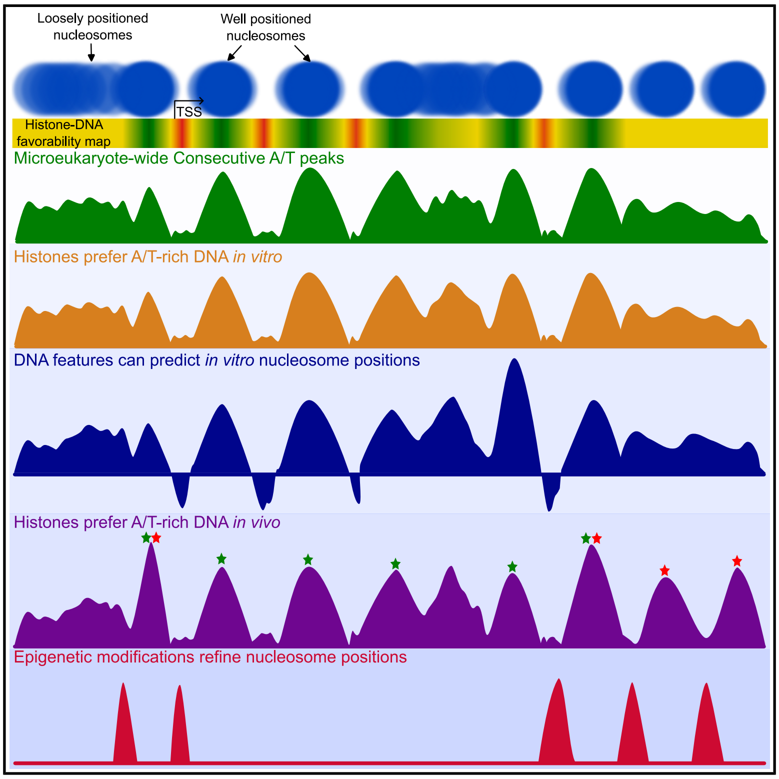
iScience

Article

Consecutive low-frequency shifts in A/T content denote nucleosome positions across microeukaryotes

Graphical abstract


Loosely positioned nucleosomes
Well positioned nucleosomes
TSS
Histone-DNA favorability map
Microeukaryote-wide Consecutive A/T peaks
Histones prefer A/T-rich DNA *in vitro*
DNA features can predict *in vitro* nucleosome positions
Histones prefer A/T-rich DNA *in vivo*
Epigenetic modifications refine nucleosome positions

Highlights

- Nucleosome-sized shifts in the A/T content are widespread across microeukaryotes
- Peaks in A/T content denote nucleosome positions across microeukaryotes
- DNA and epigenomic machinery together coordinate *in vivo* nucleosome positions
- AI model was constructed for improved prediction of nucleosome positions

Authors

Stephen J. Mondo, Guifen He,
Aditi Sharma, ..., M. Catherine Aime,
Ronan O’Malley, Igor V. Grigoriev

Correspondence

sjmondo@lbl.gov

In brief

Genomics; Chromosome organization; Molecular interaction; Model organism; Artificial intelligence

Mondo et al., 2025, iScience *28*, 112472
May 16, 2025 © 2025 The Author(s). Published by Elsevier Inc.
https://doi.org/10.1016/j.isci.2025.112472

CellPress

iScience

CellPress
OPEN ACCESS

Article

Consecutive low-frequency shifts in A/T content denote nucleosome positions across microeukaryotes

Stephen J. Mondo,[1,2,3,13,*] Guifen He,[1] Aditi Sharma,[1] Doina Ciobanu,[1] Robert Riley,[1] William B. Andreopoulos,[1,10] Anna Lipzen,[1] Alan Kuo,[1] Kurt LaButti,[1] Jasmyn Pangilinan,[1] Asaf Salamov,[1] Hugh Salamon,[1,11] Lili Shu,[4,5] John Gladden,[6] Jon Magnuson,[7] M. Catherine Aime,[8] Ronan O'Malley,[1,12] and Igor V. Grigoriev[1,2,9]

[1]US Department of Energy Joint Genome Institute, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA
[2]Environmental Genomics and Systems Biology Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA
[3]Department of Agricultural Biology, Colorado State University, Fort Collins, CO 80523, USA
[4]College of Horticulture, Shenyang Agricultural University, Shenyang 110866, P.R. China
[5]Engineering Research Center of Chinese Ministry of Education for Edible and Medicinal Fungi, Jilin Agricultural University, Changchun 130118, P.R. China
[6]Joint BioEnergy Institute, Emeryville, CA 94608, USA
[7]Pacific Northwest National Laboratory, Richland, WA 99352, USA
[8]Department of Botany and Plant Pathology, Purdue University, West Lafayette, IN, USA
[9]Department of Plant and Microbial Biology, University of California, Berkeley, Berkeley, CA 94720, USA
[10]Present address: Department of Computer Science, San José State University, San José, CA 95192, USA
[11]Present address: Exelixis, Alameda, CA 94502, USA
[12]Present address: Department of Human Genetics, University of Chicago, Chicago, IL 60637, USA
[13]Lead contact
*Correspondence: sjmondo@lbl.gov

https://doi.org/10.1016/j.isci.2025.112472

SUMMARY

Nucleosomes are the basic repeating unit, each spanning ≈150bp, that structures DNA in the nucleus and their positions have major consequences on gene activity. Here, through analyzing DNA signatures across 1117 microeukaryote genomes, we discovered ≈150bp shifts in A/T content associated with nucleosome organization. Often consecutively arrayed across the genome, A/T peaks were enriched surrounding transcriptional start sites in specific clades. Most nucleosomes (both *in vitro* and *in vivo*) across eukaryotes aligned with A/T peaks, even in the presence of DNA modifications. Using artificial intelligence-based approaches, we describe DNA features associated with nucleosomes and construct a deep learning (DL) model for improved nucleosome occupancy prediction. Using this model, we found that ≈70% of "random" transfer DNA inserts from an *in vivo* fungal RB-TDNAseq library avoided DL predicted nucleosome-bound regions. This study reveals a eukaryote-wide strategy for generating cassettes of nucleosome-favorable DNAs that has a profound impact on nucleosome organization.

INTRODUCTION

Having evolved prior to the divergence of Eukaryota and Archaea,[1] nucleosomes are a key defining feature present across nearly all eukaryotes. They are the most fundamental packaging unit of DNA within the nucleus and are extremely conserved. Almost universally, nucleosomes wrap ≈150 bp of DNA around an octamer of highly conserved histone proteins which are then connected through a variable amount (4-80bp) of linker DNA.[2–4] Since these complexes package DNA, precise control over their positioning is critical to properly phase nucleosomes, provide transcription factors access to their binding sites, and regulate gene expression.[5,6] Therefore, learning the rules that govern this complex regulatory system will be crucial for improving our broader understanding of eukaryotic chromatin, as well as for the development of more advanced capabilities to manipulate it.

Beyond their impact on chromatin structure and gene expression, nucleosomes also serve to protect DNA within the nucleus. Being physically more difficult to access, some transposons and foreign elements struggle to integrate into nucleosome-bound (closed) regions of the genome.[7,8] Indeed, this characteristic of nucleosomes has led to the innovation of new techniques for mapping open chromatin, such as ATAC-seq,[9] but also has downsides for our own genetic engineering efforts, for example, impeding Cas9 access to DNA for CRISPR-based gene editing.[10] Beyond the binding of DNA within a single nucleosome, nucleosomes can further interact with each other to form complex, 3D structures that may further bury sequences.[11] The final structure of chromatin within any given cell is orchestrated through a complex interplay of machinery acting at multiple layers, including the presence of DNA binding proteins, positions of epigenetic modifications (both upon histones and DNA), and regulation by chromatin remodelers.[11]

While the interplay between these layers is poorly understood, we know that each of them plays a crucial regulatory role. For example, through physically occupying space, non-histone DNA binding proteins can prohibit nucleosome formation and play a critical role in local structure.[6,12,13] Epigenetic modifications can also have important consequences on both chromatin structure and gene expression. For example, a variety of histone modifications have been characterized that impact nucleosome organization, alter the physical properties of nucleosomes, et cetera.[6] Meanwhile, DNA modifications such as 5-methylcytosine (5mC) reduce DNA flexibility, increasing occupancy and decreasing expression.[14,15] In contrast to 5mC, 6-methyladenine (6mA) DNA modifications have been shown in fungi, ciliates, and green algae to be enriched in linker DNA, where it is proposed that they decrease nucleosome "fuzziness,"[16–19] helping to lock them more strongly in place. Poor positioning or "fuzziness" occurs when a nucleosome is observed occupying a range of nearby positions as opposed to a single defined point.[5] Eukaryotes also deploy chromatin remodeling complexes,[20,21] which can serve multiple roles including: sliding nucleosomes, distorting histone octamers to drive chromatin compaction, and regulating higher order chromatin structure.[3,22] With all these components interacting simultaneously, it is no surprise that nucleosome positions are dynamic and can differ, even between nuclei of the same tissue type.[23,24] However, nucleosomes are also frequently observed to occupy the same position across a population of nuclei,[5,24,25] indicating that fixed nucleosome positions are both possible and critical for gene regulation.

As the substrate which physically interacts with histones to form nucleosomes, one might hypothesize that DNA itself could play an important role, either by directly specifying desirable locations for static nucleosome positioning or by altering the local consequences of regulation by any of the machinery described above. As DNA must sharply bend and twist in ∼11bp increments to wrap around histones—requiring a substantial amount of bending energy[6,26]—it is possible that certain DNA sequences might be more favorable for histone binding than others. Supporting this hypothesis, in addition to the general observation that DNA binding proteins frequently depend on local structure (shape) and cyclizability (bendability) of DNA,[27–30] several studies have described an intrinsic "DNA code" for nucleosome organization in a variety of model organisms.[31–33] This work largely described patterns such as oscillations in dinucleotide frequencies,[32,34–37] DNA flexibility[38,39] and presence of nucleosome "repelling" genomic elements.[40,41] For example, in Baker's yeast (*Saccharomyces cerevisiae*), high frequency (∼11bp) shifts in AA/TT/TA dinucleotides were reported to be highly associated with nucleosome positions,[32,33] while in fission yeast (*Schizosaccharomyces pombe*), chemical maps revealed an additional pattern of increased A/T content surrounding nucleosome center (dyad) positions.[4] Many other analyses took a "bottom up" approach, trying to find or construct optimal nucleosome positioning sequences using biophysical methods.[42,43] Possibly in conjunction with other DNA binding proteins and transcription machinery,[41] these DNA patterns have often been reported to result in "statistically positioned" nucleosomes, wherein nucleosomes are forced into a particular organization due to unfavorable flanking DNA patterns.[44,45]

While this work clearly demonstrates that histones can have preferences for certain DNA sequences, the broader significance of nucleosome positioning DNA patterns *in vivo* has been contentious, as many patterns were either not detected or poorly overlapped with *in vivo* nucleosomes in other model eukaryotes.[46,47] Furthermore, the way in which these data were analyzed could potentially lead to different interpretations, leading to confusion as to whether DNA plays any meaningful role in nucleosome organization at all.[48,49] Unfortunately, despite only analyzing a handful of model eukaryotes, we are not aware of any studies that systematically explore either i) the broader distribution of these DNA patterns across taxa or ii) survey non-model organisms for new DNA patterns that have a stronger or more consistent impact on nucleosome placement.

To more comprehensively explore this, we conducted a survey across 1117 eukaryotes, analyzing nucleotide patterns at the assembly level as well as surrounding 14 genomic features. We discovered abundant ∼150bp (roughly nucleosome-sized) shifts in A/T content in most eukaryotic genomes, the majority of which were arrayed consecutively in the genome, referred to here as Consecutive A/T Peaks (CAPs). Phylogenetically clustered across Eukaryota, CAPs were enriched surrounding specific genomic features, particularly transcriptional start sites and coding sequence start sites. Through both *in vitro* and *in vivo* assays, we confirm that these DNA signatures are related to nucleosome occupancy, revealing a eukaryote-wide strategy for generating cassettes of nucleosome-favorable DNAs at key genomic locations. *in vitro* nucleosomes tended to have AT-rich centers, flanked by peaks of GC ±37bp from the nucleosome dyad. This pattern is reminiscent of the smoothed profile reported previously in *Schizosaccharomyces pombe*[4] but differs substantially in its structure, especially at single-base resolution, suggesting a conserved ancestral pattern of AT-rich nucleosome centers that has undergone refinement in *Sc. pombe*. Through exploring the relationship between CAPs, N6-methyladenine DNA modifications (6mA), and nucleosomes in the model green alga *Chlamydomonas reinhardtii*, we reveal DNA sequence as a prominent contributor to the coordination of *in vivo* organization. Last, using artificial intelligence-based approaches, we explore DNA signatures associated with nucleosomes and develop a model to predict nucleosome positions in fungi. Beyond improved prediction of *in vitro* nucleosomes, this approach revealed that ∼70% of "random" RB-TDNA inserts[50] avoid predicted nucleosome locations, demonstrating the value of such information for designing genetic engineering studies in eukaryotes.

RESULTS

Nucleosome-sized shifts in A/T content detected across Eukaryota

To survey eukaryotes for nucleosome related DNA patterns, we explored shifts in A/T content along assembled scaffolds. 1117 diverse eukaryotes available through the DOE Joint Genome

Institute (JGI) portals MycoCosm (https://mycocosm.jgi.doe.gov/)[51] and PhycoCosm (https://phycocosm.jgi.doe.gov/)[52] as well as genomes from select taxa (Table S1) were included in this analysis (see Data S1 for more details). In these lineages, we calculated A/T frequency per-site (averaged across a ±25bp sliding window; Figure 1A), then identified A/T peaks by their prominence (vertical distance between the peak summit and its lowest point), retaining those with a prominence $\geq$2 standard deviations from the mean % A/T per scaffold. Despite this approach being distance independent, when analyzing the ~110 million peaks observed across all taxa, we found that neighboring peaks are most frequently spaced ~150 bp away from each other (Figure 1B), and this spacing is associated with increased peak prominences (indicating higher structure in these regions). ~150bp is the most common spacing observed within individual genomes as well, regardless of taxonomy (Figure 1B). We also observed many lineages where the most common spacing between peaks is ~100bp, revealing additional patterns worth exploring. Furthermore, on average, over half of all peaks within genomes are found in nucleosome-sized (100-200bp) arrays of 2 or more peaks (Data S1). Suggesting an evolutionary component to their distribution, the abundance of these Consecutive A/T content Peaks (CAPs) varied by taxonomic clade (Figures 1C and S1). Overall, our results reveal that genome-wide CAPs (gCAPs) are frequent across eukaryotes. Rather than randomly distributing across the genome, they are often organized into consecutive, structured arrays with nucleosome-sized spacing.

Cross-kingdom presence of nucleosome-sized shifts in A/T content at gene promoters

After detecting abundant genome-wide CAPs, we explored whether they accumulate preferentially surrounding specific genomic features. We profiled G/C distribution surrounding 14 features, including start and end coordinates for: 5′ and 3′ untranslated regions (UTRs), coding sequences (CDS), exons, introns, and repeats (known and *de novo* predicted). Consistent with their consecutive placement across the genome and suggesting a regulatory role, we observed that lineages dispersed across Eukaryota also display prominent, periodic shifts in nucleotide frequencies extending in both directions surrounding features (Figures 1D and S2A). These feature-specific Consecutive A/T Peaks (fCAPs) are non-random and are eliminated if feature coordinates are randomized even by only 80 bp in either direction (Figures S3A and S3B), indicating that accurate feature coordinates are crucial to fCAP detection. Some (but not all) fCAPs coincided with intron positions (Figure S2A). To calculate a feature-focused measure of this signal, we counted how many peaks were observed (minimum of 2 peaks) surrounding each feature and summed their prominence (Figure S2C), producing a "fCAP score."

fCAP scores per feature were then placed in a phylogenomic context, which revealed that i) as observed at the assembly level, lineages with high fCAP scores are phylogenetically clustered (Figures 1C and S1), and ii) fCAPs are generally more pronounced upstream of genes, specifically at transcriptional and coding sequence start sites (Figures 1C, 1D, and S2C). However, directly comparing these features reveals that scores are higher at the TSS for nearly all lineages with high fCAP scores (Figures 1C and S3B), suggesting that CAPs are particularly important surrounding TSSs. Median periodicity (distance between peaks) at the TSS and coding sequence starts (CDS)—both locations where proper nucleosome phasing is critical for gene regulation—was normally distributed around ~150bp, while periodicity was lower at other features (Figure S2D).

In fungi, the most pronounced signals are observed in the Basidiomycota. Within this phylum, members of the Microbotryomycetes, for example, *Meredithblackwellia eburnea* and *Leucosporidiella creatinivora*, showed the highest fCAP scores of any eukaryote (0.59 and 0.69, respectively), indicating that A/T is shifting dramatically surrounding TSSs in these taxa (Figures 1, S2, and S3). In addition to the Microbotryomycetes, high fCAP scores are commonly observed in Agaricomycotina (Basidiomycota), particularly in the Sebacinales and Tremellales (including the human pathogen *Cryptococcus neoformans* and close relatives) (Figure 1C). In non-fungi, fCAP scores are generally highest at TSSs in green algae (Chlorophyta). This clade includes the model green alga *Chlamydomonas reinhardtii*, which has an elevated TSS fCAP score relative to many taxa, but not as high as most other Chlorophyta. However, this was also observed in other lineages such as *Toxoplasma gondii* (Apicomplexa) and *Trichomonas vaginalis* (Figures 1C and S2A). Except for taxa with extremely high TSS fCAP scores, we did not observe a strong relationship with gCAPS (Figures 1C and S1), suggesting other mechanisms for accumulating these patterns. CAP scores surrounding all 14 features for all 1117 lineages are available in Data S1.

Consecutive A/T peaks are associated with specific genome architectures

To understand what genomic features drive the abundance of CAPs, we divided scores for fCAPs (TSS) and gCAPs (% of all peaks found in consecutive arrays) into quartiles, then explored whether higher scores were associated with a particular feature (see methods for a list of surveyed features). We found that genome size, particularly smaller and less repetitive genomes, is significantly ($p \leq 0.05$, independent T-test with Bonferroni correction for multiple comparisons) associated with higher CAP scores (Figure S4). Interestingly, higher genome-wide %GC is also significantly associated with higher CAP scores, as are higher median protein lengths. Across features, trends appear strongest for fCAPs as compared to gCAPs (Figure S4). Overall, these observations indicate that in general, smaller, higher GC genomes with larger proteins tend to harbor the most CAPs.

Consecutive A/T peaks are related to *in vitro* nucleosome organization

The observation of roughly nucleosome-sized CAPs led us to hypothesize that these patterns may play some role in nucleosome organization, especially in clades with high TSS fCAP scores. To test this, we investigated whether CAPs could directly serve as physically preferable sites for nucleosome formation by generating *in vitro* nucleosome maps (Nuc-seq) for 6 select lineages, primarily from the fungal kingdom. Lineages were chosen based on taxonomy and TSS fCAP scores, including lineages with high

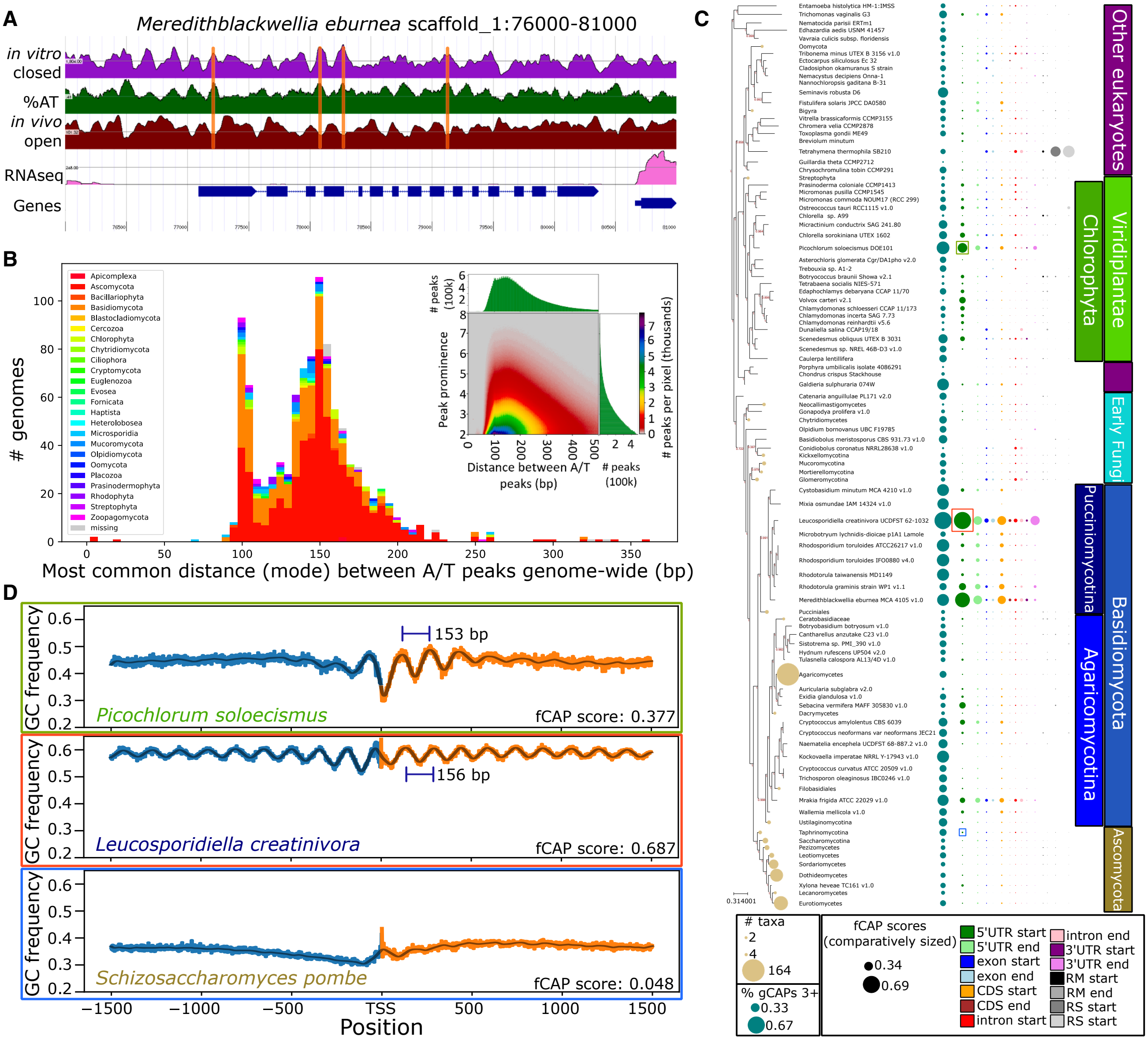


Figure 1. Consecutive A/T peaks (CAPs) are found both genome wide and at specific genomic features

(A) Example region (scaffold_1:76000-81000) from *Meredithblackwellia eburnea* showing A/T content as compared to *in vitro* and *in vivo* nucleosome data, revealing that higher AT is associated with nucleosome-bound regions. A subset of sites displaying agreement across datatypes are highlighted in orange.

(B) Analysis of genome-wide shifts in A/T content reveals that peaks are most frequently observed ∼150bp from each other, regardless of lineage. Inset shows distance between A/T peaks (x axis) for all ∼110 million peaks across eukaryotes and their observed prominence (y axis), revealing that peaks within roughly nucleosome-sized spacing from their neighbors typically have the most structure (highest prominence).

(C) Eukaryote-wide phylogeny including published and select lineages available in MycoCosm and PhycoCosm. Only lineages with UTR annotations are displayed here, see Figure S1 for an uncollapsed phylogeny of all 1117 lineages. Aligned with each leaf node, teal circles sized proportionally to the percent of gCAPs in arrays of 3 or more peaks are shown. Additionally, comparatively sized circles show fCAP scores across 7 different genomic features: 5′ UTR (greens), exons (blues), CDS (orange/yellow), introns (red/light pink), TTS/3′ UTR (purple/pink) and known/*de novo* repeats (greys). RM: RepeatMasker identified repeats, RS: *de novo* RepeatScout identified repeats. For collapsed nodes, CAP scores represent the average value across all members of that node.

(D) Average GC calculated per site (y axis) is shown surrounding surrounding transcriptional start sites (x axis) for *Picochlorum soloecismus* (top), *Leucosporidiella creatinivora* (middle), and *Schizosaccharomyces pombe* (bottom). Blue portions of the curve indicate raw GC frequency upstream of the TSS, while orange shows downstream of the TSS. A Butterworth filter was also applied to smooth raw data (black curve). Note that the y axis (percent GC) has been scaled equally across taxa to illustrate differences in peak heights across taxa. Box colors surrounding plots match those found on the phylogeny in C. All lineage names are colored within this and future figures according to the taxonomy legend in Figure 1C.

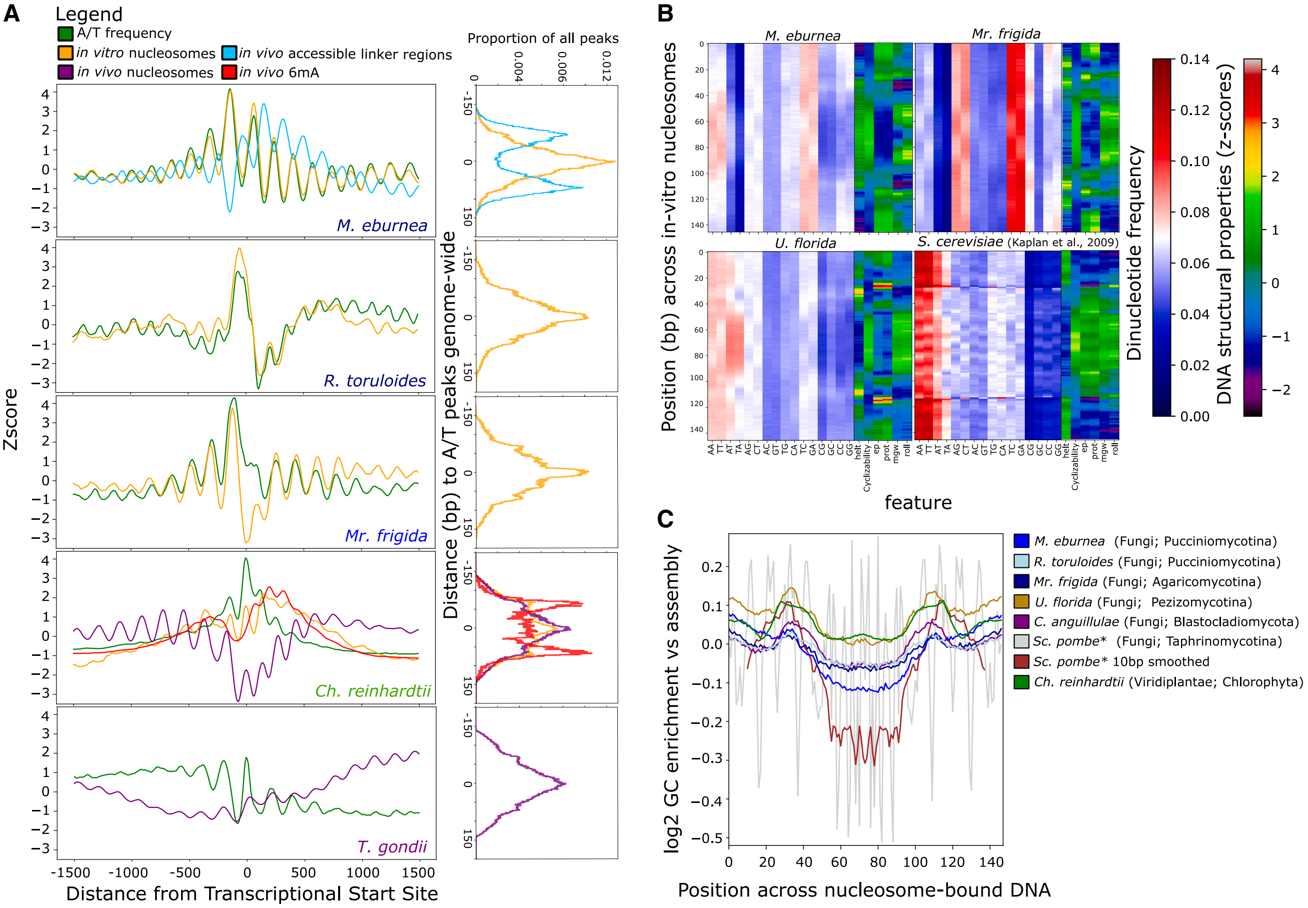


Figure 2. CAPs create sites that are physically favorable for nucleosome formation

(A) A/T frequency correlates with *in vitro* and *in vivo* nucleosome locations. *in vitro* nucleosome occupancy (orange) are plotted alongside per-site A/T frequency (green) surrounding transcriptional start sites across 3 fungi, 1 alga and 1 apicomplexan (lineage names are colored based on Figure 1C). *in vivo* Accessible Linker DNAs (ALDs) identified through ATAC-seq are plotted in light blue, while *in vivo* nucleosome occupancy identified using MNase-seq are plotted in purple. 6mA-IP results from *Ch. reinhardtii*[17] are shown in red. All curves are normalized by conversion to Z-scores (standard deviation from the mean). Z-scores (y axis) are scaled equally to illustrate differences in magnitude across taxa. Next to their TSS profiles, for each lineage, the proportion (x axis) of all genome-wide peaks and their distance to closest A/T peaks (y axis) are shown. See Figure S6D for comparison to randomized nucleosome peaks.

(B) Dinucleotide frequencies (x axis) are plotted at each position (y axis) across *in vitro* nucleosome-bound DNAs, revealing that they are lineage specific and notably different from *Sa. cerevisiae*.[33] Distribution of DNA structural properties, including DNA shape features[27] and cyclizability[30] are shown next to dinucleotide heatmaps, revealing consistency across taxa not observed for dinucleotides. See Figure S7 for all fungal profiles.

(C) Unlike dinucleotides, unsmoothed GC content (y axis) shows a conserved profile across nucleosome-bound DNA (x axis). GC content is normalized against genome-wide percent GC (y axis). Results from an *in vivo* chemical map of nucleosomes in *Sc. pombe*[4] (denoted with * in the legend) were also included, revealing a similar profile to *in vitro* results across other eukaryotes after smoothing data across 10bp windows.

(*Meredithblackwellia eburnea*: 0.591, *Mrakia frigida*: 0.202), moderate (*Rhodosporidium toruloides* ATCC26217: 0.128, *Chlamydomonas reinhardtii* [Chlorophyta]: 0.108) and low scores (*Usnea florida*: 0.002, and *Catenaria anguillulae*: 0.000). NucSeq exposes unamplified genomic DNA to histones and lets nucleosomes self-assemble prior to digestion with micrococcal nuclease (MNase), allowing us to measure the physical preferences of histones for DNA signatures in the absence of epigenomic regulators—except possible DNA modifications, which are rare in all selected lineages except *Ch. reinhardtii* and at *Ca. anguillulae* transposons.[16,17]

We found that A/T peak locations and *in vitro* nucleosome occupancy are highly correlated in all lineages, where they explain 47.6% (*Ch. reinhardtii*) to 68.79% (*M. eburnea*) of nucleosome positions within 40 bp (Figures 1A and 2A; Table S2). Proximity to A/T peaks also led to better positioning, as nucleosome "fuzziness" was lowest at these locations (Figure S5A). Additionally, CAPs are often more tightly associated with nucleosome positions than individual A/T peaks (Figure S5B). The relationship between CAPs and nucleosomes is especially strong in the Basidiomycota, where *in vitro* occupancy and A/T frequency showed extremely similar patterns surrounding genomic features, for example, TSSs (Figure 2A). This was also observed in *Ch. reinhardtii*, despite the presence of DNA modifications in this organism.[17] Interestingly, Basidiomycota also have slightly different *in vitro* occupancy profiles compared to other eukaryotes surrounding genic features such as introns and transcription termination sites (Figure S6A and S6B). Unlike basidiomycetes and Chlorophyta, lineages with low fCAP scores show little relationship between A/T and nucleosome profiles at genic features (Figure S6C), despite showing agreement at the assembly level (Figure S6D; Table S2). Importantly, none of these relationships are observed when we randomized *in vitro* peak locations, which revealed a flat distribution with a width that is simply a function of the average distance between A/T peaks (Figure S6D; Table S2).

As they were central to previous analyses of DNA-histone interactions,[32–37] we also analyzed dinucleotide frequencies per-site across nucleosome-bound DNA. As the best studied example, we included previously published *in vitro* chromatin data from *Sa. cerevisiae*,[33] where the expected ∼11bp shift in dinucleotides is clearly observed (Figure S7). This profile is similar to what was found through chemically mapping *in vivo* nucleosomes in *Sc. pombe*[4] (Figure S7). However, other taxa analyzed here did not show this pattern. While we did see some shifting dinucleotide profiles, they appear mostly lineage-specific and often larger in size (∼24bp; Figures 2B and S7). Dinucleotide patterns are also highly variable between organisms (Figures 2B and S7), even within a single class (e.g., Microbotryomycetes). Overall, these findings suggest that dinucleotides themselves may not be the main driver behind histone-DNA interactions as they do not carry information that can be generalized across taxa.

Despite these differences, all lineages (except *Sa. cerevisiae* and *Sc. pombe*) converge on a similar physical structure of DNA – calculated using DNAcycP[30] and DNAshapeR[27] – where values increase surrounding nucleosome centers for cyclizability, rolls, helical twists and minor groove widths, while values decrease for propeller twists and electrostatic potentials. Like dinucleotides, we often observe higher frequency shifts in shape profiles, but these tend to show consistency across taxa (Figure S7). A similar G/C profile across nucleosome-bound DNAs was also found, where bound DNAs harbor AT-rich centers and peaks of G/C ±37 bp from the dyad (Figure 2C). *Sc. pombe* displays a similar profile when G/C content is smoothed across 10bp windows (as demonstrated in Moyle-Heyrman et al., 2013[4]), but not at single-base resolution, as observed in all other taxa. Overall, this general conservation across divergent taxa in both structure and G/C content suggests a broadly meaningful pattern that is related to increased favorability for nucleosome formation which has potentially undergone refinement in specific taxa.

Consecutive A/T peaks are related to *in vivo* nucleosome organization

While conferring favorability for nucleosome formation, it is still unclear whether CAPs bear any meaning for *in vivo* nucleosome locations. To address this, we used all publicly available ATAC-seq and MNase-seq datasets in the NCBI GEO database as of November 1st, 2024 for Basidiomycota (2 Agaricomycetes: *Sparassis latifolia* and *Cryptococcus neoformans*[53,54]), other eukaryotes with high fCAP scores (MNase-seq datasets for *Ch. reinhardtii*,[17] *Toxoplasma gondii*[55]) and generated our own ATAC-seq data for *M. eburnea* (Pucciniomycotina). Here, we used ATAC-seq[9] data to map Accessible Linker DNA (ALD) regions while MNase-seq was used to show nucleosome-bound regions.[56] We also assessed "premium" nucleosome positions determined by Moyle-Heyrman et al., 2013[4] through chemical mapping in *Sc. pombe*.

Consistent with the presence of nucleosomes at AT-rich locations, ALDs in fungi (e.g., *M. eburnea*) show a striking inverse-correlation with A/T frequency (Figures 1A and 2A). Even in *Sp. latifolia*, a basidiomycete with relatively low CAP scores, we found that AT-rich regions are somewhat depleted in open chromatin (Figures S8A and S8B). A similar trend is observed genome-wide (Figure 2; Table S2), including in lower resolution (Figure S9) *Cr. neoformans*[54] data and across multiple genomic features. ATAC-seq also confirms the nucleosome-rich introns and transcription termination sites observed *in vitro* in Basidiomycota (Figures S6A, S6B, S8A, and S8B). More prominent A/T peaks are also more tightly associated with *in vivo* nucleosome positions (Figures S8C and S8D).

Consistently, MNase-seq datasets in non-fungi show an enrichment of nucleosomes at A/T-rich locations. Genome-wide, ∼52% of all MNase-seq dyads are within 40bp of peaks in A/T content in *T. gondii* (Apicomplexa) and *Ch. reinhardtii* (Figure 2A side panels; Table S2). This increased to 56.37% in *Ch. reinhardtii* when considering only *in vivo* peaks with prominence >2.0. Conversely, only 23.55% (*M. eburnea*) to 44.78% (*Sp. latifolia*) of ATAC-seq peaks in Basidiomycota (Figure 2) are found within 40 bp of A/T peaks, indicating less open chromatin at A/T peaks. Furthermore, 54.62% of chemically determined "premium nucleosome" positions in *Sc. pombe* overlap with A/T peaks (Figure S6D; Table S2), despite the notably different DNA profile across nucleosome-bound DNA in this organism (Figures 2C and S7). As found using *in vitro* data, this relationship is not observed when we randomized *in vivo* peak locations (Figure S6D; Table S2). Combined, our *in vitro* and *in vivo* work demonstrates that shifts in A/T frequencies (especially CAPs) consistently emerge as playing important roles in nucleosome organization across eukaryotes.

DNA and epigenomic modifications together regulate nucleosome positions

As a well-studied model eukaryote, we could leverage additional epigenomic data in *C. reinhardtii* and analyze its relationship to both A/T peaks and *in vivo* nucleosomes. Here, we took advantage of N6-methyladenine DNA modification data from.[17] Located at transcriptional start sites, 6mA is associated with active gene expression and is usually found symmetrically at ApT dinucleotides.[16–18] We found that 6mA is enriched ∼70bp from both A/T peaks and *in vivo* nucleosome centers (Figure 3), confirming 6mA enrichment within linker DNA.[17,18] As a largely similar pattern is observed across both datasets, this observation intrinsically reinforces the impact these DNA patterns have on *in vivo* chromatin organization (Figure 3). Additionally, 6mA modifications found ∼70bp away from A/T peaks and *in vivo* nucleosome dyads often display notably higher fold enrichment compared to other locations (Figure 3), indicating stronger 6mA signal in these locations. Consistent with our genome-wide analysis, we also see 6mA distributed around transcriptional start sites in regions where A/T peaks and nucleosomes are less common (Figures 2A and 3).

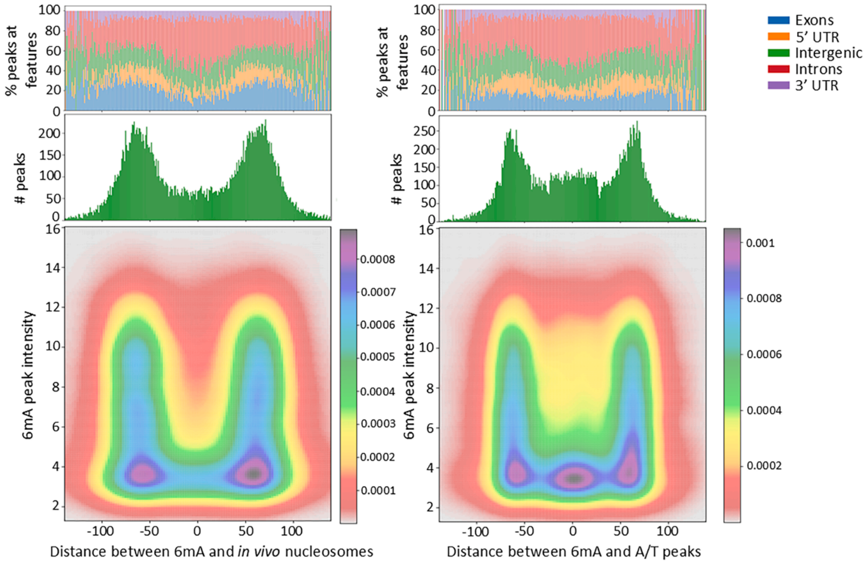


Figure 3. DNA and epigenomic modifications act in concert to organize *Ch. reinhardtii in vivo* nucleosomes

(A) Bottom: density plot showing distance between 6mA and *in vivo* nucleosomes (x axis) and 6mA peak intensity (y axis), demonstrating that most high intensity 6mA peaks are found at nucleosome boundaries. Middle: histogram showing the number of peaks observed at each position (bin size = 1bp). Top: percent of peaks at different genomic features, revealing 6mA and *in vivo* nucleosome overlaps are more common at introns (red) compared to other genomic features.

(B) A similar profile is observed when comparing 6mA positions to A/T peaks instead of nucleosome dyads. Most high intensity 6mA peaks are observed ∼70bp from peak centers, but a higher degree of peak overlap is also observed, suggesting both A/T peaks as a major player in *in vivo* nucleosome position and coordination with epigenomic machinery for final nucleosome placement.

While less frequent (and unlike *in vivo* nucleosomes), there are cases where 6mA overlapped A/T peaks, which could offer a mechanism by which this modification can override "default" nucleosome positions conferred by DNA patterns. These overlaps occurred slightly more at introns, and this pattern is further exaggerated when comparing across dark and light conditions,[17] specifically highlighting A/T peaks at introns as a possible location for condition-specific dynamic nucleosome regulation (Figure S10). Overall, our results indicate that: i) A/T peaks remain a prominent contributor to *in vivo* nucleosome organization even when epigenomic modifications are present, ii) 6mA organizes surrounding A/T peaks, which likely acts to further lock nucleosomes in place at CAPs[18] (Figure 3), and iii) epigenomic machinery can override CAPs in a condition-specific fashion, revealing that final positioning is accomplished through a coordinated effort between DNA favorability landscapes (conferred by shifting A/T frequencies) and epigenomic machinery.

DNA features can predict nucleosome positions

Given our observations that CAPs are related to *in vitro* and *in vivo* nucleosome positions, we wanted to understand what specific features of DNA are important for creating nucleosome-favorable landscapes. Therefore, for each fungus with Nuc-seq data available, we randomly sampled ten 100kb segments of DNA and used Random Forests (RF) analysis to rank the importance of 1526 DNA features (see STAR Methods) for the prediction of *in vitro* nucleosome occupancy. Both individual and summed feature importances across all positions show that in general, features that span more nucleotides simultaneously have higher importance for predicting nucleosome occupancy. Specifically, average G/C ±75bp surrounding sites was of highest importance, followed by DNA shape features (Figures 4A and S11). However, lineage-specific variability was also observed. For example, in *M. eburnea* GC in 11bp segments was notably more important than for other lineages, while shape parameters and more localized features (4mers, 3mers, and so forth) play a larger role in prediction for *C. anguillulae* (despite overall low accuracy in this lineage; Figure S11).

While shifts in A/T alone can provide important insights into nucleosome locations, RF revealed that other features (particularly local DNA shape) can also be important factors in altering the (*in vitro*) nucleosome favorability landscape. Furthermore, since a single stretch of DNA cannot be occupied by multiple nucleosomes simultaneously, we anticipate that features in neighboring regions (i.e., nearby favorable sites for histone binding) might also be important factors in local chromatin organization. Therefore, using features identified through our RF analysis plus mononucleotides, we tested the ability for deep learning to improve the computational prediction of *in vitro* nucleosomes. Our model, trained on ∼28 million sites across 4 fungi, combined multiple convolutional (CNN) and recurrent (RNN) neural network layers (following the architecture of Quang and Xie, 2016[28]) to predict nucleosome occupancy at each position in the assembly. For training, we included the 5 largest scaffolds from *M. eburnea*, *R. toruloides* ATCC26217, *U. florida* and *C. anguillulae* and tested our method on the prediction of *in vitro* nucleosome occupancy both on smaller scaffolds from the same organisms, as well as a sample never seen by the algorithm, *Mr. frigida*.

We found that our deep learning model notably improved accuracy of nucleosome prediction compared to CAPs alone *in vitro* (Figure S12A). We also observe similar or improved *in vivo* nucleosome prediction across most taxa (Table S2). Genome-wide, this method was able to predict up to 72.53% (*M. eburnea*) of *in vitro* nucleosomes within 40bp of their actual locations (Figures 4B, 4C, and S12B; Table S2), representing an increased accuracy of 1.72% (*Ch. reinhardtii*) to 7.68% (*Mr. frigida*) compared to using A/T peaks alone. In *Mr. frigida*, our method predicts *in vitro* nucleosome occupancy more reliably ±250bp surrounding the TSS than either A/T peaks or a traditional position weight matrix approach based on

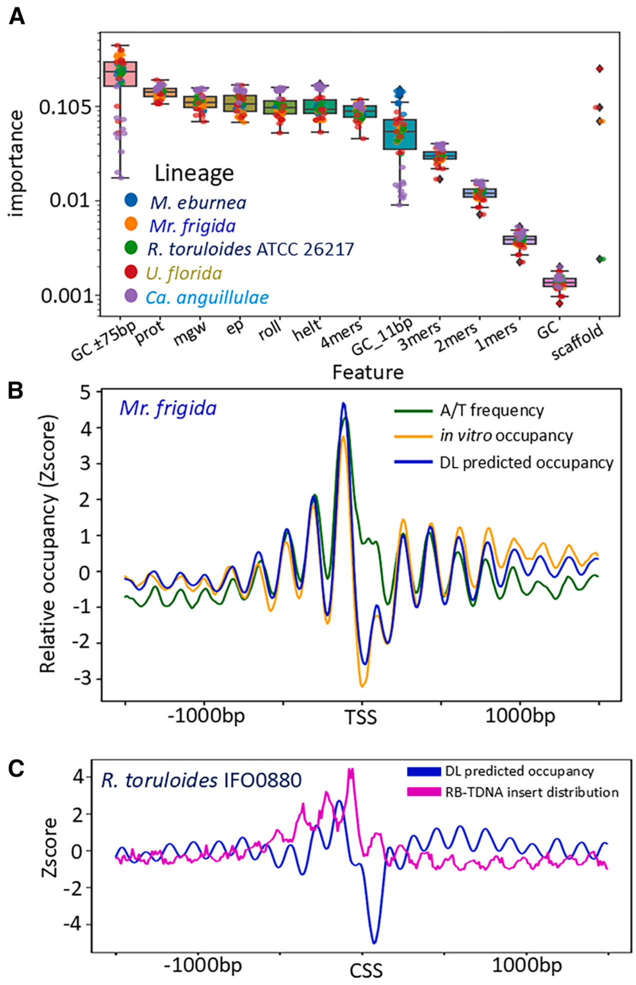


Figure 4. Machine learning methods can predict nucleosome occupancy in Basidiomycota

(A) Random Forests analysis of 1526 genomic features reveals G/C across large windows surrounding sites (±75bp) as the most important feature for predicting *in vitro* nucleosome occupancy. Swarmplots show the cumulative importance of each feature (calculated by summing importances across all position). DNA shape parameters were calculated using DNAshapeR[27] and include: propellar twist (prot), minor groove width (mgw), roll, helical twist (helt), and electrostatic potential (ep). Dots represent the feature importance calculated for each random 100kb sample, colored by lineage. See Figure S11 for the top 30 individual features.

(B) Deep learning (DL) methods can improve the prediction of *in vitro* nucleosome occupancy. *in vitro* occupancy (orange), DL predicted in silico occupancy (blue), and A/T frequency (green) surrounding transcriptional start sites in a lineage excluded from training, *Mr. frigida* (x axis).

(C) DL predicted occupancy is related to *in vivo* nucleosome occupancy and provides value for the interpretation of experimental results. DL predicted occupancy (blue) in *R. toruloides* IFO0880 surrounding coding sequence start sites (CDS) compared to counts of inserts per site (magenta) from an RB-TDNAseq library,[50] revealing that inserts avoid predicted nucleosome-bound regions.

dinucleotides (Figures 4B and S12A). Exploring the correlation between predicted and *in vitro* occupancy scores at this location, we see Pearson's correlation increase from 0.43 (CAPs) to 0.67 using our method (Figure S12A). It is interesting to note that while CAPs alone did not reliably reconstruct occupancy profiles immediately surrounding the TSS in *Mr. frigida*, our deep learning method could despite having no information on TSS locations during training.

Furthermore, through analysis of transfer DNA inserts from an *in vivo* RB-TDNAseq library in *R. toruloides* IFO0880,[50] we found that inserts avoid predicted nucleosome bound regions (Figure 4C). Comparable to what we observed when exploring ALDs in other Basidiomycota (Figure 2; Table S2), here only 32.61% of inserts are found within ±40bp of predicted nucleosome dyads, indicating that inserts strongly avoid predicted nucleosome-bound DNAs (Figure S12C). These findings again emphasize the importance of DNA features for *in vivo* chromatin organization in this clade. Combined, our analysis demonstrates that DNA features play predictable roles in organizing nucleosomes that are conserved across a wide array of eukaryotes.

DISCUSSION

Chromatin is organized through a complex interplay between multiple regulatory layers.[11] Here, our work reveals that DNA, a long-overlooked component of this system, plays an important and predictable role in innate nucleosome organization across eukaryotes. By surveying ∼1100 lineages across Eukaryota we found multiple clades displaying consecutive A/T peaks (CAPs) spaced ∼150bp from each other (roughly the size of nucleosomes), both genome-wide (gCAPs) and at specific genomic features (fCAPs), especially transcriptional start sites. Through *in vitro* analyses we found that CAPs create physically favorable sites for nucleosome formation (Figures 2, S5, and S6). We also found that *in vivo*, CAPs overlap with nucleosome positions, even when DNA modifications were present (Figures 1, 2, 3, and S8), highlighting their prominent role even in living cells. Emphasizing the conserved impact they have on nucleosome organization, machine learning approaches both i) identified conserved genomic features driving these profiles (Figure 4A) and ii) translated these observations into a method that predicts nucleosome positions across large phylogenetic distances (Figures 4B, 4C, and S12; Table S2).

When placed in a phylogenomic context, we found that CAPs are widespread across Eukaryota but that their abundance varies by clade (Figures 1C and S1). Interestingly, in many lineages gCAPs are abundant yet we observe little overlap with specific genomic features. To us, this suggests that either i) detection of fCAPs depends upon high quality gene structure prediction (which is particularly challenging for UTRs), ii) CAP positions may vary in a gene-specific manner (perhaps due to variability in TF binding sites relative to TSS locations) or iii) CAPs accumulate at features not surveyed in this analysis. While all may be true, through randomizing TSS positions (simulating poor quality 5′ UTR predictions) and comparing TSS and CDS fCAP scores (simulating no UTR predictions due to lack of RNA-seq data), it is clear that accurate gene structure prediction is crucial to detecting these patterns (Figure S3). As such, we expect that improved gene structure prediction and incorporation of additional information about promoter regions would reveal more lineages with these patterns than reported here. Similarly, as a feature related to TSS positions, future exploration into the potential application of CAPs to improve TSS prediction may prove worthwhile.

Regardless of the reason(s) for its variability, even detection of weak fCAP patterns can indicate important structure. For example, CAPs clearly play a major role *R. toruloides* IFO0880 nucleosome organization despite its modest TSS fCAP score (0.099), as ~70% of RB-TDNAseq inserts avoid predicted nucleosome-bound DNA. Furthermore, with an even lower fCAP score in *Sc. pombe* (0.048), we observe a strong overlap between A/T peak positions and high-resolution, chemically determined nucleosomes dyads[4] (Figure S6D). This is despite major differences in nucleosome-bound DNA profiles in *Sc. pombe* compared to other fungi in this study (Figure 2C and discussed later in discussion). To better understand what drives differences in CAP scores, we divided genome into quartiles based on their gCAP and TSS fCAP scores, then analyzed a variety of genomic features. This revealed that higher GC content and smaller (more gene dense) genomes are significantly associated with higher gCAP and fCAP scores. However, there are certainly exceptions, for example in the Chlorophyta where genomes are generally larger than most Fungi, but TSS fCAP scores are also high (Figures 1C and S1).

Taking a subset of fungi and algae with varying fCAP scores we found that, as suggested by their roughly nucleosome-sized spacing, these DNA signatures are strongly associated with nucleosome positions. Our *in vitro* analysis indicates that they create physically favorable sites for nucleosome formation, as highlighted by the large degree of overlap between these datasets (Figure 2). Nucleosomes at these locations are also better positioned, harboring decreased fuzziness compared to those further away from A/T peaks (Figure S5A). We also found that the DNA profile across nucleosome-bound DNA was generally consistent across taxa. In all cases, we see low frequency shifts in A/T content with G/C peaks ±37bp surrounding dyads. This profile appears to drive increased cyclizability at nucleosome centers and a consistent DNA "shape" (Figure 2C).

Of all models explored previously, the profile we observe here most broadly resembles that found in *Sc. pombe*, but with major differences. For example, unlike the lineages analyzed in this study, *Sc. pombe* shows substantial, high-frequency variation across nucleosome-bound DNA, both in G/C content and DNA structure (shape and cyclizability; Figure 2). However, when summarized at larger scales through smoothing G/C content across nucleosomes (Figure 2), or in summary plots (Figures 1D and S6D), the same association between A/T-richness and nucleosomes are observed. Such observations suggest that a refinement of histone-DNA preferences has occurred in *Sc. pombe*, shifting this lineage away from the conserved nucleosome-DNA profiles we see across other taxa.

Nucleosome-bound DNA observed in this study is also notably different from the high frequency (~11bp) shifts in AA/TT/TA dinucleotides previously reported[32,33] and confirmed here (Figures 2B and S7) in *Sa. cerevisiae*. However, other groups argue that these dinucleotide patterns have little relevance for *in vivo* nucleosome organization in other model eukaryotes,[48,49] which is consistent with our observation of diverse dinucleotide patterns even within the same phylogenetic class (*M. eburnea* and *R. toruloides* ATCC26217; Microbotryomycetes; Figures S7A and S7B). It is interesting that in some clades we see larger, ~24bp shifts in dinucleotide profiles. We anticipate that such patterns drive stronger histone-DNA interactions, perhaps through altering DNA structural properties (which show similar patterns; Figures 2B and S7). While multiple lines of evidence (Nuc-seq, ATAC-seq and MNase-seq) independently support the relationship between CAPs and nucleosomes in the lineages described here, it is important to recognize that while highly conserved, histones (and histone variants) across diverse lineages may have evolved different DNA preferences. The long evolutionary history of the Saccharomycotina, which has been characterized by pervasive gene loss,[57] may have provided space for different strategies to emerge which might partly explain the differences observed between *Sa. cerevisiae* and lineages included in our study and warrants further exploration.

Taking advantage of multiple *in vivo* approaches to map nucleosome positions (ATAC-seq and MNase-seq), CAPs again emerged as important features associated with nucleosomes (Figure 2). In fungi, we found that accessible linker DNA (ALD) predicted through ATAC-seq are depleted in regions with high A/T content, indicating increased occupancy at these locations *in vivo*. This was also observed in MNase-seq data from non-fungi. This would make sense as a natural defense against transposable elements, which tend to be A/T-rich.[58] Exploring the relationship between A/T peaks and *in vivo* positions, we find a tighter association between histones and DNA at highly prominent A/T peaks (Figures S8C and S8D), suggesting different strategies could emerge for nucleosome regulation based on the intensity of A/T content at any given genomic location.

Potentially related, in *Ch. reinhardtii*, we found even stronger agreement between CAPs and *in vivo* nucleosomes compared to *in vitro* (Table S2), suggesting either technical challenges during Nuc-seq, or additional biological mechanisms (e.g., epigenetic modifications) that help lock nucleosomes in place at some CAPs. Others have proposed that DNA modifications also aid in this effort, for example 6mA, which is found in linker DNA in green algae[17] and ciliates[18] and is proposed to help decrease nucleosome fuzziness.[18] In the model green alga, *Ch. reinhardtii*, we see CAPs and 6mA are organizing in a consistent fashion (Figure 3), where CAPs overlap most nucleosomes and 6mA is found at CAP/nucleosome borders, within the linker DNA. We did however see an elevated level of 6mA overlapping A/T peaks compared to *in vivo* nucleosomes. As 6mA is heavily enriched at ApT dinucleotides in eukaryotes,[16,17] this might be related to the increased availability at ApT sites at these locations. However, it could also offer a mechanism by which 6mA can override baseline histone-DNA preferences and drive condition-specific organization. Overall, our *in vitro* and *in vivo* analyses reveal DNA is an important driving force behind nucleosome positions, potentially providing a baseline favorability map upon which epigenomic modifications can act to dynamically tweak nucleosome positions.

If DNA were to provide a blueprint for organizing nucleosomes genome-wide, we hypothesized that conserved DNA features should emerge as consistent contributors across diverse lineages. Therefore, we employed machine learning approaches to identify common strategies across fungi. As a reasonable proxy for a physical DNA favorability map, we used Nuc-Seq data collected in this study to identify DNA features that best predicted *in vitro* occupancy scores, which consistently revealed the same important features across taxa. Namely, %GC content across large windows (±75bp surrounding sites), and local DNA shapes[27,29] (Figures 4A and S11). We also note a general decrease in importance from large-distance-spanning features such as %GC in 150bp windows to smaller-distance-spanning features (e.g., di- and mono-nucleotides). This again supports the hypothesis that instead of individual nucleotide or dinucleotide features, histone favorability is conferred on larger scales of tens to hundreds of base pairs and is possibly impacted by nucleotide profiles of neighboring sequences.

To develop more advanced methods for predicting occupancy that can take long distance interactions into account, we turned to deep learning models. Our approach combined convolutional (CNN) and recurrent (RNN) neural networks, which enabled the extraction of higher-order motifs associated with nucleosome occupancy (CNN layers) while also identifying long distance interactions between motifs (RNN layers). When applying this approach to *R. toruloides* IFO0880, a model organism for industrial biolipid production,[50] we find a non-random distribution of RB-TDNAseq inserts, where most inserts occur outside of predicted nucleosome-bound regions (Figures 4C and S12C). This reveals that: i) RB-TDNAseq,[50] such as ATAC-seq[9] and CRISPR-Cas9 based editing approaches,[10] is sensitive to nucleosome positions. ii), Features learned through training across fungi represent conserved DNA signatures meaningful for the broad prediction of fungal nucleosome positions. And iii), *R. toruloides in vivo* nucleosomes are highly structured based on DNA features. Interestingly, while we consistently see improved prediction of *in vitro* nucleosomes, agreement between *in vivo* and predicted nucleosome positions was more variable (Table S2). In some cases, we observed higher agreement (*Sp. latifolia* and *Ch. reinhardtii*), while in others we saw no difference (*Cr. neoformans* and *T. gondii*) or lower agreement (*M. eburnea*).

Experimental approaches for high quality assessment of nucleosome occupancy remain non-trivial and often require customization on a per-species basis. Our computational approach allows us to predict a baseline histone-DNA favorability map with several downstream applications. For example, in the absence of experimental data, these predictions may prove useful for improving the identification of target sites for the integration of foreign DNA and optimizing promoters for maximum access to transcription factor binding sites. Potentially useful for developing novel management strategies, we also detect notable TSS fCAPs in several important human pathogens (for example *Cryptococcus neoformans* and *Toxoplasma gondii*). Beyond applications, our work reveals a novel, conserved genomic profile that plays a major role in chromatin organization. Epigenomic modifications are expected to perturb histone-DNA interactions, but whether modifications have different responses depending on the underlying genomic context is unknown. Our work provides a foundation for beginning to address these key questions that, once understood, have the potential to transform our understanding of eukaryotic gene regulation.

Limitations of the study

As mentioned in the discussion, high-quality UTR annotations are important for fCAP detection. Given that these are not always available (or of low quality), some patterns surrounding TSSs might be missed here. Also, this study did not address histone diversity, modifications, or variants, all of which could perturb histone-DNA interactions and adjust the profiles reported here in a taxonomic or genomic-location-specific manner. Last, the Deep Learning-based method presented here was trained on 4 fungal taxa distributed broadly across the tree. While this training allowed for general prediction across microeukaryotes, it is likely that it will not be as effective outside of fungi and perhaps in specific fungal subclades that show substantially different histone-DNA signatures (e.g., *Saccharomyces cerevisiae*).

RESOURCE AVAILABILITY

Lead contact

Requests for further information and resources should be directed to and will be fulfilled by the lead contact, Stephen J. Mondo (sjmondo@lbl.gov).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- The sequencing data, assemblies, and annotations for all genomes are available from JGI Fungal and Algal Genome portals, MycoCosm (https://mycocosm.jgi.doe.gov/mycocosm/home) and PhycoCosm (https://phycocosm.jgi.doe.gov/phycocosm/home) and have been deposited in GenBank under accession numbers: JBAJAD000000000 (*Meredithblackwellia eburnea* MCA 4105), JBBAIS000000000 (*Rhodosporidium toruloides* ATCC 26217), JBAJAC000000000 (*Mrakia frigida* ATCC 22029), JBBAIT000000000 (*Usnea florida* ATCC 18376), JBAJAE000000000 (*Sparassis latifolia* CCMJ1100). See Table S1 for additional data links. JGI-generated Nuc-seq and ATAC-seq data are available through the SRA database accessions: SRR7472065 (*Mr. frigida* Nuc-seq), SRR7472154 (*U. florida* Nuc-seq), SRR7476903 (*Catenaria anguillulae* Nuc-seq), SRR28069068 (*R. toruloides* ATCC26217 Nuc-seq), SRR28069066 (*M. eburnea* Nuc-seq), SRR28069067 (*Ch. reinhardtii* Nuc-seq), SRR28069093 (*M. eburnea* heavy grind ATAC-seq), SRR28069094 (*M. eburnea* heavy grind ATAC-seq with 5 micron filter), SRR28069106 (*M. eburnea* light grind ATAC-seq), as well as made available for download and viewing on their respective MycoCosm/PhycoCosm data portals.
- Software for predicting nucleosome occupancy is available via https://github.com/sjmondo/NuclPred_v1.
- Any additional information required to reanalyze the data reported in this article is available from the lead contact upon request.

ACKNOWLEDGMENTS

We are grateful to Juna Lee for assistance in processing Nuc-Seq datasets. We are also grateful to two anonymous reviewers for their insightful feedback on this article. This work was conducted by the U.S. Department of Energy Joint Genome Institute, a DOE Office of Science User Facility, is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.

AUTHOR CONTRIBUTIONS

S.J.M. designed the study. G.H. and R.O. constructed and sequenced NucSeq libraries. A.Sh., D.C. and R.O. constructed and sequenced *M. eburnea* ATAC-seq libraries. W.A., K.L. and J.P. assembled genomes. S.J.M., R.W.R. and A.K and annotated genomes. A.Sa. generated the dinucleotide-based position weight matrix used for predicting nucleosome occupancy. S.J.M. analyzed all data. S.J.M. conducted Random Forests (with valuable input from H.S.) and Deep Learning analyses. M.C.A., J.G., J.M., L.S. and

I.V.G. coordinated genome projects. S.J.M. wrote the article, with feedback from I.V.G., M.C.A. and R.O. S.J.M and I.V.G. coordinated the project.

DECLARATION OF INTERESTS

The authors declare no competing interests.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS
- METHOD DETAILS
  - ○ Transcriptome sequencing
  - ○ Genome sequencing
  - ○ *In vitro* chromatin assembly and sequencing (Nuc-seq)
  - ○ ATAC sequencing for Meredithblackwellia eburnea
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - ○ Transcriptome assembly
  - ○ Genome assembly and annotation
  - ○ Detecting and measuring consecutive A/T peaks (CAPs)
  - ○ Nuc-seq/MNase-seq analysis
  - ○ Analysis of ATAC-seq data
  - ○ Analysis of 6mA data from *Ch. reinhardtii*
  - ○ Phylogeny reconstruction
  - ○ Machine learning to extract feature importance and predict nucleosome occupancy
- ADDITIONAL RESOURCES

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2025.112472.

Received: March 4, 2024
Revised: December 26, 2024
Accepted: April 15, 2025
Published: April 18, 2025

REFERENCES

1. Ammar, R., Torti, D., Tsui, K., Gebbia, M., Durbic, T., Bader, G.D., Giaever, G., and Nislow, C. (2012). Chromatin is an ancient innovation conserved between Archaea and Eukarya. elife *1*, e00078.
2. Cutter, A.R., and Hayes, J.J. (2015). A brief review of nucleosome structure. FEBS Lett. *589*, 2914–2922. https://doi.org/10.1016/j.febslet.2015.05.016.
3. McGinty, R.K., and Tan, S. (2015). Nucleosome Structure and Function. Chem. Rev. *115*, 2255–2273. https://doi.org/10.1021/cr500373h.
4. Moyle-Heyrman, G., Zaichuk, T., Xi, L., Zhang, Q., Uhlenbeck, O.C., Holmgren, R., Widom, J., and Wang, J.-P. (2013). Chemical map of Schizosaccharomyces pombe reveals species-specific features in nucleosome positioning. Proc. Natl. Acad. Sci. USA *110*, 20158–20163. https://doi.org/10.1073/pnas.1315809110.
5. Lai, W.K.M., and Pugh, B.F. (2017). Understanding nucleosome dynamics and their links to gene expression and DNA replication. Nat. Rev. Mol. Cell Biol. *18*, 548–562. https://doi.org/10.1038/nrm.2017.47.
6. Chereji, R.V., and Clark, D.J. (2018). Major Determinants of Nucleosome Positioning. Biophys. J. *114*, 2279–2289. https://doi.org/10.1016/j.bpj.2018.03.015.
7. Gangadharan, S., Mularoni, L., Fain-Thornton, J., Wheelan, S.J., and Craig, N.L. (2010). DNA transposon *Hermes* inserts into DNA in nucleosome-free regions in vivo. Proc. Natl. Acad. Sci. USA *107*, 21966–21972. https://doi.org/10.1073/pnas.1016382107.
8. Sultana, T., Zamborlini, A., Cristofari, G., and Lesage, P. (2017). Integration site selection by retroviruses and transposable elements in eukaryotes. Nat. Rev. Genet. *18*, 292–308. https://doi.org/10.1038/nrg.2017.7.
9. Buenrostro, J.D., Giresi, P.G., Zaba, L.C., Chang, H.Y., and Greenleaf, W.J. (2013). Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nat. Methods *10*, 1213–1218. https://doi.org/10.1038/nmeth.2688.
10. Horlbeck, M.A., Witkowsky, L.B., Guglielmi, B., Replogle, J.M., Gilbert, L. A., Villalta, J.E., Torigoe, S.E., Tjian, R., and Weissman, J.S. (2016). Nucleosomes impede Cas9 access to DNA in vivo and in vitro. Elife *5*, e12677. https://doi.org/10.7554/eLife.12677.
11. Parmar, J.J., and Padinhateeri, R. (2020). Nucleosome positioning and chromatin organization. Curr. Opin. Struct. Biol. *64*, 111–118. https://doi.org/10.1016/j.sbi.2020.06.021.
12. Mavrich, T.N., Ioshikhes, I.P., Venters, B.J., Jiang, C., Tomsho, L.P., Qi, J., Schuster, S.C., Albert, I., and Pugh, B.F. (2008). A barrier nucleosome model for statistical positioning of nucleosomes throughout the yeast genome. Genome Res. *18*, 1073–1083. https://doi.org/10.1101/gr.078261.108.
13. Chereji, R.V., Ramachandran, S., Bryson, T.D., and Henikoff, S. (2018). Precise genome-wide mapping of single nucleosomes and linkers in vivo. Genome Biol. *19*, 19. https://doi.org/10.1186/s13059-018-1398-0.
14. Ngo, T.T.M., Yoo, J., Dai, Q., Zhang, Q., He, C., Aksimentiev, A., and Ha, T. (2016). Effects of cytosine modifications on DNA flexibility and nucleosome mechanical stability. Nat. Commun. *7*, 10813. https://doi.org/10.1038/ncomms10813.
15. Breiling, A., and Lyko, F. (2015). Epigenetic regulatory functions of DNA modifications: 5-methylcytosine and beyond. Epigenetics Chromatin *8*, 24. https://doi.org/10.1186/s13072-015-0016-6.
16. Mondo, S.J., Dannebaum, R.O., Kuo, R.C., Louie, K.B., Bewick, A.J., LaButti, K., Haridas, S., Kuo, A., Salamov, A., Ahrendt, S.R., et al. (2017). Widespread adenine N6-methylation of active genes in fungi. Nat. Genet. *49*, 964–968. https://doi.org/10.1038/ng.3859.
17. Fu, Y., Luo, G.-Z., Chen, K., Deng, X., Yu, M., Han, D., Hao, Z., Liu, J., Lu, X., Doré, L.C., et al. (2015). N6-Methyldeoxyadenosine Marks Active Transcription Start Sites in Chlamydomonas. Cell *161*, 879–892. https://doi.org/10.1016/j.cell.2015.04.010.
18. Beh, L.Y., Debelouchina, G.T., Clay, D.M., Thompson, R.E., Lindblad, K. A., Hutton, E.R., Bracht, J.R., Sebra, R.P., Muir, T.W., and Landweber, L.F. (2019). Identification of a DNA N6-Adenine Methyltransferase Complex and Its Impact on Chromatin Organization. Cell *177*, 1781–1796.e25. https://doi.org/10.1016/j.cell.2019.04.028.
19. Lax, C., Mondo, S.J., Osorio-Concepción, M., Muszewska, A., Corrochano-Luque, M., Gutiérrez, G., Riley, R., Lipzen, A., Guo, J., Hundley, H., et al. (2024). Symmetric and asymmetric DNA N6-adenine methylation regulates different biological responses in Mucorales. Nat. Commun. *15*, 6066. https://doi.org/10.1038/s41467-024-50365-2.
20. Wysocka, J., Swigut, T., Xiao, H., Milne, T.A., Kwon, S.Y., Landry, J., Kauer, M., Tackett, A.J., Chait, B.T., Badenhorst, P., et al. (2006). A PHD finger of NURF couples histone H3 lysine 4 trimethylation with chromatin remodelling. Nature *442*, 86–90. https://doi.org/10.1038/nature04815.
21. Pray-Grant, M.G., Daniel, J.A., Schieltz, D., Yates, J.R., and Grant, P.A. (2005). Chd1 chromodomain links histone H3 methylation with SAGA- and SLIK-dependent acetylation. Nature *433*, 434–438. https://doi.org/10.1038/nature03242.
22. Sanulli, S., Trnka, M.J., Dharmarajan, V., Tibble, R.W., Pascal, B.D., Burlingame, A.L., Griffin, P.R., Gross, J.D., and Narlikar, G.J. (2019). HP1 reshapes nucleosome core to promote phase separation of heterochromatin. Nature *575*, 390–394. https://doi.org/10.1038/s41586-019-1669-2.
23. Jin, W., Tang, Q., Wan, M., Cui, K., Zhang, Y., Ren, G., Ni, B., Sklar, J., Przytycka, T.M., Childs, R., et al. (2015). Genome-wide detection of DNase

I hypersensitive sites in single cells and FFPE tissue samples. Nature *528*, 142–146. https://doi.org/10.1038/nature15740.

24. Buenrostro, J.D., Wu, B., Litzenburger, U.M., Ruff, D., Gonzales, M.L., Snyder, M.P., Chang, H.Y., and Greenleaf, W.J. (2015). Single-cell chromatin accessibility reveals principles of regulatory variation. Nature *523*, 486–490. https://doi.org/10.1038/nature14590.

25. Lai, B., Gao, W., Cui, K., Xie, W., Tang, Q., Jin, W., Hu, G., Ni, B., and Zhao, K. (2018). Principles of nucleosome organization revealed by single-cell micrococcal nuclease sequencing. Nature *562*, 281–285. https://doi.org/10.1038/s41586-018-0567-3.

26. Chereji, R.V., and Morozov, A.V. (2015). Functional roles of nucleosome stability and dynamics. Brief. Funct. Genomics *14*, 50–60. https://doi.org/10.1093/bfgp/elu038.

27. Chiu, T.-P., Comoglio, F., Zhou, T., Yang, L., Paro, R., and Rohs, R. (2016). DNAshapeR: an R/Bioconductor package for DNA shape prediction and feature encoding. Bioinformatics *32*, 1211–1213. https://doi.org/10.1093/bioinformatics/btv735.

28. Quang, D., and Xie, X. (2016). DanQ: a hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences. Nucleic Acids Res. *44*, e107. https://doi.org/10.1093/nar/gkw226.

29. Li, J., Sagendorf, J.M., Chiu, T.-P., Pasi, M., Perez, A., and Rohs, R. (2017). Expanding the repertoire of DNA shape features for genome-scale studies of transcription factor binding. Nucleic Acids Res. *45*, 12877–12887. https://doi.org/10.1093/nar/gkx1145.

30. Li, K., Carroll, M., Vafabakhsh, R., Wang, X.A., and Wang, J.-P. (2022). DNAcycP: a deep learning tool for DNA cyclizability prediction. Nucleic Acids Res. *50*, 3142–3154. https://doi.org/10.1093/nar/gkac162.

31. Thåström, A., Loward, P.T., Widlund, H.R., Cao, H., Kubista, M., and Widom, J. (1999). Sequence motifs and free energies of selected natural and non-natural nucleosome positioning DNA sequences. J. Mol. Biol. *288*, 213–229. https://doi.org/10.1006/jmbi.1999.2686.

32. Segal, E., Fondufe-Mittendorf, Y., Chen, L., Tha, A., Wang, J.-P.Z., and Widom, J. (2006). A genomic code for nucleosome positioning. Nature *442*, 772–778.

33. Kaplan, N., Moore, I.K., Fondufe-Mittendorf, Y., Gossett, A.J., Tillo, D., Field, Y., LeProust, E.M., Hughes, T.R., Lieb, J.D., Widom, J., and Segal, E. (2009). The DNA-encoded nucleosome organization of a eukaryotic genome. Nature *458*, 362–366. https://doi.org/10.1038/nature07667.

34. Satchwell, S.C., Drew, H.R., and Travers, A.A. (1986). Sequence periodicities in chicken nucleosome core DNA. J. Mol. Biol. *191*, 659–675. https://doi.org/10.1016/0022-2836(86)90452-3.

35. Liu, H., Lin, S., Cai, Z., and Sun, X. (2011). Role of 10–11bp periodicities of eukaryotic DNA sequence in nucleosome positioning. Biosystems *105*, 295–299. https://doi.org/10.1016/j.biosystems.2011.05.016.

36. Bettecken, T., and Trifonov, E.N. (2009). Repertoires of the Nucleosome-Positioning Dinucleotides. PLoS One *4*, e7654. https://doi.org/10.1371/journal.pone.0007654.

37. Tillo, D., and Hughes, T.R. (2009). G+C content dominates intrinsic nucleosome occupancy. BMC Bioinf. *10*, 442. https://doi.org/10.1186/1471-2105-10-442.

38. Ngo, T.T.M., Zhang, Q., Zhou, R., Yodh, J.G., and Ha, T. (2015). Asymmetric Unwrapping of Nucleosomes under Tension Directed by DNA Local Flexibility. Cell *160*, 1135–1144. https://doi.org/10.1016/j.cell.2015.02.001.

39. Drew, H.R., and Travers, A.A. (1985). DNA bending and its relation to nucleosome positioning. J. Mol. Biol. *186*, 773–790. https://doi.org/10.1016/0022-2836(85)90396-1.

40. Grishkevich, V., Hashimshony, T., and Yanai, I. (2011). Core promoter T-blocks correlate with gene expression levels in *C. elegans*. Genome Res. *21*, 707–717. https://doi.org/10.1101/gr.113381.110.

41. Parmar, J.J., Marko, J.F., and Padinhateeri, R. (2014). Nucleosome positioning and kinetics near transcription-start-site barriers are controlled by interplay between active remodeling and DNA sequence. Nucleic Acids Res. *42*, 128–136. https://doi.org/10.1093/nar/gkt854.

42. Hatakeyama, A., Hartmann, B., Travers, A., Nogues, C., and Buckle, M. (2016). High-resolution biophysical analysis of the dynamics of nucleosome formation. Sci. Rep. *6*, 27337. https://doi.org/10.1038/srep27337.

43. Scipioni, A., and De Santis, P. (2011). Predicting nucleosome positioning in genomes: Physical and bioinformatic approaches. Biophys. Chem. *155*, 53–64. https://doi.org/10.1016/j.bpc.2011.03.006.

44. Kornberg, R.D., and Stryer, L. (1988). Statistical distributions of nucleosomes: nonrandom locations by a stochastic mechanism. Nucl. Acids Res. *16*, 6677–6690. https://doi.org/10.1093/nar/16.14.6677.

45. Milani, P., Chevereau, G., Vaillant, C., Audit, B., Haftek-Terreau, Z., Marilley, M., Bouvet, P., Argoul, F., and Arneodo, A. (2009). Nucleosome positioning by genomic excluding-energy barriers. Proc. Natl. Acad. Sci. USA *106*, 22257–22262. https://doi.org/10.1073/pnas.0909511106.

46. Valouev, A., Ichikawa, J., Tonthat, T., Stuart, J., Ranade, S., Peckham, H., Zeng, K., Malek, J.A., Costa, G., McKernan, K., et al. (2008). A high-resolution, nucleosome position map of *C. elegans* reveals a lack of universal sequence-dictated positioning. Genome Res. *18*, 1051–1063. https://doi.org/10.1101/gr.076463.108.

47. Valouev, A., Johnson, S.M., Boyd, S.D., Smith, C.L., Fire, A.Z., and Sidow, A. (2011). Determinants of nucleosome organization in primary human cells. Nature *474*, 516–520.

48. Zhang, Y., Moqtaderi, Z., Rattner, B.P., Euskirchen, G., Snyder, M., Kadonaga, J.T., Liu, X.S., and Struhl, K. (2009). Intrinsic histone-DNA interactions are not the major determinant of nucleosome positions in vivo. Nat. Struct. Mol. Biol. *16*, 847–852. https://doi.org/10.1038/nsmb.1636.

49. Stein, A., Takasuka, T.E., and Collings, C.K. (2010). Are nucleosome positions in vivo primarily determined by histone–DNA sequence preferences? Nucleic Acids Res. *38*, 709–719. https://doi.org/10.1093/nar/gkp1043.

50. Coradetti, S.T., Pinel, D., Geiselman, G.M., Ito, M., Mondo, S.J., Reilly, M. C., Cheng, Y.-F., Bauer, S., Grigoriev, I.V., Gladden, J.M., et al. (2018). Functional genomics of lipid metabolism in the oleaginous yeast Rhodosporidium toruloides. Elife *7*, e32110. https://doi.org/10.7554/eLife.32110.

51. Grigoriev, I.V., Nikitin, R., Haridas, S., Kuo, A., Ohm, R., Otillar, R., Riley, R., Salamov, A., Zhao, X., Korzeniewski, F., et al. (2014). MycoCosm portal: gearing up for 1000 fungal genomes. Nucl. Acids Res. *42*, D699–D704. https://doi.org/10.1093/nar/gkt1183.

52. Grigoriev, I.V., Hayes, R.D., Calhoun, S., Kamel, B., Wang, A., Ahrendt, S., Dusheyko, S., Nikitin, R., Mondo, S.J., Salamov, A., et al. (2021). PhycoCosm, a comparative algal genomics resource. Nucleic Acids Res. *49*, D1004–D1011. https://doi.org/10.1093/nar/gkaa898.

53. Yang, C., Ma, L., Xiao, D., Ying, Z., Jiang, X., and Lin, Y. (2019). Integration of ATAC-Seq and RNA-Seq Identifies Key Genes in Light-Induced Primordia Formation of Sparassis latifolia. Int. J. Mol. Sci. *21*, 185. https://doi.org/10.3390/ijms21010185.

54. Lin, J., Zhao, Y., Ferraro, A.R., Yang, E., Lewis, Z.A., and Lin, X. (2019). Transcription factor Znf2 coordinates with the chromatin remodeling SWI/SNF complex to regulate cryptococcal cellular differentiation. Commun. Biol. *2*, 412. https://doi.org/10.1038/s42003-019-0665-2.

55. Nardelli, S.C., Silmon de Monerri, N.C., Vanagas, L., Wang, X., Tampaki, Z., Sullivan, W.J., Angel, S.O., and Kim, K. (2022). Genome-wide localization of histone variants in Toxoplasma gondii implicates variant exchange in stage-specific gene expression. BMC Genom. *23*, 128. https://doi.org/10.1186/s12864-022-08338-6.

56. Cui, K., and Zhao, K. (2012). Genome-Wide Approaches to Determining Nucleosome Occupancy in Metazoans Using MNase-Seq. In Chromatin Remodeling Methods in Molecular Biology, R.H. Morse, ed. (Humana Press), pp. 413–419. https://doi.org/10.1007/978-1-61779-477-3_24.

57. Shen, X.-X., Opulente, D.A., Kominek, J., Zhou, X., Steenwyk, J.L., Buh, K. V., Haase, M.A.B., Wisecaver, J.H., Wang, M., Doering, D.T., et al. (2019).

Tempo and Mode of Genome Evolution in the Budding Yeast Subphylum, *46*.

58. Boissinot, S. (2022). On the Base Composition of Transposable Elements. Int. J. Mol. Sci. *23*, 4755. https://doi.org/10.3390/ijms23094755.

59. Gnerre, S., MacCallum, I., Przybylski, D., Ribeiro, F.J., Burton, J.N., Walker, B.J., Sharpe, T., Hall, G., Shea, T.P., Sykes, S., et al. (2011). High-quality draft assemblies of mammalian genomes from massively parallel sequence data. Proc. Natl. Acad. Sci. USA *108*, 1513–1518. https://doi.org/10.1073/pnas.1017351108.

60. Lam, K.-K., LaButti, K., Khalak, A., and Tse, D. (2015). FinisherSC: a repeat-aware tool for upgrading *de novo* assembly using long reads. Bioinformatics *31*, 3207–3209. https://doi.org/10.1093/bioinformatics/btv280.

61. Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat. Methods *9*, 357–359. https://doi.org/10.1038/nmeth.1923.

62. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R.; 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079. https://doi.org/10.1093/bioinformatics/btp352.

63. (2019). Picard Toolkit. Broad Institute, GitHub Repository (Broad Institute). https://broadinstitute.github.io/picard/;.

64. Chen, K., Xi, Y., Pan, X., Li, Z., Kaestner, K., Tyler, J., Dent, S., He, X., and Li, W. (2013). DANPOS: Dynamic analysis of nucleosome position and occupancy by sequencing. Genome Res. *23*, 341–351. https://doi.org/10.1101/gr.142067.112.

65. Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B. E., Nusbaum, C., Myers, R.M., Brown, M., Li, W., and Liu, X.S. (2008). Model-based Analysis of ChIP-Seq (MACS). Genome Biol. *9*, R137. https://doi.org/10.1186/gb-2008-9-9-r137.

66. Steinegger, M., and Söding, J. (2017). MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nat. Biotechnol. *35*, 1026–1028. https://doi.org/10.1038/nbt.3988.

67. Edgar, R.C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. BMC Bioinf. *5*, 113. https://doi.org/10.1186/1471-2105-5-113.

68. Castresana, J. (2000). Selection of Conserved Blocks from Multiple Alignments for Their Use in Phylogenetic Analysis. Mol. Biol. Evol. *17*, 540–552. https://doi.org/10.1093/oxfordjournals.molbev.a026334.

69. Price, M.N., Dehal, P.S., and Arkin, A.P. (2010). FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. PLoS One *5*, e9490. https://doi.org/10.1371/journal.pone.0009490.

70. Huerta-Cepas, J., Serra, F., and Bork, P. (2016). ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. Mol. Biol. Evol. *33*, 1635–1638. https://doi.org/10.1093/molbev/msw046.

71. Martin, J., Bruno, V.M., Fang, Z., Meng, X., Blow, M., Zhang, T., Sherlock, G., Snyder, M., and Wang, Z. (2010). Rnnotator: an automated de novo transcriptome assembly pipeline from stranded RNA-Seq reads. BMC Genom. *11*, 663. https://doi.org/10.1186/1471-2164-11-663.

72. Zerbino, D.R., and Birney, E. (2008). Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. *18*, 821–829. https://doi.org/10.1101/gr.074492.107.

73. Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., et al. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat. Biotechnol. *29*, 644–652. https://doi.org/10.1038/nbt.1883.

74. Smit, A.F.A., Hubley, R., and Green, P. (1996-2010). *RepeatMasker Open-3.0*.. http://www.repeatmasker.org.

75. Price, A.L., Jones, N.C., and Pevzner, P.A. (2005). De novo identification of repeat families in large genomes. Bioinformatics *21* (*Suppl 1*), i351–i358. https://doi.org/10.1093/bioinformatics/bti1018.

76. Yan, F., Powell, D.R., Curtis, D.J., and Wong, N.C. (2020). From reads to insight: a hitchhiker's guide to ATAC-seq data analysis. Genome Biol. *21*, 22. https://doi.org/10.1186/s13059-020-1929-3.

77. Mondo, S.J., Kuo, R.C., and Singan, V.R. (2018). Fungal Epigenomics: Detection and Analysis. In Fungal Genomics: Methods and Protocols Methods in Molecular Biology, R.P. de Vries, A. Tsang, and I.V. Grigoriev, eds. (Springer), pp. 155–170. https://doi.org/10.1007/978-1-4939-7804-5_14.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Critical commercial assays		
Chromatin Assembly Kit	Active Motif	Cat#53500
Illumina Tagment DNA Enzyme and Buffer	Illumina	Cat#20034197
Deposited data		
Meredithblackwellia eburnea MCA 4105 genome	This study	GenBank: JBAJAD000000000
Rhodosporidium toruloides ATCC 26217 genome	This study	GenBank: JBBAIS000000000
Mrakia frigida ATCC 22029 genome	This study	GenBank: JBAJAC000000000
Usnea florida ATCC 18376 genome	This study	GenBank: JBBAIT000000000
Sparassis latifolia CCMJ1100 genome	This study	GenBank: JBAJAE000000000
Mrakia frigida NucSeq	This study	NCBI SRA: SRR7472065
Usnea florida ATCC 18376 Nuc-seq	This study	NCBI SRA: SRR7472154
Catenaria anguillulae Nuc-seq	This study	NCBI SRA: SRR7476903
Rhodosporidium toruloides ATCC26217 Nuc-seq	This study	NCBI SRA: SRR28069068
Meredithblackwella eburnea MCA 4105 Nuc-seq	This study	NCBI SRA: SRR28069066
Chlamydomonas reinhardtii Nuc-seq	This study	NCBI SRA: SRR28069067
Meredithblackwella eburnea MCA 4105 heavy grind ATAC-seq	This study	NCBI SRA: SRR28069093
Meredithblackwella eburnea MCA 4105 heavy grind ATAC-seq with 5 micron filter	This study	NCBI SRA: SRR28069094
Meredithblackwella eburnea MCA 4105 light grind ATAC-seq	This study	NCBI SRA: SRR28069106
Toxoplasma gondii MNase-seq	Nardelli et al.[55]	NCBI SRA: SRR4216894
Chlamydomonas reinhardtii MNase-seq	Fu et al.[17]	NCBI SRA: SRR1994849 and SRR1994850
Saccharomyces cerevisiae in vitro nucleosomes	Kaplan et al.[33]	NCBI SRA: SRR023798 and SRR023799
Chlamydomonas reinhardtii 6mA-IP-seq	Fu et al.[17]	NCBI SRA: SRR1994831, SRR1994832 and SRR1994833
Sparassis latifolia ATAC-seq	Yang et al.[53]	NCBI SRA: SRR8924630, SRR8924631, SRR8924633, SRR8924632
Cryptococcus neoformans ATAC-seq	Lin et al.[54]	NCBI SRA: SRR10097538, SRR10097539, SRR10097540, SRR10097541
Experimental models: Organisms/strains		
Meredithblackwellia eburnea MCA4105	NRRL	NRRL Y-48821 = ATCC MYA-4884 = CBS 12589 = MCA4105
Rhodosporidium toruloides ATCC 26217	ATCC	ATCC 26217
Mrakia frigida ATCC 22029	ATCC	ATCC 22029
Usnea florida ATCC 18376	ATCC	ATCC 18376
Catenaria anguillulae PL171	NA	PL171
Sparassis latifolia CCMJ1100	Jilin Agricultural University	CCMJ1100
Chlamydomonas reinhardtii CC-503	Chlamydomonas Resource Center	CC-503

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Software and algorithms		
AllPathsLG (versions R47710 and R49403)	Gnerre et al.[59]	https://www.broadinstitute.org/computational-rd/computational-research-and-development
Falcon version 0.4.2	PacBio	https://github.com/PacificBiosciences/FALCON
FinisherSC version 2.0	Lam et al.[60]	http://kakitone.github.io/finishingTool/
Quiver version smrtanalysis_2.3.0.140936.p5	PacBio	https://github.com/PacificBiosciences/GenomicConsensus
BBtools version 38.69	JGI	https://sourceforge.net/projects/bbmap/
Bowtie1 version 1.2.2	Langmead et al.[61]	https://github.com/BenLangmead/bowtie
Bowtie2 version 2.2.5	Langmead et al.[61]	https://github.com/BenLangmead/bowtie
Samtools version 1.6	Li et al.[62]	http://www.htslib.org/
Picard version 2.27.4	Broadinstitute/picard et al.[63]	https://github.com/broadinstitute/picard
Danpos v2.2.2	Chen et al.[64]	https://sites.google.com/site/danposdoc/
DNAshapeR v1.34.0	Chiu et al.[27]	https://www.bioconductor.org/packages/release/bioc/html/DNAshapeR.html
DNAcycP version 0.0.1dev1	Li et al.[30]	https://github.com/jipingw/DNAcycP
Macs2 version 2.1.4	Zhang et al.[65]	https://github.com/macs3-project/MACS/wiki/Install-macs2
mmseqs2 release 14-7e284	Steinegger et al.[66]	https://github.com/soedinglab/MMseqs2
Muscle version 3.8.1551	Edgar et al.[67]	https://github.com/rcedgar/muscle
Gblocks version 0.91b	Castresana et al.[68]	https://www.biologiaevolutiva.org/jcastresana/Gblocks.html
FastTree version 2.1.10 SSE3	Price et al.[69]	https://morgannprice.github.io/fasttree/
ete3 toolkit	Huerta-Cepas et al.[70]	https://etetoolkit.org/
NuclPred v1.0	This study	https://github.com/sjmondo/NuclPred_v1

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

Alongside comparative analysis across published eukaryotic genomes (see Data S1 for lineage information), sequencing and experimental work was conducted using DNA extracted from six lineages as part of this study, *Meredithblackwellia eburnea* MCA4105, *Rhodosporidium toruloides* ATCC 26217, *Mrakia frigida* ATCC 22029, *Usnea florida* ATCC 18376, *Catenaria anguillulae* PL171, and *Chlamydomonas reinhardtii* CC-503, including Nuc-Seq and ATAC-seq (*M. eburnea*). See method details for more information.

METHOD DETAILS

Transcriptome sequencing

For all lineages sequenced as part of this study (*M. eburnea*, *R. toruloides* ATCC26217, *Mr. frigida*, *Sp. latifolia* and *U. florida*), stranded cDNA libraries were generated using the Illumina Truseq Stranded mRNA Library Prep kit. mRNA was purified from 1 µg (899ng from *U. florida*) of total RNA using magnetic beads containing poly-T oligos. mRNA was fragmented and reversed transcribed using random hexamers and SSII (Invitrogen) followed by second strand synthesis. The fragmented cDNA was treated with end-pair, A-tailing, adapter ligation, and 8 cycles of PCR. The prepared libraries were then quantified using KAPA Biosystem's next-generation sequencing library qPCR kit (Roche) and run on a Roche LightCycler 480 real-time PCR instrument. The quantified library was then multiplexed with other libraries, and the pool of libraries was then prepared for sequencing on the Illumina HiSeq sequencing platform utilizing a TruSeq paired-end cluster kit, v4, and Illumina's cBot instrument to generate a clustered flow cell for sequencing. Sequencing of the flow cell was performed on the Illumina HiSeq 2500 sequencer using HiSeq TruSeq SBS sequencing kits, v4, following a 2 × 150 indexed run recipe.

Genome sequencing

The genomes of *M. eburnea*, *R. toruloides* ATCC26217, *U. florida* and *Mr. frigida* were sequenced using the Illumina technology. These included both short read (2 × 150bp for *M. eburnea*, 2 × 151bp for others) and long mate pair libraries for all lineages except

R. toruloides ATCC26217 (2 × 100bp, 5.5kb for *M. eburnea*, 2 × 151bp 5.3 kb and 6.1kb for *Mr. frigida* and *U. florida*, respectively). For short (300bp) read libraries, 100 ng (98.8 ng for *U. florida*) of DNA was sheared to 300 bp using the Covaris LE220 and size selected using SPRI beads (Beckman Coulter). The fragments were treated with end-repair, A-tailing, and ligation of Illumina compatible adapters (IDT, Inc) using the KAPA-Illumina library creation kit (KAPA biosystems). For long mate pair libraries, 5 μg of DNA was sheared using the Covaris g-TUBE (Covaris) and gel size selected for 4kb. The sheared DNA was treated with end repair and ligated with biotinylated adapters containing *loxP*. The adapter ligated DNA fragments were circularized via recombination by a Cre excision reaction (NEB). The circularized DNA templates were then randomly sheared using the Covaris LE220 (Covaris). The sheared fragments were treated with end repair and A-tailing using the KAPA-Illumina library creation kit (KAPA biosystems) followed by immobilization of mate pair fragments on strepavidin beads (Invitrogen). Illumina compatible adapters (IDT, Inc) were ligated to the mate pair fragments and 8 cycles of PCR (12 for *M. eburnea*) was used to enrich for the final library (KAPA Biosystems). All prepared libraries were then quantified using KAPA Biosystem's next-generation sequencing library qPCR kit (Roche) and run on a Roche LightCycler 480 real-time PCR instrument. The quantified library was then multiplexed with other libraries, and the pool of libraries was then prepared for sequencing on the Illumina HiSeq sequencing platform utilizing a TruSeq paired-end cluster kit, either v3 or v4, and Illumina's cBot instrument to generate a clustered flow cell for sequencing. Sequencing of the flow cell was performed on the Illumina HiSeq 2500 sequencer using HiSeq TruSeq SBS sequencing kits, either v3 (short read libraries for *Mr. frigida*, *U. florida* and *R. toruloides* ATCC26217) or v4, following a 2 × 150 indexed run recipe.

For *Sp. latifolia*, 5 μg of genomic DNA was sheared to >10kb using Covaris g-Tubes. The sheared DNA was treated with exonuclease to remove single-stranded ends and DNA damage repair mix followed by end repair and ligation of blunt adapters using SMRTbell Template Prep Kit 1.0 (Pacific Biosciences). The library was purified with AMPure PB beads. PacBio Sequencing primer was then annealed to the SMRTbell template library and Version P6 sequencing polymerase was bound to them. The prepared SMRTbell template libraries were then sequenced on a Pacific Biosciences RSII sequencer using Version C4 chemistry and 1 × 240 sequencing movie run times.

In vitro chromatin assembly and sequencing (Nuc-seq)

In vitro nucleosome sequencing, conducted at the DOE Joint Genome Institute (JGI), was generated using the Illumina technology. For all samples (*M. eburnea*, *R. toruloides* ATCC26217, *Mr. frigida*, *U. florida*, *Ca. anguillulae* and *Ch. reinhardtii*), 1 μg of genomic DNA was chromatin-assembled using Chromatin Assembly Kit (Active Motif). Samples were partially digested with 0.5 μL of Enzymatic Shearing Cocktail from the kit for 2 min and deproteinated with proteinase K. The fragments were purified by phenol/chloroform and analyzed by BioAnalyzer. 100 ng of fragments were treated with end-repair, A-tailing, and ligation of Illumina compatible adapters (IDT, Inc) using the KAPA HyperPrep library creation kit (KAPA biosystems). The prepared libraries were then quantified using KAPA Biosystem's next-generation sequencing library qPCR kit (Roche) and run on a Roche LightCycler 480 real-time PCR instrument. The quantified library was then multiplexed with other libraries, and pool was then loaded and sequenced on the Illumina NextSeq 500 (except *Ch. reinhardtii*) sequencing platform utilizing a NextSeq Mid-Output Reagent Kit, v2 300 cycle, following a 2 × 150 indexed run recipe. For *Ch. reinhardtii*, the sample was loaded and sequenced on the Illumina MiSeq sequencing platform utilizing a MiSeq Reagent Kit, v2 300 cycle, following a 2 × 150 indexed run recipe.

ATAC sequencing for Meredithblackwellia eburnea

Sequence data for *M. eburnea* ATAC-seq was generated at the DOE Joint Genome Institute (JGI) using Illumina technology. For all libraries, cells were grown in potato dextrose broth and approximately 5,000 fresh cells were extracted, cryoground in liquid nitrogen and filtered through 35 micron Cell Strainer (Corning). The flow through was washed three times with the cold extraction buffer with 0.25% Triton X-100 and the pellet was resuspended in Tagment DNA Buffer from Nextera DNA Library Prep kit (Illumina). Tagmentation Enzyme was then added and incubated at 37C for 30 min. The tagmented materials were directly amplified using Nextera PCR mix and indexed primers at 12 cycles. The amplified products were purified with 1.4x SPRI beads (Omega Bio-Tek, TotalPure NGS beads). The prepared libraries were then quantified using KAPA Biosystem's next-generation sequencing library qPCR kit (Roche) and run on a Roche LightCycler 480 real-time PCR instrument. The quantified library was then then multiplexed with other libraries, and the pool of libraries was then prepared for sequencing on the Illumina NovaSeq 6000 sequencing platform using NovaSeq XP v1 reagent kits, S4 flow cell, following a 2 × 150 indexed run recipe.

QUANTIFICATION AND STATISTICAL ANALYSIS

Transcriptome assembly

Raw RNAseq fastq file reads were filtered and trimmed using the JGI QC pipeline. Briefly, using BBDuk, raw reads were evaluated for artifact sequence by kmer matching (kmer = 25), allowing 1 mismatch and detected artifact was trimmed from the 3′ end of the reads. RNA spike-in reads, PhiX reads and reads containing any Ns were removed. Quality trimming was performed using the phred trimming method set at Q6. Finally, following trimming, reads under the length threshold were removed (minimum length 25 bases or 1/3 of the original read length - whichever is longer). Filtered fastq files were used as input for *de novo* assembly of RNA contigs. For *M. eburnea*, filtered reads were assembled into consensus sequences using Rnnotator (v. 3.4.0).[71] Rnnotator was used for assembly and post-processing of contigs. Assembly was completed through eight runs of velveth (v. 1.2.07),[72] performed in parallel, once for each hash length for the De Bruijn graph. Minimum contig length was set at 100. The read depth minimum was set to 3 reads. Redundant contigs were removed using Vmatch (v. 2.2.0) and contigs with significant overlap were further assembled using Minimus2 with a minimum overlap of 40. Contig postprocessing included splitting misassembled contigs, contig extension and polishing using the strand information of the reads. Single base errors were corrected by aligning the reads back to each contig with BWA to generate a consensus nucleotide sequence. Post-processed contigs were clustered into loci and putative transcript precursors were identified. In creation of the contigs, all the reads were used not just those which uniquely mapped. This enabled the generation of isoforms as the reads forked off each other to create v1, v2, v3, etc. Individual reads were then aligned uniquely to all the isoforms and "v1″ was always the most highly expressed transcript. For *R.* toruloides ATCC26217, *U. florida*, *Mr. frigida* and *Sp. latifolia*, reads were assembled into consensus sequences using Trinity (ver. 2.1.1).[73] Trinity was run with the –normalize_reads (in-silico normalization routine) and –jaccard_clip (minimizing fusion transcripts derived from gene dense genomes) options.

Genome assembly and annotation

Illumina-sequenced genomes were assembled using AllPathsLG[59] (version R47710 for *M. eburnea*, version R49403 for others). For *R. toruloides* ATCC26217, each short read fastq file was QC filtered for artifact/process contamination and subsequently assembled together with Velvet. The resulting assembly was used to create a long mate-pair library with insert 3000 ± 300 bp which was then assembled with the original Illumina library with AllPathsLG. For *Sp. latifolia* (sequenced on the PacBio RSII platform), filtered subread data were assembled with Falcon version 0.4.2 to generate an initial assembly. An automated attempt was made to reassemble any potential organelle (mitochondrion) from the Falcon pre-assembled reads and used to filter organelle out of the preads with an in-house tool (assemblemito.sh). A secondary Falcon assembly was generated using the filtered preads with Falcon version 0.4.2, improved with finisherSC version 2.0[60] and polished with Quiver version smrtanalysis_2.3.0.140936.p5. Contigs less than 1000 bp were excluded from the final assembly. All assemblies were then annotated using the JGI fungal genome annotation pipeline.[51]

Detecting and measuring consecutive A/T peaks (CAPs)

In addition to the five lineages sequenced, assembled, and annotated as part of this study, we also included 1112 published genomes available through the MycoCosm[51] and PhycoCosm[52] web portals. See Data S1 for a list of all genomes included and associated publications. In these lineages, we first calculated A/T frequency per-site (averaged across a ±25bp sliding window). A/T content was then smoothed using a Butterworth digital filter (as implemented in Scipy with parameters $N = 2$, Wn = 0.2, analog = False) and peaks in A/T content were extracted using the Scipy signal package find_peaks, retaining only those were peak prominence (vertical distance between the peak summit and its lowest point) was ≥ 2 standard deviations from the mean A/T content per scaffold. This distance-independent method was selected to provide an estimate of typical spacing between prominent peaks across eukaryotes. Genomic Consecutive A/T Peaks (gCAPs) were identified by summing the number of peaks detected within 100-200bp of each other. If distance between peaks was not within this range, we treated the subsequent peak as independent of the preceding peak/gCAP region.

To calculate feature Consecutive A/T Peak (fCAP) scores, for each genome we first calculated average G/C per-site ±1500bp surrounding start and end positions for 7 features: 5′ untranslated regions (UTRs), 3′ UTRs, exons, introns, coding sequences (CDS), and known repeats and *de novo* predicted repeats. For example, there are 11135 genes in *M. eburnea* and consequently 11135 coding sequence start sites, so average G/C at any given position is sum(G or C)/11135. To explore the relationship between GC curves and intron locations (Figure S2A), intronic sequence was assigned a value of 1 (0 for non-intron sequence), then the frequency of introns was calculated the same way as above for ±1500bp surrounding each feature. Repeats were identified using RepeatMasker,[74] which identifies known repeats, and RepeatScout,[75] which identifies *de novo* repeats. RepeatMasker and RepeatScout data were analyzed separately.

Following calculation of per-site GC frequencies, a Butterworth digital filter was applied for data smoothing. Smoothed data were then scanned for peaks using the Scipy signal package find_peaks function, with parameters prominence = 0 (minimum prominence) and wlen = 300. For each feature, fCAP scores were then calculated as the sum of all identified peak prominences within 100-200bp of each other inside the ±1500bp window surrounding features.

For exploring the impact of 5′ UTR annotation quality on detection of fCAPs (Figure S3), 5′ UTR start site coordinates were randomly shifted within an 80bp window surrounding the actual site (using the python3 random.randint function) for the 100 lineages with the highest TSS gCAP scores, then scores were calculated using the same methods described above. Median periodicity (spacing) between peaks (Figure S2D) was analyzed for the 200 lineages within the top fCAP score at each feature. To analyze the relationship between CAPs and various genomic features, we split genomes with UTR annotations ($n = 554$) into quartiles based on fCAP and gCAP (% 2+ consecutive peaks) scores, then analyzed quartiles with respect to various features extracted from MycoCosm and PhycoCosm, including: assembly length, repeat coverage length, percent of assembly comprised of repeats, % GC, # genes, gene density, mean/median exon, gene, intron, transcript and protein lengths and mean/median # exons and introns per gene. Differences between quartiles were assessed using the independent T-test with Bonferroni correction for multiple comparisons (Figure S4).

Nuc-seq/MNase-seq analysis

For Nuc-seq data, BBDuk (version 37.90) was used to remove contaminants, trim reads that contained adapter sequence and right quality trim reads where quality drops to 0. BBDuk was also used to remove reads that contained 1 or more "N" bases, had an average quality score across the read less than 13 or had a minimum length $\leq$ 41 bp or 33% of the full read length. Reads mapped with BBMap to microbial contaminants, masked human, cat, dog and mouse references at 93% identity were also removed. Reads were then aligned to the genome assembly using bowtie1[61] with parameters -S -t -m 1 -v 2. Results were sorted using samtools,[62] then mapped reads were deduplicated using the Picard MarkDuplicates tool,[63] with parameter REMOVE_DUPLICATES = True. Subsequently, danpos v2.2.2[64] using parameters -m 1 -a 1 was used to identify nucleosome bound regions, nucleosome summits/dyads and calculate nucleosome occupancy values for each position in the assembly. Occupancy values calculated by Danpos were also used in subsequent machine learning analyses.

For publicly available *in vitro* and *in vivo* MNase-seq datasets, reads were downloaded from the SRA database. These included MNase seq runs: SRR4216894 for *Toxoplasma gondii*,[55] and both SRR1994849 and SRR1994850 for *Chlamydomonas reinhardtii*.[17] *In vitro* chromatin assembly data for *Sa. cerevisiae* are from SRR023798 and SRR023799.[33] Data were collected using the sratoolkit.2.9.2 fastq-dump tool, then and processed/analyzed in the same way as above.

To calculate dinucleotide and GC frequencies across nucleosome bound DNAs (Figures 2B, 2C, and S7), summit points identified by danpos were collected, then underlying DNA sequence ±75 surrounding summits were used to calculate the average frequency of G/C and all 16 dinucleotides at each position across nucleosome-bound DNA. For *Sc. pombe*, the same procedure was followed, but we instead used "premium" nucleosome dyad positions published in Moyle-Heyrman et al., 2013.[4] In addition to dinucleotides, DNAshapeR[27] was used to calculate 5 different DNA shape parameters per-site, including: propeller twist (ProT), helix twist (HelT), minor groove width (MGW), Roll, and electrostatic potential (EP) per-site. DNA cyclizability was also calculated using DNAcycP[30] with default parameters. Like dinucleotides, these additional measurements were then summarized ±75 surrounding nucleosome dyad positions. Per-site G/C frequencies across nucleosome-bound DNAs were normalized against the average genome-wide GC. Nucleosome occupancy surrounding genomic features was visualized using the danpos profile tool, targeting the following features using the –genomic_sites parameter: TSS,TTS,CSS,CTS,ESS,ETS.

As danpos was designed for analysis of MNase-seq data and here we are analyzing multiple different data types (Nuc-seq, MNase-seq, ATAC-seq, 6mA, and gCAPs), to ensure consistency across methods, for analysis of distance between CAPs and experimental data all peaks were identified using the Scipy find_peak function with minimum peak distance set to 140bp (except for "premium" nucleosome positions previously determined by Moyle-Heyrman et al., 2013[4]). A distance-based method was used here instead of the prominence-based one for identifying gCAPs because we found that while effective at capturing high confidence locations, we were missing peak locations that overlap *in vivo* nucleosomes using a prominence-based approach, which manifests as increased observations of A/T peaks and *in vivo* nucleosomes being separated by ~300 and ~450 bp (multiples of 150bp). This is not surprising, as we could anticipate that even if no sites with prominence $\geq$2 exist, nucleosomes will naturally form at the most favorable location within a given region. Importantly, direct comparison of scipy find_peak and danpos peaks revealed that >80% of these perfectly overlap between methods. Distance between Nuc-Seq/MNase-seq peaks and A/T peaks was calculated by identifying the closest peak summit locations between datasets, then calculating the distance between them. To generate a random distribution of NucSeq, MNase and ATAC-seq peaks, we took all peak coordinates on each scaffold and randomized their positions using the python random.sample function, which randomly selects scaffold coordinates without replacement, then compared those positions to the closest A/T peaks. To assess how nucleosome "fuzziness" changed as a function of distance to A/T peaks (Figure S5), for all fungi with available NucSeq data, we calculated fuzziness using danpos, then compared the distance between danpos-determined summit positions and closest A/T peaks using the same approach as described above.

Analysis of ATAC-seq data

For M. eburnea ATAC-seq data generated at the JGI, BBDuk (version 38.69) was used to remove contaminants, trim reads that contained adapter sequence and right quality trim reads where quality drops below 6. BBDuk was used to remove reads that contained 2 or more "N" bases, had an average quality score across the read less than 10 or had a minimum length $\leq$ 35 bp or 20% of the full read length. Reads mapped with BBMap to masked microbial contaminants, human, cat, dog and mouse references at 93% identity were additionally removed. For *Cr. neoformans* and *Sp. latifolia*, reads were downloaded from SRA. For *Cr. neoformans*,[54] this included SRA IDs: SRR10097538 (YPD replicate 1), SRR10097539 (YPD replicate 2), SRR10097540 (V8 replicate 1), SRR10097541 (V8 replicate 2). For *Sp. latifolia*,[53] this included SRA IDs: SRR8924630 (dark condition replicate 1), SRR8924631 (dark condition replicate 2), SRR8924633 (light condition replicate 1) and SRR8924632 (light condition rep 2).

Reads were aligned to the genome assembly using bowtie2[61] with parameters –very-sensitive -k 1 and coordinates were adjusted as described in.[76] Mapped reads were then sorted using samtools (*34*) and deduplicated using Picard[63] MarkDuplicates tool, with parameter REMOVE_DUPLICATES = True. Accessible Linker DNAs were then extracted for further analysis by selecting mapped reads with insert size $\leq$100bp. Following read mapping and size selection, coverage data were converted to wig format using bedtools and smoothed using a Butterworth digital filter. Distance between nucleosome-free regions and A/T peaks were calculated as described in the "Nuc-seq/MNase-seq analysis" STAR Methods section. Danpos v2.2.2 profile was then used to visualize ATAC-seq coverage surrounding genomic features. For Figure 3 and Table S2, although all replicates showed similar patterns (Figure S8), ATAC-seq data from *Sp. latifolia* light condition replicate 2, *Cr. neoformans* V8 replicate 1 and *M. eburnea* "light grind" were used as representatives, since resolution was higher in these libraries compared to others. However, quality was low overall across *Cr. neoformans* datasets (Figure S9) and therefore it was excluded from analyses surrounding gene features.

Analysis of 6mA data from *Ch. reinhardtii*

For exploration of the relationship between epigenomic modifications and A/T peaks, we leveraged previously published 6mA IP-seq data from *Ch. reinhardtii*.[17] This included data from 2 conditions: 6mA deposition in *Ch. reinhardtii* grown under constant light (SRR1994831 and SRR1994832), which matches the growth condition used in the same study for *in vivo* nucleosome mapping via MNase-seq, as well as growth in constant darkness (SRR1994833). Reads from both replicates in light conditions were merged prior to analysis. Following Mondo et al., 2018,[77] reads were then mapped to the *Ch. reinhardtii* v5.6 assembly using bwa, sorted using samtools, then 6mA peaks were identified using macs2[65] callpeak function, including the -q 0.01 and –nomodel parameters. Macs2 Identified 6mA peak positions and fold enrichment were then compared to A/T peaks detected using the Scipy find_peak function as described above.

Phylogeny reconstruction

To put our findings within a phylogenetic context, we first clustered protein sequences across all 1117 genomes to identify orthologs using mmseqs2[66] with default parameters. To identify informative clusters for phylogeny reconstruction, we retained any cluster with <40% of lineages missing data. Given the large phylogenetic distances spanned here, we also searched multigene clusters, where we retained a single representative from each lineage within the cluster. Clusters were excluded from analysis if representation from any individual lineage was >6x the average number of copies – this was done to eliminate potential transposable elements while still allowing representatives from diploid genomes to be considered. This yielded 2256 clusters which were used for phylogeny reconstruction. Each cluster was then individually aligned using Muscle[67] and trimmed only to informative sites using Gblocks[68] with parameters: -t = p -e = .gb -b4 = 5 -b5 = h, resulting in 274,069 sites for tree building. After Gblocks trimming, all clusters were concatenated and phylogeny was reconstructed with FastTree[69] using the -gamma and -wag parameters. The phylogeny and associated CAP scores were then visualized using the ete3 toolkit[70] for all lineages with UTR predictions (617 genomes) so fCAPs across all features could be visually compared (Figure 1). In cases where lineages were collapsed at a particular node, the displayed CAP scores represent the average of all scores within that clade. Figure S1 includes all lineages, regardless of UTR annotations.

Machine learning to extract feature importance and predict nucleosome occupancy

To analyze and rank features important for predicting nucleosome occupancy, for each site in *M. eburnea*, *R. toruloides* ATCC26217_1, *Ca. anguillulae*, *U. florida*, and *Mr. frigida*, we gathered per-site information ±75bp from the target position on: G/C (0 or 1), mononucleotide, dinucleotide, trinucleotide, tetranucleotide and 5 different DNA shape features calculated using DNAshapeR.[27] These included: propeller twist (ProT), helix twist (HelT), minor groove width (MGW), Roll, and electrostatic potential (EP). Additionally, we provided %GC spanning the entire ±75bp window, GC content spanning 11bp segments (distance between major grooves) from start to end of the ±75bp window, and scaffold identity (in case there were chromosome-level differences). This amounted to 1526 features for each position in the genome. The ability of these features to predict *in vitro* nucleosome occupancy was then assessed using Random Forests, as implemented in the scikit-learn python package, with parameters: –num_trees = 4000, –max_depth = 15. For each lineage, 100kb of sequence was randomly sampled for training 10 times. Feature importance was then analyzed per position (Figure S11A shows the top 30), as well as summarized across all positions for the same feature type (Figure 4A).

As nucleosome positions are likely dependent upon both local (i.e., GC content and DNA structural features/shapes) and long distance features (i.e., positions of nearby nucleosomes), we developed a deep learning method aimed at predicting *in vitro* nucleosome occupancy across the genome. The most prominent features identified through Random Forests, specifically surrounding GC across ±75bp windows and DNA shape features, as well as mononucleotides were provided as the input data matrix. For training, we included the largest 5 scaffolds from 4 fungal lineages with *in vitro* chromatin data, specifically *M. eburnea*, *Ca. anguillulae*, *R. toruloides* ATCC26217, and *U. florida* (27,947,593 sites). These were selected for their phylogenetic breadth, variability in features important for nucleosome prediction, and difference in CAP scores. As this approach takes information spanning large ranges, it is possible that any given window will span both low and high coverage (i.e., repeat) regions, which could impact learning in unintended ways. To account for this, occupancy scores were scaled using the SciKitLearn StandardScaler function in 10kb windows.

To improve training performance, nucleotides were one-hot encoded and shape parameters were scaled (0–1 scale) using the SciKitLearn MinMaxScaler function. These data were provided in ±80 bp windows (to expand the search window slightly beyond the size of nucleosomes), with a batch size of 15000 to a deep neural network comprised of both convolutional and recurrent layers. Network architecture is a modification of that presented in Quang and Xie 2016,[28] and structured as follows: two 128 neuron 1D convolutional layers with kernel sizes of 33 followed by a 1D 11 × 11 pooling layer, then 2 256 neuron 1D convolutional layers with kernel sizes of 16 followed by a 1D 2 × 2 pooling layer, then 2 bidirectional LSTM layers of 40 and 60 neurons, respectively, followed by 2 dense layers of 200 and 1 neurons, respectively. The Relu activation function was used for all layers. The model was compiled using the Adam optimizer and loss function was set to mean square error. Training was conducted on each scaffold, where 20% of each scaffold was split from the training set to use for validation, and dropout was set to 0.4 to avoid overfitting. Training was allowed to continue for 13 epochs, at which point the network showed no further reduction in loss.

After training was complete, the best model and weights were used to predict occupancy per nucleotide on scaffolds not included in training for *M. eburnea*, *Ca. anguillulae*, *R. toruloides* ATCC26217 and *U. florida*. We also predicted occupancy in *Mr. frigida*, which was not included in training, to assess accuracy of *in vitro* occupancy predictions. In addition to calculating distance between predicted and *in vitro* peaks (Table S2), we also analyzed improvement in predicting occupancy per-site compared to other methods (CAPs and position weight matrix-based) using Pearson's correlation (Figure S12B).

The position weight matrix was generated based on *M. eburnea* dinucleotide frequencies at each position across nucleosome-bound DNA, then each position in the *Mr. frigida* assembly was scored for agreement with this profile. After analysis of predictions compared to *in vitro* nucleosome occupancy, we predicted occupancy in smaller scaffolds from the same genomes included in training, as well as *Cr. neoformans*, *Sp. latifolia*, *R. toruloides* IFO0880, *Ch. reinhardtii* and *T. gondii*. Results were then compared with experimental (*in vitro* and *in vivo*) profiles. Analysis of ATAC-seq and MNase-seq datasets followed the same workflow as described in those STAR Methods sections, except instead of comparing to A/T peaks, comparisons were made to DL predicted nucleosome dyads. For analysis of RB-TDNAseq,[50] the number of inserts at each position in the genome assembly were summed, then plotted with respect to *in silico* DL predicted nucleosome positions using the danpos profile tool surrounding coding sequence start sites. Distance between each insert location and A/T peaks was also calculated for genome wide analysis (Figure S12C).

ADDITIONAL RESOURCES

Meredithblackwellia eburnea MycoCosm homepage: https://mycocosm.jgi.doe.gov/Mereb1.

Rhodosporidium toruloides ATCC26217 MycoCosm homepage: https://mycocosm.jgi.doe.gov/Rhoto_ATCC26217_1.

Mrakia frigida MycoCosm homepage: https://mycocosm.jgi.doe.gov/Mrafri1.

Usnea florida MycoCosm homepage: https://mycocosm.jgi.doe.gov/Usnflo1.

Catenaria anguillulae MycoCosm homepage: https://mycocosm.jgi.doe.gov/Catan2.

Chlamydomonas reinhardtii v5.6 PhycoCosm homepage: https://phycocosm.jgi.doe.gov/Chlre5_6.

Nucleosome prediction software generated as part of this study is available via https://github.com/sjmondo/NuclPred_v1.