bioRxiv preprint doi: https://doi.org/10.1101/2023.09.18.558303; this version posted September 22, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Genome-wide distribution of 5-hydroxymethyluracil and chromatin accessibility in the *Breviolum minutum* genome

Georgi K. Marinov[1,*,#], Xinyi Chen[2,*], Matthew P. Swaffer[3], Tingting Xiang[4], Arthur R. Grossman[4], and William J. Greenleaf[1,5,6,7,#]

[1] *Department of Genetics, Stanford University, Stanford, California 94305, USA*
[2] *Department of Bioengineering, Stanford University, Stanford, California 94305, USA*
[3] *Department of Biology, Stanford University, Stanford, CA 94305, USA*
[4] *Department of Plant Biology, Carnegie Institution for Science, Stanford, California 94305, USA*
[5] *Center for Personal Dynamic Regulomes, Stanford University, Stanford, California 94305, USA*
[6] *Department of Applied Physics, Stanford University, Stanford, California 94305, USA*
[7] *Chan Zuckerberg Biohub, San Francisco, California, USA*
[*] *These authors contributed equally to this work*
[#] *Corresponding author*

Abstract

In dinoflagellates, a unique and extremely divergent genomic and nuclear organization has evolved. The highly unusual features of dinoflagellate nuclei and genomes include permanently condensed liquid crystalline chromosomes, primarily packaged by proteins other than histones, genes organized in very long unidirectional gene arrays, a general absence of transcriptional regulation, high abundance of the otherwise very rare DNA modification 5-hydroxymethyluracil (5-hmU), and many others. While most of these fascinating properties were originally identified in the 1970s and 1980s, they have not yet been investigated using modern genomic tools. In this work, we address some of the outstanding questions regarding dinoflagellate genome organization by mapping the genome-wide distribution of 5-hmU (using both immunoprecipitation-based and basepair-resolution chemical mapping approaches) and of chromatin accessibility in the genome of the Symbiodiniaceae dinoflagellate *Breviolum minutum*. We find that the 5-hmU modification is preferentially enriched over certain classes of repetitive elements, often coincides with the boundaries between gene arrays, and is generally correlated with decreased chromatin accessibility, the latter otherwise being largely uniform along the genome. We discuss the potential roles of 5-hmU in the functional organization of dinoflagellate genomes and its relationship to the transcriptional landscape of gene arrays.

Introduction

Dinoflagellates are perhaps the most remarkable lineage within the spectrum of known eukaryote diversity, with numerous extreme deviations from the genomic and cellular organization of other eukaryotes, especially regarding their highly unusual nuclei[1–6]. They are also a very diverse, successful, and ecologically important group that includes numerous photosynthetic lineages, free living heterotrophs, and even parasites, playing a major ecological role in marine ecosystems. The best known such example is the endosymbiotic association of Symbiodiniaceae dinoflagellates[7] with reef-building corals. The photosynthetic capability of the dinoflagellate symbionts provides the metabolic foundation for the highly biologically diverse reef ecosystems[8], and the expulsion of these symbionts from their host cells upon heat stress causes coral "bleaching" and the eventual death of coral reefs[9], an increasingly acute problem in the modern world due to the effects of global climate change[10].

The list of unorthodox features of dinoflagellate nuclei is long[1,3,4,11–14]. Dinoflagellate chromosomes exist in a permanently condensed liquid crystalline state throughout most of the cell cycle, and are characterized by an unusually low protein-to-DNA ratio (1:10, compared to 1:1 in other eukaryotes[2,15]). This condensed, protein-poor structure is caused by the loss of nucleosomal histones as the main packaging component of chromatin. This role has instead been taken over by a distinct set of proteins – small dinoflagellate-specific virus-derived nucleoproteins (DVNP) and histone-like proteins (HLPs)[16–24] – that appear to have been acquired through horizontal gene transfer from viruses and bacteria, respectively[25,26]. Such chromatin composition is

bioRxiv preprint doi: https://doi.org/10.1101/2023.09.18.558303; this version posted September 22, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

an extreme departure from the norm for an eukaryote, as nucleosomal chromatin is otherwise universal[27]. Furthermore, the four core histones – H2A, H2B, H3 and H4 – are the most conserved of all eukaryotic proteins, a result due not just to their role in packaging DNA, but also because they, especially in their N-terminal tails, are subject to extensive posttranslational modifications (PTMs) that serve as the basis of the so-called "histone code"[28]. This code plays important roles in all chromatin-related biological processes, such as transcription, gene regulation, replication, DNA repair and others. Dinoflagellates have not lost histones – in fact multiple and highly diverse histone genes are retained in all dinoflagellates for which genomic data is available[29] – but these histone proteins are highly divergent from those of other eukaryotes and such divergence makes it unclear what role these proteins might play in dinoflagellate nuclei. DVNPs and HLPs appear to have been acquired at different points in dinoflagellate evolution[26] – DVNPs are found in all dinoflagellates, but HLPs are absent in early branching lineages and common only to core dinoflagellates. Another curious feature of dinoflagellate nuclei is the high abundance of divalent cations such as Ca^{2+} and Mg^{2+}, which have been proposed to play a role in the stabilization of chromosome structure[30–33].

Genome organization in dinoflagellates also represents a highly derived state, as their genes are organized into long unidirectional gene arrays[34–37], presumably transcribed as a single unit. *Trans*-splicing is ubiquitous in the group, with mature mRNAs generated through the addition of a spliced leader (SL) sequence[34,38,39]. Transcriptional regulation is thought to be largely absent, with all genes transcribed at all times. The primary mode of gene regulation is presumed to be at the level of translation and/or RNA stability.

We still know very little about the inner workings of these remarkable nuclei, and most of the numerous fascinating questions regarding how eukaryotic transcription, replication and DNA repair occur and are regulated in the absence of histones and within permanently condensed liquid crystalline chromosomes remain unanswered. Recently, we and others[40,41] began to unravel some of these mysteries by applying three-dimensional genome conformation mapping using Hi-C[42] to the members of Symbiodiniaceae *Breviolum minutum* and *Symbiodinium microadriaticum*, showing that the genome is folded into distinct topologically associating domains coinciding with pairs of divergent gene arrays and separated by the points where convergent gene arrays meet (termed "dinoTADs"). These domains appear to be the product of strong transcription-induced supercoiling in a context of extremely long transcriptional units and the absence of histones.

Earlier, de Mendoza et al.[43] mapped the distribution of the frequently found in eukaryotes 5-mehylcytosine (5mC) modification in two members of the Symbiodiniaceae – *Fugacium kawagutii* and *Breviolum (Symbiodinium) minutum*. An unusual compared to other eukaryotes pattern of uniform hypermethylation throughout the genome was observed.

Two other still unanswered questions concern the potential roles and genomic distribution of the 5-hmU modification in dinoflagellate genomes and the characteristics of the chromatin accessibility landscape. While it appears that nucleosomes are not the main packaging component of dinoflagellate chromatin, it is not known whether DVNPs and/or HLPs might provide similar levels of physical protection of DNA. Past studies have suggested that DVNPs bind to DNA with similar affinity to histones[25] but HLPs, although they can compact DNA in a concentration-dependent manner, have weaker affinity than histones[44–46]. However, whether and how these proteins confer distinct chromatin states through their association with DNA is not known. Thus, whether dinoflagellate genomes are uniformly accessible given the lack of histones, whether distinct open chromatin region exist as in conventional eukaryotes, and whether perhaps the inverse phenomenon is observed – localized areas of decreased accessibility – is an open question.

That dinoflagellates contain the otherwise highly unusual for eukaryotes 5-hmU modification (present in abundance and discovered originally in some phages[47]) was first realized in the 1970s[48–50]. It was found that unexpectedly large fractions of thymines (T) in the genome of various species are replaced by 5-hmU – 12% in *Exuviaella cassubica*[50], 12% in *Symbiodinium microadriaticum*[50], 37-38% in *Crypthecodinium cohnii*[50,51], 62% in *Amphidinium carterae*[50], 62.8% in *Prorocentrum micans*[52], and 68% in *Peridinium triquetrum*[50]. What functions 5-hmU might have is not known, but it has been suggested that it enhances the flexibility and hydrophilicity of double-stranded DNA[53], especially in some sequence contexts[54–56].

Curiously, 5-hmU has also been observed in non-negligible quantities[57] in another major lineage of eukaryotes – the important parasitic clade Kinetoplastida. There are many parallels between the genomic organization of dinoflagellates and kinetoplastids[58] – although kinetoplastids have conventional nucleosomal chromatin, they too have lost transcriptional regulation as a primary mechanism for controlling gene expression and their genes are also organized into long arrays[59–62], with mature mRNAs being the product of *trans*-splicing[63–67]. These properties are shared with other members of the larger Euglenozoa lineage that have been studied, such as *Euglena gracilis*[68]. However, in kinetoplastids 5-hmU appears to be simply a precursor to the synthesis of the larger modification β-D-Glucopyranosyloxymethyluracil, better known as base J[69–71], which does play a significant role in their genomes. Base J replaces about 1% of thymines and is predominantly found in repetitive DNA, especially in telomeric regions[72–74], but more importantly, it also demarcates the boundaries between gene arrays[75] and likely prevents transcriptional readthrough events[76,77]. The free living *Euglena* also has base J[68], and thus 5-hmU too. It is likely that so do all members of the larger Euglenozoa grouping (comprising the kinetoplastids, euglenids, diplonemids and a few smaller

bioRxiv preprint doi: https://doi.org/10.1101/2023.09.18.558303; this version posted September 22, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

clades), but that has not yet been directly demonstrated.

To answer these questions, we mapped the 5-hmU distribution in the genome of *B. minutum*, finding that it is enriched over certain repetitive element classes and often around the boundaries between gene arrays. In contrast, chromatin accessibility is anti-correlated with elevated 5-hmU levels; this inverse relationship is specifically strong around gene array/dinoTAD boundaries, pointing to potential localization of histones (or other proteins that protect DNA) to regions enriched for 5-hmU (and thus conferring them greater protection from transposase insertion). We do not detect increased accessibility associated with transcription start sites (TSSs), and generally we do not observe strongly localized DNA accessibility peaks in the genome comparable to those in metazoans. These results provide a foundation for the future detailed understanding of the organization of transcription in dinoflagellates and its interplay with DNA modifications.

Results

Mapping 5-hmU and chromatin accessibility in *B. minutum*

In order the map the distribution of 5-hmU in the *B. minutum* genome, we first adapted the MeDIP (Methylated DNA ImmunoPrecipitation) protocol[78] for mapping DNA methylation using high-throughput sequencing (MeDIP-seq[79]). We used an antibody specific to 5-hmU (see Methods) and a spike-in control to confirm specific enrichment of 5-hydoxymethyluracil. As mammalian genomes do not contain appreciable amounts of 5-hmU, we used a mixture of human and *B. minutum* genomic DNA (gDNA) as input to the MeDIP procedure, and we also sequenced three different controls – input DNA, "depleted" DNA (the supernatant remaining after the immunoprecipitation step), and an IgG control (using only beads with no primary antibody). We observed that the fraction of human reads decreased ∼2× after 5-hmU MeDIP relative to controls (Figure 1A), confirming the specific enrichment of dinoflagellate DNA. We also made an interesting observation – 5-hmU MeDIP is also ∼2× enriched for multimapping reads compared to the controls (Figure 1B), hinting that 5-hmU might be preferentially associated with repetitive elements.

We did not observe enrichment or 5-hmU around the starting positions of genes (Figure 1C) – in fact we observe a slight depletion ±1-kb around gene starts (note that the three spikes observed in the plot are an artefactual result due to the presence of collapsed repeats in the current *B. minutum* assembly; see further discussion on this topic below).

We also deployed an orthogonal method for mapping 5-hmU at base-pair resolution using chemical conversion of 5-hmU into cytosine C (see the Methods section for details). In contrast to the 5mC modification in mammals, which is found specifically in a CpG context, we do not find any sequence preference for T bases modified into 5-hmU in *B. minutum* (Figure 1D). We note that early studies from the 1980s reported that 5-hmU preferentially replaces thymines in TA and TC sequence contexts[51], and we do not recover such a preference in our datasets; it is possible that such preferences indeed exist in *Crypthecodinium cohnii*, which was assayed by those studies, but not in Symbiodiniaceae, or that the discrepancy is due to methodological differences. Because the current chemical mapping protocol does not provide for 100% conversion rate of 5-hmU modified bases it is not possible to estimate the absolute levels of 5-hmU in the *B. minutum* genome based on chemical mapping data alone.

To map the *B. minitum* chromatin accessibility landscape, we utilized ATAC-seq[80] (Assay of Transposase-Accessible Chromatin using sequencing), specifically in its omniATAC[81] modification (see Methods). We generated a very deeply sequenced (∼130 million mapped reads) library from actively growing cells as well as two replicates each for cells grown at the usual temperature of 27 °C and heat-stressed cells that were incubated at 34 °C.

In eukaryotes with nucleosomal chromatin, ATAC-seq libraries sequenced in a paired-end format display a characteristic nucleosomal signature in their fragment length distribution, with a subnucleosomal peak at ≤∼120 bp, a prominent mononucleosomal peak, and a weaker dinucleosomal peak. *B. minutum* ATAC-seq only displays a peak at short fragment lengths (∼ 60 bp), with no nucleosomal peaks (Figure 1E). Thus, we conclude that wherever they are found in the genome, nucleosomes apparently are of too low abundance to substantially affect the overall fragment length distribution, while DVNPs and HLPs do not form structures consisting of multiple closely positioned proteins that strongly protect against transposition. We also observe a modest depletion of multimapping reads in ATAC-seq libraries relative to a matched naked gDNA control (Figure 1F), i.e. the opposite trend of that observed for MeDIP-seq. ATAC-seq signal is also not enriched around gene start positions (Figure 1G).

Genome browser inspection of ATAC-seq and gDNA controls (Supplementary Figures 1 and 2) revealed that the available *B. minutum* assembly includes multiple collapsed repeats, i.e. regions that exist in multiple copies in the actual genome but are only present in the assembly as a single copy (or as many fewer copies than their actual abundance in the genome). This complicates the interpretation of sequencing datasets as these regions appear as artificial "peaks" if analysis is not carried out against a proper control. Therefore, we performed all subsequent analysis as a comparison against matched input or negative gDNA controls. The regions of enrichment over gDNA that we identified did not show a concentration around gene starts/TSSs (Figure 1H)., and they show overall lower enrichment over background/controls than ATAC-seq peaks in human datasets[82] (Supplementary Figure 3), i.e. we do not really observe strongly localized chromatin accessibility as in other eukaryote genomes. Comparing the heat

bioRxiv preprint doi: https://doi.org/10.1101/2023.09.18.558303; this version posted September 22, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

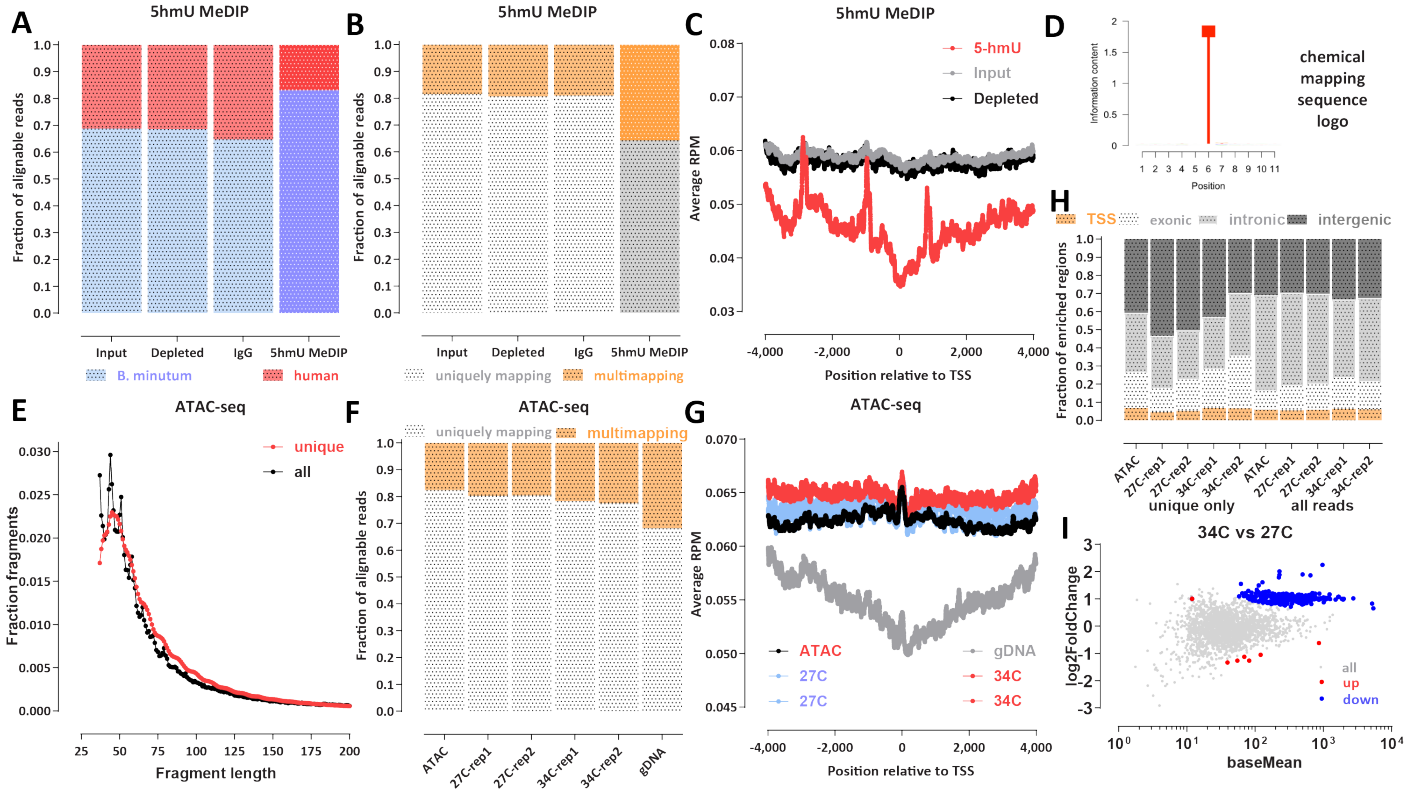


Figure 1: Mapping the 5-hmU and chromatin accessibility landscape in *B. minutum*. (A) Proportions of human and *B. minutum* gDNA in 5-hmU Methylated DNA immunoprecipitation sequencing (MeDIP-seq) and control libraries. A mixture of human and dinoflagellate gDNA was used as input to MeDIP-seq experiments, and the fraction of reads that map to each genome is shown. The 5-hmU MeDIP-seq library is enriched for dinoflagellate reads confirming the specificity of 5-hmU pull down. (B) Proportion of multimapping reads in 5-hmU MeDIP-seq and control libraries. The 5-hmU MeDIP-seq library exhibits a higher fraction of multimapping reads, suggesting that 5-hmU is enriched over repetitive elements. (C) Metaprofiles of 5-hmU and control libraries signal over *B. minutum* transcription start sites/gene starts. (D) Basepair-resolution chemical mapping of 5-hmU does not reveal a sequence motif associated with the modification in *B. minutum*. (E) Fragment length distribution of *B. minutum* ATAC-seq datasets. Shown are uniquely mapping reads alone as well as all reads that can be mapped. (F) Proportion of multimapping reads in *B. minutum* ATAC-seq datasets as well as a control genomic DNA (gDNA) library. (G) Metaprofiles of ATAC-seq signal over *B. minutum* transcription start sites/gene starts as well as the gDNA control. (H) Distribution of ATAC-seq regions of enrichment relative to annotated genomic features. (I) Differential accessibility analysis for the 27 °C and 34 °C conditions.

stressed (34 °C) and normal temperature (34 °C) conditions did not reveal large scale changes in the chromatin accessibility landscape (Figure 1I).

Heterologous expression of DVNPs has a modest effect on chromatin organization in the yeast *S. cerevisiae*

Previous studies had examined the effect of DVNPs on chromatin structure by expressing a DVNP (*Hematodinium sp.* DVNP.5) in the yeast *Saccharomyces cerevisiae*[83]. The resulting changes in the chromatin landscape (measured using MNase-seq) were reported to reveal nucleosome disruption, while overall the expression of the DVNPs had a negative effect on cell growth, likely because it impaired transcription. We sought to replicate and expand on these results by expressing several DVNPs in *S. cerevisiae* and carrying out ATAC-seq as well as single-molecule footprinting (SMF[84,85]; providing information about the absolute levels of accessibility/protection along the genome associated with DVNP occupancy).

We hetorologously expressed (see Methods) three different DVNPs in yeast – the previously assayed *Hematodinium sp.* DVNP.5 as well as *Hematodinium sp.* DVNP.12 and *B. minutum* DVNP symbB.v1.2.006931. *Hematodinium sp.* DVNP.5 had a negative effect on growth, as previously reported, but much more strongly so than the other two. We carried out ATAC-seq and SMF using an internal control in all experiments – *Candida glabrata* cells, which we used to account for experimental variation, as previously described[86].

bioRxiv preprint doi: https://doi.org/10.1101/2023.09.18.558303; this version posted September 22, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

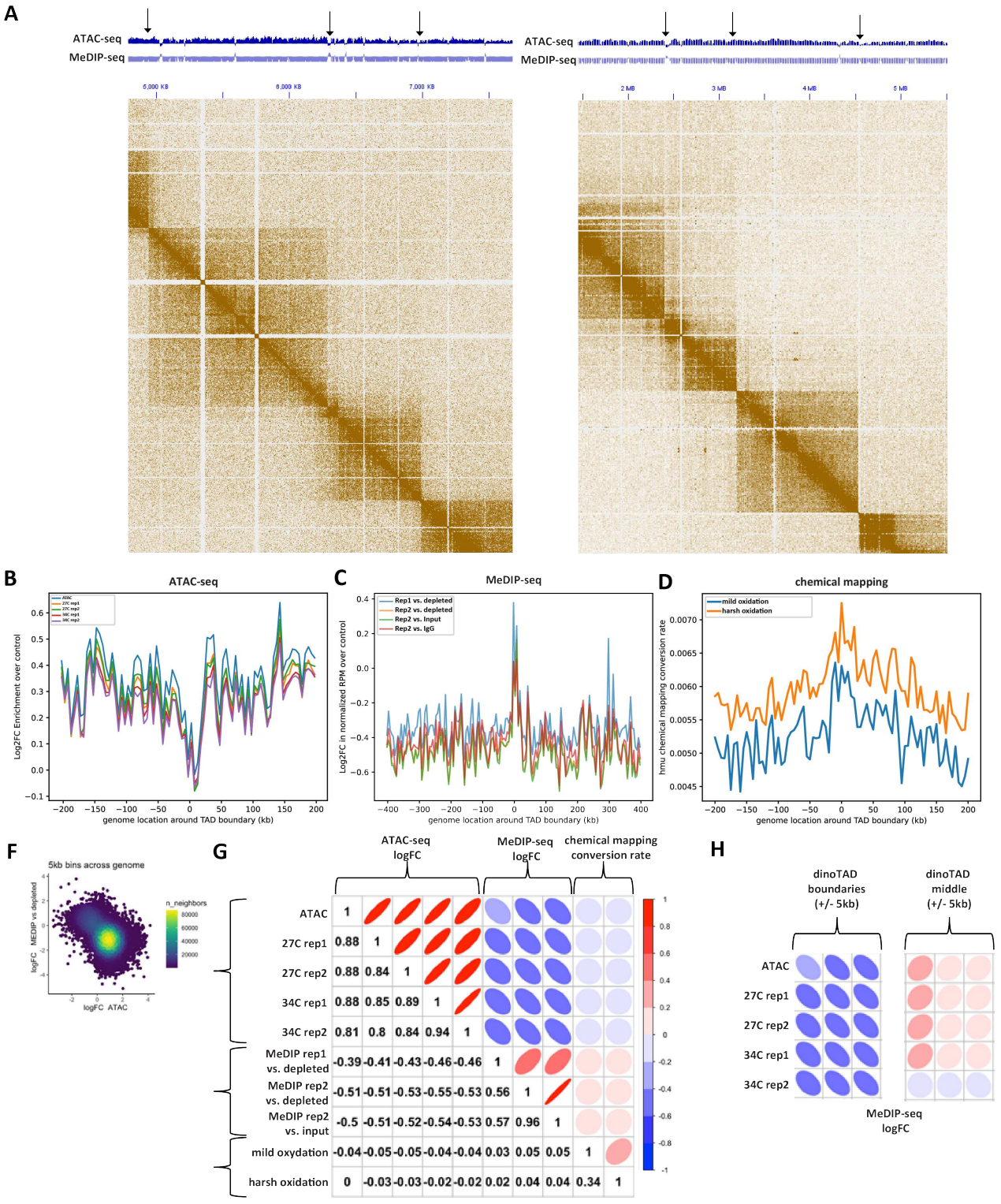


Figure 2: Inverse correlation between 5-hmU and chromatin accessibility and association with dinoTADs boundaries in the *B. minutum* genome. (A-B) Representative snapshots of the distribution of 5-hmU enrichment and decreased chromatin accessibility relative to dinoTAD boundaries. (C) Depletion of ATAC-seq signal around dinoTAD boundaries. (D) Enrichment of MeDIP-seq signal around dinoTAD boundaries. (E) Increased 5-hmU chemical mapping conversion rate around dinoTAD boundaries. (F-G) ATAC-seq and MeDIP-seq are generally anti-correlated (calculated for 5-kbp bins over the whole genome) (H) ATAC-seq and MeDIP-seq are specifically strongly anti-correlated around dinoTAD boundaries.

bioRxiv preprint doi: https://doi.org/10.1101/2023.09.18.558303; this version posted September 22, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

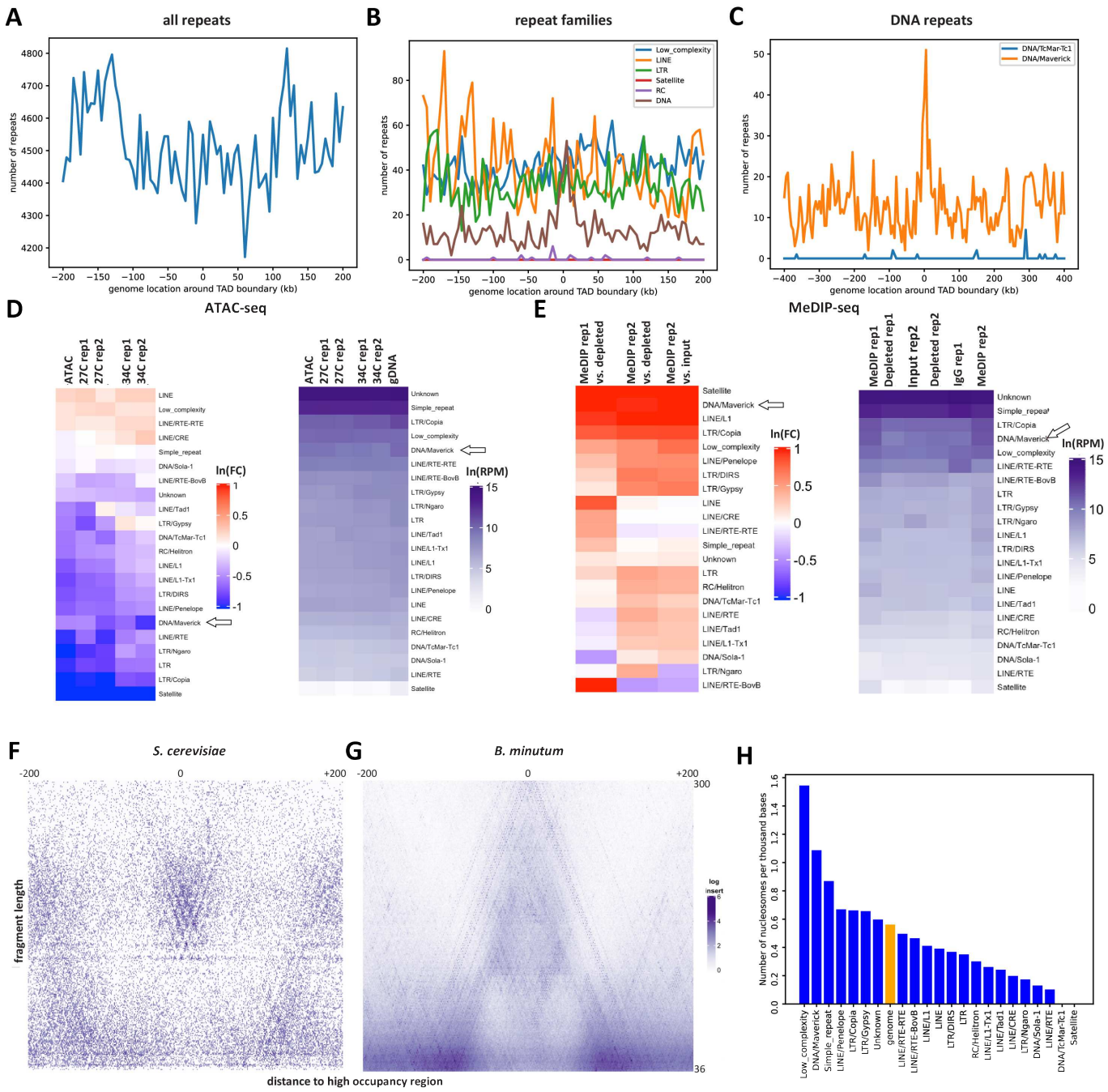


Figure 3: Association of 5-hmU and chromatin accessibility with repetitive elements in the *B. minutum* genome. (A-C) Distribution of all repeats, individual repeat families, and some DNA elements around dinoTAD boundaries. (D) ATAC-seq enrichment/depletion over repetitive elements. (E) MeDIP-seq enrichment/depletion over repetitive elements. (F-G) V-plot[94] around positioned nucleosomes in S. cerevisiae (for comparison) and *de novo* identified putative positioned nucleosomes in *B. minutum*. (H) Enrichment/depletion of positioned nucleosomes over repetitive elements.

ATAC-seq did not reveal dramatic changes in the accessibility landscape upon DVNP expression (Supplementary Figures 5 and 4) except perhaps for a slight decrease in the height of some peaks. On the other hand, SMF data showed a decrease in accessibility around TSSs and reduced strength of nucleosome positioning (Supplementary Figure 5), broadly consistent with the previous MNase-seq results suggesting that nucleosome disruption is induced by DVNPs[83]. This disruption is, however, apparently not sufficient to dramatically reshape the accessibility landscape and represents a moderate quantitative rather than a major qualitative alteration.

bioRxiv preprint doi: https://doi.org/10.1101/2023.09.18.558303; this version posted September 22, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Inverse correlation between 5-hmU and chromatin accessibility

Next, we examined the distribution of 5-hmU and chromatin accessibility around other available genomic features. We noticed that in many cases (although this is not an exclusive association) 5-hmU is enriched around the boundaries of dinoTADs while ATAC-seq shows decreased accessibility in those same regions (Figure 2A). We generalized this observation by evaluating the global ATAC-seq distribution around dinoTAD boundaries (Figure 2C-E) and found that indeed ATAC-seq is globally depleted nearby these locations, while MeDIP-seq is enriched and 5-hmU chemical conversion rate is also elevated. The observed anti-correlation between chromatin accessibility and 5-hmU is specifically strong around dinoTAD boundaries while we do not observe substantial inverse correlation between the two in the middle of dinoTAD domains (Figure 2H). However, these observations are trends and not universal patterns, as a number of gene array boundaries do not show strong MeDIP enrichment and ATAC-seq depletion.

Association of 5-hmU and chromatin accessibility with repetitive elements

Because of the previously noted enrichment and depletion of multimapping reads in 5-hmU and ATAC libraries, respectively, we next aimed to identify which, if any, repetitive elements might be specifically associated with 5-hmU and/or ATAC. We first examined the distribution of annotated repetitive elements (see Methods for details) around dinoTAD boundaries (Figure 3A-C), and found no specific preference at dinoTAD boundaries neither for repeats as a whole, nor for any specific repeat family. However, Maverick DNA elements did exhibit strong enrichment around the edges of dinoTADs (Figure 3C). Maverick elements are also known as Polintons, are typically 15-40 kbp in size, and often encode putative viral capsid proteins, suggesting that they might form virions under some conditions[87–91], a view supported by the large abundance of Polinton-like viruses reported in aquatic ecosystems[92]. Maverick does not account for all dinoTAD boundaries though – while a majority of TAD boundaries show 5-hmU enrichment, Maverick elements are found around only ∼11% of them (Supplementary Figure 6).

Global analysis of ATAC-seq depletion/enrichment over repetitive elements (Figure 3D) showed that most repeats are depleted for accessibility, with Copia LTR and Maverick DNA elements most highly abundant in the gDNA control relative to ATAC-seq sample. The exceptions are CRE and RTE-RTE LINE elements. MeDIP-seq data reveals a generally inverse picture – most repeats are enriched in the MeDIP libraries, apart from some LINE elements (Figure 3E), with Maverick/Polinton DNA repeats most strongly enriched for the 5-hmU modification.

These results point to increased protein occupancy and elevated 5-hmU levels over repetitive elements such as Maverick. We therefore asked whether we can specifically find nucleosomes corresponding to certain repeat classes. We utilized the nucleoATAC algorithm[93] to identify positioned nucleosomes genome-wide in the *B. minutum* genome (see Methods; with the caveat that the nucleoATAC algorithm was designed to look for eukaryote-like nucleosomes and we do not know if this is still the case in dinoflagellates). We identified 30,107 low-resolution and 2,166 high-resolution putative positioned nucleosomes; these are overall not preferentially located to well defined general genomic features such as dinoTAD boundaries (Supplementary Figure 7). V-plot analysis[94] of the fragment distribution around positioned nucleosomes revealed an A-shaped structure, with a peak in the 120-160 bp range (this fragment length is higher for the smaller set of high-resolution nucleosomes; Supplementary Figure 7), flanked by very short fragments. This observation is distinct from what is observed in other eukaryotes such as yeast (Figure 3F-G), where multiple nearby nucleosomes are visible. We interpret these structures as arising from a single positioned protective feature, quite possibly a histone-based nucleosome, without other strongly positioned nearby nucleosomes. We note that these observations are not explainable by mappability biases (i.e. only a single nucleosome is observed because all adjacent sequences are not mappable), as we carried out this analysis while allowing for multimapping reads and the center point of the putative positioned nucleosomes is in fact slightly less uniquely mappable than the flanks (Supplementary Figure 8).

Strikingly, Maverick DNA elements are preferentially enriched for positioned nucleosomes, at ∼2× the genomic average (Figure 3H). This observation corroborates the depletion of ATAC-seq signal observed over these elements.

Discussion

In this study, we provide the first global maps of the distribution of the 5-hmU modification and chromatin accessibility in a dinoflagellate species (*B. minutum* in the Symbiodiniaceae clade). Our results point to the preferential enrichment for 5-hmU over certain classes of repetitive elements and also around the boundaries of the previously identified dinoTAD topologically associating domains that also coincide with the points of convergence of the long unidirectional arrays into which dinoflagellate genes are organized. In contrast, chromatin accessibility is depleted in those same areas and is generally anti-correlated with high levels of 5-hmU. We do not observe strong accessibility peaks as seen in eukaryotes with conventional nucleosomal chromatin, nor do we see any preferential accessibility around transcription start sites, suggesting that most of the dinoflagellate genome is not protected by strings of nucleosomes and is generally uniformly physically accessible. We also do not see evidence for large-scale domains of preferential accessibility/protection; previously ATAC-see[95] was used to visualize through microscopy the accessi-

bioRxiv preprint doi: https://doi.org/10.1101/2023.09.18.558303; this version posted September 22, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

ble chromatin in the cells of the basal dinoflagellate *Hematodinium*[13], which suggested the existence of large-scale "open" and "closed" compartments, perhaps encompassing whole chromosomes. However, no ATAC sequencing data is available for *Hematodinium*, thus we do not know how these open/closed compartments translate into genome-wide sequencing profiles. It is possible that the open compartment is fully open and the closed one is fully closed, i.e. without distinct loci of open chromatin, resulting in a similar average genome-wide profile to what we observe. Alternatively, chromatin structure in basal and core dinoflagellates (Symiodiniaceae belonging to the latter) may differ substantially. The application of single-molecule footprinting techniques to dinoflagellates should be able to resolve these possibilities in the future[85,96] though at present it is hampered by the high sequencing coverage over the very large dinoflagellate genomes that is required. We do identify several thousand putative positioned nucleosomes; however, these, if they are confirmed to be indeed histone-based nucleosomes, appear to be isolated and not parts of larger-order structures. An interesting trend that emerges is the association of elevated 5-hmU, decreased chromatin accessibility and increased frequency of positioned nucleosomes over certain repetitive elements, in particular Maverick/Polinton DNA elements, which are also enriched over dinoTAD boundaries. This is, however, by no means an absolute rule as not all dinoTAD boundaries are associated with such features.

Nevertheless, it is tempting to draw parallels between these initial observations in dinoflagellates and what is known in much more details for kinetoplastids[58]. The latter group shares the same general loss of transcriptional regulation as a primary mechanism for modulating gene regulation and the organization of the genome into long unidirectional gene arrays, and in the kinetoplastids base J demarcates the regions between these arrays. Base J is synthesized through 5-hmU as an intermediate, and thus 5-hmU also is localized to the same regions of the genome in the cases where it has been measured (e.g. in *Leishmania*[97,98]). It is possible that it may play an analogous role in dinoflagellates, even though they have not evolved the further chemical elaboration needed that for base J synthesis.

However, such a speculation still leaves many unanswered questions. What is the mechanistic role of 5-hmU? In our previous work in which dinoTAD structures were discovered[40], we showed them to depend on transcriptional activity and to disappear upon blocking transcription, i.e. they are most likely the product of extreme transcription induced DNA supercoiling. At the same time, 5-hmU has been reported to increase the flexibility of the DNA double helix[53–56], which may suggest a possible role for 5-hmU in alleviating the supercoiling stress under which dinoflagellate genomes appear to exist. However, the mechanistic details of such a link are currently unclear.

Why does 5-hmU vary so much between different dinoflagellate species – from 12% to 68% where it has been assayed – and where is it located in the genome in the extreme cases? The preferential localization to dinoTAD boundaries suggested from our work is consistent with genome-wide rates of 5-hmU on the lower end of this spectrum (which is also what the available data for other Symbiodiniaceae points to[50]), as array boundaries are fairly localized and encompass a minor fraction of the whole genome. However, the *B. minutum* genome is relatively small for a dinoflagellate – only on the order of 1 Gbp – while other species have much larger and more repeat-rich genomes, which might be related to higher overall 5-hmU levels given the preference of 5-hmU for repetitive elements that we observe. To better understand its properties and functions, it will be important to assay 5-hmU in a wide variety of dinoflagellate species with diverse genomic characteristics.

Similarly, it will also be vital to obtain very high-quality genome assemblies to work with. For example, in our current work we have not been able to test whether 5-hmU is strongly associated with telomeres the way base J is in kinetoplastids, as the currently available *B. minutum* assembly is of too poor quality to allow such analysis.

What is the precise role of histones, DVNPs and HLPs in dinoflagellate genomes? Here, we demonstrate decreased chromatin accessibility over certain regions in the genome as well as the existence of nucleosome-like structures, which suggests the presence of nucleosomes along the genome. However, mapping of histone/DVNP/HLP occupancy using chromatin immunoprecipitation (ChIP) will be needed to generate direct genome-wide profiles of the distribution of these proteins along the genome. Currently this is precluded due to the extreme sequence divergence of dinoflagellate histone proteins[29], which makes existing anti-histone antibodies unreliable reagents for carrying out such experiments in dinoflagellates. Establishing the absolute levels of protection/occupancy in dinoflagellate genomes, through the application of methylation-based (especially long read-based) and enzymatic approaches[99] will also be highly valuable.

Methods

B. minutum cell culture

The clonal axenic *Symbiodinium/Breviolum minutum* strain SSB01 was used in all experiments. Stock cultures were grown as previously described[100,101] in Daigo's IMK medium for marine microalgae (Wako Pure Chemicals) supplemented with casein hydrolysate (IMK+Cas) at 27 °C at a light intensity of 10 μmol photons m^{-2} s^{-1} from Philips ALTO II 25-W bulbs on a 12-h-light:12-h-dark cycle. The medium was prepared in artificial seawater (ASW).

Genomic DNA isolation

B. minutum genomic DNA was isolated as previously described[100]. Briefly, cells were centrifuged at 1,000 *g* for 5 minutes, then resuspended in 500 μL 1× Cell Lysis

bioRxiv preprint doi: https://doi.org/10.1101/2023.09.18.558303; this version posted September 22, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Buffer (prepared by mixing equal volumes of 2× Cell Lysis Buffer – 2% SDS, 400 mM NaCl, 40 mM EDTA, 100 mM Tris-HCl, pH 8.0 – and H_2O) and vortexed. The lysed cells were mixed with an equal 500 μL volume Phenol:Chloroform:Isoamyl alcohol (25:24:1), and mixed well by inverting a few times. The phases were centrifugation at 13,000 g for 5 minutes, then the top phase was transferred to a new tube and treated with 4 μL Ribonuclease A (20 mg/mL) by incubating for 30 minutes at 37 °C.

DNA was purified by adding an equal volume Phenol:Chloroform:Isoamyl alcohol (25:24:1), mixing well and centrifuging at 13,000 g for 5 minutes, then transferring the top layer to a new tube, to which Phenol:Chloroform:Isoamyl alcohol (25:24:1) was added again, and the centrifugation and top phase isolation was repeated. Then 2.5× volumes of 100% EtOH were added and the mixture was incubated on ice for 30 minutes or at -20 °C overnight. The solution was then centrifuged at 13,000 g at room temperature for 20 minutes, the pellet was washed with 70% EtOH, dried on air and resuspended in 50 μL H_2O.

ATAC-seq experiments

ATAC-seq experiments were performed following the omni-ATAC protocol[81].

Briefly, ∼100K *B. minutum* cells were centrifuged at 1,000 g, then resuspended in 500 μL 1× PBS and centrifuged again. Cells were then resuspended in 50 μL ATAC-RSB-Lysis buffer (10 mM Tris-HCl pH 7.4, 10 mM NaCl, 3 mM $MgCl_2$, 0.1% IGEPAL CA-630, 0.1% Tween-20, 0.01% Digitonin) and incubated on ice for 3 minutes. Subsequently 1 mL ATAC-RSB-Wash buffer (10 mM Tris-HCl pH 7.4, 10 mM NaCl, 3 mM $MgCl_2$, 0.1% Tween-20, 0.01% Digitonin) were added, the tubes were inverted several times, and nuclei were centrifuged at 500 g for 5 min at 4 °C.

Transposition was carried out by resuspending nuclei in a mix of 25 μL 2× TD buffer (20 mM Tris-HCl pH 7.6, 10 mM $MgCl_2$, 20% Dimethyl Formamide), 2.5 μL transposase (custom produced) and 22.5 μL nuclease-free H_2O, and incubating at 37 °C for 30 min in a Thermomixer at 1000 RPM.

Transposed DNA was isolated using the MinElute PCR Purification Kit (Qiagen Cat# 28004/28006), and PCR amplified as previously described[81]. Libraries were purified using the MinElute kit, then sequenced on a Illumina NextSeq 550 instrument as 2×36mers or as 2×75mers.

ATAC-seq control experiments

Genomic DNA controls for ATAC-seq were generated by transposing purified gDNA. Briefly, 100 ng of gDNA were mixed with 2 μL Tn5, 25 μL 2×TD buffer and H_2O for a total volume of 50 μL, then incubated at 55 °C for 5 minutes. The reaction was stopped by immediately proceeding with DNA isolation using the MinElute kit. Libraries were generated as described above for ATAC-seq.

Genome assemblies

Datasets were processed against either the original *B. minutum* assembly[35] or against the Hi-C scaffolded assembly for *B. minutum* previously described[40], which is based on the original fragmented assembly for this species[35] and scaffolded into chromosome-level contigs using Hi-C data following established protocols[102].

General analysis procedures

Browser tracks generation, fragment length estimation, and other analyses were carried out using custom-written Python scripts (`https://github.com/georgimarinov/GeorgiScripts`).

Mappability track generation

Mappability tracks were generated as by tiling the whole genome with reads of length RL starting at each position. These reads were the mapped back to the genome using the same settings used for processing real datasets. Average mappability over each position was calculated as the ratio RC/RL between its read coverage RC and the read length RL.

ATAC-seq data processing

Demultipexed FASTQ files were mapped as 2×36mers using Bowtie[103] with the following settings: `-v 2 -k 2 -m 1 --best --strata -X 1000`. Duplicate reads were removed using `picard-tools` (version 1.99). This mapping generated a set of uniquely mapping alignments only.

For the purpose of the analysis of multimappers, alignments were generated with unlimited alignment multiplicity with the following settings: `-v 2 -a --best --strata -X 1000`.

Normalization of multimappers was performed using the previously described[104,105] method of dividing each alignment by its read multiplicity, i.e:

$$S_{c,i} = \frac{\sum_{R \in R_{c,i}} \frac{1}{NH_R}}{\frac{|R|}{10^6}} \quad (1)$$

Where $S_{c,i}$ is the signal score for position i on chromosome c (in RPM, or Read Per Million mapped reads units), $|R|$ is the total number of mapped reads, $|R_{c,i}|$ is the number of reads covering position i on chromosome c, and NH_R is the number of locations in the genome a read maps to.

bioRxiv preprint doi: https://doi.org/10.1101/2023.09.18.558303; this version posted September 22, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

ATAC-seq peak calling

Peak calling was carried out using MACS2[106], with the gDNA library as a control, and with the following settings: `-g 569785352 -f BAMPE`. Differentially accessible regions were identified using DESeq2[107].

Analysis of positioned nucleosomes

The analysis of positioned nucleosomes was carried out using NucleoATAC[93]. We used the low resolution nucleosome calling program `nucleoatac occ` with default parameters that requires ATAC-seq data and genomic windows of interest, and returns a list of nucleosome positions based on the distribution of ATAC-seq fragment lengths centered at these positions. Sliding windows of 1 kbp in steps of 500 bp were taken as inputs, and redundant nucleosome positions were eventually discarded. V-plots were made by aggregating unique-mapping ATAC-seq reads centered around the positioned nucleosomes, and mapping the density of fragment sizes versus fragment center locations relative to the positioned nucleosomes as previously described[93,94].

MeDIP-seq experiments

To prepare inputs for MeDIP-seq experiments, gDNA was first sonicated using a Qsonica S-4000 with a 1/16" tip for 3 minutes, with 10 second pulses at intensity 3.5, and 20 seconds rest between pulses. The IP procedure was adapted from the protocol for ChIP-seq as previously described[108].

For each reaction, 100 μL of Protein A Dynabeads (ThermoFisher Cat # 10002D) were washed 3 times with a 5 mg/mL BSA solution. Beads were then resuspended in 1 mL BSA solution and 5 μL of α-5-hmU antibody (Abcam Cat # ab19735) were added. Coupling of antibodies to beads was carried out overnight on a rotator at 4 °C. Beads were again washed 3 times with BSA solution and resuspended in 100 μL of BSA solution.

Sheared genomic DNA (~1 μg 1:1 mix of *B. minutum* and *Homo sapiens*) was end repaired and adapters were ligated to it following the procedure of the NEBNext Ultra II DNA Library Prep Kit for Illumina (NEB, E7645S), purified using AMPure XP beads and eluted in 50 μL of H_2O, and then denatured at 98 °C for 10 minutes. DNA was then immediately placed on ice, resuspended in 850 μL RIPA buffer (1× PBS, 1% IGEPAL, 0.5% Sodium Deoxycholate, 0.1% SDS, Roche Protease Inhibitor Cocktail) and added to the beads, then incubated overnight on a rotator at 4 °C.

Beads were washed 5 times with LiCl buffer (10 mM Tris-HCl pH 7.5, 500 mM LiCl, 1% NP-40/IGEPAL, 0.5% Sodium Deoxycholate) by incubating for 10 minutes at 4 °C on a rotator, then rinsed once with 1× TE buffer. Beads were then resuspended in 200 μL IP Elution Buffer (1% SDS, 0.1 M $NaHCO_3$) and incubated at 65 °C in a Thermomixer (Eppendorf) with interval mixing to dissociate antibodies. Beads were separated from the DNA solution by centrifugation, and DNA was purified using the MinElute kit.

Library generation was completed by carrying out PCR following the rest of the steps of the NEBNext Ultra II DNA Library Prep Kit protocol, using 15 cycles of amplification. Final libraries were purified using AMPure XP beads.

Several control libraries were prepared – "Input" from the gDNA that was used as input to the immunoprecipitation, "Depleted" from the supernatant from the first bead separation after the incubation of DNA with beads, and "IgG", generated from a parallel immunoprecipitation reaction that used only Protein A beads (without a primary antibody)

MeDIP-seq data processing

MeDIP-seq libraries processing was carried out in the same way as that of ATAC-seq datasets.

5-hmU chemical mapping experiments

Chemical mapping of 5-hmU as carried out following the previously described by Kawasaki et al. chemical conversion method[98].

Briefly, sheared genomic DNA was used as input and end prep and adapter ligation were carried out using the NEBNext Ultra II DNA Library Prep Kit. After the ligation step, DNA was purified using AMPure XP beads and eluted in 50 μL of H_2O. DNA denaturation was performed by adding NaOH to a final concentration of 0.05M and incubating at 37 °C for 30 minutes. Oxidation was carried out by adding 2 μL of $KRuO_4$ solution (15 mM in 0.05 M NaOH) and incubating for 30 minutes at room temperature. Oxidized DNA was purified using AMPure XP beads and extension was carried out by mixing 13.5 μL DNA, 1.6 μL 100 mM $MgSO_4$, 2 μL NEB Index Primer, 2 μL 10× ThermoPol Reaction Buffer (NEB), 0.5 μL 10 mM dNTP mix, and 0.4 μL Bst DNA Polymerase, Large Fragment (NEB), then incubating for 1 hour at 37 °C. PCR amplification was carried out using the NEB Ultra DNA Library Prep Kit, with 12 cycles of PCR. Final libraries were purified using AMPure XP beads.

Processing of 5-hmU chemical mapping datasets

The `slamdunk` package[109] (`https://t-neumann.github.io/slamdunk/`), which was originally developed for the analysis of SLAM-seq[110] datasets (the SLAM-seq protocol also generates T → C conversions) was adapted to estimate 5-hmU conversion levels.

First, the genome was tiled into 500-bp bins starting every 100 bp. Second, sequencing reads were trimmed of adaptors using Trim Galore, and used as input to `slamdunk` together with the genome tiling with the following settings: `--max-read-length 75 -5 9 -n 1000000 -m --skip-sam`.

bioRxiv preprint doi: https://doi.org/10.1101/2023.09.18.558303; this version posted September 22, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Repeat annotation

Repeats were identified de novo from the scaffolded assembly using RepeatModeler-2.0.1 with default parameters. Repeat annotations were subsequently generated using RepeatMasker-4.1.1[111] with RMBlast-2.10.0 as the sequence search engine.

Analysis of ATAC-seq and MeDIP-seq data in repeat space

Sequencing datasets were analyzed in repeat space as previously described[105]. Briefly, reads were mapped to consensus repeat sequences with relaxed settings (`-e 200` instead of `- v 2`) and with unlimited multimappers. Normalization of multimapping reads was carried out as above.

Heterologous expression of DVNPs in yeast

The MS46 *S. cerevisiae* strain[112] was used for all experiments.

For the experiments in Supplementary Figure 4, each DVNP was expressed from a SIVu-WTC846::TetPr-DVNP-3xNLS-linker-3PK construct that was integrated in a single copy at the URA locus of the MS46 stain. The WTC846::TetPr promoter is reported previously[113]. Cells were grown in YPD media at 30 °C overnight and expression was induced by the addition of 200 nM anhydrotetracycline. Cells were collected 5 hours after induction. Untransformed MS46 cells were used as control.

For the experiments in Supplementary Figure 5, each DVNP was expressed from a multicopy pRS416-GAL1pr-DVNP-3xHA-NLS plasmid, as first reported in by Irwin et al.[83], that was transformed into MS46. Cells were grown in synthetic media lacking uracil + 2% raffinose at 30 °C and cultured overnight before expression was induced by the addition of 2% galactose to the media. Cells were collected 7 hours after induction. MS46 transformed with an empty pRS416-GAL1pr construct was used as control.

Yeast SMF experiments

Yeast SMF experiments were carried out as previously described[96,114–116]

A 1:1 mixture of *S. cerevisiae* cells expression DVNPs and *Candida glabrata* cells (used as a control for normalization, as previously described[86]) amounting to a total of 2.5×10^8 cells was used as input. Cells in log phase ($OD_{660} \leq 1.0$) were first centrifuged at 13,000 rpm for 1 minute, then washed with 100 μL Sorbitol Buffer(1.4 M Sorbitol, 40 mM HEPES-KOH pH 7.5, 0.5 mM $MgCl_2$), and centrifuged again at 13,000 rpm for 1 minute. Cells were then spheroplasted by resuspending in 200 μL Sorbitol Buffer with DTT added at a final concentration of 10 mM and 0.5 mg/mL 100T Zymolase, followed by incubating for 5 minutes at 30 °C at 300 rpm in a Thermomixer. The pellet was centrifuged for 2 minutes at 5,000 rpm, washed in 100 μL Sorbitol Buffer, and centrifuged again at 5,000 rpm for 2 minutes.

Cells were then resuspended in 100 μL ice-cold Nuclei Lysis Buffer (10 mM Tris pH 7.4, 10 mM NaCl, 3 mM $MgCl_2$, 0.1 mM EDTA, 0.5% NP-40) and incubated on ice for 10 minutes. Nuclei were then centrifuged at 5000 rpm for 5 min at 4 °C, resuspended in 100 μL cold Nuclei Wash Buffer (10 mM Tris pH 7.4, 10 mM NaCl, 3 mM $MgCl_2$, 0.1 mM EDTA), and centrifuged again at 5,000 rpm for 5 min at 4 °C. Finally, nuclei were resuspended in 100 μL M.CviPI Reaction Buffer (50 mM Tris-HCl pH 8.5, 50 mM NaCl, 10 mM DTT).

Nuclei were then first treated with M.CviPI (GpC methyltransferase) by adding 200 U of M.CviPI (NEB), SAM at 0.6 mM and sucrose at 300 mM, and incubating at 30 °C for 7.5 min. After this incubation, 128 pmol SAM and another 100 U of enzymes were added, and a further incubation at 30 °C for 7.5 min was carried out. Immediately after, M.SssI treatment (CpG methyltransferase) followed, by adding 60 U of M.SssI (NEB), 128 pmol SAM, $MgCl_2$ at 10 mM and incubation at 30 °C for 7.5 min.

The reaction was stopped by adding an equal volume of Stop Buffer (20 mM Tris-HCl pH 8.5, 600 mM NaCl, 1% SDS, 10 mM EDTA).

HMW DNA was isolated using the MagAttract HMW DNA Kit (Qiagen; cat # 67563) following the manufacturer's instructions.

Enzymatically labeled DNA was then sheared on a Covaris E220 and converted into sequencing libraries following the EM-seq protocol, using the NEBNext Enzymatic Methyl-seq Kit (NEB, Cat # E7120L).

Yeast SMF data processing

Adapters were trimmed from reads using Trimmomatic[117] (version 0.36). Trimmed reads were aligned against a combined *S. cerevisiae sacCer3* plus *Candida glabrata* `C_glabrata_CBS138` genome index using `bwa-meth` with default settings. Duplicate reads were removed using `picard-tools` (version 1.99). Methylation calls were extracted using `MethylDackel` (`https://github.com/dpryan79/MethylDackel`). Additional analyses were carried out using custom-written Python scripts (`https://github.com/georgimarinov/GeorgiScripts`).

Chemically mapped nucleosome positions in *S. cerevisiae* were obtained from Brogaard et al. 2012[118] as previously described[96].

Yeast ATAC-seq experiments

Yeast ATAC-seq experiments were carried out as previously described[96,119].

Briefly, ATAC-seq was carried out on the same nuclei isolated for SMF as described above (before resuspension in M.CviPI Reaction Buffer), by resuspending nuclei with 25 μL 2× TD buffer (20 mM Tris-HCl pH 7.6, 10 mM $MgCl_2$, 20% Dimethyl Formamide), 2.5 μL transposase (custom

bioRxiv preprint doi: https://doi.org/10.1101/2023.09.18.558303; this version posted September 22, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

produced) and 22.5 μL nuclease-free H_2O, and incubating at 37 °C for 30 min in a Thermomixer at 1000 RPM. Transposed DNA was isolated using the DNA Clean & Concentrator Kit (Zymo, cat # D4014) and PCR amplified as described before[81]. Libraries were then sequenced on a Illumina NextSeq instrument as 2×36mers or as 2×75mers.

ATAC-seq data processing

FASTQ files were mapped against a combined *S. cerevisiae sacCer3* plus *Candida glabrata* `C_glabrata_CBS138` genome index as 2×36mers using Bowtie[103] with the following settings: `-v 2 -k 2 -m 1 --best --strata`. Duplicate reads were removed using `picard-tools` (version 1.99). Additional analysis was carried out as previously described[120].

Author contributions

G.K.M. conceptualized the study, and carried out ATAC-seq, MeDIP and 5-hmU chemical mapping experiments. X.C. analyzed data. M.P.S. carried yeast DVNP expression experiments. T.X. carried out cell culture and DNA isolation. A.R.G, A.K. and W.J.G. supervised the study. G.K.M. and X.C. wrote the manuscript with input from all authors.

Acknowledgements

This work was supported by NIH grants (P50HG007735, RO1 HG008140, U19AI057266 and UM1HG009442 to W.J.G., 1UM1HG009436 to W.J.G. and A.K., 1DP2OD022870-01 and 1U01HG009431 to A.K.), the Rita Allen Foundation (to W.J.G.), the Baxter Foundation Faculty Scholar Grant, and the Human Frontiers Science Program grant RGY006S (to W.J.G). W.J.G. is a Chan Zuckerberg Biohub investigator and acknowledges grants 2017-174468 and 2018-182817 from the Chan Zuckerberg Initiative. Fellowship support provided by the Stanford School of Medicine Dean's Fellowship (G.K.M.). This work is also supported by NSF-IOS EDGE Award 1645164 to A.R.G. and Carnegie Venture grant 10907 (to T.X. and G.K.M.).

The authors would like to thank Alexandro Trevino for supplying the α-5-hmU antibody, Nicholas Irwin and Patrick Keeling for providing the construct for expressing *Hematodinium* DVNP.5, as well as members of the Greenleaf, Kundaje, Pringle and Grossman laboratories for helpful discussion and suggestions regarding this work.

Data Availability

Data associated with this manuscript have been submitted to GEO under accession number GSE241969.

Code Availability

Custom code used to process the data is available at `https://github.com/georgimarinov/GeorgiScripts`.

Competing Interests

The authors declare no competing interests.

References

1. Rizzo PJ. 1991. The enigma of the dinoflagellate chromosome. *J Protozool* **38**(3):246–52.
2. Wargo MJ, Rizzo PJ. 2001. Exception to eukaryotic rules. *Science* **294**:2477.
3. Rizzo PJ. 2003. Those amazing dinoflagellate chromosomes. *Cell Res* **13**:215–217.
4. Hackett JD, Anderson DM, Erdner DL, Bhattacharya D. 2004. Dinoflagellates: a remarkable evolutionary experiment. *Am J Bot* **91**:1523–1534.
5. Lin S. 2011. Genomic understanding of dinoflagellates. *Res Microbiol* **162**(6):551–569.
6. Wisecaver JH, Hackett JD. 2011. Dinoflagellate genome evolution. *Annu Rev Microbiol* **65**:369–387.
7. LaJeunesse TC, Parkinson JE, Gabrielson PW, Jeong HJ, Reimer JD, Voolstra CR, Santos SR. 2018. Systematic Revision of Symbiodiniaceae Highlights the Antiquity and Diversity of Coral Endosymbionts. *Curr Biol* **28**(16):2570–2580.e6.
8. Trench RK. 1993. Microalgal–invertebrate symbiosis: a review. *Endocyt Cell Res* **9**:135–175
9. Hoegh–Guldberg O. 1999. Climate change, coral bleaching and the future of the world's coral reefs. *Mar Freshw Res* **50**:839–866
10. Hoegh–Guldberg O, Mumby PJ, Hooten AJ, Steneck RS, Greenfield P, Gomez E, Harvell CD, Sale PF, Edwards AJ, Caldeira K, Knowlton N, Eakin CM, Iglesias–Prieto R, Muthiga N, Bradbury RH, Dubi A, Hatziolos ME. 2007. *Science* **318**(5857):1737–1742. Coral reefs under rapid climate change and ocean acidification.
11. Rizzo PJ, Nooden LD. 1972. Chromosomal proteins in the dinoflagellate alga *Gyrodinium cohnii*. *Science* **176**:796–797.
12. Rizzo PJ. 1987. Biochemistry of the dinoflagellate nucleus. In Taylor FJR (Ed.) *The Biology of Dinoflagellates*. Blackwell, Oxford, pp.143–173.
13. Gornik SG, Hu I, Lassadi I, Waller RF. 2019. The Biochemistry and Evolution of the Dinoflagellate Nucleus. *Microorganisms* **7**(8):245
14. Riaz S, Sui Z, Niaz Z, Khan S, Liu Y, Liu H. 2018. Distinctive Nuclear Features of Dinoflagellates with A Particular Focus on Histone and Histone-Replacement Proteins. *Microorganisms* **6**(4):128.
15. Herzog M, Soyer MO. 1981. Distinctive features of dinoflagellate chromatin. Absence of nucleosomes in a

bioRxiv preprint doi: https://doi.org/10.1101/2023.09.18.558303; this version posted September 22, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

primitive species *Prorocentrum micans* E. *Eur J Cell Biol* **23**(2):295–302.

16. Dodge JD. 1964. Chromosome structure in the Dinophyceae. *Arch Microbiol* **48**:66–80.
17. Rizzo P, Nooden L. 1974. Partial characterization of dinoflagellate chromosomal proteins. *Biochim Biophys* **349**:415–427.
18. Rizzo PJ. 1981. Comparative aspects of basic chromatin proteins in dinoflagellates. *Biosystems* **14**(3–4):433–443
19. Rizzo PJ, Burghardt RC. 1982. Histone-like protein and chromatin structure in the wall-less dinoflagellate *Gymnodinium nelsoni*. *Biosystems* **15**(1):27–34.
20. Rizzo PJ, Morris RL. 1983-1984. Some properties of the histone-like protein from *Crypthecodinium cohnii* (HCc). *Biosystems* **16**(3-4):211–216.
21. Wargo MJ, Rizzo PJ. 2000. Characterization of *Gymnodinium mikimotoi* (Dinophyceae) nuclei and identification of the major histone-like protein, Hgm. *J Phycol* **36**:584–589.
22. Chudnovsky Y, Li JF, Rizzo PJ, Hastings J, Fagan TF. 2002. Cloning, expression, and characterization of a histone-like protein from the marine dinoflagellate *Lingulodinium polyedrum* (Dinophyceae). *J Phycol* **38**:543–550.
23. Chan YH, Kwok AC, Tsang JS, Wong JT. 2006. Alveolata histone-like proteins have different evolutionary origins. *J Evol Biol* **19**(5):1717–1721.
24. Wong JT, New DC, Wong JC, Hung VK. 2003. Histone-like proteins of the dinoflagellate *Crypthecodinium cohnii* have homologies to bacterial DNA-binding proteins. *Eukaryot Cell* **2**(3):646–650.
25. Gornik SG, Ford KL, Mulhern TD, Bacic A, McFadden GI, Waller RF. 2012. Loss of nucleosomal DNA condensation coincides with appearance of a novel nuclear protein in dinoflagellates. *Curr Biol* **22**(24):2303–2312.
26. Janouškovec J, Gavelis GS, Burki F, Dinh D, Bachvaroff TR, Gornik SG, Bright KJ, Imanian B, Strom SL, Delwiche CF, Waller RF, Fensome RA, Leander BS, Rohwer FL, Saldarriaga JF. 2017. Major transitions in dinoflagellate evolution unveiled by phylotranscriptomics. *Proc Natl Acad Sci U S A* **114**(2):E171–E180.
27. Talbert PB, Meers MP, Henikoff S. 2019. Old cogs, new tricks: the evolution of gene expression in a chromatin context. *Nat Rev Genet* **20**(5):283–297.
28. Jenuwein T, Allis CD. 2001. Translating the histone code. *Science* **293**(5532):1074–1080.
29. Marinov GK, Lynch M. 2015. Diversity and Divergence of Dinoflagellate Histone Proteins. *G3 (Bethesda)* **6**(2):397–422.
30. Kearns LP, Sigee DC. 1980. The occurrence of period IV elements in dinoflagellate chromatin: An X-ray microanalytical study. *J Cell Sci* **46**:113–127.
31. Herzog M, Soyer MO. 1983. The native structure of dinoflagellate chromosomes and their stabilization by Ca^{2+} and Mg^{2+} cations. *Europ J Cell Biol* **30**:33–41.
32. Strick R, Strissel PL, Gavrilov K, Levi-Setti R. 2001. Cation–chromatin binding as shown by ion microscopy is essential for the structural integrity of chromosomes. *J Cell Biol* **155**:899–910.
33. Levi-Setti R, Gavrilov K, Rizzo P. 2008. Divalent cation distribution in dinoflagellate chromosomes imaged by high-resolution ion probe mass spectrometry. *Europ J Cell Biol* **87**:963–976.
34. Bachvaroff TR, Place AR. 2008. From stop to start: tandem gene arrangement, copy number and trans-splicing sites in the dinoflagellate *Amphidinium carterae*. *PLoS One* **3**(8):e2929.
35. Shoguchi E, Shinzato C, Kawashima T, Gyoja F, Mungpakdee S, Koyanagi R, Takeuchi T, Hisata K, Tanaka M, Fujiwara M, Hamada M, Seidi A, Fujie M, Usami T, Goto H, Yamasaki S, Arakaki N, Suzuki Y, Sugano S, Toyoda A, Kuroki Y, Fujiyama A, Medina M, Coffroth MA, Bhattacharya D, Satoh N. 2013. Draft assembly of the *Symbiodinium minutum* nuclear genome reveals dinoflagellate gene structure. *Curr Biol* **23**(15):1399–1408.
36. Aranda M, Li Y, Liew YJ, Baumgarten S, Simakov O, Wilson MC, Piel J, Ashoor H, Bougouffa S, Bajic VB, Ryu T, Ravasi T, Bayer T, Micklem G, Kim H, Bhak J, LaJeunesse TC, Voolstra CR. 2016. Genomes of coral dinoflagellate symbionts highlight evolutionary adaptations conducive to a symbiotic lifestyle. *Sci Rep* **6**:39734.
37. Lin S, Cheng S, Song B, Zhong X, Lin X, Li W, Li L, Zhang Y, Zhang H, Ji Z, Cai M, Zhuang Y, Shi X, Lin L, Wang L, Wang Z, Liu X, Yu S, Zeng P, Hao H, Zou Q, Chen C, Li Y, Wang Y, Xu C, Meng S, Xu X, Wang J, Yang H, Campbell DA, Sturm NR, Dagenais-Bellefeuille S, Morse D. 2015. The *Symbiodinium kawagutii* genome illuminates dinoflagellate gene expression and coral symbiosis. *Science* **350**(6261):691–694.
38. Zhang H, Hou Y, Miranda L, Campbell DA, Sturm NR, Gaasterland T, Lin S. 2007. Spliced leader RNA trans-splicing in dinoflagellates. *Proc Natl Acad Sci U S A* **104**(11):4618–4623.
39. Slamovits CH, Keeling PJ. 2008. Widespread recycling of processed cDNAs in dinoflagellates. *Curr Biol* **18**(13):R550–552.
40. Marinov GK, Trevino AE, Xiang T, Kundaje A, Grossman AR, Greenleaf WJ. 2021. Transcription–dependent domain-scale three-dimensional genome organization in the dinoflagellate *Breviolum minutum*. *Nat Genetics* **53**:613–617.
41. Nand A, Zhan Y, Salazar OR, Aranda M, Voolstra CR, Dekker J. 2021. Genetic and spatial organization of the unusual chromosomes of the dinoflagellate *Symbiodinium microadriaticum*. *Nat Genet* **53**(5):618–629.

bioRxiv preprint doi: https://doi.org/10.1101/2023.09.18.558303; this version posted September 22, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

42. Lieberman–Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO, Sandstrom R, Bernstein B, Bender MA, Groudine M, Gnirke A, Stamatoyannopoulos J, Mirny LA, Lander ES, Dekker J. 2009. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**(5950):289–293.
43. de Mendoza A, Bonnet A, Vargas-Landin DB, Ji N, Li H, Yang F, Li L, Hori K, Pflueger J, Buckberry S, Ohta H, Rosic N, Lesage P, Lin S, Lister R. 2018. Recurrent acquisition of cytosine methyltransferases into eukaryotic retrotransposons. *Nat Commun* **9**(1):1341.
44. Vernet G, Salarovira M, Maeder M, Jacques F, Herzog M. 1990. Basic Nuclear Proteins of the Histone-Less Eukaryote *Crypthecodinium cohnii* (Pyrrhophyta) – 2-dimensional electrophoresis and DNA-binding properties. *Biochim Biophys Acta* **1048**:281–289.
45. Chan YH, Wong JT. 2007. Concentration-dependent organization of DNA by the dinoflagellate histone-like protein HCc3. *Nucleic Acids Res* **35**(8):2573–2583
46. Sun S, Liu M, Dong F, Fan S, Yao Y. 2013. A histone-like protein induces plasmid DNA to form liquid crystals in vitro and gene compaction in vivo. *Int J Mol Sci* **14**(12):23842–23857.
47. Kallen RG, Simon M, Marmur J. 1962. The occurrence of a new pyrimidine base replacing thymine in a bacteriophage DNA: 5-hydroxymethyl uracil. *J Mol Biol.* **5**:248–250.
48. Rae PM. 1973. 5-Hydroxymethyluracil in the DNA of a dinoflagellate. *Proc Natl Acad Sci U S A* **70**(4):1141–1145.
49. Rae PM. 1976. Hydroxymethyluracil in eukaryote DNA: a natural feature of the pyrrophyta (dinoflagellates). *Science* **194**(4269):1062–1064
50. Rae PM, Steele RE. 1978. Modified bases in the DNAs of unicellular eukaryotes: an examination of distributions and possible roles, with emphasis on hydroxymethyluracil in dinoflagellates. *Biosystems* **10**(1–2):37–53.
51. Steele RE, Rae PM. 1980. Ordered distribution of modified bases in the DNA of a dinoflagellate. *Nucleic Acids Res* **8**(20):4709–4725.
52. Herzog M, Soyer MO, Daney de Marcillac G. 1982. A high level of thymine replacement by 5-hydroxymethyluracil in nuclear DNA of the primitive dinoflagellate *Prorocentrum micans* E. *Eur J Cell Biol* **27**(2):151–155.
53. Carson S, Wilson J, Aksimentiev A, Weigele PR, Wanunu M. 2016. Hydroxymethyluracil modifications enhance the flexibility and hydrophilicity of double-stranded DNA. *Nucleic Acids Res* **44**(5):2085–2092.
54. Davies W, Jakobson KS, Norby Ø. 1988. Characterization of DNA from the dinoflagellate *Woloszynskia bostoniensis*. *J Protozool* **35**:418–422.
55. Pasternack LB, Bramham J, Mayol L, Galeone A, Jia X, Kearns DR. 1996. 1H NMR studies of the 5-(hydroxymethyl)-2'-deoxyuridine containing TF1 binding site. *Nucleic Acids Res* **24**(14):2740–2745.
56. Vu HM, Pepe A, Mayol L, Kearns DR. 1999. NMR-derived solution structure of a 17mer hydroxymethyluracil-containing DNA. *Nucleic Acids Res* **27**(21):4143–4150.
57. Gommers-Ampt JH, Teixeira AJ, van de Werken G, van Dijk WJ, Borst P. 1993. The identification of hydroxymethyluracil in DNA of *Trypanosoma brucei*. *Nucleic Acids Res* **21**(9):2039–2043
58. Lukes J, Leander BS, Keeling PJ. 2009. Cascades of convergent evolution: the corresponding evolutionary histories of euglenozoans and dinoflagellates. *Proc Natl Acad Sci U S A* **106** Suppl 1:9963–9970.
59. Ivens AC, Peacock CS, Worthey EA, Murphy L, Aggarwal G, Berriman M, Sisk E, Rajandream MA, Adlem E, Aert R, Anupama A, Apostolou Z, Attipoe P, Bason N, Bauser C, Beck A, Beverley SM, Bianchettin G, Borzym K, Bothe G, Bruschi CV, Collins M, Cadag E, Ciarloni L, Clayton C, Coulson RM, Cronin A, Cruz AK, Davies RM, De Gaudenzi J, Dobson DE, Duesterhoeft A, Fazelina G, Fosker N, Frasch AC, Fraser A, Fuchs M, Gabel C, Goble A, Goffeau A, Harris D, Hertz-Fowler C, Hilbert H, Horn D, Huang Y, Klages S, Knights A, Kube M, Larke N, Litvin L, Lord A, Louie T, Marra M, Masuy D, Matthews K, Michaeli S, Mottram JC, Müller-Auer S, Munden H, Nelson S, Norbertczak H, Oliver K, O'neil S, Pentony M, Pohl TM, Price C, Purnelle B, Quail MA, Rabbinowitsch E, Reinhardt R, Rieger M, Rinta J, Robben J, Robertson L, Ruiz JC, Rutter S, Saunders D, Schäfer M, Schein J, Schwartz DC, Seeger K, Seyler A, Sharp S, Shin H, Sivam D, Squares R, Squares S, Tosato V, Vogt C, Volckaert G, Wambutt R, Warren T, Wedler H, Woodward J, Zhou S, Zimmermann W, Smith DF, Blackwell JM, Stuart KD, Barrell B, Myler PJ. 2005. The genome of the kinetoplastid parasite, *Leishmania major*. *Science* **309**(5733):436–442.
60. El-Sayed NM, Myler PJ, Blandin G, Berriman M, Crabtree J, Aggarwal G, Caler E, Renauld H, Worthey EA, Hertz-Fowler C, Ghedin E, Peacock C, Bartholomeu DC, Haas BJ, Tran AN, Wortman JR, Alsmark UC, Angiuoli S, Anupama A, Badger J, Bringaud F, Cadag E, Carlton JM, Cerqueira GC, Creasy T, Delcher AL, Djikeng A, Embley TM, Hauser C, Ivens AC, Kummerfeld SK, Pereira-Leal JB, Nilsson D, Peterson J, Salzberg SL, Shallom J, Silva JC, Sundaram J, Westenberger S, White O, Melville SE, Donelson JE, Andersson B, Stuart KD, Hall N. 2005. Comparative genomics of trypanosomatid parasitic protozoa. *Science* **309**(5733):404–409.
61. El-Sayed NM, Myler PJ, Bartholomeu DC, Nilsson D, Aggarwal G, Tran AN, Ghedin E, Worthey EA, Delcher AL, Blandin G, Westenberger SJ, Caler E,

bioRxiv preprint doi: https://doi.org/10.1101/2023.09.18.558303; this version posted September 22, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Cerqueira GC, Branche C, Haas B, Anupama A, Arner E, Aslund L, Attipoe P, Bontempi E, Bringaud F, Burton P, Cadag E, Campbell DA, Carrington M, Crabtree J, Darban H, da Silveira JF, de Jong P, Edwards K, Englund PT, Fazelina G, Feldblyum T, Ferella M, Frasch AC, Gull K, Horn D, Hou L, Huang Y, Kindlund E, Klingbeil M, Kluge S, Koo H, Lacerda D, Levin MJ, Lorenzi H, Louie T, Machado CR, McCulloch R, McKenna A, Mizuno Y, Mottram JC, Nelson S, Ochaya S, Osoegawa K, Pai G, Parsons M, Pentony M, Pettersson U, Pop M, Ramirez JL, Rinta J, Robertson L, Salzberg SL, Sanchez DO, Seyler A, Sharma R, Shetty J, Simpson AJ, Sisk E, Tammi MT, Tarleton R, Teixeira S, Van Aken S, Vogt C, Ward PN, Wickstead B, Wortman J, White O, Fraser CM, Stuart KD, Andersson B. 2005. The genome sequence of *Trypanosoma cruzi*, etiologic agent of Chagas disease. *Science* **309**(5733):409–415.

62. Berriman M, Ghedin E, Hertz-Fowler C, Blandin G, Renauld H, Bartholomeu DC, Lennard NJ, Caler E, Hamlin NE, Haas B, Böhme U, Hannick L, Aslett MA, Shallom J, Marcello L, Hou L, Wickstead B, Alsmark UC, Arrowsmith C, Atkin RJ, Barron AJ, Bringaud F, Brooks K, Carrington M, Cherevach I, Chillingworth TJ, Churcher C, Clark LN, Corton CH, Cronin A, Davies RM, Doggett J, Djikeng A, Feldblyum T, Field MC, Fraser A, Goodhead I, Hance Z, Harper D, Harris BR, Hauser H, Hostetler J, Ivens A, Jagels K, Johnson D, Johnson J, Jones K, Kerhornou AX, Koo H, Larke N, Landfear S, Larkin C, Leech V, Line A, Lord A, Macleod A, Mooney PJ, Moule S, Martin DM, Morgan GW, Mungall K, Norbertczak H, Ormond D, Pai G, Peacock CS, Peterson J, Quail MA, Rabbinowitsch E, Rajandream MA, Reitter C, Salzberg SL, Sanders M, Schobel S, Sharp S, Simmonds M, Simpson AJ, Tallon L, Turner CM, Tait A, Tivey AR, Van Aken S, Walker D, Wanless D, Wang S, White B, White O, Whitehead S, Woodward J, Wortman J, Adams MD, Embley TM, Gull K, Ullu E, Barry JD, Fairlamb AH, Opperdoes F, Barrell BG, Donelson JE, Hall N, Fraser CM, Melville SE, El-Sayed NM. The genome of the African trypanosome *Trypanosoma brucei*. *Science* **309**(5733):416–422
63. Boothroyd JC, Cross GA. 1982. Transcripts coding for variant surface glycoproteins of *Trypanosoma brucei* have a short, identical exon at their 5' end. *Gene* **20**(2):281–289.
64. Nelson RG, Parsons M, Barr PJ, Stuart K, Selkirk M, Agabian N. 1983. Sequences homologous to the variant antigen mRNA spliced leader are located in tandem repeats and variable orphons in *Trypanosoma brucei*. *Cell* **34**(3):901–909.
65. De Lange T, Michels PA, Veerman HJ, Cornelissen AW, Borst P. 1984. Many trypanosome messenger RNAs share a common 5' terminal sequence. *Nucleic Acids Res* **12**(9):3777–3790.
66. De Lange T, Berkvens TM, Veerman HJ, Frasch AC, Barry JD, Borst P. 1984. Comparison of the genes coding for the common 5' terminal sequence of messenger RNAs in three trypanosome species. *Nucleic Acids Res* **12**(11):4431–4443.
67. Sutton RE, Boothroyd JC. 1986. Evidence for trans splicing in trypanosomes. *Cell* **47**(4):527–535.
68. Dooijes D, Chaves I, Kieft R, Dirks-Mulder A, Martin W, Borst P. 2000. Base J originally found in kinetoplastida is also a minor constituent of nuclear DNA of *Euglena gracilis*. *Nucleic Acids Res* **28**(16):3017–3021.
69. Gommers-Ampt J, Lutgerink J, Borst P. 1991. A novel DNA nucleotide in *Trypanosoma brucei* only present in the mammalian phase of the life-cycle. *Nucleic Acids Res* **19**(8):1745–151.
70. Gommers-Ampt JH, Van Leeuwen F, de Beer AL, Vliegenthart JF, Dizdaroglu M, Kowalak JA, Crain PF, Borst P. 1993. beta-D-glucosyl-hydroxymethyluracil: a novel modified base present in the DNA of the parasitic protozoan *T. brucei*. *Cell* **75**(6):1129–1136.
71. Borst P, Sabatini R. 2008. Base J: discovery, biosynthesis, and possible functions. *Annu Rev Microbiol* **62**:235–251.
72. van Leeuwen F, Taylor MC, Mondragon A, Moreau H, Gibson W, Kieft R, Borst P. 1998. β-D-glucosyl-hydroxymethyluracil is a conserved DNA modification in kinetoplastid protozoans and is abundant in their telomeres. *Proc Natl Acad Sci U S A* **95**(5):2366–2371.
73. van Leeuwen F, Wijsman ER, Kieft R, van der Marel GA, van Boom JH, Borst P. 1997. Localization of the modified base J in telomeric VSG gene expression sites of *Trypanosoma brucei*. *Genes Dev* **11**(23):3232–3241.
74. van Leeuwen F, Wijsman ER, Kuyl-Yeheskiely E, van der Marel GA, van Boom JH, Borst P. 1996. The telomeric GGGTTA repeats of *Trypanosoma brucei* contain the hypermodified base J in both strands. *Nucleic Acids Res* **24**(13):2476–2482..
75. Cliffe LJ, Siegel TN, Marshall M, Cross GA, Sabatini R. 2010. Two thymidine hydroxylases differentially regulate the formation of glucosylated DNA at regions flanking polymerase II polycistronic transcription units throughout the genome of *Trypanosoma brucei*. *Nucleic Acids Res* **38**(12):3923–3935.
76. van Luenen HG, Farris C, Jan S, Genest PA, Tripathi P, Velds A, Kerkhoven RM, Nieuwland M, Haydock A, Ramasamy G, Vainio S, Heidebrecht T, Perrakis A, Pagie L, van Steensel B, Myler PJ, Borst P. 2012. Glucosylated hydroxymethyluracil, DNA base J, prevents transcriptional readthrough in *Leishmania*. *Cell* **150**(5):909–921
77. Reynolds D, Cliffe L, Förstner KU, Hon CC, Siegel TN, Sabatini R. 2014. Regulation of transcription termination by glucosylated hydroxymethyluracil, base J, in *Leishmania major* and *Trypanosoma brucei*. *Nucleic Acids Res* **42**(15):9717–9729.

bioRxiv preprint doi: https://doi.org/10.1101/2023.09.18.558303; this version posted September 22, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

78. Weber M, Davies JJ, Wittig D, Oakeley EJ, Haase M, Lam WL, Schübeler D. 2005. Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed human cells. *Nat Genet* **37**(8):853–862.
79. Down TA, Rakyan VK, Turner DJ, Flicek P, Li H, Kulesha E, Gräf S, Johnson N, Herrero J, Tomazou EM, Thorne NP, Bäckdahl L, Herberth M, Howe KL, Jackson DK, Miretti MM, Marioni JC, Birney E, Hubbard TJ, Durbin R, Tavaré S, Beck S. 2008. A Bayesian deconvolution strategy for immunoprecipitation-based DNA methylome analysis. *Nat Biotechnol* **26**(7):779–785.
80. Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ. 2013. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat Methods* **10**(12):1213–1218.
81. Corces MR, Trevino AE, Hamilton EG, Greenside PG, Sinnott-Armstrong NA, Vesuna S, Satpathy AT, Rubin AJ, Montine KS, Wu B, Kathiria A, Cho SW, Mumbach MR, Carter AC, Kasowski M, Orloff LA, Risca VI, Kundaje A, Khavari PA, Montine TJ, Greenleaf WJ, Chang HY. 2017. An improved ATAC-seq protocol reduces background and enables interrogation of frozen tissues. *Nat Methods* **14**(10):959–962.
82. ENCODE Project Consortium, Moore JE, Purcaro MJ, Pratt HE, Epstein CB, Shoresh N, Adrian J, Kawli T, Davis CA, Dobin A, Kaul R, Halow J, Van Nostrand EL, Freese P, Gorkin DU, Shen Y, He Y, Mackiewicz M, Pauli-Behn F, Williams BA, Mortazavi A, Keller CA, Zhang XO, Elhajjajy SI, Huey J, Dickel DE, Snetkova V, Wei X, Wang X, Rivera-Mulia JC, Rozowsky J, Zhang J, Chhetri SB, Zhang J, Victorsen A, White KP, Visel A, Yeo GW, Burge CB, Lécuyer E, Gilbert DM, Dekker J, Rinn J, Mendenhall EM, Ecker JR, Kellis M, Klein RJ, Noble WS, Kundaje A, Guigó R, Farnham PJ, Cherry JM, Myers RM, Ren B, Graveley BR, Gerstein MB, Pennacchio LA, Snyder MP, Bernstein BE, Wold B, Hardison RC, Gingeras TR, Stamatoyannopoulos JA, Weng Z. 2020. Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature* **583**(7818):699–710.
83. Irwin NAT, Martin BJE, Young BP, Browne MJG, Flaus A, Loewen CJR, Keeling PJ, Howe LJ. 2018. Viral proteins as a potential driver of histone depletion in dinoflagellates. *Nat Commun* **9**(1):1535.
84. Kelly TK, Liu Y, Lay FD, Liang G, Berman BP, Jones PA. 2012. Genome-wide mapping of nucleosome positioning and DNA methylation within individual DNA molecules. *Genome Res* **22**(12):2497–2506.
85. Krebs AR, Imanci D, Hoerner L, Gaidatzis D, Burger L, Schübeler D. 2017. Genome-wide Single-Molecule Footprinting Reveals High RNA Polymerase II Turnover at Paused Promoters. *Mol Cell* **67**(3):411–422.e4.
86. Swaffer MP, Kim J, Chandler-Brown D, Langhinrichs M, Marinov GK, Greenleaf WJ, Kundaje A, Schmoller KM, Skotheim JM. 2021. Transcriptional and chromatin-based partitioning mechanisms uncouple protein scaling from cell size. *Mol Cell* **81**(23):4861–4875.e7
87. Pritham EJ, Putliwala T, Feschotte C. 2007. Mavericks, a novel class of giant transposable elements widespread in eukaryotes and related to DNA viruses. *Gene* **390**(1–2):3–17.
88. Krupovic M, Koonin EV. 2015. Polintons: a hotbed of eukaryotic virus, transposon and plasmid evolution. *Nat Rev Microbiol* **13**(2):105–115.
89. Krupovic M, Bamford DH, Koonin EV. 2014. Conservation of major and minor jelly-roll capsid proteins in Polinton (Maverick) transposons suggests that they are bona fide viruses. *Biol Direct* **9**:6.
90. Haapa-Paananen S, Wahlberg N, Savilahti H. 2014. Phylogenetic analysis of Maverick/Polinton giant transposons across organisms. *Mol Phylogenet Evol* **78**:271–274.
91. Bellas C, Hackl T, Plakolb MS, Koslová A, Fischer MG, Sommaruga R. 2023. Large-scale invasion of unicellular eukaryotic genomes by integrating DNA viruses. *Proc Natl Acad Sci U S A* **120**(16):e2300465120
92. Bellas CM, Sommaruga R. 2021. Polinton-like viruses are abundant in aquatic ecosystems. *Microbiome* **9**(1):13. doi: 10.1186/s40168-020-00956-0.
93. Schep AN, Buenrostro JD, Denny SK, Schwartz K, Sherlock G, Greenleaf WJ. 2015. Structured nucleosome fingerprints enable high-resolution mapping of chromatin architecture within regulatory regions. *Genome Res* **25**(11):1757–1770.
94. Henikoff JG, Belsky JA, Krassovsky K, MacAlpine DM, Henikoff S. 2011. Epigenome characterization at single base-pair resolution. *Proc Natl Acad Sci U S A* **108**:18318–18323
95. Chen X, Shen Y, Draper W, Buenrostro JD, Litzenburger U, Cho SW, Satpathy AT, Carter AC, Ghosh RP, East-Seletsky A, Doudna JA, Greenleaf WJ, Liphardt JT, Chang HY. 2016. ATAC-see reveals the accessible genome by transposase-mediated imaging and sequencing. *Nat Methods* **13**(12):1013–1020.
96. Shipony Z, Marinov GK, Swaffer MP, Sinnott-Armstrong NA, Skotheim JM, Kundaje A, Greenleaf WJ. 2020. Long-range single-molecule mapping of chromatin accessibility in eukaryotes. *Nat Methods* **17**(3):319–327.
97. Kawasaki F, Beraldi D, Hardisty RE, McInroy GR, van Delft P, Balasubramanian S. 2017. Genome-wide mapping of 5-hydroxymethyluracil in the eukaryote parasite *Leishmania*. *Genome Biol* **18**(1):23
98. Kawasaki F, Martínez Cuesta S, Beraldi D, Mahtey A, Hardisty RE, Carrington M, Balasubramanian S. 2018. Sequencing 5-Hydroxymethyluracil at

bioRxiv preprint doi: https://doi.org/10.1101/2023.09.18.558303; this version posted September 22, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

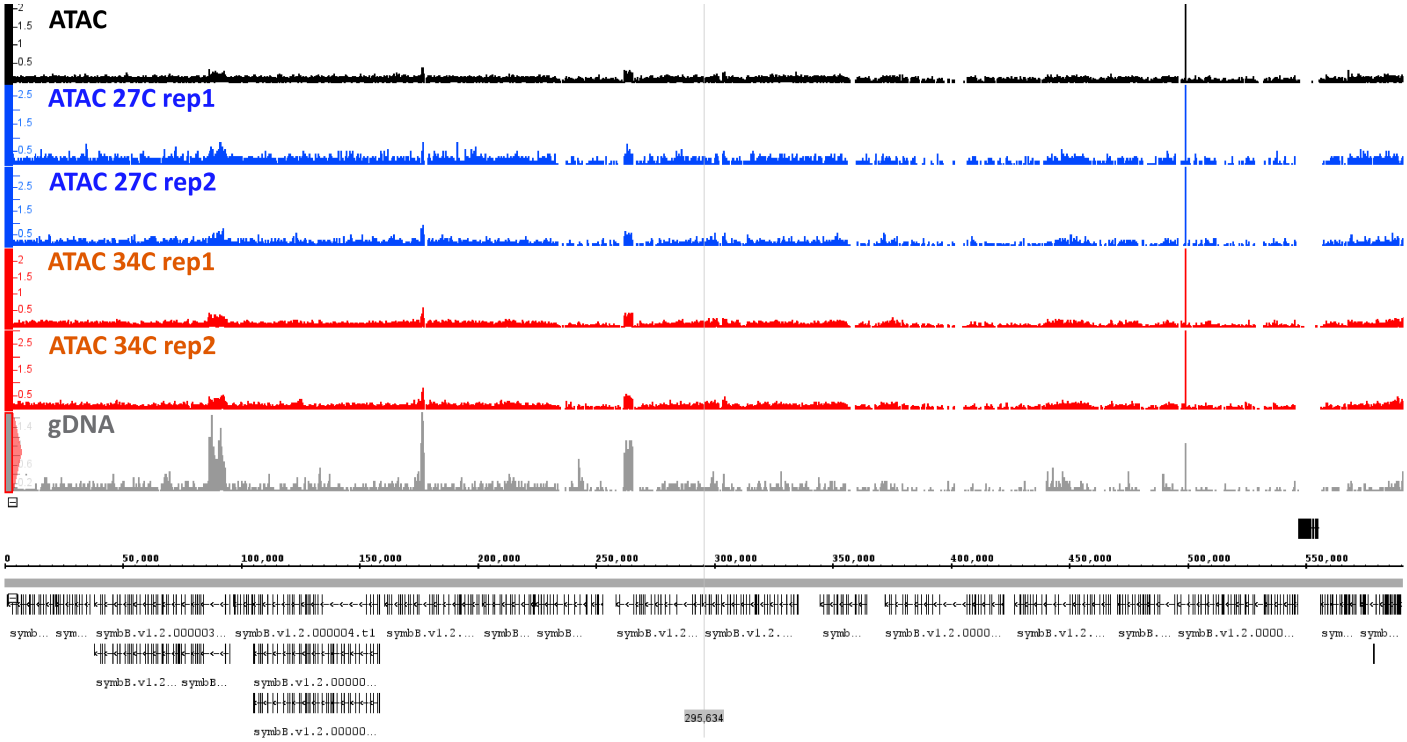
Single-Base Resolution. *Angew Chem Int Ed Engl* **57**(31):9694–9696.

99. Oberbeckmann E, Wolff M, Krietenstein N, Heron M, Ellins JL, Schmid A, Krebs S, Blum H, Gerland U, Korber P. 2019. Absolute nucleosome occupancy map for the Saccharomyces cerevisiae genome. *Genome Res* **29**(12):1996–2009.
100. Xiang T, Hambleton EA, DeNofrio JC, Pringle JR, Grossman AR. 2013. Isolation of clonal axenic strains of the symbiotic dinoflagellate *Symbiodinium* and their growth and host specificity. *J Phycol* **49**(3):447–458.
101. Xiang T, Nelson W, Rodriguez J, Tolleter D, Grossman AR. 2015. *Symbiodinium* transcriptome and global responses of cells to immediate changes in light intensity when grown under autotrophic or mixotrophic conditions. *Plant J* **82**(1):67–80.
102. Dudchenko O, Batra SS, Omer AD, Nyquist SK, Hoeger M, Durand NC, Shamim MS, Machol I, Lander ES, Aiden AP, Aiden EL. 2017. De novo assembly of the Aedes aegypti genome using Hi-C yields chromosome-length scaffolds. *Science* **356**(6333):92–95.
103. Langmead B, Trapnell C, Pop M, Salzberg SL. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10**(3):R25.
104. Marinov GK, Wang J, Handler D, Wold BJ, Weng Z, Hannon GJ, Aravin AA, Zamore PD, Brennecke J, Toth KF. 2015. Pitfalls of mapping high-throughput sequencing data to repetitive sequences: Piwi's genomic targets still not identified. *Dev Cell* **32**(6):765–771
105. Le Thomas A, Rogers AK, Webster A, Marinov GK, Liao SE, Perkins EM, Hur JK, Aravin AA, Tóth KF. 2013. Piwi induces piRNA-guided transcriptional silencing and establishment of a repressive chromatin state *Genes Dev* **27**(4):390–399.
106. Feng J, Liu T, Qin B, Zhang Y, Liu XS. 2012. Identifying ChIP-seq enrichment using MACS. *Nat Protoc* **7**(9):1728–1740.
107. Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15**(12):550.
108. Marinov GK. 2017. ChIP-seq for the Identification of Functional Elements in the Human Genome. *Methods Mol Biol* **1543**:3–18.
109. Neumann T, Herzog VA, Muhar M, von Haeseler A, Zuber J, Ameres SL, Rescheneder P. 2019. Quantification of experimentally induced nucleotide conversions in high-throughput sequencing datasets. *BMC Bioinformatics* **20**(1):258.
110. Herzog VA, Reichholf B, Neumann T, Rescheneder P, Bhat P, Burkard TR, Wlotzka W, von Haeseler A, Zuber J, Ameres SL. 2017. Thiol-linked alkylation of RNA to assess expression dynamics. *Nat Methods* **14**(12):1198–1204.
111. Smit AFA, Hubley R, Green P. *RepeatMasker Open-4.0*. 2013-2015 `http://www.repeatmasker.org`.
112. Swaffer MP, Marinov GK, Zheng H, Tsui CY, Jones AW, Greenwood J, Kundaje A, Greenleaf WJ, Reyes-Lamothe R, Skotheim JM. 2021. RNA polymerase II dynamics and mRNA stability feedback scale mRNA in proportion to cell size. *bioRxiv* 2021.09.20.461005
113. Azizoglu A, Brent R, Rudolf F. 2021. A precisely adjustable, variation-suppressed eukaryotic transcriptional controller to enable genetic discovery. *Elife* **10**:e69549
114. Marinov GK, Shipony Z, Kundaje A, Greenleaf WJ. 2023. Genome-Wide Mapping of Active Regulatory Elements Using ATAC-seq. *Methods Mol Biol* **2611**:3–19.
115. Hinks M, Marinov GK, Kundaje A, Bintu L, Greenleaf WJ. 2023. Single-Molecule Mapping of Chromatin Accessibility Using NOMe-seq/dSMF. *Methods Mol Biol* **2611**:101–111.
116. Marinov GK, Shipony Z, Kundaje A, Greenleaf WJ. 2022. Single-Molecule Multikilobase-Scale Profiling of Chromatin Accessibility Using m6A-SMAC-Seq and m6A-CpG-GpC-SMAC-Seq. *Methods Mol Biol* **2458**:269–298.
117. Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**(15):2114–2120.
118. Brogaard K, Xi L, Wang JP, Widom J. 2012. A map of nucleosome positions in yeast at base-pair resolution. *Nature* **486**(7404):496–501.
119. Marinov GK, Shipony Z, Kundaje A, Greenleaf WJ. 2022. Single-Molecule Multikilobase-Scale Profiling of Chromatin Accessibility Using m6A-SMAC-Seq and m6A-CpG-GpC-SMAC-Seq. *Methods Mol Biol* **2458**:269–298.
120. Marinov GK, Shipony Z. 2020. Interrogating the accessible chromatin landscape of eukaryote genomes using ATAC-seq. *Methods Mol Biol* **2243**:183–226.

bioRxiv preprint doi: https://doi.org/10.1101/2023.09.18.558303; this version posted September 22, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

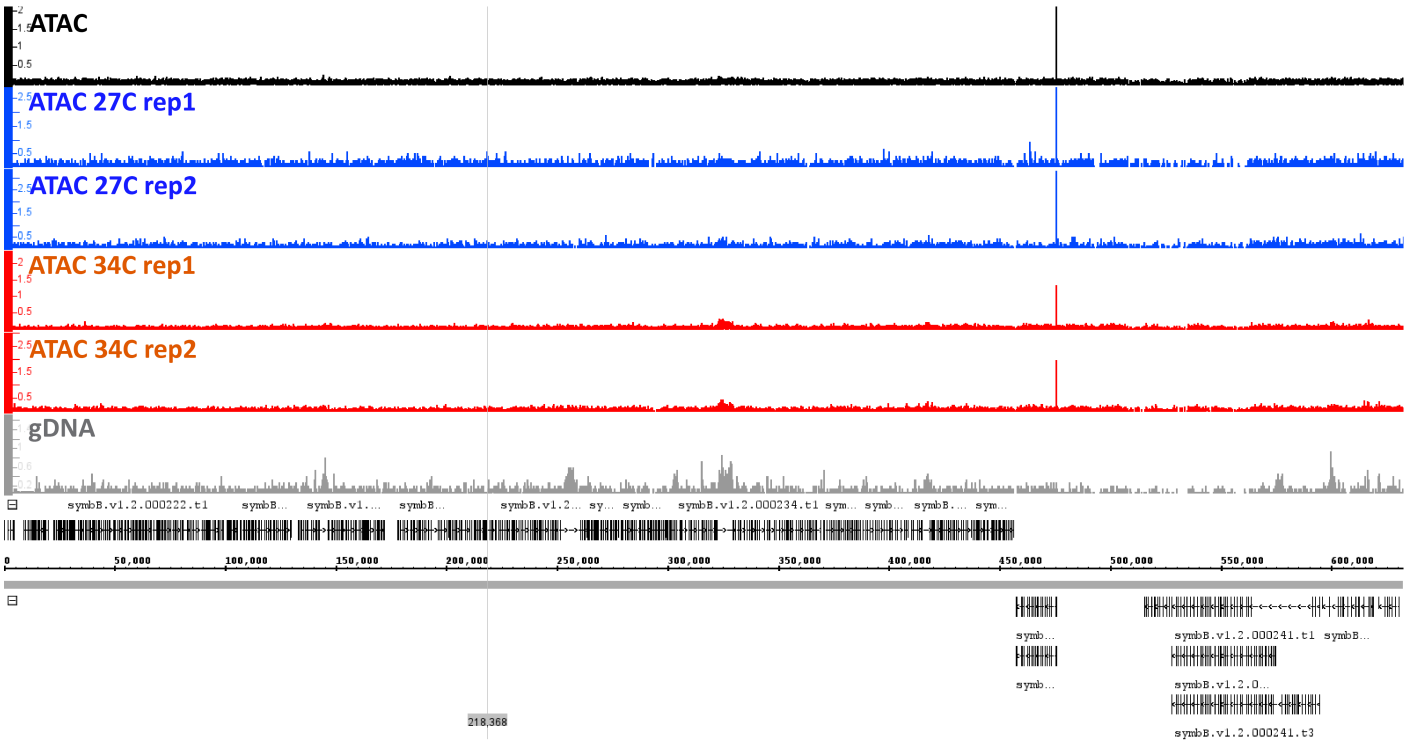
Supplementary Materials

Supplementary Figures

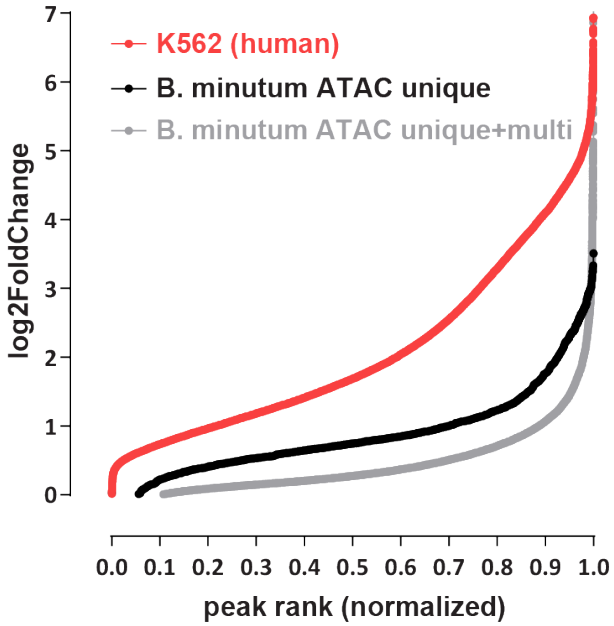


Supplementary Figure 1: Representative genome browser view of ATAC-seq and gDNA control signal in the *B. minutum* genome.

bioRxiv preprint doi: https://doi.org/10.1101/2023.09.18.558303; this version posted September 22, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

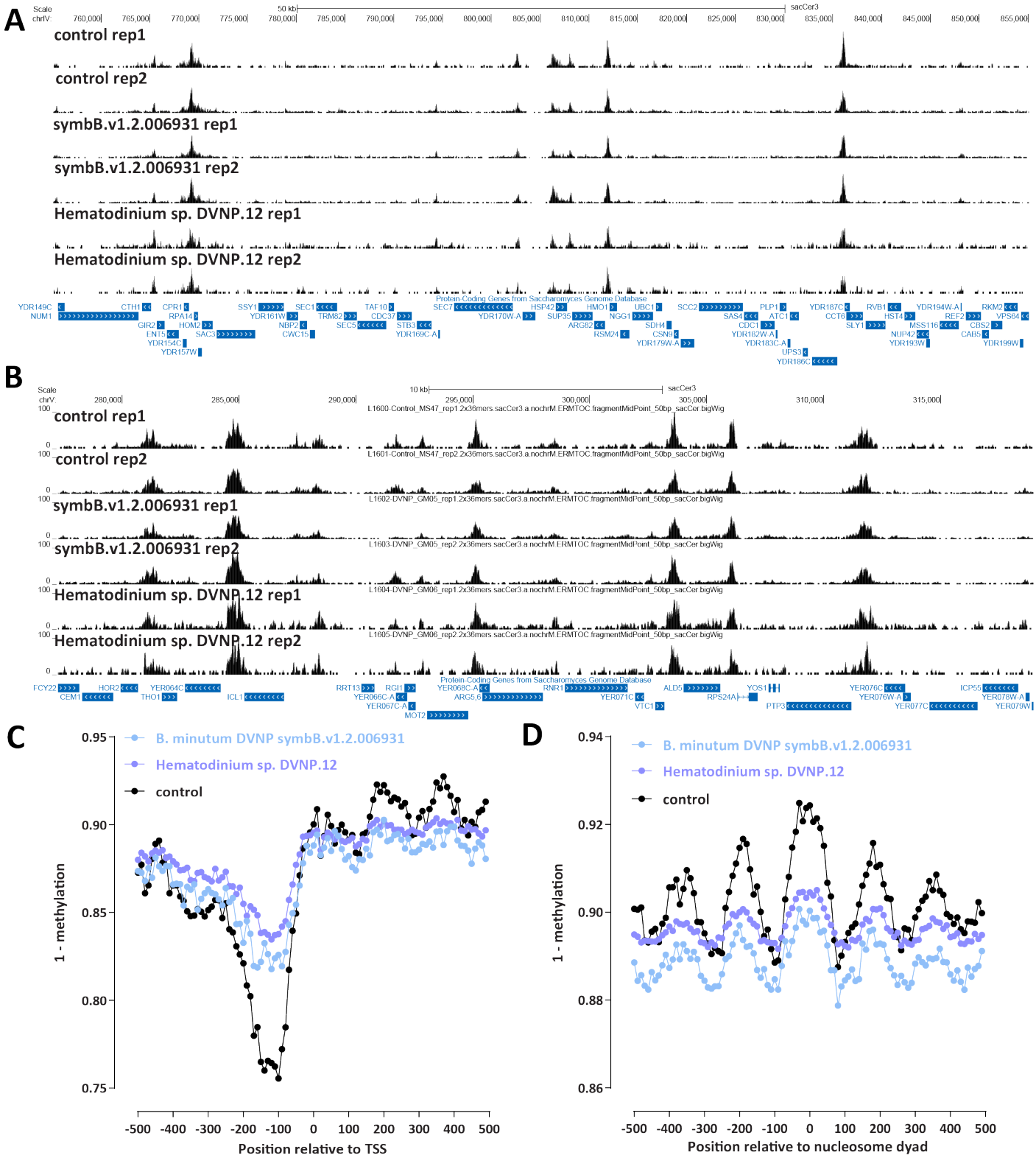


Supplementary Figure 2: Representative genome browser view of ATAC-seq and gDNA control signal in the *B. minutum* genome.



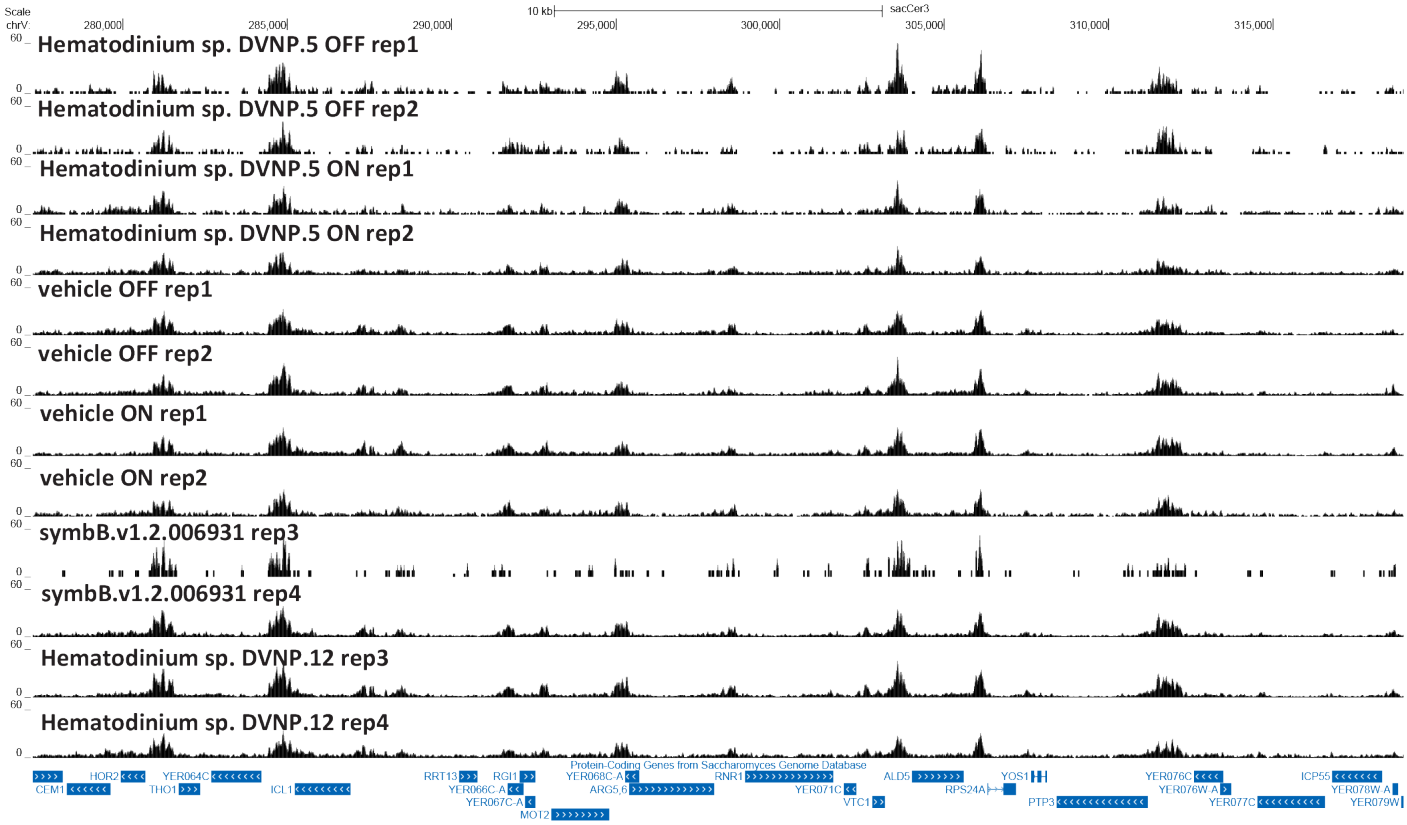
Supplementary Figure 3: Relative degree of ATAC-seq enrichment in *B. minutum* versus a representative mammalian genome sample. Shown is the log_2(fold change) ratio of ATAC-seq signal versus a negative control for a representative human ATAC-seq sample (K562 cell line from the ENCODE Project Consortium[82]; dataset ID ENCFF512VEZ was used for ATAC and dataset ID ENCFF285UKJ – a whole genome bisulfite sequencing library – as a negative control, over peaks from dataset ID ENCFF695IGF). A separately sequenced gDNA control was generated for the *B. minutum* ATAC.

bioRxiv preprint doi: https://doi.org/10.1101/2023.09.18.558303; this version posted September 22, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



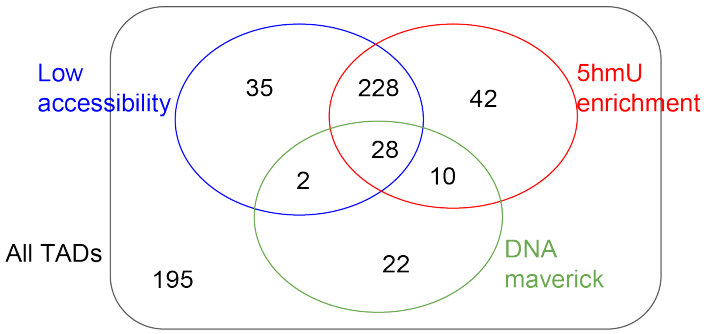
Supplementary Figure 4: Effects of exogenous expression dinoflagellate DVNPs on chromatin accessibility in the yeast *S. cerevisiae*. (A-B) ATAC-seq profiles of *S. cerevisiae* expressing *B. minutum* DVNP symbB.v1.2.006931 and *Hematodinium* sp. DVNP.12 and control samples. (C) SMF profiles (corrected using average SMF methylation from the *Candida* internal control) over *S. cerevisiae* TSSs in *S. cerevisiae* expressing *B. minutum* DVNP symbB.v1.2.006931 and *Hematodinium* sp. DVNP.12 and control samples. (D) SMF profiles (corrected using average SMF methylation from the *Candida* internal control) over positioned *S. cerevisiae* nucleosomes in *S. cerevisiae* expressing *B. minutum* DVNP symbB.v1.2.006931 and *Hematodinium* sp. DVNP.12 and control samples.

bioRxiv preprint doi: https://doi.org/10.1101/2023.09.18.558303; this version posted September 22, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

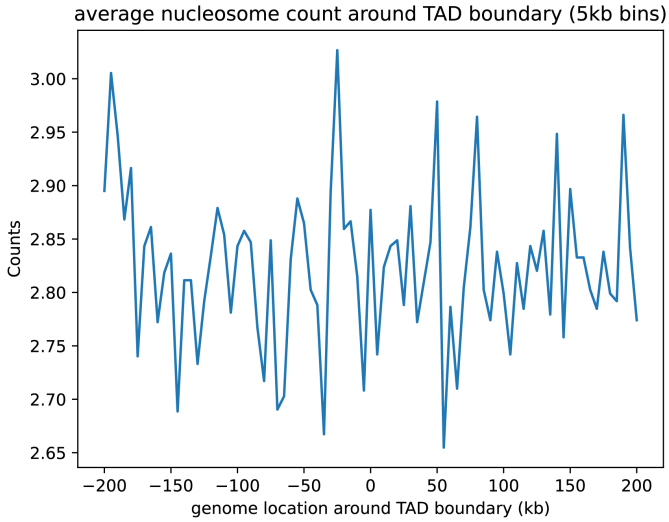


Supplementary Figure 5: Effects of exogenous expression of dinoflagellate DVNPs on chromatin accessibility in the yeast *S. cerevisiae*. ATAC-seq profiles of *S. cerevisiae* expressing *Hematodinium* sp. DVNP.5 (from Irwin et al. 2018[83]) and a vehicle control, as well as additional replicates for *B. minutum* DVNP symbB.v1.2.006931 and *Hematodinium* sp. DVNP.12 and control samples. "OFF" and "ON" refer to cells in which the expression of *Hematodinium* sp. DVNP.5 is induced or not.

bioRxiv preprint doi: https://doi.org/10.1101/2023.09.18.558303; this version posted September 22, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

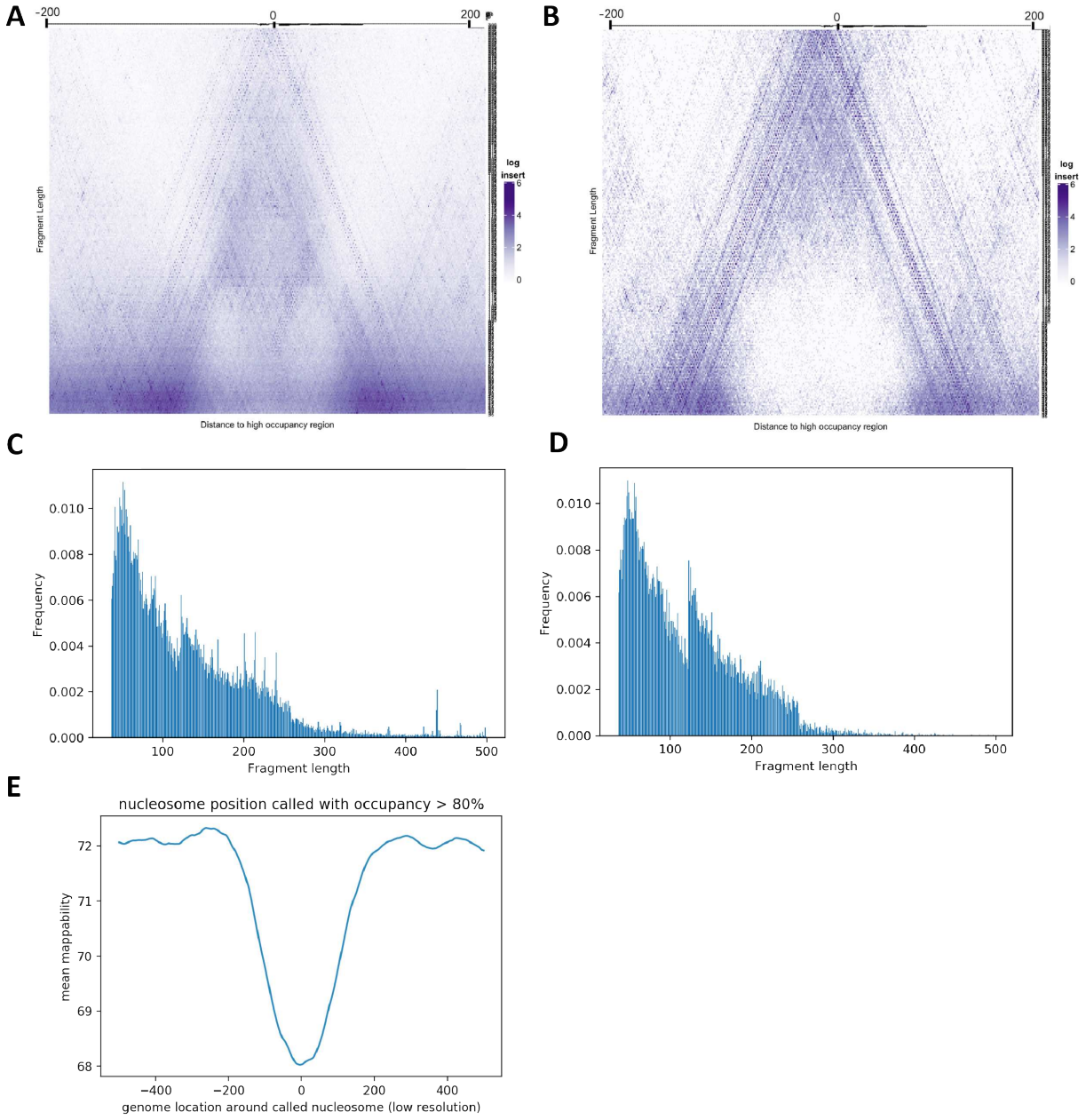


Supplementary Figure 6: Overlap between dinoTAD boundaries, regions of low accessibility, and regions of high 5-hmU.



Supplementary Figure 7: Positioned nucleosomes as a whole are not strongly enriched around dinoTAD boundaries.

bioRxiv preprint doi: https://doi.org/10.1101/2023.09.18.558303; this version posted September 22, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



Supplementary Figure 8: Properties of putative positioned nucleosomes in the *B. minutum* genome. (A) V-plot of low-resolution positioned nucleosomes (n=30,107) (B) V-plot of low-resolution positioned nucleosomes with minimum occupancy cutoff of 0.8 (n=2,166) (C) Fragment distribution over low-resolution positioned nucleosomes (D) Fragment distribution over high-resolution positioned nucleosomes with minimum occupancy cutoff of 0.8 (E) Average mappability (for reads of length 75 bp) over positioned nucleosomes.