

Supplementary material

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

James M. Gahan^{1,2,3*}, Lily W. Helfrich^{4, 5}, Laura A. Wetzel^{4, 6}, Natarajan V. Bhanu⁷, Zuo-Fei Yuan⁸, Benjamin A. Garcia⁷, Robert J. Klose², Alex de Mendoza^{9,10}, David S. Booth^{1, 11,*}

¹Department of Biochemistry and Biophysics, University of California, San Francisco, San Francisco, CA 94158, USA

²Department of Biochemistry, University of Oxford, Oxford, UK

³Present Address: Centre for Chromosome Biology, School of Biological and Chemical Sciences, University of Galway, Galway, Ireland

⁴Howard Hughes Medical Institute / University of California, Berkeley, Department of Molecular and Cell Biology, Berkeley, CA 94720

⁵Present Address: Benchling, San Francisco, USA

⁶Present Address: BioMarin Pharmaceutical Inc., San Francisco, USA

⁷Department of Biochemistry and Molecular Biophysics, Washington University School of Medicine, St Louis, MO, USA

⁸Center for Proteomics and Metabolomics, St. Jude Children's Research Hospital, Memphis, TN 38105, USA

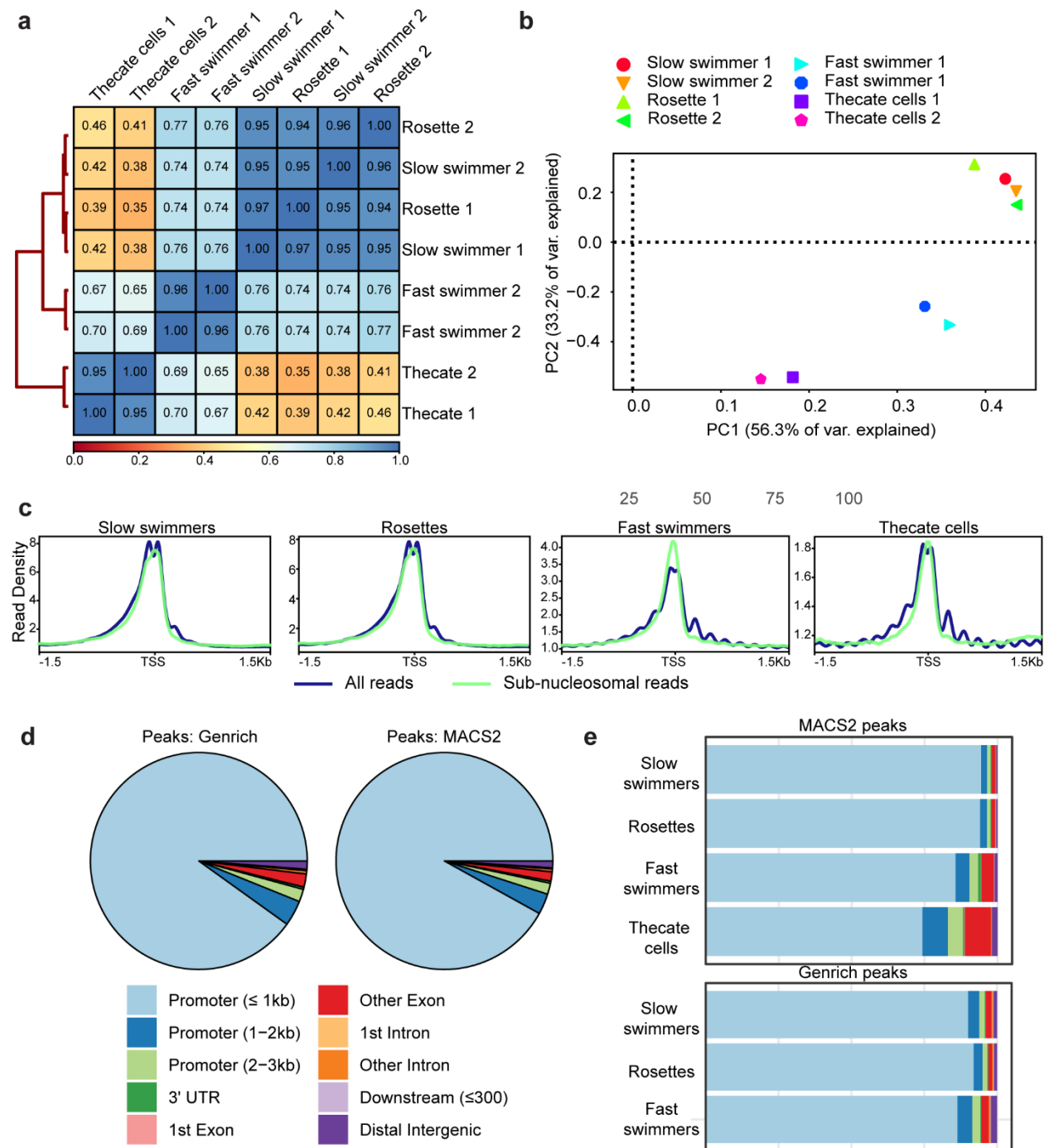
⁹School of Biological and Behavioural Sciences, Queen Mary University of London, Mile End Road, London, E1 4NS, UK

¹⁰Centre for Epigenetics, Queen Mary University of London, London, UK

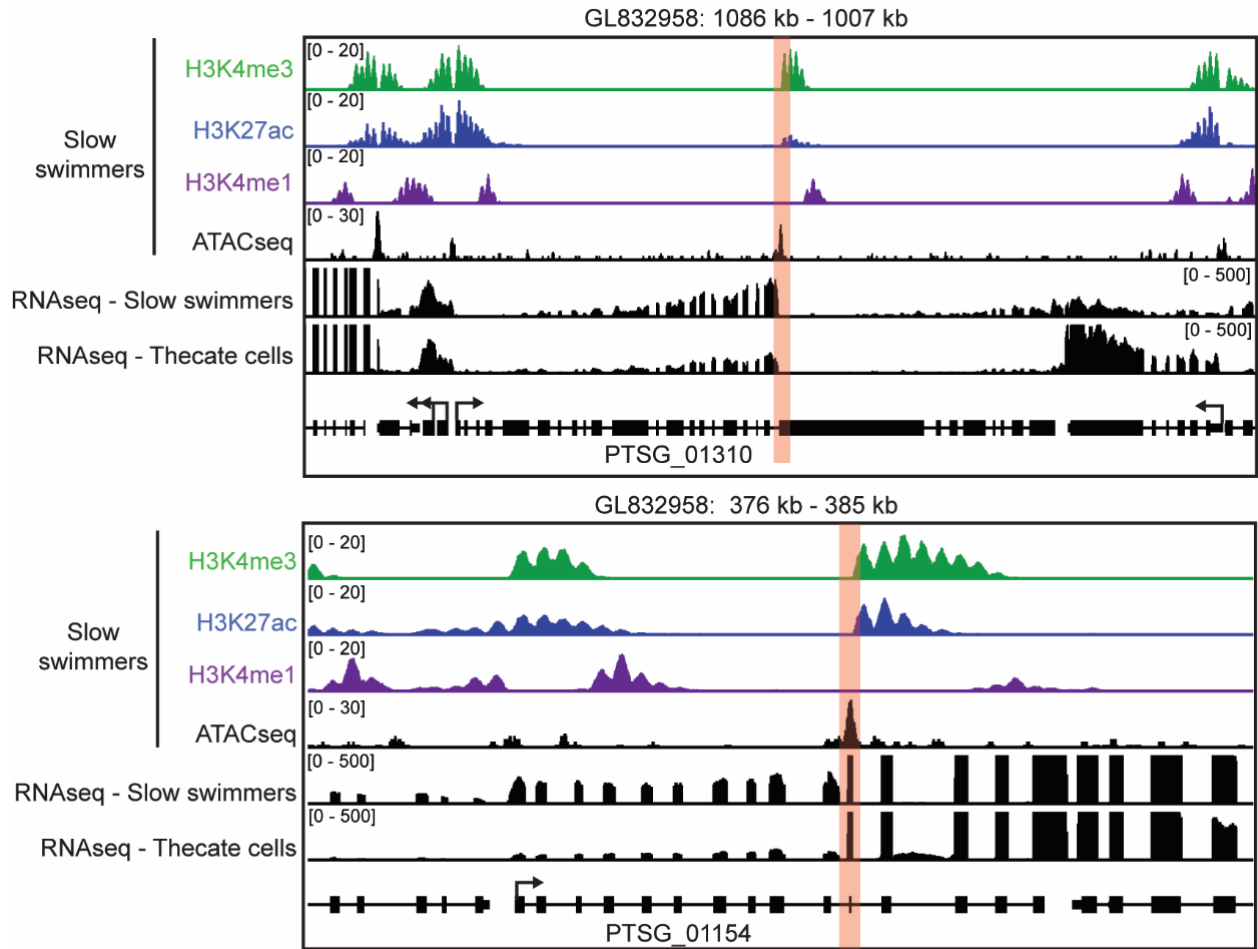
¹¹Chan Zuckerberg Biohub, San Francisco, CA 94158, USA

*Co-corresponding authors: james.gahan@universityofgalway.ie, David.Booth@ucsf.edu

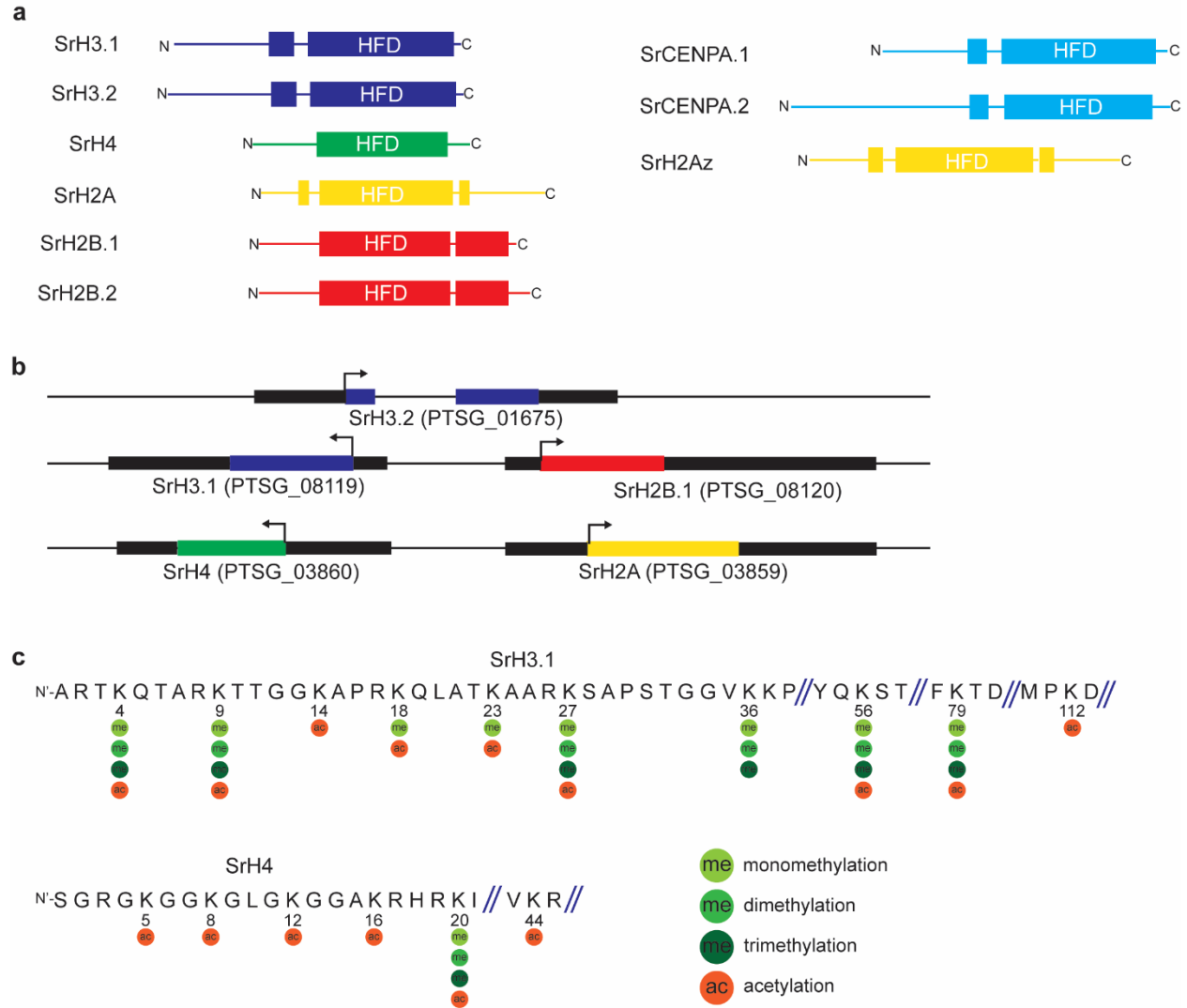
Supplementary figures



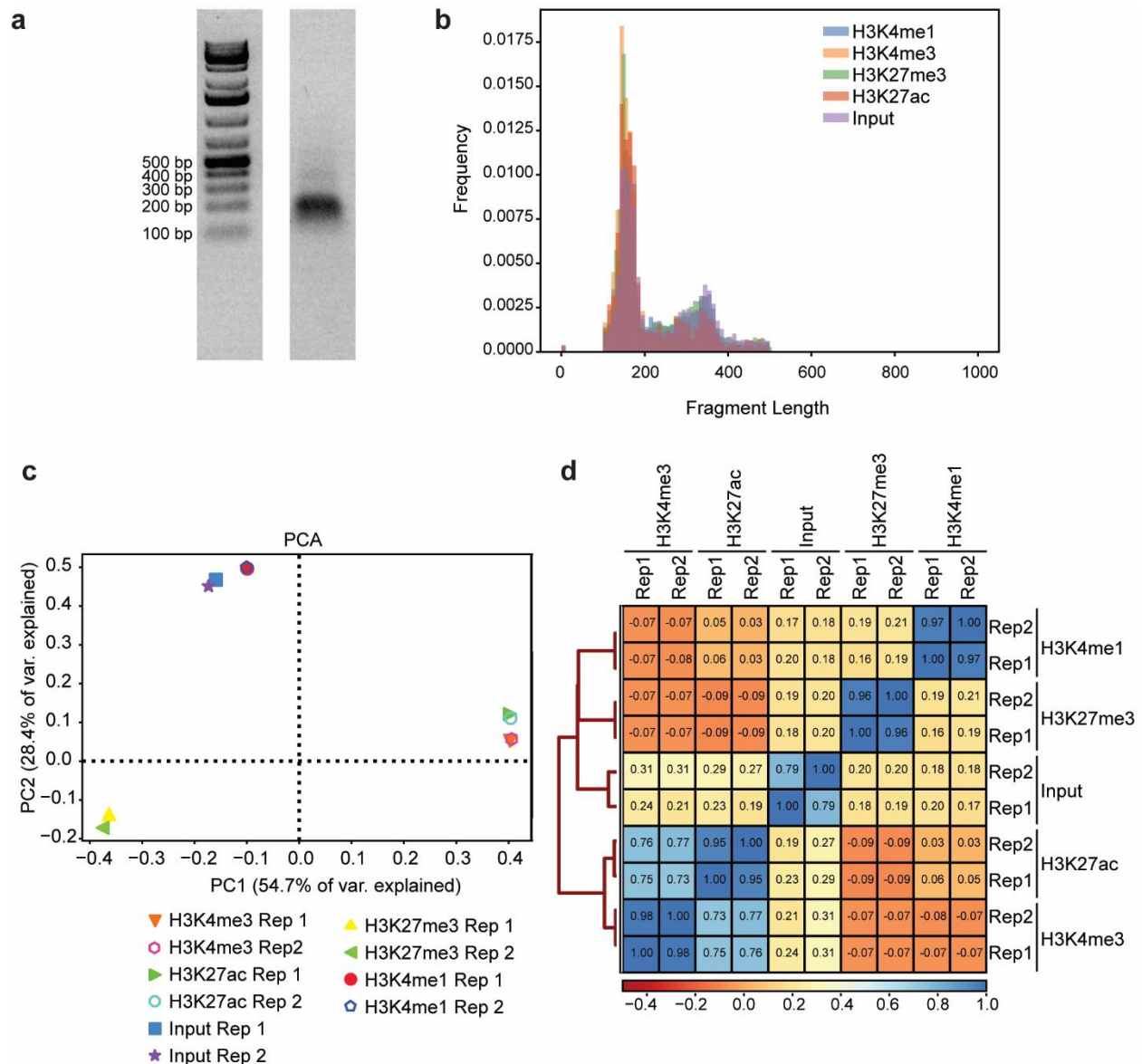
Supplementary Fig. 1. Additional ATAC-seq analysis. (a) Clustered heatmap showing the correlation coefficients between samples based on the Pearson method. (b) PCA plot on all samples showing the two first principal components. (c) Metaplot of ATAC-seq read density around the TSS's of all genes comparing all reads to sub-nucleosome sized fragments. (d) ChIPseeker annotation of peaks from either Genrich or MACS2. For MACS2 peaks from all cell types are shown while for Genrich all cell types except thecate cells were used (e) ChIPseeker annotation of peaks showing each cell type separately.



Supplementary Fig. 2. Examples of putative distal intragenic ATAC-seq peaks. Genome browser snapshots showing two examples of annotated distal intragenic peaks showing ChIP-seq with the indicated antibodies along with ATAC-seq in slow swimmers and RNA-seq in both slow swimmers and thecate cells. The putative distal peaks are highlighted in red. The RNA-seq data shows strong differences in expression on either side of the peaks and, in combination with the histone PTM profiles, indicates that these are likely un-annotated TSSs. The genomic scaffolds and positions are shown on top. Annotated genes are shown along the bottom.

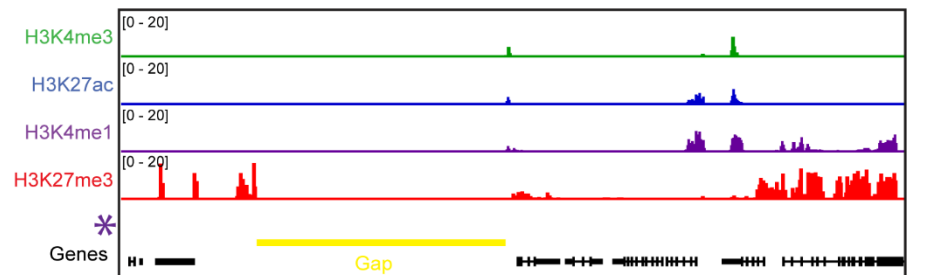
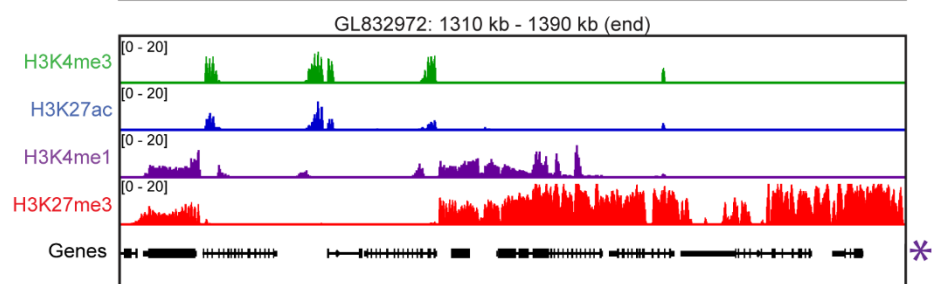
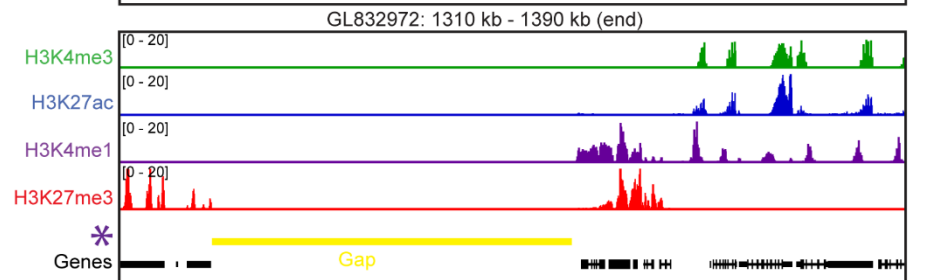
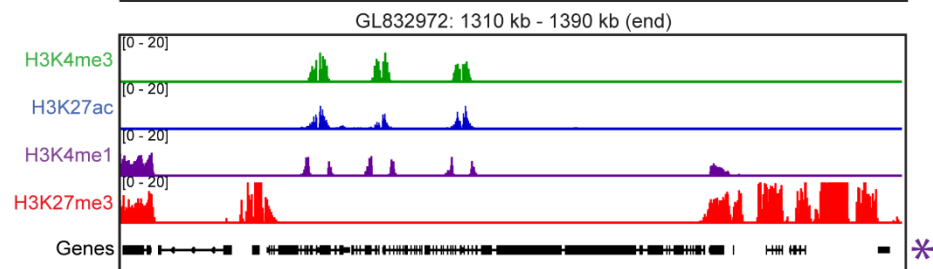
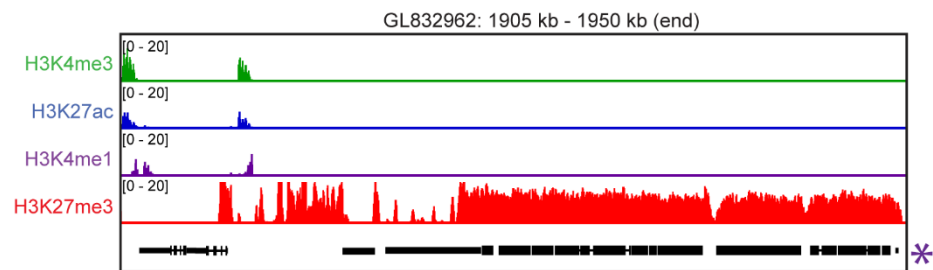
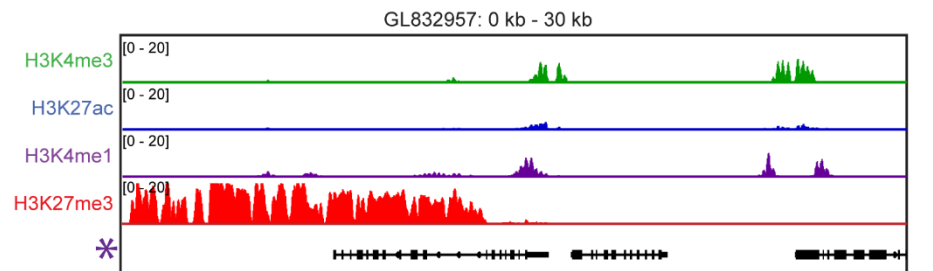


Supplementary Fig. 3. The histone PTM complement of *S. rosetta*. (a) Schematic models of all histone proteins annotated in *S. rosetta*. We annotated the histone complement using blast searches as well as previously published phylogenetic analyses¹. *S. rosetta* has two H3 variants which are putatively annotated as a replication-dependent SrH3.1 and a replication-independent SrH3.2. This is due to SrH3.1 being in a cluster with SrH2A as expected for a canonical replication-dependent histone while SrH3.2 is a single gene which also contains an intron characteristic of a replication-independent variant (see b)². *S. rosetta* has a single H4 protein, two H2B variants with a variable c-terminal length and a single H2A which is a H2Ax based on the presence of characteristic SQEY amino-acid motif in the C-terminal region. In addition to the canonical histones, *S. rosetta* possesses two centromeric histone H3 variants, SrCENPA.1 and SrCENPA.2 and an H2Az homolog, SrH2Az. (b) Genomic organization of the single *SrH3.2* gene and examples of *SrH3.1*, *SrH4*, *SrH2B.1* and *SrH2A*. *SrH3.2* is a single copy in the genome and contains an intron. SrH3.1 and SrH2B are in a cluster as are SrH4 and SrH2A. (c) Schematic of part of the SrH3.1 and SrH4 protein sequence showing positions where lysine methylation and acetylation were identified by mass-spectrometry.

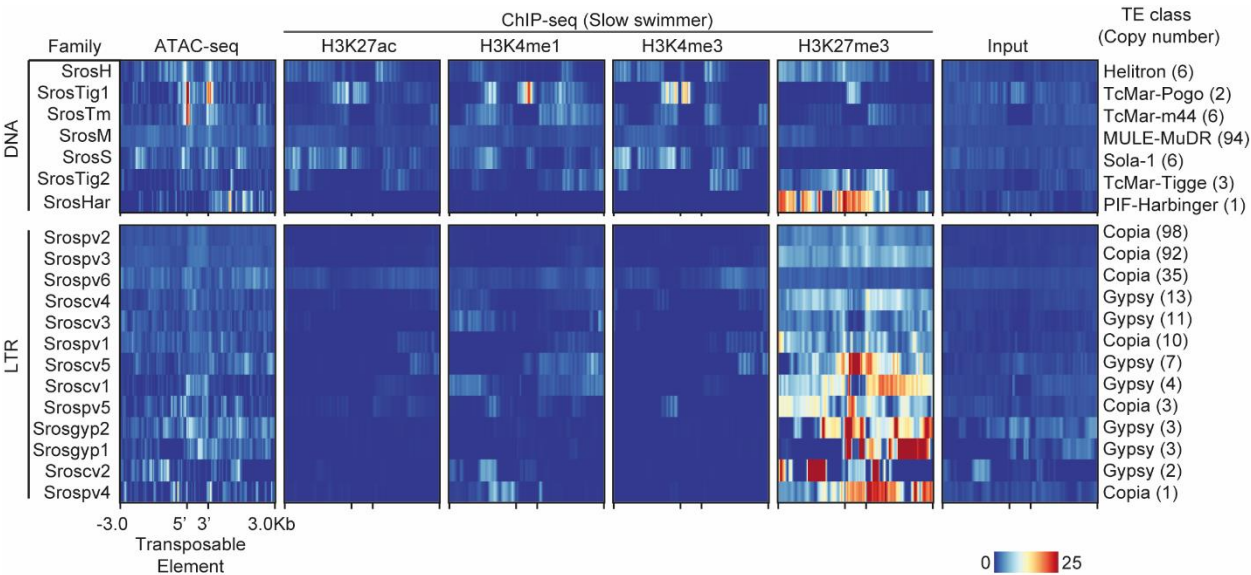


Supplementary Fig. 4. Additional ChIP-seq data. (a) Gel electrophoresis of DNA from a nucleosome extraction used for ChIP-seq. Most fragments are approximately mono-nucleosome size (~140bps). (b) Frequency plot showing the size of sequenced fragments from ChIP-seq data. A bin size of 10bp was used. In all 4 antibodies and input the majority of fragments are below 200bps, as expected for mono-nucleosome fragments. (c) PCA plot on all ChIP-seq samples showing the high correlation between replicates. (d) Clustered heatmap showing the correlation coefficients between samples based on the Pearson method.

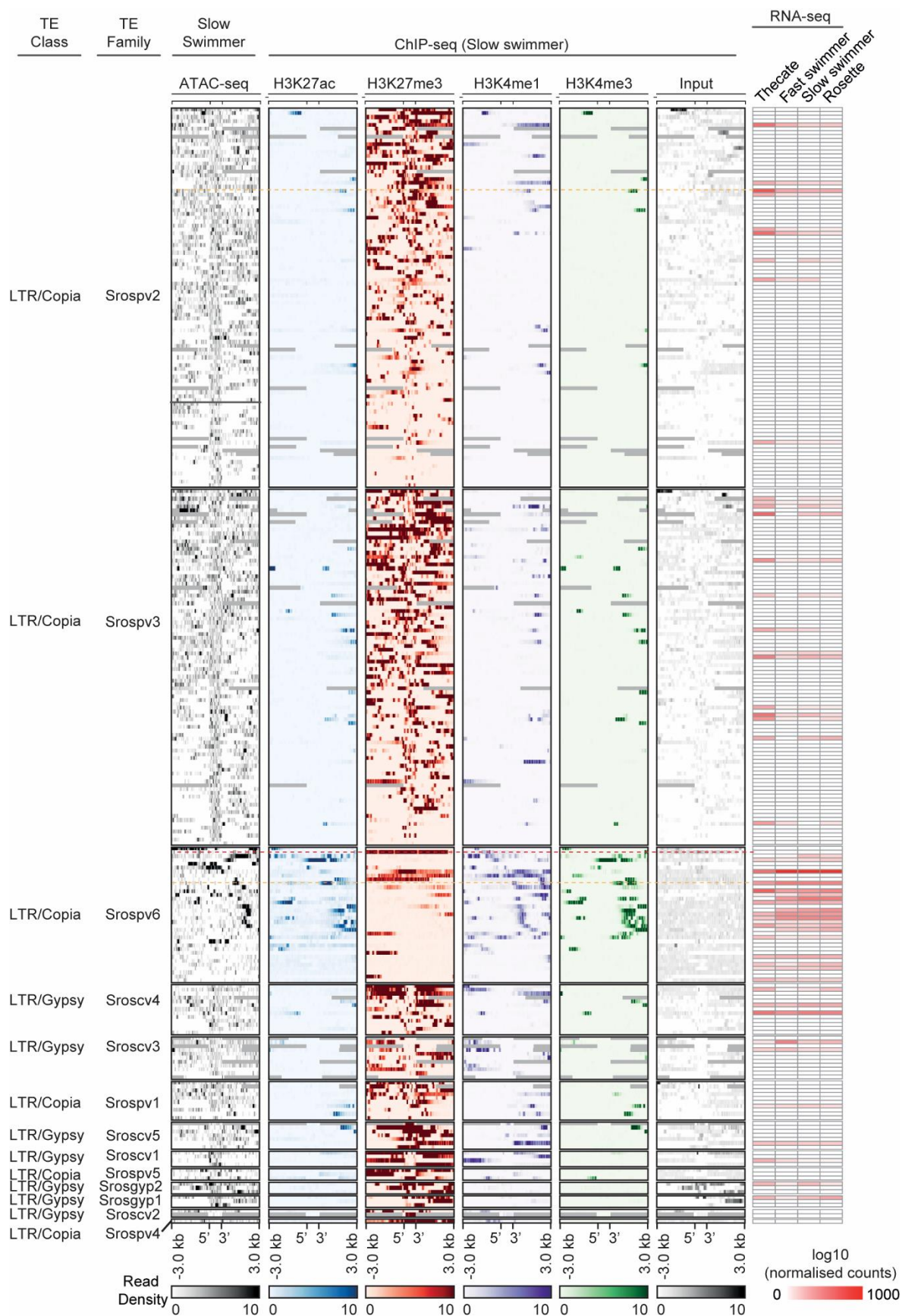
Supplementary Fig. 5. Cluster 2 genes are more highly upregulated between swimming cells and thecate cells than other genes. (a) Left panel shows a Venn diagram of cluster 2 genes and genes upregulated in thecate cells. The right panel shows a heatmap and metaplot of the overlapping genes showing ChIP-seq data with the indicated antibodies and ATAC-seq from swimming cells. Clustering was performed by k-means clustering (b) Density plot showing the log₂ fold changes from RNA-seq between thecate cells and slow swimmers for all genes upregulated in thecate cells, separated based on whether they overlap with Cluster 2 genes or not. (c) Density plot showing the log₂ fold change from RNA-seq between thecate cells and slow swimmers for all genes and for the Cluster 2 genes. Log₂ fold change was calculated using DESeq2. (d) Genome browser snapshots of examples of genes from Cluster 2A and 2B showing ChIP-seq with the indicated antibodies along with ATAC-seq in slow swimmers and RNA-seq in both slow swimmers and thecate cells. The genomic scaffolds and positions are shown on top. Annotated genes are shown at the bottom along with the cluster to which they belong (from a).



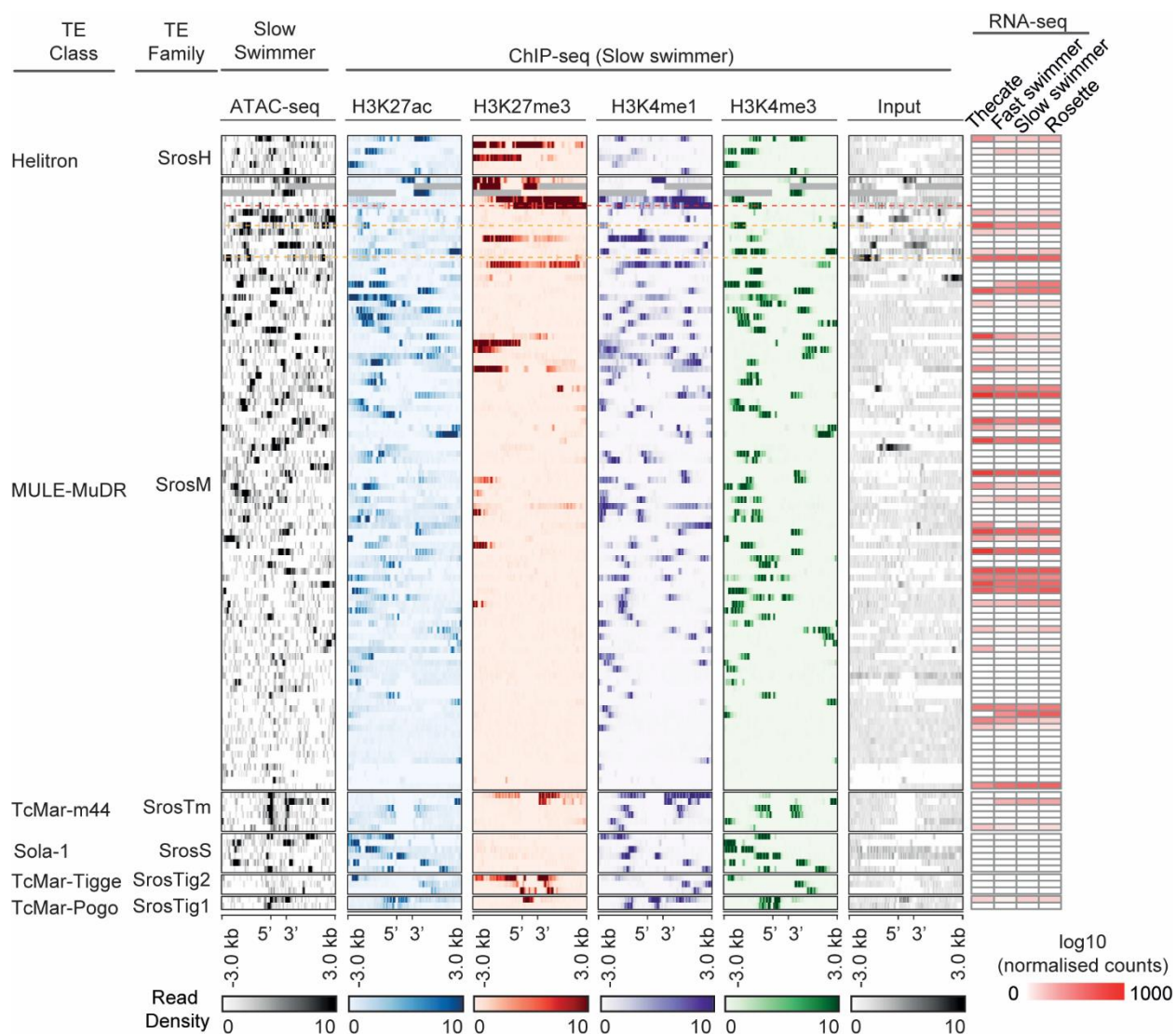
Supplementary Fig. 6. H3K27me3 is enriched in sub-telomeric regions. Genome browser snapshots showing ChIP-seq data from slow swimmers with the indicated antibodies for all 6-scaffolds containing telomeric repeats. The genomic scaffolds and positions are shown on top and annotated genes are shown on the bottom. Gaps in the sequencing data are shown in yellow. The position of the end of the scaffold with telomeric repeats is annotated with an Asterisk.



Supplementary Fig. 7. H3K27me3 is enriched over LTR retrotransposons. Heatmap showing average ATAC-seq and ChIP-seq in slow swimmers over annotated *Salpingoeca rosetta* transposable element families³, divided by Class. Antibodies used for ChIP-seq are indicated on top and reads used were filtered for a MAPQ score of 30. The TE family is shown on the left and the Class and number of copies of each are shown on the right, heatmap scale bottom right. Only inserts spanning at least 50% of the consensus sequence were used.



Supplementary Fig. S8. H3K27me3 is enriched over repressed LTR retrotransposons. Heatmaps of ATAC-seq and ChIP-seq data from slow swimmers are shown on the left for all annotated LTR retrotransposons. Transposable elements are grouped based on a previous annotation³ with the specific class and family shown on the left. Each line corresponds to an individual TE copy in the genome. Antibodies used for ChIPseq are shown on top and reads used were filtered for a MAPQ score of 30. RNA-seq data from all 4 cell types for each copy is shown on the right-hand side, with values showing DEseq2 normalised counts.



Supplementary Fig. 9. H3K27me3 is largely absent on DNA transposons. Heatmaps of ATAC-seq and ChIP-seq data from slow swimmers are shown on the left for all annotated DNA transposons in the *S. rosetta* genome. Transposons are grouped based on a previously published annotation³ the specific class and family shown on the left. Each line corresponds to an individual TE copy in the genome. Antibodies used for ChIPseq are shown on top and reads used were filtered for MAPQ score of 30. RNA-seq data from all 4 cell types for each transposon are shown on the right-hand side with values showing DEseq2 normalised counts.

Supplementary Table 1. Histone genes present in *S. rosetta* and their genomic locations.

Histone (protein)	Gene names	Genomic location (Supercontig:position)
SrH3.1	PTSG_08119	GL832975:510871-511825
	PTSG_10057	GL832986:418356-419309
SrH3.2	PTSG_01675	GL832957:484880-486601
SrH4	PTSG_03860	GL832962:1811404-1812225
	PTSG_00674	GL832956:439532-440249
SrH2A	PTSG_03859	GL832962:1810216-1811282
	PTSG_00675	GL832956:440469-441465
SrH2Az	PTSG_11229	GL832996:49505-51660
SrH2B	PTSG_08120	GL832975:512187-513213
	PTSG_10056	GL832986:416738-417364
SrCENPA.1	PTSG_09298	GL832982:51979-53872
SrCENPA.2	PTSG_08448	GL832977:831645-832995

Supplementary Table 2. List of antibodies used in this study.

Target	Source	Concentration used
H3K27me3	Rose et al., 2016. ⁴	5ul per ChIP
H3K4me3	Farcas et al., 2012 ⁵	5ul per ChIP
H3K4me1	Cell Signalling Technologies D1A9	5ul per ChIP
H3K27ac	Cell Signalling Technologies D5E4	5ul per ChIP

Supplementary Table 3. List of software used.

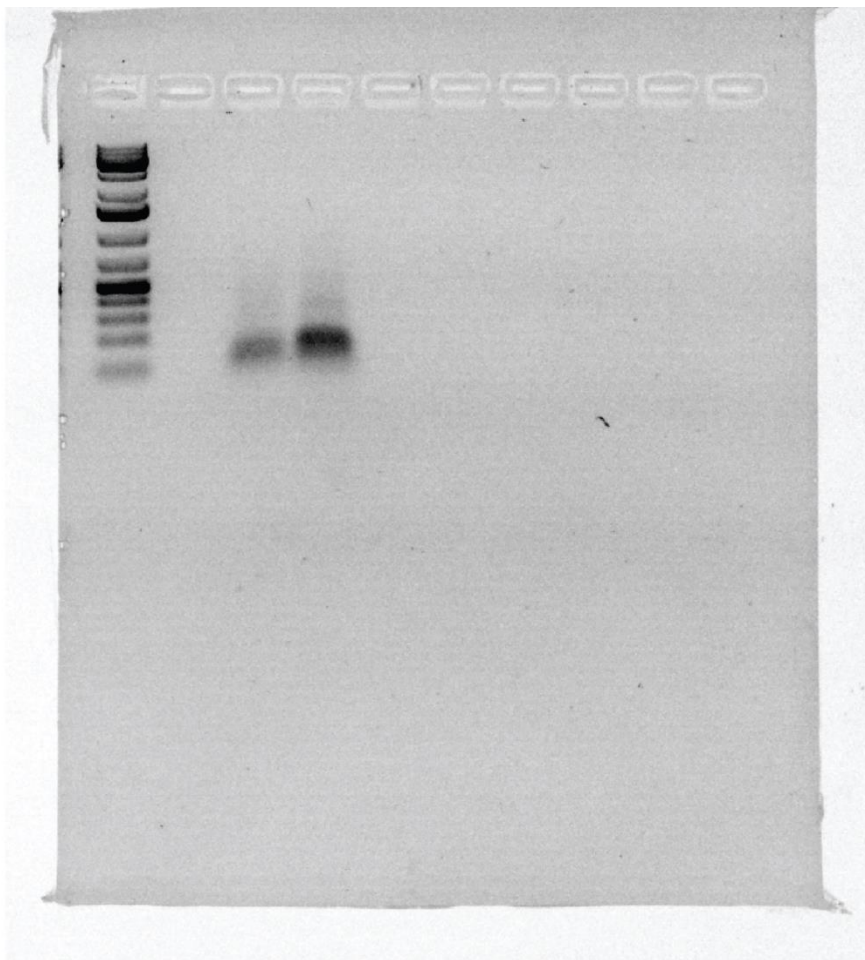
Software	Reference	Website
Illumina BCL Convert v4		https://support-docs.illumina.com/SW/BCL_Convert_v4.0/Content/SW/BCLConvert/BCLConvert.htm
Bedtools v2.26.0	⁶	https://bedtools.readthedocs.io/en/latest/
Trimmomatic v0.39	⁷	http://www.usadellab.org/cms/?page=trimmomatic
Bowtie2 v2.3.4.1	⁸	http://bowtie-bio.sourceforge.net/bowtie2/index.shtml
Picard v2.18.26		https://broadinstitute.github.io/picard/
SAMtools v1.7	⁹	https://www.htslib.org/
deepTools v3.4.3 or v3.5.0	¹⁰	https://deeptools.readthedocs.io/en/develop/
HiSAT2	¹¹	https://daehwankimlab.github.io/hisat2/
Macs2 v2.1.1	¹²	https://pypi.org/project/MACS2/
Genrich v0.6.2		https://github.com/jsh58/Genrich
STAR v020101	¹³	https://github.com/alexdobin/STAR
ChIPseeker	¹⁴	https://www.bioconductor.org/packages/release/bioc/html/ChIPseeker.html

DESeq2 v1.38.3		https://bioconductor.org/packages/release/bioc/html/DESeq2.html
GenomicRanges	15	https://bioconductor.org/packages/release/bioc/html/GenomicRanges.html
RepeatMasker v4.1.2-p1	16	https://www.repeatmasker.org/
IGV	17	https://igv.org/
ggplot2	18	https://ggplot2.tidyverse.org/
TElocal 1.1.1		https://github.com/mhammell-laboratory/TElocal
EpiProfile2.0	19	https://github.com/zfyuan/EpiProfile2.0_Family
pheatmap 1.0.12		https://rdr.io/cran/pheatmap/

Supplementary references

1. Grau-Bove, X. *et al.* A phylogenetic and proteomic reconstruction of eukaryotic chromatin evolution. *Nat Ecol Evol* (2022).
2. Henikoff, S. & Smith, M.M. Histone variants and epigenetics. *Cold Spring Harb Perspect Biol* **7**, a019364 (2015).
3. Southworth, J., Grace, C.A., Marron, A.O., Fatima, N. & Carr, M. A genomic survey of transposable elements in the choanoflagellate *Salpingoeca rosetta* reveals selection on codon usage. *Mobile DNA* **10**, 1-19 (2019).
4. Rose, N.R. *et al.* RYBP stimulates PRC1 to shape chromatin-based communication between Polycomb repressive complexes. *Elife* **5**(2016).
5. Farcas, A.M. *et al.* KDM2B links the Polycomb Repressive Complex 1 (PRC1) to recognition of CpG islands. *Elife* **1**, e00205 (2012).
6. Quinlan, A.R. & Hall, I.M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841-842 (2010).
7. Bolger, A.M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114-2120 (2014).
8. Langmead, B. & Salzberg, S.L. Fast gapped-read alignment with Bowtie 2. *Nature methods* **9**, 357-359 (2012).
9. Li, H. *et al.* The sequence alignment/map format and SAMtools. *bioinformatics* **25**, 2078-2079 (2009).
10. Ramírez, F., Dündar, F., Diehl, S., Grüning, B.A. & Manke, T. deepTools: a flexible platform for exploring deep-sequencing data. *Nucleic acids research* **42**, W187-W191 (2014).
11. Kim, D., Paggi, J.M., Park, C., Bennett, C. & Salzberg, S.L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat Biotechnol* **37**, 907-915 (2019).
12. Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome biology* **9**, 1-9 (2008).
13. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15-21 (2013).
14. Yu, G., Wang, L.-G. & He, Q.-Y. ChIPseeker: an R/Bioconductor package for ChIP peak annotation, comparison and visualization. *Bioinformatics* **31**, 2382-2383 (2015).
15. Lawrence, M. *et al.* Software for computing and annotating genomic ranges. *PLoS Comput Biol* **9**, e1003118 (2013).
16. Smit, A., Hubley, R & Green, P. RepeatMasker at <http://repeatmasker.org>. (2013-2015).
17. Robinson, J.T. *et al.* Integrative genomics viewer. *Nature biotechnology* **29**, 24-26 (2011).
18. Wickham, H. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag, New York. ISBN 978-3-319-24277-4. (2016).
19. Yuan, Z.F. *et al.* EpiProfile 2.0: A Computational Platform for Processing Epi-Proteomics Mass Spectrometry Data. *J Proteome Res* **17**, 2533-2541 (2018).

Source Data



Original blot for Supplementary Fig. 4.a.