et al., 2024).

Related with the above: There is poor correlation in the level of explanation along the text, the plots in the figures and the corresponding figure legends. This together with the fact that not all experiments have been done with the same cell types, makes the interpretation of the data quite confusing. For example, from line 162 to line 168, the results from ATAC-seq (that are done in all cell types) are related to chromatin marks (H3K27ac, H3K4 me1/2, and later H3K27me3) which are done only in slow swimmers, and then conclusions are taken. Even looking at the figures, one has to concentrate in which type of cells each assay is performed and therefore try to understand from where the general conclusion comes from. We think the data shows interesting indications but does not provide proofs

We have added some additional explanations in the text and also have added annotation to the figures to make it clearer which cell types we are discussing.

2) The authors also claim that the decoration of H3K27me3 over Transposable Elements (TEs) correlates with the specificity of this mark seen in in TEs other organisms. The authors localize retrotransposons at the subtelomeric regions and after that they further look to identify and annotate other TEs.

This is incorrect, we do not show any evidence of enrichment of any TEs at sub-telomeric regions. We have shown that 1. H3K27me3 is found in a subset of sub-telomeric regions. These 6 regions are the only scaffolds where the telomeric repeats are assembled in the current genome assembly. 2. It is enriched on transcriptionally silent LTR retrotransposons. We did not claim to have localized retrotransposons at sub-telomeric regions. We have edited the text to make this clearer.

2a) It is not clear where the other two main superfamilies identified, LTR/gypsy-like and MULE are also located at the subtelomeric regions or elsewhere in the genome.

As above, we do not state that any TEs are located in sub-telomeric regions. We see that TEs are spread throughout the genome with some possible enrichment around the centromeres (we have just begun to investigate the centromeres of *S. rosetta* and this is therefore very preliminary). A new *S. rosetta* genome will be necessary in order to fully understand the sub-telomeric regions as they are missing on most chromosome ends in the current assembly.

2b) The authors use another study (Southworth, J et al.) to better annotate the TEs. In (Southworth, J et al.) it was demonstrated that several families of TEs in *S. rosetta* are active and are being expressed. Nevertheless, the authors do not mention or discuss this piece of data,. It would be important to know if Gaham et al. are referring to only the ones located at the subtelomeric regions or also elsewhere in the genome.

Again, we do not mention sub-telomeric TEs in our manuscript. With regard to expression, please see our reply to reviewer number 1. Briefly, we have included an expanded description as well as including the RNAseq data to compare expressed vs non-expressed TEs.

2c) The authors claim that TEs in *S. rosetta* are likely silenced by H3K27me3 as has been demonstrated in other eukaryotes. Nevertheless, it has also been demonstrated in several organisms that subtelomeric regions are often decorated by H3K27me3 and PRC2 has been shown to be important for repressing expression in this telomeric compartment. This presents the opportunity to clarify if H3K27me3 is marking and silencing TEs or subtelomeric regions, or both. Again, as in point 1, it would be desirable to have individual Chip-seq in thecate cells for some TEs known to be in the subtelomeric regions or otherwise

As we have not shown any TEs to be in sub-telomeric regions this is not a point we can address.

2d) Lines 204 to 209 are really confusing. It is hard to navigate the results explained in this paragraph with the different sections of the different figures cited. I think some reorganization and better explanation correlating results and figures could help.

We apologise for this somewhat confusing section. As we have now radically changed our TE analysis, this section is no longer present in the revised manuscript.

These referees do not have any comment of the results of the ATAC-seq experiments and the conclusions that mostly cis-regulatory elements in *S. rosetta* seem to be close to TSS and the actual data suggests that, as in *C. owczarzaki*, enhancers or long-distance regulatory elements are not present in this genome.

About the text and bibliography

-The manuscript is missing some key references that are directly related to the topic. One of them is Grau-Bové et al. 2022. In this publication there is information regarding PRC2 and PRC1 in other unicellular Holozoans and also in two Choanoflagellate species. The authors cite the paper in the Supplementary section (the old bioRxiv version) regarding the methods used, but it is not cited in the main text, and it should be there. Also Mikulski et al 2017 is missing, where they characterise H3K27me3 in unicellular algae.

We thank the reviewer for pointing out these omissions. We have added citations.

- In the introduction (lines 86-90), the authors claim that “despite their importance, repressive PTMs have thus far not been studied in unicellular holozoans”. This is not true, they are studied in the forementioned paper (Grau-Bové et al. 2022) where there is information on *Capsaspora owczarzaki* (Filasterea), *Creolimax fragrantissima* (Ichthyosporea), *Corallochytrium limacisporum* (Corallochytreia), and two Choanoflagellates (*S. rosetta* and *Monosiga brevicollis*).

We apologise for not being more specific in how we describe this. We have now changed the text to reflect the fact that their presence/absence has been studied in some of these species

but their genome-wide patterns are unknown (lines 88-90). There is no data in Grau-Bové et al. on any hPTMS in *S. rosetta* and *Monosiga brevicollis* although the paper does describe the presence/absences of different histone modifying enzymes.

-Related with the above. There is no discussion on the other unicellular Holozoans at the end of the discussion (line 276), and this is needed. The discussion, in general, could be much more elaborated, having into account previous works. Again in (lines 243-246), reference to the other unicellular organisms studied in other works, such as Grau-Bové et al 2022 is missing.

With regard to H3K27me3 there is actually very little to compare/discuss. It is absent in *Capsaspora* and although it is present in some others we know nothing of its localization and/or function.

-Also in the discussion when they comment on the results of the different clusters where they only observe modifications downstream of the TSS, they must mention that this was also observed in *Capsaspora* in Sebé-Pedrós et al. 2016.

We have now added a line to the discussion where we compare directly to repressed genes in *Capsaspora* and note the potential of a shared mechanism for H3K4me1 in these species (lines 311-315).

-There are many references to future work to be done, but some of it could have been done for this paper. For instance, screening the other Choanoflagellates taking advantage of the data already available in the search for PRC1 or the variant PRC1 complex that is found in *S. rosetta*.

We do not suggest searching for homologs of variant PRC1 in other choanoflagellates as something we would do in the future. In fact, de Potter et al., have already performed this analysis and confirmed our previous findings that choanoflagellates possess a variant PRC1 and that this is the more evolutionarily ancient form of PRC1. We already cite this in the manuscript. (de Potter et al., 2023)

Comments on methodology

-What was the criteria to decide the number of cells used for each experiment? 500 million cells for ChIP-seq experiments seems too much. Was there a particular reason for that?

The number of cells used in ChIPseq experiments is very variable across species. In our case we adapted a protocol for 50 million human cells and given the small genome size in *S. rosetta* we optimized using much higher cell numbers. We are not sure what would constitute “too much” as long as the ChIP protocol works. We have optimized ChIP with this cell number, and it work extremely well.

-The protocol used to treat the cells before extracting the nuclei seems too aggressive. Why didn't the authors use a less stressful technique, such as digitonin permeabilization?

Again, these protocols vary widely between species. We do not know on what basis the reviewer thinks this is too aggressive. In our hands this protocol works very well to extract nuclei which we used for multiple downstream analyses.

-Was any filtering applied for contamination in the raw RNAseq data? Please, cite it.

No there was no filtering applied. The authors who developed this protocol optimized the RNA extraction in order to remove most bacterial contamination at the pre-library stage (Leon et al., 2024, Coyle et al., 2023).

Minor comments

-One reference is missing in line 303.

We have removed this error.

-There is a typo in line 134, where Fig 1D is cited instead of Fig 2D.

We have fixed this error in the text

-Again and in general, it is not clear what experiments are the authors referring to in the text. While explaining the results, the text jumps from slow swimmers to fast swimmers or thecates. There are no results shown or discussed about the rosettes. It is important to be clearer on which experiments are being described in the results section, and that they are also further discussed and put in context with the literature.

As discussed above we have made it clearer in the figures which cell types are being used for each experiment and where appropriate have changed the text to better explain this.

-In line 158 the authors describe results about slow swimmers and thecate cells, but they don't mention fast swimmers, why?

We only discuss slow swimmers and thecates in the majority of the text because on the transcriptional level there is no difference between slow swimmers and fast swimmers (Leon et al., 2024). Therefore, comparing gene expression between these cells would not be informative. We add a line to the text to explain this (lines 163-165).

-Clusters1, 2, 3 ,4 5 are cited in different ways (cluster 2, cluster2, Cluster 2...)

We apologise for this oversight and have now changed it so that we use the same style throughout.

Reviewer #4 (Remarks to the Author):

I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

We thank this reviewer for taking the time to co-review our manuscript.

Reviewer #5 (Remarks to the Author):

Gahan and colleagues have produced an excellent manuscript on DNA modifications in the genome of the choanoflagellate *Salpingoeca rosetta*. The work appears to have been performed to a high standard and is an important study, as information on gene regulation in the closest relatives of animals remains limited. The authors show a second holozoan protist that lacks distal gene enhancers, however promoter sequences are characterized by histone modifications. The conclusions drawn by the authors are appropriate and appear justified on the presented results. I should state at this point I have no past experience in working with ATAC-seq.

I believe the manuscript will have broad appeal to readers from a variety of research fields. There would be a clear interest from the choanoflagellate community, but the research will also be of considerable interest to workers studying gene regulation, both in animals and other model organisms.

We thank the reviewer for taking the time to evaluate the manuscript and for their positive overall evaluation of the paper.

One issue is the emphasis on transposons in the title, which did not appear justified in the manuscript. A simple solution to this would be for the authors to remove transposon (which is not used in the technically correct manner) from the title, so that the emphasis is on host gene regulation. A more satisfying approach would be to expand their analyses on DNA modifications in transposable element sequences in the *S. rosetta* genome.

We have taken on board the reviewers' comments and now include an expanded analysis of TEs in the manuscript and have modified the title.

Major Comments

Transposons are mentioned before host genes in the title of the manuscript, so it is a little surprising that they only appear in one paragraph at the end of the Results section. As full-length annotations of 20 *S. rosetta* transposable element families are available (Southworth et al. 2019, Mobile DNA, 10:44), it would improve the manuscript if the authors could perform a more detailed analysis of the chromatin modification and the different TEs. The Southworth study analyses showed that individual families have quite different population dynamics and it would improve the submitted manuscript if the authors could show whether these differences are reflected in the DNA modifications present in TE DNA.

The authors have analysed LTR/gypsy-like retrotransposons and report on them as a group, however there are a minimum of seven distinct gypsy-like families within the *S. rosetta* genome, with both chromoviral and non-chromoviral gypsy-like families present. Published RNA-Seq data indicate considerable variation in the expression levels between families. It would improve the paper if they could report on analysis of individual gypsy-like families.

Although not commented upon here in the current manuscript, the Southworth study provided multiple lines of evidence (e.g. identical paralogous copy number, RNA-Seq read number and strong selection on codon usage) to show that copia-like families, specifically Srospv2 and Srospv3, rather than gypsy-like elements or transposons, are the most active TE families in the *S. rosetta* genome. It feels an omission on behalf of the authors to not report on DNA modification within those two highly active families and the copia-like families in general. Furthermore, the analyses of transposons would be improved if the authors looked beyond the MULE-like family and also analysed the other transposon families known to be present in the genome.

We thank the reviewer for their recommendation. As outlined in our response to reviewer number 1 we have now included an expanded analysis of all TEs in the *S. rosetta* genome and have looked at histone PTMs, ATACseq and RNAseq.

Minor Comments

1. General: The term transposon should really only be used for Class II, DNA-based, transposable elements and should not be used as a generic term for all transposable elements. Ideally, the authors should replace the term transposon with transposable element, or TE, to avoid confusion, unless they are specifically discussing Class II transposable elements (transposons).

We totally agree with the reviewer and have now replaced transposon with transposable elements for the majority of the paper except when we talk about specific families.

2. General: A brief explanation of ATAC-sequencing would assist the generalist reader.

We have added a line on what ATACseq does (line 127-129).

3. Introduction, Page 1, lines 61-63: The absence of enhancer sequences in *Capsaspora* is thin evidence for their metazoan origin. The absence is also consistent with an earlier holozoan origin and their subsequent loss in the *Capsaspora* lineage. The authors should tone down, or reword, this statement.

We agree with the reviewer but, in this case, we are summarizing current thinking in the field. We have edited the text to be clearer that we are referring to the common hypothesis in the field rather than stating our own interpretation (line 63).

References

- COYLE, M. C., TAJIMA, A. M., LEON, F., CHOKSI, S. P., YANG, A., ESPINOZA, S., HUGHES, T. R., REITER, J. F., BOOTH, D. S. & KING, N. 2023. An RFX transcription factor regulates ciliogenesis in the closest living relatives of animals. *Current Biology*, 33, 3747-3758. e9.
- DE POTTER, B., RAAS, M. W., SEIDL, M. F., VERRIJZER, C. P. & SNEL, B. 2023. Uncoupled evolution of the Polycomb system and deep origin of non-canonical PRC1. *Communications Biology*, 6, 1144.
- LEON, F., ESPINOZA-ESPARZA, J. M., DENG, V., COYLE, M. C., ESPINOZA, S. & BOOTH, D. S. 2024. Cell-type-specific expression of a DCYTB ortholog enables the choanoflagellate *Salpingoeca rosetta* to utilize ferric colloids. *bioRxiv*, 2024.05.25.595918.
- LIN, S., WEIN, S., GONZALES-COPE, M., OTTE, G. L., YUAN, Z. F., AFJEHI-SADAT, L., MAILE, T., BERGER, S. L., RUSH, J., LILL, J. R., ARNOTT, D. & GARCIA, B. A. 2014. Stable-isotope-labeled histone peptide library for histone post-translational modification and variant quantification by mass spectrometry. *Mol Cell Proteomics*, 13, 2450-66.

Dear Reviewers,

We would like to thank the reviewers for taking the time to review our manuscript again. We are very happy to see that all reviewers are now very positive, and we agree with the reviewers that the process has led to a much-improved manuscript. We have addressed the additional comments from this round and lay out a detailed response below.

Reviewer #1 (Remarks to the Author):

Most comments have been fully addressed and we congratulate the authors on the much-improved revised manuscript.

There are still some points listed below that should be addressed to improve clarity:

1. The authors argue that intergenic distal ATAC-seq peaks correspond to unannotated TSSs, and this is nicely illustrated in Supplementary Fig. 2, yet these correspond to intra and not inter-genic peaks. This should be corrected.

We apologise for this mistake and have corrected the figure legend to replace “intergenic” with “distal intragenic” peaks.

2. The fact that the mass spectrometry data cannot be used to determine stoichiometry of individual histone modifications is worth mentioning somewhere (perhaps in the Materials and Methods section). Same comment for the lack of detection of H2AK119ub.

We have added a line in the methods to point out that the data was not used to look at stoichiometry and why (line 541/542).

We have also added a line in the discussion to note we did not detect H2A monoubiquitylation but for technical not biological reasons (Line 302/303).

3. Regarding H3K27me3, there seem to be three “flavors” of H3K27me3 domains that would be worth displaying in the main Figure 6 as screenshots. In addition to the snapshot of the cluster of LTR retrotransposons, the authors could display an inactive gene body and a gap-less sub-telomeric region (one shown in Supplementary Fig. 6).

We have now incorporated the genome browser snapshot from that figure into the new and simplified Figure 5. Since we already show a repressed gene body in Figure 4, we think it would be unnecessary to show this again in Figure 5. In addition, we are happy with keeping the sub-telomeric figures in the supplementary

material.

4. The expanded ATAC-seq analyses are very welcome. One major concern about these new heatmaps based on histone PTM enrichment over individual TEs is whether individual TEs are unique in sequence in rosetta. i.e. are they mappable? Are the MAPQ scores for aligned reads over individual TEs above 10 for example? If not, perhaps a “small family”-based analysis is more appropriate (breaking down TEs into SrosH, SrosM, etc).

We have now generated a radically changed figure for the TEs which is now Figure 5. We use a heatmap as a simplified way to show the data at the family level. We also include the same analysis as a new supplementary figure where we show the same data but filtered for a MAPQ score of 30. We have moved the heatmaps to the supplementary material and they show the filtered reads. This is now indicated in the figure legends. The overall picture from analysis of both sets of reads is the same. In the unfiltered set you see much more enrichment over the TEs themselves and this is to be expected as in the larger families there is quite a lot of similar sequence. We discuss this in the results (Lines 216-221). Overall, however we are happy that message is the same regardless of which set of reads are used.

Assuming TEs in rosetta are mappable, I still have a minor concern about the legibility of the revised heatmaps. For both Fig. 5 and 7, I suggest you shuttle the families with a few individual elements to the supplemental, resulting in simplified and more legible main figures. Metaplots (tops of heatmaps) are uninterpretable with these many classes. You could apply a cutoff of 20 elements for example. I realize this may more or less recreate the TE heatmap from the initial version with the addition of copia elements. However, the full heatmaps, as currently presented in the updated Fig5 and 7, would still of interest as supplemental figures.

We have now simplified our analysis and made a new Figure 5 which we believe address these concerns. We have kept the original heatmaps and they are now in the supplementary material.

Alternatively, and this is a non-issue, only a suggestion, instead of heatmaps, 2D scatterplots showing individual element expression (x-axis) versus H3K27me3 enrichment over the TSS (y-axis) would be clearer. You could have one scatterplot for each TE family (small or large), or one big plot with data points colored by family. Expression vs K27me3 is the main comparison anyway, the other PTMs and accessibility could still be kept as supplementary information.

We think with the new visualization in Figure 5 and the supplementary material make the message very clear and we therefore do not think that this would add to the story.

Reviewer #2 (Remarks to the Author):

I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

Reviewer #3 (Remarks to the Author):

Second revision of the manuscript entitled: Chromatin profile identifies putative dual roles H3K27me3 in regulating cell type-specific genes and Transposable Elements in choanoflagellates. Authored by James Gahan et al.

In this new version the text has substantially gained clarity and key references for the conceptual framework have been added improving the overall quality of the manuscript.

We agree that the authors demonstrate the role of the H3K27me3 chromatin modification as likely the on and off switch for some developmental genes in *S. rosetta*. We think this is a sound and important claim which the scientific community will receive as important knowledge concerning evolution of genome regulation and developmental control. Nevertheless, we have some important comments to different aspects of this work.

Major

- The authors have made an effort to expand their work on TEs and have included three figures (5, 6 and 7) where they combine their data with previous RNA-seq data (Southworth et al 2019 Mobile DNA). The insight from these figures is just reflected in a total of four lines in the results (216 to 220). These three figures contain quite a substantial amount of data that is poorly reflected, in our opinion, in the results and in the discussion sections.

We now have much less emphasis on TEs in terms of figures (just one figure: Figure 5) and have moved a lot to the supplementary material. We have also expanded the description in the text somewhat in both the results and the discussion.

- Moreover, figure legends of these figures are not sufficiently explanatory. For example, for Figures 5 and 7, what does the code color for the different subfamilies of TEs (right above corner) mean? It is unclear whether these colors represent the different TEs in the different profiles (ATAC-seq, Chip_seq)? If that is so, at the scale that the profiles are done is impossible to see anything.

These colours are no longer a part of the newly generated figures which we hope are much more easily understood.

- Also for Figures 5 and 7, very important, are the horizontal lines inside the box of each TE subfamily, corresponding to different copies of each specific subfamily, so that one can follow the amount of H3K27me3 and expression in the different cell types for each copy?

Yes, each line corresponds to an individual copy. We have included this in the figure legend (now Supplementary figures 8/9) to make it easier to follow.

- If, in both Figures 5 and 7, the horizontal lines from the RNA seq data and histone modifications correspond to individual copies of a specific TE there are several occasions where a specific copy is both highly expressed in slow swimmers and strongly decorated with H3K27me3. If this interpretation is correct, they should be consistently mentioned as “copies” rather than “each transposon” throughout the figure legends to avoid confusion.

We have changed it to copies.

- Checking the data from Southworth et al., Mobile DNA (2019), Srospv2 has 17 paralogous copies, and in the Figure 5 of Gahan et al., there are many more horizontal lines. Same for Figure 7. These are important figures for understanding the extent to which H3K27me3 and TE transcription support one of the main claims of the manuscript. Right now, these numbers are unclear, and the figure legends are minimal.

We have included a count of the number of copies for each TE family in the new Figure 5.

In Table 1 of Southworth et al., the authors define 17 full-length copies of **Srospv2**, using a strict criterion where both LTRs within a given copy must be more than 99% identical. However, they also note that Srospv2 is among the most abundant transposable element families in *S. rosetta*.

In our study, we adopted a more inclusive definition. Specifically, we used *Additional file 2* from their publication and processed it with RepeatMasker under default parameters. We then filtered the resulting hits, retaining only those insertions that were at least 50% of the length of the consensus sequence described in their work. This approach allows us to capture not only very recent and intact copies, but also older insertions or those that may have accumulated deletions / recombination over time.

Using a strict definition that limits analysis to nearly identical, recent insertions would exacerbate mappability issues. Moreover, our interest lies in understanding the role of H3K27me3, which is known to mark both old and new transposable element

insertions. The distinction between a “true” transposable element and a long-dead remnant is often ambiguous. While Southworth et al. aimed to catalogue full-length, potentially active elements, this represents only part of the contribution of TEs to genome regulation. Even truncated or catalytically inactive elements can play important regulatory roles, which justifies our broader criteria.

Finally, it is important to consider that the *S. rosetta* genome is still in draft form. Future resequencing with long-read technologies may further refine the number and structure of TE insertions.

- While the Figure legends mention the presence or absence of H3K27me3 on different classes of transposable elements (TEs), they do not explain why this is biologically meaningful. Consider briefly describing somewhere in the text the functional significance—how does H3K27me3 relate to the transcriptional regulation or silencing of LTR retrotransposons versus DNA transposons?

We have added some additional discussion on this, particularly on H3K9me3. (Line 226 and Lines 270-278)

- Ensure the description of ATAC-seq, ChIP-seq, and RNA-seq data is uniform across both figures. In Figure 5, RNA-seq is described as “DESeq2 normalized counts,” while the corresponding description in Figure 7 is slightly different. Consistent phrasing will enhance clarity. Please, clarify how TE subfamilies are defined (e.g., based on prior annotations) and whether copy numbers were filtered or normalized in the presentation of the heatmaps.

We have now ensured that the descriptions are uniform across figures. We have included a reference in each figure legend to the paper in which the annotations were originally performed. There was no filtering for copy number.

- Related to the above point, Line 217-218: The authors state that H3K27me3 levels drop in LTRs that are substantially expressed. Since ChIP-seq data is only available for slow swimmers, it is unclear whether this statement refers specifically to expression during this stage or to highly expressed LTRs across all stages. Clarifying whether these observations reflect increased expression in slow swimmers or in other stages would strengthen the argument. We trust that further examination of this relationship could provide stronger support for the claim in Line 222 regarding H3K27me3 regulating TE expression. Maybe this result is already in Figure 5, but the current figure legend is not self-explanatory to make this interpretation clear.

We have removed any reference to levels dropping and we now refer to TEs having higher or lower levels relative to each other. As we do not have ChIPseq in other cell

types we cannot compare whether changes in expression across cell types is related to H3K27me3 levels.

- Also Figure 7 shows that H3K27me3 is not significantly present on DNA transposons, but there are signals of H3K4me1 and H3K4me3 that the authors could discuss in the discussion. Exploring whether these histone marks suggest active regulatory regions or other functional implications would provide a more comprehensive interpretation of the chromatin landscape surrounding DNA transposons.

We have added a discussion of this to the manuscript (Lines 227-231 and 271-273.)

- Also, it is necessary to have the reference from Southworth et al Mobile DNA (2019) correctly cited in the figure legends where the data is used and again in the M&M section, where it is only cited as Southworth et al.

We have now also cited this paper in all the relevant figure legends

- IMPORTANT: Also, in Figures 5 and 7 legends: TEs are classified in families and Superfamilies, not “bigger” families.

We have now annotated the figures correctly and included the “family” and “class” of TEs.

- Figure 6 legend should contain the name or code for the “other genes” shown in the figure.

We have added the codes of all of the other genes to the figure legend (now Figure 5).

- Last but not least, and very importantly, the final message is not clear. Is the main conclusion that H3K27me3 marks developmental genes, LTR transposable elements, and sub-telomeric regions? If so, which additional histone or chromatin marks would enable developmental genes to be turned on and off when needed? In other words, which chromatin factors keep LTRs and sub telomeric regions repressed? Since the authors have detected H3K9me3 in *S. rosetta*, I think it is likely and worth speculating on a possible role of H3K9me3 as this possible additional mark for repressing sub telomeric and TEs sequences. This speculation is particularly relevant since H3K9me3 is well-known to mediate transposon and heterochromatin (including subtelomeric sequences) silencing in other organisms.

We have added a line on the potential role of H3K9me3 (Lines 273-278).

- Related to the above, it would be important to mention in the discussion the lack of experiments regarding H3K9me3 exposing the particular experimental difficulties. Any reader with knowledge in genome regulation would wonder why this is not inspected, especially after mentioning that the authors have actually identified its presence in *S. rosetta*. Moreover, the particular difficulty with the antibody for this modification will be of interest for the scientific community working on non-model organisms which might present the same threonine substitution.

We have added a line on this to the discussion (Lines 273-278)..

Minor comments

- Finally, the authors devote quite a substantial portion of the manuscript discussing enhancers, which *S. rosetta* does not seem to have, at least with the current dataset (Intro; line 57 to 67, Results; lines 122 to 141 Discussion; 225-243). While we understand that it is important to determine the existence of distal regulatory elements in unicellular holozoans, since choanoflagellates are the closest relatives of animals, and only in *C. owczarzaki* the presence of enhancers has been inspected, we feel that the discussion of enhancers may be disproportionate to the available evidence. This contrasts with the emphasis placed on TEs, which is highlighted in the title. It would be helpful to achieve a better balance between these two aspects, aligning the title's focus on TEs with an in-depth discussion of their role in *S. rosetta*.

We believe that the discussion is well balanced given what we present in the paper and what is known already. In the case of distal-enhancers this is a very active field which still has some controversies and therefore a long discussion is justified. In the case of TEs we believe we discuss sufficiently what we have shown. We have however added a few extra lines, particularly on the role of H3K9me3 and also to discuss the presence of active modification on some DNA transposons.

- Line 64 "...enhancer are truly animal-specific". We understand the intention of this definition might be to distinguish metazoans from unicellular holozoans since the authors are working inside the Holozoa branch, but enhancers are also present in plants, and therefore we think this sentence could be misleading.

Since this round of revision even more data has been published reinforcing this concept and this is the current view in the field. We have cited this work throughout the paper and edited that section to make our point more clear.

- References not cited in the text: 44 and 47, which is a pity since they look interesting.

These references were cited in the manuscript. They are still cited in the same position (line 90)

- Results from line 148 to 151 miss the figure citation

We have added a figure citation here.

- Line 873-874 seems to contain an error. The term "grouped" should be replaced with "groups" to maintain grammatical consistency.

We have changed this.

Reviewer #4 (Remarks to the Author):

I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

Reviewer #5 (Remarks to the Author):

Gahan and co-workers have undertaken many of my original suggestions, which I feel has improved the resulting manuscript. Where they have disagreed with my comments, I view their rebuttals to be fair and reasonable. I consider the resubmitted manuscript an excellent piece of work, which should be of considerable interest to researchers in a broad range of fields. As a result, I recommend it is accepted for publication in Nature Communications.

My only minor suggestion is that the authors tighten up their word formatting a little. Slow swimmers are consistently written with an upper case "S", whilst fast swimmers are always written with a lower case "f". Neither term appears to be a formal noun, so lower case letters would be more appropriate. This also applies to thecate cells, which is written as Thecate suggesting it is a formal, proper noun rather than a descriptive term. In addition, binomial species/genus names should be written consistently in italics, whereas at present there are a number of examples where names are presented in regular font (for example, on page 13 of the main manuscript, *Capsaspora owczarzaki* is written with the genus name in regular font, but species name in italics).

We have changed all instances of slow swimmers and thecate cells to have lower case letters. We have also checked all genus/species names are in italics throughout.