Fiber-seq reveals the single-molecule architecture of nuclear transcription and the mitochondrial genome

HARVARD
Kenneth C. Griffin

GRADUATE SCHOOL
OF ARTS AND SCIENCES

DISSERTATION ACCEPTANCE CERTIFICATE

The undersigned, appointed by the

Division of Medical Sciences

in the subject of Biological and Biomedical Sciences

have examined a dissertation entitled

Fiber-seq reveals the single-molecule architecture of nuclear transcription and the mitochondrial genome

presented by Thomas William Tullius

candidate for the degree of Doctor of Philosophy and hereby certify that it is worthy of acceptance.

Signature: _____

Typed Name: Dr. Stephen Buratowski

Signature: Fred Winston (Nov 3, 2023 16:19 EDT)

Typed Name: Dr. Fred Winston

Signature: Jason Buenrostro (Nov 3, 2023 16:28 EDT)

Typed Name: Dr. Jason Buenrostro

Signature: Seychelle Vos (Nov 3, 2023 16:05 EDT)

Typed Name: Dr. Seychelle Vos

Date: November 03, 2023

Fiber-seq reveals the single-molecule architecture of nuclear transcription and the mitochondrial genome

A dissertation presented

by

Thomas William Tullius

to

The Division of Medical Sciences

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

in the subject of

Biological and Biomedical Sciences

Harvard University

Cambridge, Massachusetts

November 2023

Dissertation Advisor: Stirling Churchman Thomas William Tullius

Fiber-seq reveals the single-molecule architecture of nuclear transcription and the mitochondrial genome

Abstract

Gene regulation is driven by the interplay of numerous features including regulatory factors, genome packaging, and the transcriptional machinery. Established approaches to characterize these features on a genomic scale have largely relied on short-read sequencing, providing an averaged, aggregate view across a large population of cells and obfuscating the underlying single molecule dynamics at play. Meanwhile, complementary single-molecule approaches tend to be limited in scope and scale, even as they provide a more granular view. However, the recent development of long-read single-molecule footprinting approaches like Fiber-seq has provided the potential to bridge the genome-scale breadth of sequencing and the single-molecule resolution of microscopy. In this dissertation, I describe how we applied Fiber-seq to characterize two distinct realms of gene regulation— nuclear transcription and mitochondrial genome organization.

First, we used Fiber-seq to visualize RNA polymerases within their native chromatin context at single-molecule and near single-nucleotide resolution along up to 30 kb chromatin fibers. We found that RNA Polymerase II (Pol II) pausing destabilized downstream nucleosomes, with frequently paused genes maintaining a short-term memory of these destabilized nucleosomes. Furthermore, we demonstrated pervasive direct coordination between nearby Pol II genes, Pol III genes, and transcribed enhancers. Overall, we illustrated that transcription

initiation mediates both competition and coordination with nucleosomes and nearby transcriptional machinery along individual chromatin fibers.

Second, we used Fiber-seq to measure the packaging of individual full-length mtDNA molecules at nucleotide resolution, and found that, unlike the nuclear genome, human mtDNA largely undergoes all-or-none global compaction. In addition, we showed that the primary nucleoid-associated protein TFAM directly modulates the fraction of inaccessible nucleoids both in vivo and in vitro, acting consistently with a nucleation-and-spreading mechanism to coat and compact mitochondrial nucleoids. Together, these findings revealed the primary architecture of mtDNA packaging and regulation in human cells. In combination with our characterization of nuclear transcription, this work demonstrates the capability of Fiber-seq to capture the complexity of gene regulation in distinct cellular contexts.

Table of Contents

Listing of Figures

Acknowledgements

I first thank my dissertation advisor, Dr. Stirling Churchman, for her unwavering support of my development as a scientist through the years of my PhD. Stirling's willingness to allow me to pursue independent ideas helped me grow immensely as a scientist, and her exceptional care and eye for detail in science pushed me to apply the same rigor to my work. The research group and program that Stirling has built over the years is a testament to her ability to build a strong community and foster bold ideas, and I hope to emulate her in my future career.

I would also like to thank my coadvisor on this dissertation work, Dr. Andrew Stergachis. Andrew's mentorship on my Fiber-seq work was invaluable, both in his understanding of the challenges and potential of Fiber-seq and in his patience spending hours on Zoom chatting about ideas, nitpicking figures, and editing the paper. He is a truly thoughtful and dedicated mentor and an exceptionally creative and effective scientist. I cannot wait to see what comes out of his research group over the next years and I hope to continue to collaborate with him in the future.

Next, I would like to thank my Churchman labmates for their friendship and scientific help over the years. I want to specially acknowledge Hope Merens, Dr. Blake Tye, Dr. Erik McShane, Dr. Dan Davidi, Dr. Danya Martell-Smart, and Dr. Robert Ietswaart, who have been there for me as both friends and collaborators since I set foot in the Churchman Lab. I have enjoyed every late night with everyone in lab commiserating about experiments after beer hour or chatting on Basecamp to procrastinate from analysis. I would like to especially thank Dr. Stefan Isaac, a postdoc without whom the mitochondrial work in this dissertation would have been impossible, and who has been a huge source of support in both friendship and science over the years.

Graduate school would have been near-impossible without my fellow students and friends. I thank my BBS classmates and all the other graduate students I have known throughout my PhD. I am particularly grateful for Dr. Pauline Schmit, Dr. Greg Babunovic, Dr. Brandon Sit, and Dr.

Olivia Foster Rhoades who have been close friends throughout my time in the BBS program. There are too many other classmates to list, but the memories, especially from the early years of BBS, will stick with me forever. To my friends outside of BBS, especially Bert, Maura, Dave, Morgan, Wellern, Heather, Sandhya, and Jordon: thank you for the constant group chat notifications while I'm trying to pipette or code, the pandemic dinners, the late night gaming sessions, and above all reminding me that there is life outside of science.

I would not be in this position without the support of the labs where I have previously worked. I am indebted to Drs. Peter Chivers, Hiroshi Sugiyama, Peter Burbach, Gary Ruvkun, Chelsea Martinez, Aaron Goldman, and Roderic Guigo for their willingness to open their doors to an undergraduate research student and introduce me to the vast world of biological research. I thank my mentors and friends in and around each of those labs, including Grace, Peter, Daegan, Yousuke, Morinaga, Marty, and Swapan who made my early time as a scientist so rewarding.

To Mom and Dad: thank you from the bottom of my heart. Your support through my life has allowed me to pursue my dreams and passions, and I owe my success to you. Thank you for being role models for me. Thank you for showing me the value of hard work, and always seeing the best in me.

At the heart of my gratitude is my partner, Zoe. Your support, love, and patience, especially chatting science after a long day at the computer and bench, have been invaluable. Graduate school and research had its challenges but having you by my side going through it with me made all the difference. I'm truly grateful, and I cherish our journey and future life together.

This thesis would not have been possible without the additional contributions of dozens of unnamed friends, family, collaborators, and colleagues. I am truly thankful for all of you and your support.

Chapter 1: Introduction

Single-molecule sequencing approaches as a new paradigm

Single-molecule experimental approaches in biology have existed for decades but have increasingly come to the forefront on a genomic scale with advances to long-read, spatial, and single-cell sequencing. As these approaches become more widespread and practical, they have required a paradigm shift in analytical and experimental design, and more generally in the questions that can be asked about biological systems. Here I will discuss the development of single-molecule techniques from a sequencing perspective, with special attention to long-read sequencing and Fiber-seq, the approach that has served as the basis of my dissertation work. I will also lay out the questions in the realm of nuclear and mitochondrial gene regulation that drove the work presented in this dissertation, especially in relation to how Fiber-seq was key to their resolution.

Short-read sequencing and the birth of genomics

The rapid growth of the field of genomics since the early 2000s has led to an unprecedented leap in the understanding of the complex systems underlying gene expression, a revolution up until recently based almost entirely on short-read sequencing techniques. Short-read sequencing, as epitomized by Illumina's Next Generation Sequencing (NGS), allows for the parallel readout by synthesis of millions of individual 150-300bp sequences of DNA or reverse-transcribed RNA (1). This technology has proved to be a wildly adaptable approach, allowing not just for the sequencing of genomes and transcriptomes, but in combination with chemical and biological labeling and enrichment strategies, has served as the basis for an innumerable amount of so-called "-seq" methods. These "-seq" methods expand on the basic toolset of short-read sequencing to probe everything from the specific positions of transcription factors, polymerases, or nucleosomes on chromatin, to the modified bases found on RNA or DNA, to the accessibility of regulatory elements.

What unifies these non-single-cell and non-spatial short-read “-seq” approaches, though, is that they lack single-molecule information beyond the length of each read, which is miniscule in comparison to even the smallest genome or most RNA transcripts. As such, the analysis of short-read approaches requires the pooling of these short segments of DNA from millions of individual cells. The resulting output can generally be boiled down to a distribution of enrichment of these reads at each given position in the genome, showing population-scale trends. These reads are of course still rich with information, especially in combination with perturbations to drive broad changes across the cell population, but they lack the capability to capture temporal and spatial relationships between single chromatin fibers or transcripts.

As a metaphor to understand the shortcomings of short-read sequencing approaches, one can imagine them as akin to characterizing the motion of a running man as a series of overlaid, blended photographs of thousands of men. From such a visualization, one could ascertain that arms and legs are both in motion during running, and that they undergo motion within a given arc with respect to the more static body. Further, one could probe the importance of various observed components, such as by restricting the legs or arms of the test subjects, which would reveal the importance of legs over arms. However, determining the actual simultaneous positions of arms and legs that exist to allow running would not be possible: for example, whether the legs swing in opposite directions or in the same direction. Only a single-molecule approach (or in this case single-man), capturing individual snapshots of the running men at various points in time would allow for a clearer picture of the relationship between each limb in the mechanical process. With this frame of reference, we can now turn towards single-molecule sequencing.

Towards single-molecule genomics

Single-molecule, genome-scale sequencing approaches have become increasingly common in recent years thanks to major steps forward in long-read sequencing technology. Long-read sequencing is exactly what it sounds like—sequencing that generates reads that are orders

of magnitude longer than those generated by short-read sequencers. Long-read sequencing is currently dominated by the competing yet complementary technologies of Oxford nanopore sequencing and PacBio sequencing. Specifically, nanopore sequencing captures up to megabase-scale reads, though typically in the range of 10-100kb (2,3). PacBio sequencing captures up to 25kb reads, with consistently increasing average read lengths (4). Each approach provides certain benefits and limitations, being built on distinct technologies.

Oxford nanopore sequencing is based on changes to current across a membrane as DNA or RNA is pulled through a protein pore, with each individual base providing a unique signature. Due to the speed of translocation, nanopore sequencing has issues with accuracy, though, which is its main Achilles heel. However, the technology has major advantages: it allows for direct RNA sequencing, uses tiny sequencers that individual labs can purchase and run themselves, and has the potential to detect almost any modified base with sufficient training (5,6,7,8). In fact, it may soon prove possible to use nanopore-based approaches to directly sequence the amino-acids of proteins (9,119).

In contrast, PacBio sequencing is based on the incorporation of labeled nucleotides by a polymerase, read out with pulses of light. PacBio sequencing is limited to DNA sequencing and requires expensive equipment. But it has several major advantages over nanopore sequencing. These largely come down to accuracy: DNA is circularized and undergoes multiple cycles of sequencing, giving the method high accuracy and the ability to reliably identify certain DNA modifications based on the speed of translocation (10,**Error! Reference source not found.**). This accuracy has proven essential to approaches like Fiber-seq which rely on extremely accurate and sensitive methylation detection.

Long-read sequencing in general provides numerous benefits over short read sequencing. For one, the mapping difficulties faced by short-read approaches are largely avoided, with long reads able to be uniquely mapped in all but the lengthiest repetitive regions. This has allowed for the sequencing, assembly, and study of previously difficult to map regions of the genome like the

centromere and telomere, and has allowed sequencing of a complete telomere-to-telomere human genome for the first time (11). In addition, each long read represents a single-molecule state of RNA or chromatin, and can inherently capture long range, single-molecule variation in processes like splicing, which otherwise have required complex analysis approaches using short-read sequencing (13).

The difficulties of long read "-seq" experiments

Increasingly, much like short-read sequencing, long-read sequencing has been adapted into "-seq" experiments. However, this adaptation has often required a distinct approach from any short-read "-seq" method. This is because it is essential to preserve the integrity of the DNA or RNA being sequenced to produce long reads, in contrast to short-read approaches which usually require that the DNA or RNA is fragmented to be sequenced. As a result, one of the most common strategies for "-seq" approaches, targeted degradation by nucleases, is nonviable, such that workhorse short-read approaches like MNAse-seq and DNAse-seq are entirely incompatible with long-read sequencing (14,15). This requirement also invalidates a host of chemical marking approaches, like bisulfite sequencing or DMS sequencing (16,17), which degrade DNA or RNA to a degree that is acceptable for short-read sequencing but not for long reads.

Beyond the necessity of preserving the length of the sequenced molecule, there is a further requirement to take advantage of the single-molecule nature of long-read sequencing. To put it simply, each individual read must have the potential to contain multiple observations of whatever feature is being observed. If individual reads cannot show multiple observations, then they cannot capture single-molecule relationships between observations on a given chromatin fiber or transcript. Many strategies that work well for short-read, population-scale methods do not fulfill this requirement. For example, methods relying on the enrichment of DNA or RNA using antibodies (ChIP-seq, RIP-seq) or chemical modifications or the interruption of reverse transcription (DMS-seq, Pseudo-seq) (18,19,20,21). These strategies only capture a single

observation across a massive molecule, and in the case of enrichment would make the location of the observation in long-read sequencing extremely inaccurate due to the length of the read.

Methyltransferase footprinting and Fiber-seq

One of the most fruitful approaches to long-read "-seq" experiments has been to employ nonspecific methyltransferases to mark bases, either throughout the entire genome or based on proximity to a targeted protein of interest. The general concept is to use a methyltransferase to mark accessible chromatin with methylated bases, allowing for a footprinting approach that is essentially the inverse of nuclease-based approaches like MNAse-seq. This approach was implemented initially using short-read sequencing (22) but has since been adapted to both nanopore and PacBio sequencing by several groups in the form of such techniques as Fiber-seq, SAMOSA and DiMeLo-seq (23,24,132).

Fiber-seq, the main technique used in this dissertation, uses a nonspecific bacterial methyltransferase, Hia5, to mark accessible chromatin using m6A residues (23). Long chromatin fibers are purified and sequenced using PacBIO sequencing, which allows for high accuracy and direct, sensitive readout of m6A residues based on the interpulse distance (IPD) of the Fiber-seq read using the fibertools package (25). The result is on average 15-20 kb reads with methylation marks at accessible positions that can then be used to call inaccessible footprints of the chromatin-bound proteome using FiberHMM, a hidden Markov model-based footprint caller we developed to support the work outlined in this dissertation. FiberHMM considers the sequence biases of Hia5 and of the IPD from the PacBio sequencer to account for variation in the methylation rate across the genome in calling footprints.

Applying Fiber-seq to transcription

One of the key difficulties in working with Fiber-seq data is in identifying questions about a given biological system that can only be addressed using a single-molecule approach like Fiber-seq. In the remaining portion of Chapter 1, I will describe the general questions and corresponding background information relating to the two very distinct systems of gene regulation that make up this dissertation. The first system is transcription in the *Drosophila melanogaster* nucleus. The focus of this work, outlined in Chapter 2, is on how different polymerase complexes coordinate with each other and with the nucleosome landscape to coexist on chromatin and produce the transcriptome.

Targeting transcription with footprinting approaches

The key initial challenge we were faced with as we set out to use Fiber-seq to characterize transcription was the need to identify Pol II footprints within the data. As such, we chose to focus our work on the stages of transcription that were most likely to be identifiable with high confidence. Specifically, to be confidently identifiable in Fiber-seq at minimum a footprint in question must be found in a consistent, predictable location, be stable enough to be footprinted during the 10-minute methylation step, and have corresponding orthogonal short-read datasets to confirm its location. Taking these factors into account we decided to focus on initiation and pausing as the steps of transcription to target, given their consistent positions, importance in gene regulation, the success of other footprinting approaches in identifying these footprints, and the numerous corresponding short-read datasets in our system of choice, *Drosophila melongaster* S2 cells (22,26).

Transcription begins with the clearing of nucleosomes from the promoter by various factors including chromatin remodelers like SWI/SNF and ISWI and pioneer transcription factors like GAF, forming a nucleosome depleted region (NDR) (27). Once the promoter region is accessible, the preinitiation complex (PIC) forms with general transcription factors like TFIID and the mediator complex binding around the TSS and recruiting RNA polymerase II (Pol II) (28).

TFIIH binding to the PIC and phosphorylation of the Pol II C-terminal domain leads to opening of the transcription bubble which initiates transcription and RNA synthesis. After initiation and before productive elongation, Pol II often pauses around 25-50bp downstream of the transcription start site (TSS) (28). This pause, the promoter proximal pause (PPP), varies in stability and duration from gene to gene. (29). Pausing is mediated by the factors DSIF and NELF, and upon phosphorylation by the kinase P-TEFb, NELF is released and polymerase proceeds into productive elongation (28).

Pausing was first discovered in the context of heat shock in Drosophila (30,31), and has since been shown to be particularly prevalent in genes associated with tightly regulated transcriptional programs like development (32,33,34) and responses to environmental perturbations (30,31). As such, pausing is thought to represent a checkpoint that allows cells to synchronize transcription, integrate regulatory signals, and maintain an accessible promoter, particularly at genes which require a rapid, regulated transcriptional response (28). Pausing is not necessarily associated with high expression– in fact, it often is associated with reduced expression as it serves as a block both to further elongation and to further initiation (35,36,37). Both the PPP and PIC proved to be stable enough after nuclear isolation to be readily visible in footprinted Fiber-seq data, allowing us to ask specific questions about the relationship between these footprints and the broader chromatin landscape. The first was testing the steric mechanism of the inhibition of initiation by pausing at individual loci, a mechanism we directly validated using these footprints in Chapter 2.

RNA Pol II and the nucleosome landscape

Transcription occurs in the context of a chromatin fiber packaged by nucleosomes, which provide a barrier to initiation, pausing, and elongation. Transcription, in turn, has been shown to influence nucleosome organization as well. Initiation is affected by nucleosome positioning for reasons discussed in the previous section, given that it requires an accessible TSS. The promoter

proximal pause has also been associated with changes to the nucleosome landscape both at the promoter and surrounding it. Highly paused genes are associated with low promoter accessibility, suggesting a role for pausing in maintaining accessibility of otherwise occluded promoters. Beyond the immediate vicinity of the TSS, disrupted upstream and downstream nucleosomes are found at genes with high levels of pausing suggesting a possible influence of pausing on the downstream nucleosome landscape (38).

Transcription elongation has also been associated more directly with disruption of nucleosomes, however the barrier posed by nucleosomes to elongation is stronger in *in vitro* reconstitution experiments than is seen *in vivo* (39). This is at least in part due to histone chaperones like SPT6 and FACT, which are responsible for facilitating the traversal and reassembly of nucleosomes during transcription (40). Traversal of nucleosomes during transcription elongation involves the generation of detectable hexasomal intermediates (41,42), nucleosome complexes lacking a single histone H2A-H2B dimer. With high levels of transcription, as in heat shock, nucleosomes can be depleted significantly, likely by a combination of supercoiling from transcription itself (43) and chromatin remodelers (44).

The overall interactions between transcription and the surrounding nucleosome landscape are a prime target for a single-molecule approach like Fiber-seq, given that they represent changes to nucleosomes, the footprints of which are readily identifiable in Fiber-seq data. These changes can be directly observed in combination with a particular transcriptional state on a given single chromatin fiber, as defined by the existence of a particular Pol II footprint. Indeed, as we lay out in the results of Chapter 2 of this dissertation, we can demonstrate that pausing drives transient changes to nucleosomes downstream that are otherwise impossible to capture and characterize using aggregated short-read methods.

Genome organization and transcription

Beyond the immediate chromatin environment, genomes contain a tuned spatial organization of gene position, both in their proximity along the linear map of the genome and between distal portions of the genome organized in three-dimensional space in the nucleus (45,46). Transcription has been shown through microscopy to often occur in foci, with multiple genomic regions and numerous polymerases and regulatory factors clustered in a tight area (47,48,49). Pairs of genes in *Drosophila* especially are also often linked based on their proximity, with correlated expression and function drawing parallels to prokaryotic genome organization (50,51,52).

Transcription is regulated by enhancers, short regions of DNA bound by transcription factors which facilitate the expression of target genes via binding of regulatory proteins like transcription factors. These targets can be found either in close proximity or across long distances, with looping of chromatin bringing distant enhancers into proximity across 3D space, further stressing the importance of spatial proximity in gene-regulatory relationships (53,54,55). Many enhancers are also transcribed themselves, producing short, unstable transcripts called eRNAs (56,57). eRNA production is generally correlated with target gene expression and has been thought to play roles in facilitating looping and recruiting relevant factors to the target promoter, although the exact mechanism by which they mediate expression of their targets is yet to be fully resolved (57,58).

Eukaryotic chromosomes are packaged into self-interacting regions called topologically associated domains (TADs), detectable via chromosome conformation capture approaches like Hi-C (59,60). Chromatin contacts are interrupted by insulators, architectural regions bound by a factors including CTCF, SuHW, Mod, and CP190 in Drosophila, which block enhancer-promoter relationships (61,62,63). In Drosophila tethering elements have been shown to serve the opposite purpose, mediating long-range interactions between groups of genes and enhancers (63).

What is unclear about these spatial-transcriptional relationships between pairs of promoters and enhancer-promoter pairs, though, is if they are representative of correlation alone,

or if they represent direct coactivation of transcription of pairs of genes or of genes and transcribed enhancers. Further, it is unclear how exactly eRNA transcription increases the expression of its target gene, and if such a mechanism is related to the correlation in expression observed between nearby coding genes. As set forth in Chapter 2, Fiber-seq proved to be capable of testing for direct coordination in transcription initiation, given its ability to capture snapshots across ~20kb distances of the co-occupancy of multiple promoters and enhancers by polymerase footprints. Our work outlined in Chapter 2 reveals pervasive, distance-dependent transcriptional coactivation of pairs of Pol II genes and enhancer-promoter pairs, modulated by genome architecture.

RNA Polymerase III

Pol II is not the only polymerase in eukaryotic genomes—in fact the majority of the total output of RNA is generated by RNA Polymerase I (Pol I) and RNA Polymerase III (Pol III). Pol III transcribes several classes of RNAs, most prominently tRNAs and the 5S rRNA, with the help of general transcription factors including TFIIIA, TFIIIB and TFIIIC (64). As these RNAs are required in a massive abundance to allow for translation, Pol III must transcribe rapidly, being recycled after each round of transcription (65). Pol III genes are often found in a clustered arrangement, in the case of the 5S rRNA in an array of over 100 copies, and in the case of tRNAs in smaller clusters throughout the genome. tRNA genes are also often found in proximity to Pol II genes, with evidence that they can function as insulators in yeast and human cells (66). This result suggested a connection between these two polymerases and prompted interest as an extension of our work on coordination of Pol II transcription.

Pol III transcription represents a difficult system for transcriptional studies, given that there are multiple sequence-identical copies of most tRNA genes in the Drosophila genome, and given that the reverse transcriptases required for RNA-seq experiments have difficulty proceeding through the strong secondary structure and numerous modifications of tRNA transcripts. 5S rRNA genes are more intractable, being identical in sequence and found in a highly repetitive region of

tandem repeats that makes distinguishing between them on a transcript level impossible. Fiber-seq, given the ease of alignment of its long reads, proved to be an excellent tool to characterize the expression patterns of tRNA and 5S rRNA genes, both within a family of tRNAs and 5S rRNAs and between different tRNA and 5S rRNA genes.

Capturing the mitochondrial genome in a single read

The other major focus of this dissertation, with work laid out in Chapter 3, is the architecture of the mitochondrial genome. Derived from an endosymbiotic relationship in the early stages of eukaryotic evolution, mitochondria are responsible for the energy production of the cell via oxidative phosphorylation (OXPHOS) (67). While most mitochondrial proteins are encoded in the nuclear genome, a subset of essential OXPHOS components, rRNAs and tRNAs are encoded in the mitochondrial genome (68). The human mitochondrial genome, the subject of Chapter 3, is circular and only 16,569 bp, present in hundreds to thousands of copies across the mitochondrial network of a given cell (69).

Mitochondrial genomes are an ideal size for Fiber-seq, as the short length of the genome compared to the nuclear genome allowed each Fiber-seq read to capture an entire mitochondrial genome, resulting in a coverage of hundreds of thousands reads. In contrast, we were only able to achieve a ~1000-fold lower coverage across the Drosophila nuclear genome with an order of magnitude more overall reads per treatment condition, given its much larger size. These data were exceptionally rich, capturing with single-molecule resolution the wide variation in accessibility levels and patterns of individual mitochondrial genomes. In contrast to nuclear transcription, though, the components of which have been characterized *in vivo* to an extensive degree, mitochondrial genome regulation remains more of a mystery. As a result, applying Fiber-seq provided an almost diametrically opposed set of challenges and advantages compared to the nuclear genome, with incredibly deep coverage but a lack of orthogonal *in vivo* datasets to complement the Fiber-seq data.

Mitochondrial genome packaging

Mitochondria lack histones, which means that unlike the nuclear genome their genomes are not packaged into nucleosomes. This lack of nucleosomes does not mean that mitochondrial genomes are not packaged or compacted, however. The primary candidate for a histone equivalent in mitochondria is TFAM. TFAM binds promiscuously and cooperatively to DNA and in vitro can alter DNA compaction based on its concentration in solution (70). TFAM's compaction ability is aided by the sharp turns it induces in DNA, leading to looping (71). Due to its broadly nonspecific binding preferences, characterizing the pattern of TFAM binding on the mitochondrial genome in vivo has proven difficult, as it does not bind in the periodic regular pattern seen with nucleosomes on the nuclear genome (72). This means that a short-read approach, like the ATAC-seq data shown in Chapter 3, provides essentially uninterpretable, largely uniform results. As such, understanding the variance in TFAM binding between individual genomes was an ideal application of a single-molecule approach like Fiber-seq, allowing us to query rarer states of the mitochondrial genome that were otherwise overwhelmed by what proved to be a vast majority of highly compacted mitochondrial genomes.

TFAM and mitochondrial genome regulation

Mitochondrial genomes contain a single major regulatory region, the noncoding region (NCR). The NCR generally contains a structure known as a displacement loop, or D-Loop, where a ~650 bp region is copied by the mitochondrial polymerase, Polɣ with the assistance of TWINKLE, a helicase, starting at the heavy-strand origin of replication. This results in the 7S DNA, the newly replicated DNA copy, bound to the template with a displaced single-stranded loop of DNA. In most cases Polɣ then stops at the termination associated sequence (TAS), with

TWINKLE dissociating. Polγ is at this point poised to continue elongation if TWINKLE associates with it once again, suggesting that this termination step is the main regulatory checkpoint for mitochondrial genome replication. (73,74,75)

Mitochondrial replication is directly connected to mitochondrial transcription, with transcription producing a primer necessary for replication to proceed from the heavy-strand origin of replication (76). Mitochondrial transcription is carried out by the polymerase PolRMT, which transcribes long polycistronic transcripts from light strand and heavy strand promoters (77). The long transcripts produced are then cleaved and processed into individual transcripts, with most of these transcripts punctuated by mitochondrial tRNAs. As such, regulation of gene expression inherently cannot be driven by transcription levels, as the transcripts are produced in groups. There has been speculation based on *in vitro* experiments that MTERF1, a putative termination factor bound at the border of the rRNA and coding transcripts in the genome, serves to modulate the relative levels of rRNA and coding transcripts produced. However, *in vivo* evidence points towards a role in preventing interference with transcription initiation at the light strand promoter instead (78,79).

Mitochondrial transcription requires TFAM– in fact TFAM was originally discovered as a transcription-stimulating mitochondrial protein in extracts (80). TFAM and TFB2M, an initiation factor, work together to recruit POLRMT, melt the promoter, and enable RNA elongation (81). As such, TFAM is required for all of the processes described above: genome architecture, transcription, and replication (since mitochondrial replication requires the transcription of a primer by POLRMT) (82). This highlights not just the importance of TFAM in the mitochondrial genome, but also the interconnectedness of each of these distinct processes. As a result, a major goal of the work outlined in Chapter 3 was to take advantage of the single-molecule nature of Fiber-seq to capture these processes on single mitochondrial genomes, to directly characterize both their states on individual genomes and their connection to TFAM-driven mitochondrial genome architecture.

Chapter 2: RNA Polymerase initiation and pausing alter nucleosome landscape and coordinate transcription activity on single chromatin fibers

<u>This chapter has been adapted from a manuscript in preparation:</u> Thomas W. Tullius, R. Stefan Isaac, Jane Ranchalis, Danilo Dubocanin, L. Stirling Churchman, Andrew B. Stergachis. **RNA Polymerase initiation and pausing alter nucleosome landscape and coordinate transcription activity on single chromatin fibers**

Author Contributions

Conceptualization, T.W.T., L.S.C., and A.B.S.; Methodology, T.W.T. (lead), R.S.I., A.B.S., and L.S.C.; Investigation, T.W.T. (lead), R.S.I., and J.R.; Formal Analysis, T.W.T. (lead), D.D.; Writing - Original Draft, T.W.T..; Writing - Review & Editing, T.W.T., R.S.I., A.B.S, and L.S.C.; Funding Acquisition, A.B.S. and L.S.C.; Supervision, A.B.S. and L.S.C.

Abstract

During eukaryotic transcription, RNA polymerases must initiate and pause within a crowded, complex environment, surrounded by nucleosomes and other transcriptional activity. This environment creates a spatial arrangement along individual chromatin fibers ripe for both competition and coordination, yet the basic principles driving this relationship inside the nucleus remain largely unknown owing to the inherent limitations of traditional structural and sequencing methodologies. To address these limitations, we employ single-molecule long-read chromatin fiber sequencing (Fiber-seq) to visualize RNA polymerases within their native chromatin context at single-molecule and near single-nucleotide resolution along up to 30 kb chromatin fibers. We demonstrate that Fiber-seq enables the identification of single-molecule RNA Pol II and Pol III transcription associated footprints, which in aggregate mirror bulk short-read sequencing-based measurements of transcription. We show that Pol II pausing destabilizes downstream nucleosomes, with frequently paused genes maintaining an epigenetic short-term memory of these destabilized nucleosomes. Furthermore, we demonstrate pervasive direct coordination

between nearby Pol II genes, Pol III genes, and transcribed enhancers. This coordination is largely limited to spatially organized elements within 1 kb of each other, implicating short-range chromatin environments as a predominant determinant of coordinated polymerase initiation. In contrast, we find marked anti-coordination between transcription initiation at neighboring Pol II and Pol III genes, indicating that tRNA mediated silencing of Pol II-transcribed genes is largely occurring at the level of transcription initiation. Overall, we show that transcription initiation directly mediates both competition and coordination with nucleosomes and nearby transcriptional machinery along individual chromatin fibers.

Introduction

Eukaryotic RNA Polymerase II (Pol II) transcription initiates in a highly regulated manner along a genome that is co-bound by numerous chromatin-associated proteins, including nucleosomes, transcription factors, chromatin remodelers, and other RNA polymerase complexes. Initiation begins with the assembly of the preinitiation complex (PIC), consisting of Pol II and an array of general transcription factors which work in unison to initiate transcription. Following PIC assembly and Pol II release, Pol II generally transcribes 25-50 bp and pauses at the promoter-proximal pause site, a critical step in transcription which allows for the integration of regulatory signals before proceeding into productive elongation (28,83,84).

Pol II initiation takes place amidst nucleosomes, accessible enhancer elements, and neighboring transcribed genes, creating a spatial arrangement ripe for both competition and coordination along individual chromatin fibers. Specifically, Pol II must contend with nucleosomes which serve as a physical barrier to both transcription initiation, pausing, and elongation (28,40,42,85,86,87). Genes with higher rates of Pol II pausing are known to be associated with distinct promoter-proximal nucleosome patterns (38). However, it is unclear whether these associations reflect paused Pol II directly modulating the positioning of assembled nucleosomes

that it is competing with along an individual fiber. Furthermore, transcription initiation occurs in spatial proximity to nearby transcriptionally active promoters and enhancers - a configuration that is potentially conducive for coordination between these neighboring elements (50,52,53). For example, neighboring promoters are known to have correlated expression patterns, transcriptional activity at enhancers is often associated with the expression of neighboring genes, and transcriptionally active genes are known to co-localize at specific foci within the nucleus (47,48,49,89,90). It is unknown, though, if these associations represent direct coordination in transcription initiation along individual chromatin fibers within the nucleus.

Although pioneering cryo-EM and biophysical studies have provided a detailed single-molecule understanding of how transcription initiation occurs along individual DNA templates (36, 42,85,86,87,97), our understanding of this process along intact multi-kilobase chromatin fibers within the nucleus is limited. Emerging footprinting methods (22,23,24,91,92) have the potential to overcome these limitations, with single-molecule chromatin fiber sequencing (Fiber-seq) specifically providing near single-nucleotide resolution maps of protein occupancy along individual multi-kilobase fibers within the nucleus (23). Fiber-seq uses a nonspecific N6-adenine methyltransferase (m6A-MTase) to stencil protein footprints onto their underlying DNA templates, which are directly read using PacBio SMRT sequencing, producing single-molecule maps of 100s of individual protein footprints along 15-20 kb chromatin fibers (23).

In this study we leverage Fiber-seq to directly characterize both competition and coordination between RNA polymerases and surrounding chromatin proteins along individual chromatin fibers. We present an HMM-based footprint caller (FiberHMM) that enables the identification and quantification of single-molecule paused Pol II, PIC, nucleosome, and Pol III transcription-associated footprints. Using this single-molecule data in combination with targeted perturbation assays, we characterize the dynamics of transcription initiation, the mechanism for pause-inhibition of initiation, pause-driven changes to nucleosome architecture and direct

coordination between nearby Pol II and Pol III genes, and transcribed enhancers within spatially organized chromatin domains.

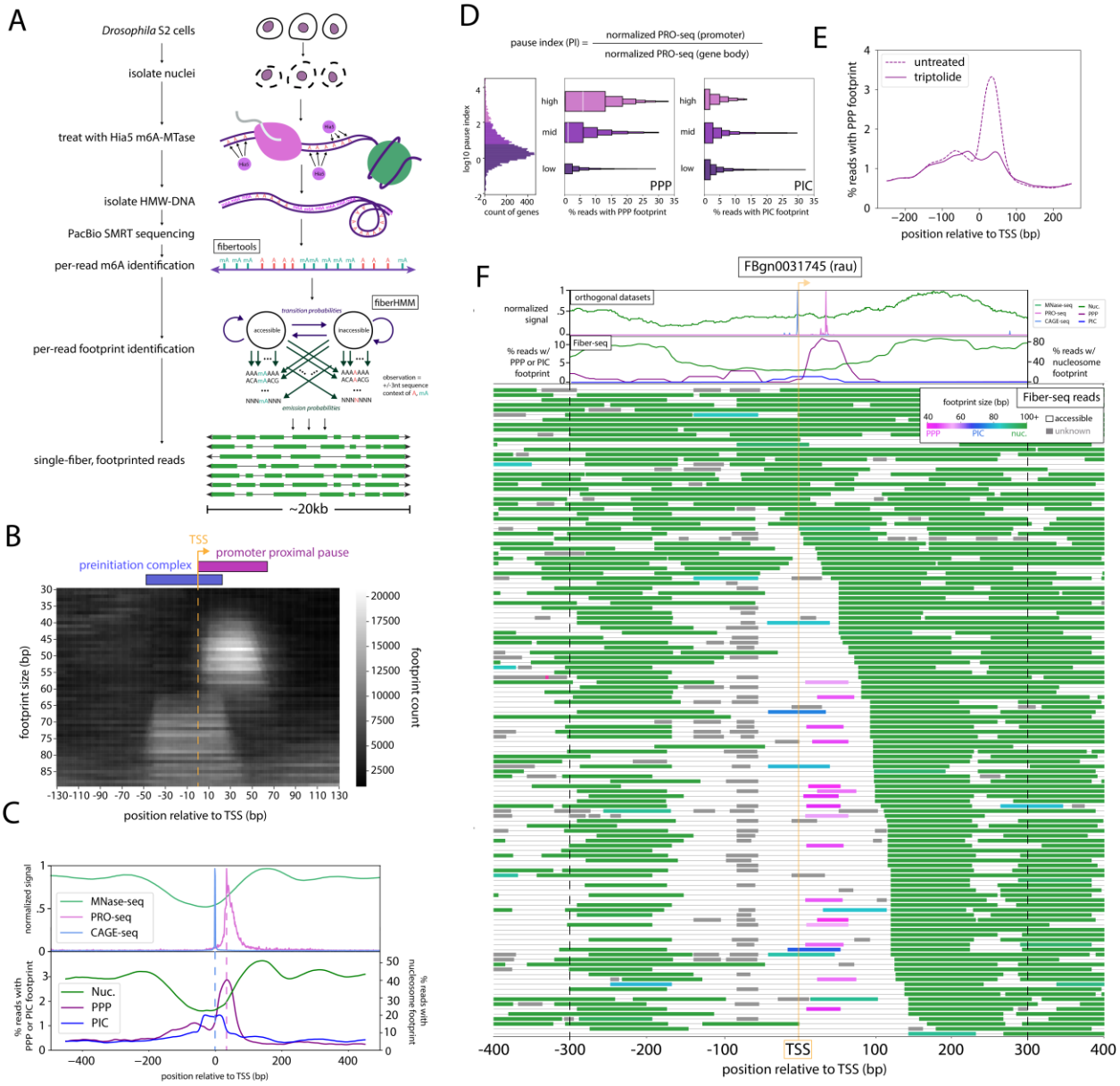
Results

Identification of single-molecule RNA polymerase II footprints via FiberHMM

To identify single-molecule RNA polymerase-related footprints, we first applied Fiber-seq to *Drosophila* S2 cells, resulting in an average sequencing coverage of 160x across the genome (12kb average read length). Next, we developed a hidden Markov model based footprint caller (FiberHMM) to identify unmethylated patches of adenines along each fiber that correspond to single-molecule RNA polymerase-related footprints. Specifically, FiberHMM incorporates observed adenine methylation patterns as well as false negative and false positive methylation probabilities for each base, which are derived from sequencing data from Hia5-treated dechromatinized genomic DNA (gDNA) as well as untreated genomic DNA, respectively. These false negative and positive methylation rates are used as fixed emission probability parameters within the HMM, while the starting and transition probabilities are trained using a subset of the Fiber-seq data. This trained model is then used to identify methylation footprints within each read at single-molecule and near single-base resolution **(Fig. 2.1A)**.

Promoter-proximal paused RNA polymerase II (PPP) and preinitiation complex (PIC) footprints have been previously observed using short-read MTase-based approaches (24). As such, we sought to determine whether FiberHMM could similarly identify PICs and PPPs at genomic sites known to harbor these complexes. When we aligned Fiber-seq footprints on a metagenomic scale around transcription start sites (TSSs), two distinct populations of footprints were visible: the first directly overlapping the TSS and a second located 0-60 bp downstream of the TSS **(Fig. 2.1B)**. This first footprint population matches the previously observed location of

Fig. 2.1: Identification of Pol II footprints in Fiber-seq
(A) Schematic of the Fiber-seq experiment. (From top to bottom) *Drosophila Melanogaster* S2 cells were subjected to nuclear permeabilization and subsequently treatment with Hia5 m6A-methyltransferase. The resulting HMW-DNA were then sequenced via PacBio sequencing, with methylated As subsequently called via the fibertools package. Footprints were called using FiberHMM, schematized in the diagram shown. The result was 15-20kb footprinted Fiber-seq reads representing individual chromatin fibers **(B)** Heatmap depicting the enrichment of differently sized Fiber-seq footprints with respect to the nearest transcription start site (TSS) from a gene with a pause index >=10. The expected position of the promoter proximal pause (PPP) and preinitation complex (PIC) are shown via rectangles above the plot. There is a strong enrichment of 40-60bp footprints at the expected location of the PPP and 60-80bp footprints at the expected location of the PIC. **(V)** Plot depicting the overall enrichment of (Top) MNAse-seq, PRO-seq, and CAGE-seq and (Bottom) Nucleosome-, PPP-, and PIC-sized footprints with respect to the TSS of all genes with a pause index >=10. The y-axis for the lower plot is split between (Left) PPP and PIC footprints and (Right) Nucleosome footprints. A dashed line is drawn from the peak of the CAGE-seq and PRO-seq peaks to demonstrate the alignment with the PPP and PIC footprints. **(D)** A set of plots, showing a (Left) histogram of the distribution of pause index values for genes colored based on binning into “high” (>=100), “mid” (>=10, <100), and “low” (<10) values, with pause index defined based on PRO-seq coverage in the promoter divided by PRO-seq coverage in the gene body (Above). Boxenplots showing the enrichment of (Center) PPP and (Right) PIC footprints in Fiber-seq reads from genes binned by pause index. PPP footprints show a strong enrichment with higher pause index, while PIC footprints are not dependent on pause index. **(E)** Plot showing the % reads with PPP-sized footprints in Fiber-seq from S2 cells either untreated (dashed) or after 30 minutes treatment with 10uM triptolide (solid) with respect to the TSS of genes with a pause index >=10. Fewer PPP footprints are seen at the promoter proximal pause site with triptolide treatment. **(F)** Fiber-seq reads at an example locus demonstrating variation in nucleosome positioning and PPP and PIC occupancy on a single-molecule basis. The top two tracks plot (Above) MNAse-seq, PRO-seq, and CAGE-seq signal, or (Below) percentage of reads with nucleosome-, PPP-, or PIC-sized footprints, showing their similar patterns of enrichment to the orthogonal short-read datasets. (Bottom) Individual Fiber-seq reads are plotted on each line with footprints represented by colored blocks and accessible regions by a thin black line. Footprints are colored based on their size, with nucleosome-, PPP-, and PIC-sized footprints being green, pink, and blue respectively. Footprints smaller than a nucleosome, but not overlapping PRO-seq or CAGE-seq signal are colored grey to indicate that their identity is unknown. Reads are sorted by the position of the +1 nucleosome.


A
Drosophila S2 cells
isolate nuclei
treat with Hia5 m6A-MTase
isolate HMW-DNA
PacBio SMRT sequencing
per-read m6A identification
per-read footprint identification
single-fiber, footprinted reads
fibertools
fiberHMM
~20kb
B
TSS
promoter proximal pause
preinitiation complex
footprint size (bp)
footprint count
position relative to TSS (bp)
C
normalized signal
MNase-seq
PRO-seq
CAGE-seq
% reads with PPP or PIC footprint
Nuc.
PPP
PIC
% reads with nucleosome footprint
position relative to TSS (bp)
D
pause index (PI) = normalized PRO-seq (promoter) / normalized PRO-seq (gene body)
log10 pause index
count of genes
high
mid
low
PPP
PIC
% reads with PPP footprint
% reads with PIC footprint
E
untreated
triptolide
% reads with PPP footprint
position relative to TSS (bp)
F
FBgn0031745 (rau)
orthogonal datasets
normalized signal
Fiber-seq
% reads w/ PPP or PIC footprint
% reads w/ nucleosome footprint
footprint size (bp)
Fiber-seq reads
accessible
unknown
TSS
position relative to TSS (bp)

PICs, whereas the latter matches the previously observed location of PPPs. The size ranges of these putative PIC and PPP footprints were 60-80 bp and 40-60 bp respectively, mirroring the expected sizes from previous biochemical and structural studies (86,93).

Validating single-molecule transcription complex occupancy

We next sought to validate the positioning and enrichment of these FiberHMM-derived putative PIC and PPP footprints using a combination of orthologous datasets, as well as chemical perturbations. First, we compared the positioning of these footprints to short-read sequencing derived transcript initiation and pausing data from CAGE-seq and PRO-seq, respectively (94,95). As expected, putative PIC footprints are selectively positioned directly over CAGE-seq peaks, and putative PPP footprints are selectively positioned directly over PRO-seq peaks **(Fig. 2.1C)**. Second, we evaluated the enrichment of PPP footprints within TSSs for highly paused genes. As expected, putative PPP footprints were significantly enriched at genes with a high PRO-seq derived pause index, and gene pause index and PPP footprint enrichment showed a monotonic relationship. In contrast, the enrichment of putative PIC footprints was independent of gene pause index, as expected **(Fig. 2.1D).**

Finally, we performed Fiber-seq on cells treated with two compounds known to disrupt the transcriptional machinery, triptolide and α-amanitin. First, treatment with triptolide, an initiation inhibitor that blocks progression from the PIC to the promoter proximal pause (96), significantly reduced the levels of putative PPP footprints, while having no effect on the levels of putative PIC footprints **(Fig. 2.1E)**. Second, treatment with α-amanitin, a transcription elongation inhibitor that blocks progression of both the PIC and PPP to elongating RNA pol II (97), significantly increased the levels of both putative PPP and PIC footprints **(Fig. S2.1A,B)**.

In addition to PIC and PPP footprints, FiberHMM also identified nucleosome-sized footprints that are preferentially localized to sites that mirror bulk MNase-seq data, as well as smaller footprints upstream of the TSS that likely correspond to single-molecule transcription

factor (TF) occupancy events (Fig. 2.1F). Overall, each Fiber-seq molecule on average contains ~70 protein occupancy events defined using FiberHMM, with the majority of these being nucleosome footprints, enabling us to evaluate the single-molecule co-dependency between PIC, PPP, and nucleosome occupancy at individual loci genome-wide.

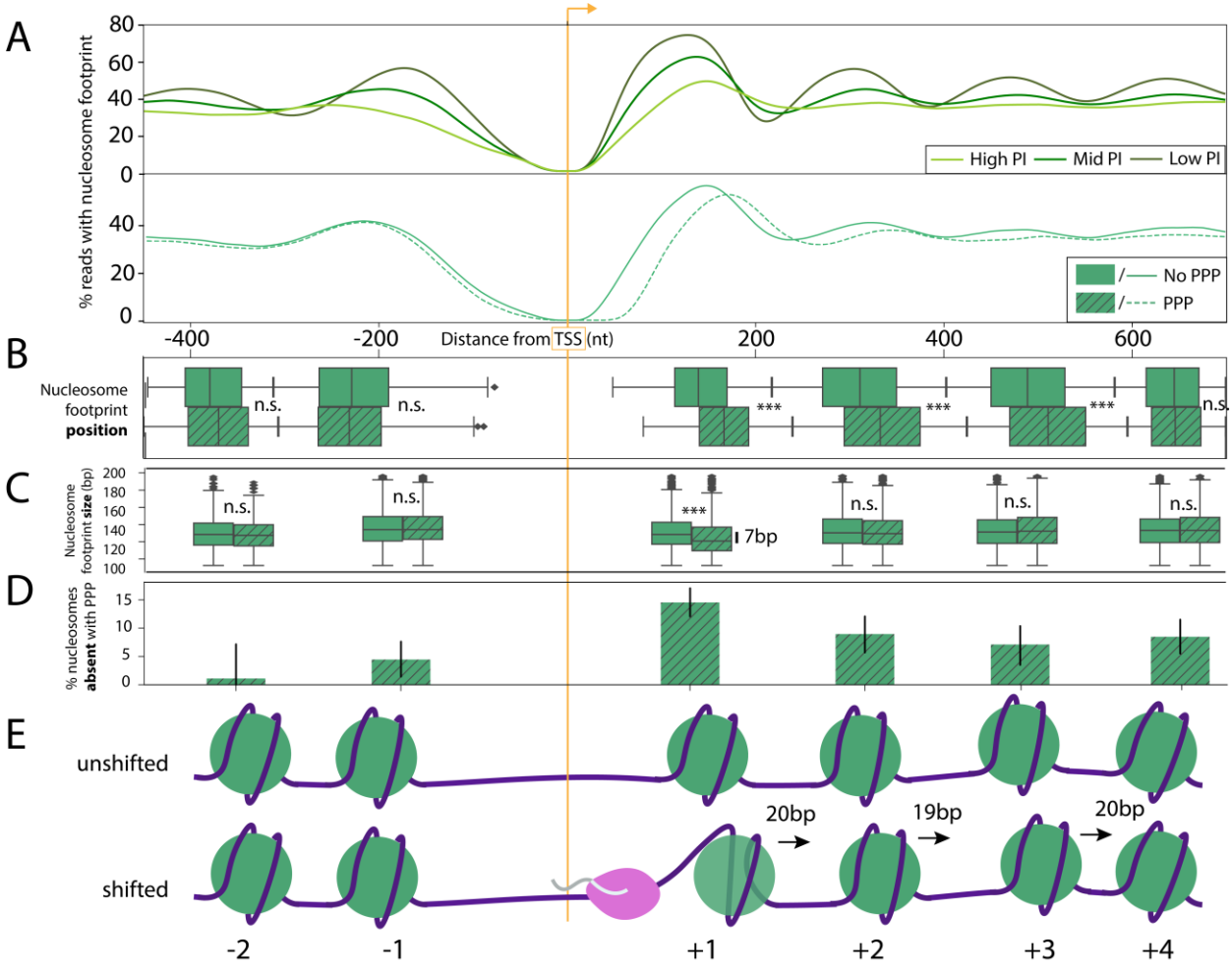
Pausing is associated with changes in nucleosome architecture

Genes with high pausing are also known to be associated with disrupted chromatin patterns. Specifically, studies using MNAse-seq have reported that highly paused genes tend to have shifted upstream and downstream nucleosomes and decreased promoter accessibility (38). However, directly disentangling the relative contribution of promoter accessibility and PPP occupancy to these neighboring chromatin changes has been stymied by the lack of single-molecule resolution. Taking advantage of the ability of Fiber-seq to capture the occupancy of hundreds of proteins along a single chromatin fiber, we set out to determine the degree to which pause index-associated changes to nucleosome architecture are directly explained by paused Pol II itself.

In aggregate, we found that Fiber-seq mirrored the MNase-seq derived patterns of nucleosome enrichment seen at genes with different pausing levels **(Fig. S2.3B)**. (**Error! Reference source not found.**). Similarly, we found that genes with higher levels of pausing had more fibers with inaccessible promoters **(Fig. S2.3D)**. Importantly, we observed that chromatin fibers with inaccessible versus accessible promoters had markedly distinct nucleosome architectures, which are unavoidably aggregated together when analyzing MNase-seq data **(Fig. S2.3E)**. To account for the potential artifactual impact these inaccessible promoters have on the overall pattern of nucleosome enrichment, we selectively evaluated nucleosome patterns on chromatin fibers containing accessible promoters. This demonstrated that genes with high pause

indices are truly associated with shifted +1 and -1 nucleosomes **(Fig. 2.2A)**. In contrast, controlling for promoter

Fig. 2.2: Pause-associated changes to nucleosomes
(A) (Top) Plot depicting the enrichment of nucleosome footprints from genes binned by pause index, using the same bins as Figure 2.1D: “high” (light green, PI >= 100), “mid” (mid green, 100>PI>=10), “low” (dark green, 10>PI). Reads are filtered to only include those without a nucleosome overlapping the TSS. Genes with a higher pause index have shifted nucleosomes both upstream and downstream of the TSS. (Bottom) Plot depicting the enrichment of nucleosome footprints from reads with (solid) or without (dashed) a PPP footprint. Reads are filtered to only include those without a nucleosome overlapping the TSS and are sampled to include an equal count of reads with or without a PPP footprint from each gene. Reads with a PPP footprint have shifted downstream nucleosomes, but unchanged upstream nucleosomes, showing a directional effect associated with the PPP itself. **(B)** Boxplot showing the distribution of the position of each nucleosome upstream and downstream of the TSS between the same set of reads with (dashed) or without (solid) a PPP footprint. The nucleosome footprints were segmented using a gaussian mixture model (GMM) based on a 95% posterior probability, and the distributions were compared with a T-test (n.s. signifies p>.05, *** signifies p-value < 0.001). The +1, +2, and +3 nucleosomes are significantly shifted by on average 20 bp, 19 bp, and 20 bp respectively, with other nucleosomes showing no significant change. **(C)** Boxplot showing a comparison of the footprint size of the same segmented sets of nucleosomes as in (B) with (dashed) or without (solid) a PPP footprint. Size distributions were compared using a T-test (n.s. signifies p>.05, *** signifies p-value < 0.001). The +1 nucleosome alone showed a significant difference, with a 7nt smaller footprint suggesting a partially unwound footprint. **(D)** Bar plot showing the percentage of additional nucleosome footprints absent in reads with a PPP footprint compared to the baseline frequency of absence found in reads without a PPP footprint in the same segmented sets of nucleosomes as in (B). Error bars represent the confidence interval derived from the bootstrapped distribution of the ratio of nucleosome footprint counts in PPP and no PPP samples. **(E)** Cartoon summarizing the changes to nucleosome position and stability associated with the PPP. Compared to the unshifted state, reads with a PPP footprint and shifted nucleosomes have a partially unwound +1 nucleosome, shifted +1, +2, and +3 nucleosomes, and often an absent +1 nucleosome.


A
% reads with nucleosome footprint
High PI
Mid PI
Low PI
No PPP
PPP
Distance from TSS (nt)
B
Nucleosome footprint position
n.s.
***
C
Nucleosome footprint size (bp)
7bp
D
% nucleosomes absent with PPP
E
unshifted
shifted
20bp
19bp
20bp
-2
-1
+1
+2
+3
+4

accessibility, we found no difference in nucleosome architecture based on expression levels of genes, suggesting a unique role for pausing in altering the nucleosome landscape **(Fig. S2.3G)**. Based on this analysis alone, though, it was unclear whether these pause index associated changes to nucleosome architecture simply reflected sequence differences of highly paused genes, as opposed to the impact of the PPP itself.

To identify specifically PPP-associated changes, we further divided Fiber-seq reads containing an accessible promoter based on whether they also contained a PPP footprint. We then compared the positioning of surrounding nucleosome footprints based on whether that fiber contained a PPP footprint by subsampling an equal number of fibers with and without a PPP at individual promoters. We found that PPP footprints across all genes were associated with disrupted downstream nucleosome positioning, similar to the global effects seen associated with pause index. However, there was no significant difference seen for upstream nucleosomes, meaning that these previously observed differences were independent of the PPP **(Fig. 2.2B)**. This strong directionality was consistent with the lack of antisense transcription in *Drosophila* (56) and suggested an active role for the PPP in defining the nucleosome landscape.

Overall, we observed that chromatin fibers containing a PPP footprint had multiple changes in the organization of the +1 to +4 nucleosomes. First, the +1 nucleosome was shifted downstream by ~20 bp and had a footprint ~7 bp smaller, indicating that this nucleosome is both displaced and partially unwound **(Fig. 2.2B, Fig. 2.1C)**. The +2 and +3 nucleosomes also were significantly shifted downstream, ~19 bp and ~20 bp respectively, but showed no significant size difference **(Fig. 2.2B)**. Second, we observed that a significant number of chromatin fibers containing a PPP footprint were missing one of the +1 to +4 nucleosomes, indicating that PPP footprints are associated with displacement of downstream nucleosomes **(Fig. 2.2D)**. Notably, these effects were not seen upstream of the promoter or further than 1 kb downstream of the promoter and were similarly not seen in response to PIC occupancy within the promoter **(Fig.**

S2.3H). Taken together, these observations illustrate a nearby downstream nucleosome landscape transiently and significantly altered during promoter proximal pausing **(Fig. 2.2E)**.

Pause-associated disruption of downstream nucleosomes is stably maintained

Notably, we observed numerous fibers containing a shifted +1 nucleosome, yet lacking a PPP footprint **(Fig. 2.3A)**. These fibers were primarily observed at genes with a high pause index, with low pause index genes containing few fibers with a shifted +1 nucleosome and no PPP footprint **(Fig. 2.3B)**. This enrichment at highly paused genes demonstrated an association between a gene's overall PPP occupancy and the propensity of a reads from that gene to contain a shifted +1 nucleosome. To disentangle the causal relationship between PPP formation and the shifted +1 nucleosome, we performed Fiber-seq after significantly reducing the formation of PPPs using triptolide. As expected, triptolide treatment markedly reduced the proportion of reads containing a PPP footprint, yet for reads that retained a PPP footprint the absolute shift of the +1 nucleosome was unchanged **(Fig. 2.3B)**. In contrast, reads that lacked a PPP footprint had a dramatically reduced shift in the +1 nucleosome **(Fig. 2.3B)**. This reduction in the shift of the +1 nucleosome was particularly enriched at highly paused genes **(Fig. 2.3B)**, which also showed the largest reduction in PPP footprints. Notably, reads from highly paused genes that lack PPP footprints now displayed a +1 nucleosome distance that mirrored that observed on lowly paused genes in the untreated sample. Together, these data demonstrate PPP formation is directly linked to a shift in the +1 nucleosome, and that this shifted +1 nucleosome is stably maintained on that chromatin fiber after the PPP has been released.

Disrupted downstream nucleosomes are associated with active transcription

We next set out to test whether reads that lacked a PPP footprint yet retained a shifted +1 nucleosome represented reads for which Pol II had shifted into transcriptional elongation. To

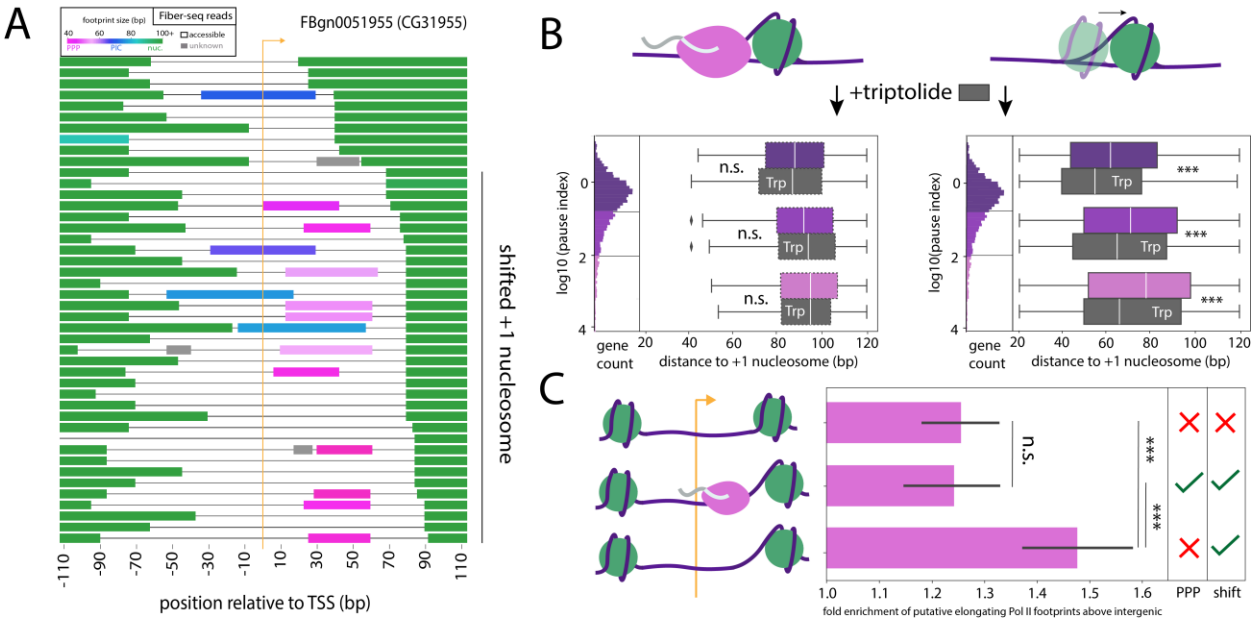


Fig. 2.3: Pausing drives disruption of downstream nucleosomes
(A) Sample of Fiber-seq reads from a representative example locus (CG31955), with footprints colored based on predicted identify (PPP=pink, PIC=blue, nucleosome=green, unknown=grey). Reads are sorted by distance to the +1 nucleosome. A line is added to indicate reads with the shift of the +1 nucleosome described in Figure 2.2b. Shifted +1 nucleosomes are defined as starting between 70 and 150bp downstream of the TSS, unshifted +1 nucleosome are defined as starting between 10 and 70bp downstream of the TSS. Note that at this locus the PPP appears to require a shifted +1 nucleosome, but the +1 nucleosome shift includes reads with no PPP. **(B)** Boxplots showing the distance to the +1 nucleosome footprint across reads with (Left) or without (Right) a PPP footprint. Reads are divided by pause index, as schematized in the associated histogram, using the same bins as in Figure 2.1d: “high” (pink, PI >= 100), “mid” (purple, 100>PI>=10), “low” (dark purple, 10>PI). +1 nucleosome distances from cells treated with triptolide are plotted in paired boxplots below the colored boxplots corresponding to untreated cells. +1 nucleosome distances for reads with a PPP (Left) do not vary based on pause index or triptolide treatment (T-test, n.s. signifies p>.05). In contrast, for reads with no PPP (Right) +1 nucleosome distances increase with pause index, and decrease with triptolide treatment (T-test, *** signifies p-value < 0.001), demonstrating that the +1 shift is dependent on and driven by pausing even on fibers with no PPP footprint. **(C)** Bar plot showing the fold enrichment of putative elongating Pol II footprints within gene bodies compared to the level of equivalently sized footprints in intergenic regions, divided by the proposed order of pausing and the shift of the +1 nucleosome. These groups are reads with no PPP and no +1 nucleosome shift (Top), a PPP and a +1 nucleosome shift (Mid), or no PPP and a +1 nucleosome shift (Bottom). Putative elongating Pol II footprints are identically enriched in reads with an unshifted +1 nucleosome or with a PPP footprint but are significantly more enriched in reads with no PPP and a shifted +1 nucleosome (T-test, n.s. signifies p>.05, *** signifies p-value < 0.001).

accomplish this, we first needed to determine whether Fiber-seq could identify putative elongating RNA Pol II footprints, which we expected should be a similar size to the PPP given their similar chromatin contacts and structure (116). Putative elongating Pol II footprints (i.e., 40-60bp footprints found within a gene body) occurred at a rate of ~1.5 per 10kb on average within gene bodies, and were markedly enriched within gene bodies for highly expressed genes based on

PRO-seq **(Fig. S2.5A)**. Furthermore, gene bodies along fibers lacking corresponding promoter accessibility were depleted in these putative elongating Pol II footprints **(Fig. S2.5B)** and reducing the ability of PICs to progress into elongation using triptolide resulted in a concomitant decrease in the rate of putative elongating Pol II footprints **(Fig. S2.5C)**. Overall, these findings suggest that the density of putative elongating Pol II footprints within a gene body is directly related to the number of expected elongating Pol II proteins for a gene.

We then set out to determine if reads with a shifted +1 nucleosome, but no PPP footprint, represented fibers undergoing elongating transcription. Notably, compared to reads without a recently released PPP (i.e., an unshifted +1 nucleosome, or a PPP footprint), reads with a recently released PPP (i.e., no PPP and a shifted +1 nucleosome) had a significantly higher density of putative elongating Pol II footprints **(Fig. 2.3C)**. Consequently, pausing establishes a state of disrupted downstream nucleosomes, which upon pause release is maintained and is associated with transcriptional elongation.

Pausing sterically inhibits initiation

Having demonstrated a coordination between pausing and elongation, we next sought to address how initiation and pausing are coordinated at individual loci. It is well established that pausing inhibits initiation, which is thought based on PPP and PIC structures to be at least partially explained by steric interference (35,36,37). Overall, we found that PPPs and PICs rarely co-occurred on the same chromatin fiber **(Fig. 2.4A)**. Notably, chromatin fibers containing simultaneous PPP and PIC footprints were observed as frequently as expected at loci with primary CAGE-seq and PRO-seq peaks separated by more than 75 bp, a distance that would always allow for unobstructed binding of both footprints **(Fig. 2.4B)**. In contrast, genes with primary CAGE-seq and PRO-seq peaks separated by fewer than 75 bp, and especially fewer than 50 bp, were significantly depleted in chromatin fibers with simultaneous PPP and PIC footprints

(Fig. 2.4B). This strong distance-dependent effect supports steric interference as the primary mechanism of pause-inhibited initiation.

Coordinated transcription initiation at neighboring genes

Having demonstrated the direct influence of transcription initiation on surrounding chromatin architectures at individual promoters, we next set out to assess if and how initiation can influence the activity of neighboring gene promoters. RNA expression studies have shown that transcription at nearby pairs of genes is correlated in *Drosophila* and other eukaryotes, with this effect predominantly seen at gene pairs located less than 1 kb from each other (50,51,52). Furthermore, imaging studies have demonstrated that RNA polymerase localization within the nucleus is often clustered (47,48,49,89,90)., suggesting a model whereby spatially proximal pairs of genes may obtain correlated expression via the coordinated occupancy of spatially clustered RNA polymerase complexes at both promoters simultaneously.

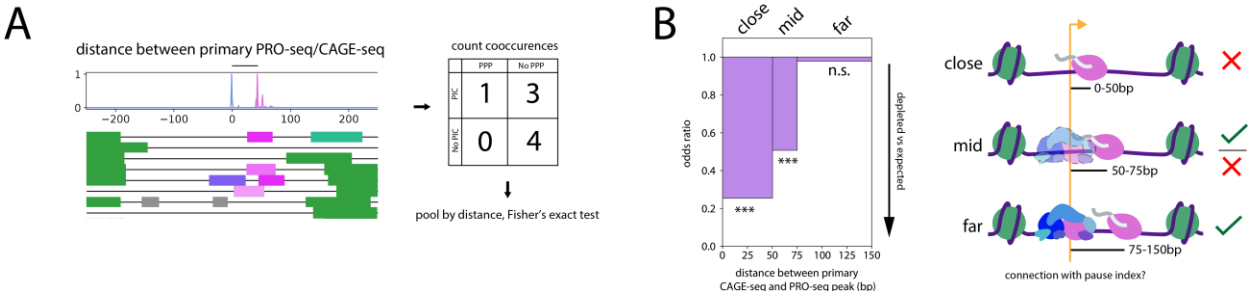


Fig. 2.4: Pausing sterically inhibits initiation
(A) (Left, Bottom) Sample of Fiber-seq reads from a representative example locus with footprints colored based on predicted identify (PPP=pink, PIC=blue, nucleosome=green, unknown=grey). (Left, Top) corresponding PRO-seq and CAGE signal at the locus. As an explanation of the methodology used in this figure, an example contingency table is depicted to the right summarizing the cooccurrence of PPP and PIC footprints at the locus. These counts are pooled between genes binned by the distance between the tallest CAGE-seq and PRO-seq peaks at the locus. An odds ratio and p-value is then determined based on Fisher's exact test carried on the pooled contingency table. **(B)** Bar plot, horizontally scaled on distance, showing the Fisher's exact test odds ratio of PPP and PIC footprint cooccupancy for each group of genes. The bins are schematized (Right) and labeled as close (0-50bp separation), mid (50-75bp separation), and far (75-150bp separation). PPP and PIC footprints cooccur on a single read at a single locus less frequency than expected compared to their overall frequency (Fisher's exact test, n.s. signifies p>.05, *** signifies p-value < 0.001) at the close and mid distances, but as frequently as expected at the far distance (Fisher's exact test, p>.05), demonstrating that inhibition of initiation by the PPP only occurs at distances where steric blocking can occur. PPP and PIC cooccurrences at close- or mid- range that do occur are explained by alternative TSSs or pause sites.

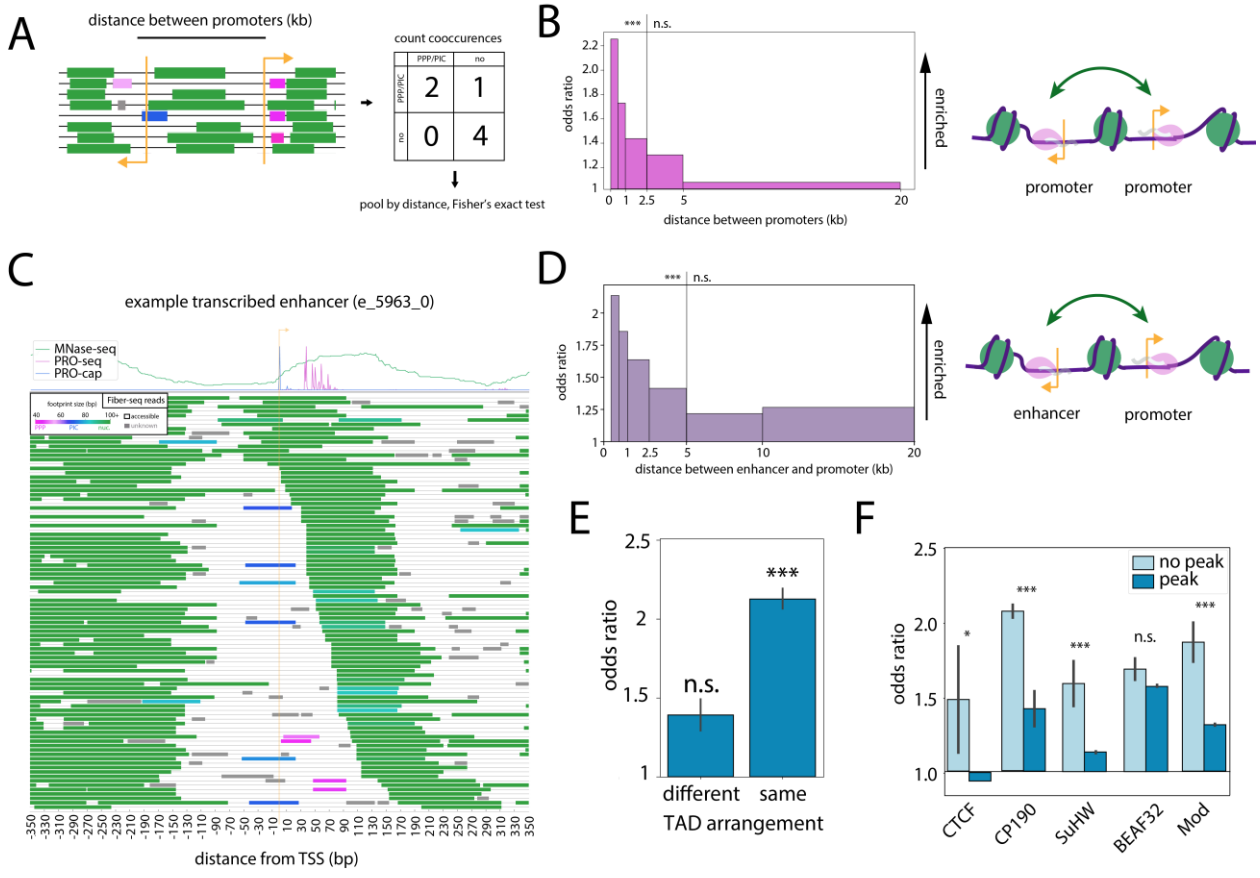
Of note, we observed that spatially proximal transcribed gene promoters were preferentially accessible along the same chromatin fiber, suggesting coordination in the chromatin architectures at neighboring gene promoters **(Fig. S2.6A)**. To directly test whether transcription initiation is coordinated between neighboring gene promoters, we specifically evaluated for transcription initiation footprints along chromatin fibers for which both promoters are accessible, thereby controlling for any confounding effects that chromatin accessibility may have had on this measurement. Overall, we found that chromatin fibers overlapping pairs of accessible promoters located within 5 kb of each other were preferentially co-bound by PPP and PIC footprints significantly more frequently than expected. Furthermore, we observed a strong distance-dependent relationship in this effect, with the strongest co-occupancy seen at paired promoters located within 1kb of each other **(Fig. 2.5B)**. These global effects were recapitulated on a per-gene basis **(Fig. S2.6B-C)**. Together, these findings provide evidence supporting the model that correlated expression of spatially proximal gene pairs is at least partially driven via the coordinated occupancy of RNA polymerase complexes at both promoters simultaneously.

Coordinated transcription initiation between enhancers and nearby genes

Many eukaryotic enhancers are transcribed by RNA polymerase II to form enhancer RNAs (eRNAs), the production of which is strongly associated with enhanced transcription at nearby genes (56,57,58). Given the coordination of transcriptional machinery between neighboring protein-coding gene promoters, we hypothesized that transcription initiation at eRNA-promoter pairs may be similarly coordinated. Using *Drosophila* S2 cell STARR-seq and START-seq data (54,56,98), we identified ~4000 enhancers that appeared to produce eRNAs. Overall, eRNA PPP and PIC footprints mirrored those at transcribed promoters, consistent with them being initiated by RNA polymerase II **(Fig. S2.7D-G)**. Like paired promoters, we found that chromatin fibers

Fig. 2.5: Distance-dependent and chromatin architecture-dependent coordination in transcription initiation

(A) Sample of Fiber-seq reads from a representative example pair of loci with footprints colored based on predicted identify (PPP=pink, PIC=blue, nucleosome=green, unknown=grey). As an explanation of the methodology used in this figure, an example contingency table is depicted to the right summarizing the cooccurrence of PPP or PIC footprints between the two example loci. These counts are pooled by the distance between the pairs of genes and an odds ratio and p-value are determined based on Fisher's exact test on the pooled contingency table. **(B)** Bar plot, horizontally scaled on distance, showing the Fisher's exact test odds ratio of PPP/PIC footprint cooccupancy on reads overlapping promoter pairs binned by distance. PPP/PIC footprints cooccur on a single read at a both loci more frequently than expected (*** signifies p-value < 0.001) at up to 2.5kb separation, showing evidence of coordinated initiation. **(C)** (Bottom) Fiber-seq reads at an example transcribed enhancer, with footprints colored based on predicted identify (PPP=pink, PIC=blue, nucleosome=green, unknown=grey). (Top) corresponding MNAse-seq, PRO-seq and START-seq signal at the locus. **(D)** Bar plot, horizontally scaled on distance, showing the Fisher's exact test odds ratio of PPP/PIC footprint cooccupancy on reads overlapping enhancer-promoter pairs binned by distance. PPP/PIC footprints cooccur on a single read at both loci more frequently than expected (p<.0001) at up to 5kb separation, showing evidence of coordinated initiation. **(E)** Bar plot depicting the coordination of transcription initiation at pairs of genes in either different or the same topologically associating domain (TAD), with all pairs being found within 2.5kb of each other. Reads from both groups are sampled to capture an identical count of reads at each distance from each group to account for the strong distance dependence shown in 2.3B and 2.3D, with error bars corresponding the confidence interval calculated from 10,000x sampling iterations. Pairs of genes in different TADs do not show significant coordination but pairs in the same TAD do show strong coordination (Fisher's exact test, n.s. signifies p>.05, *** signifies p-value < 0.001). **(F)** Bar plot depicting the coordination between pairs of genes separated (dark blue) or not separated (light blue) by a ChIP-seq peak from the corresponding factor. Distance dependence is controlled for via sampling as described in 2.3E, with error bars corresponding the confidence interval calculated from 10,000x resampling iterations. All factors except BEAF32 show a significant reduction in coordination (Fisher's exact test, *** = p<.0001, * = p <.05).


A
distance between promoters (kb)
count cooccurences
PPP/PIC
no
2
1
0
4
pool by distance, Fisher's exact test
B
***
n.s.
odds ratio
enriched
distance between promoters (kb)
promoter
promoter
C
example transcribed enhancer (e_5963_0)
MNase-seq
PRO-seq
PRO-cap
Fiber-seq reads
distance from TSS (bp)
D
***
n.s.
odds ratio
enriched
distance between enhancer and promoter (kb)
enhancer
promoter
E
odds ratio
***
n.s.
different
same
TAD arrangement
F
odds ratio
no peak
peak
*
***
***
n.s.
***
CTCF
CP190
SuHW
BEAF32
Mod

overlapping eRNA-producing enhancers and a neighboring promoter were preferentially co-bound by initiating polymerase footprints significantly more frequently than expected, with a strong distance-dependent relationship in this effect. Notably, the overall frequency of initiating polymerase footprint occupancy was similar between the eRNA-producing enhancer and the coordinated promoter, implying that neither one of them was preferentially driving this coordination **(Fig. S2.5A,B)**. Together, these findings are consistent with a model whereby the spatial organization of an eRNA-producing enhancer adjacent a promoter enables the coordinated occupancy of RNA polymerase complexes at both simultaneously, thereby potentiating the transcription of both the enhancer and gene.

Coordinated initiation is enriched within TADs and blocked at TAD boundaries

Given the importance of 3D genome architecture in spatially connecting promoter and enhancer activity (53,54,63,98), we set out to determine if the coordination in transcription initiation between pairs of genes and enhancer-promoter pairs based on 2D proximity would be affected by 3D chromatin structure. Eukaryotic genomes are partitioned into topologically associating domains (TADs), over 4,000 of which have been delineated in *Drosophila* S2 cells (59,60,61,100). We divided the pairs of genes and gene-enhancer pairs from the previous coordination analysis based on if they were both contained in the same TAD, or if they were located in separate TADs based on Hi-C data (59). Controlling for differences in the distribution of inter-promoter distances in the same-TAD and different-TAD groups, we found that fibers overlapping pairs of genes split by TAD boundaries lost all evidence of transcriptional coordination **(Fig. 2.5E).** TAD boundaries are generally defined in *Drosophila* by insulator proteins including BEAF-32, CP190, CTCF, SuHW, and Mod, and we observed that pairs of genes separated by the binding sites of these factors showed significantly lower than expected transcriptional coordination **(Fig. 2.5F)** (101,**Error! Reference source not found.**,**Error! Reference source not found.**). Together, these results demonstrate that coordination, both enhancer-promoter and

promoter-promoter, is disrupted by TAD boundaries and insulator elements, demonstrating an additional layer of regulation beyond proximity.

Identification of RNA polymerase III-associated footprints via FiberHMM

Given the ubiquity of RNA polymerase II initiation coordination, we next sought to determine whether this coordination was a feature of other RNA polymerases, specifically RNA polymerase III (64). RNA polymerase III is primarily responsible for the transcription of tRNA and 5S rRNA genes, which are often highly expressed and spatially clustered together along the genome. We first sought to determine whether FiberHMM could identify footprints at tRNA genes associated with RNA polymerase III transcription, which is mediated by TFIIIB and TFIIIC occupancy at specific sequence elements upstream of and within the tRNA gene (102,**Error! Reference source not found.**). Consistent with our expectations, we identified several distinct populations of footprints enriched at tRNA genes corresponding to the known binding sites and expected footprint sizes of TFIIIB and TFIIIC **(Fig. 2.6A)**. Additionally, the size and position of these putative Pol III transcription associated footprints was consistent across all tRNA genes **(Fig. 2.6B)**.

We next sought to determine whether the Pol III transcription associated footprints are related to tRNA expression. However, determining the expression level of individual tRNA genes using transcript-based sequencing is largely impossible as most tRNA genes in *Drosophila* have multiple copies that are identical in sequence at the transcript level. Furthermore, tRNA transcripts form strong secondary structures and undergo numerous modifications that impair standard reverse transcription-based transcript profiling methods (105,107,108,**Error! Reference source not found.**). To determine whether these Pol III transcription-associated footprints reflected the transcriptional output of each tRNA gene, we calculated a ‘transcription frequency’ for each tRNA gene, which corresponds to the fraction of reads mapping to that gene occupied by Pol III transcription associated footprints **(Fig. 2.6A)**. Notably, we observed that individual copies of the

same tRNA gene can have drastically different transcription activity scores, suggesting a high degree of variability between

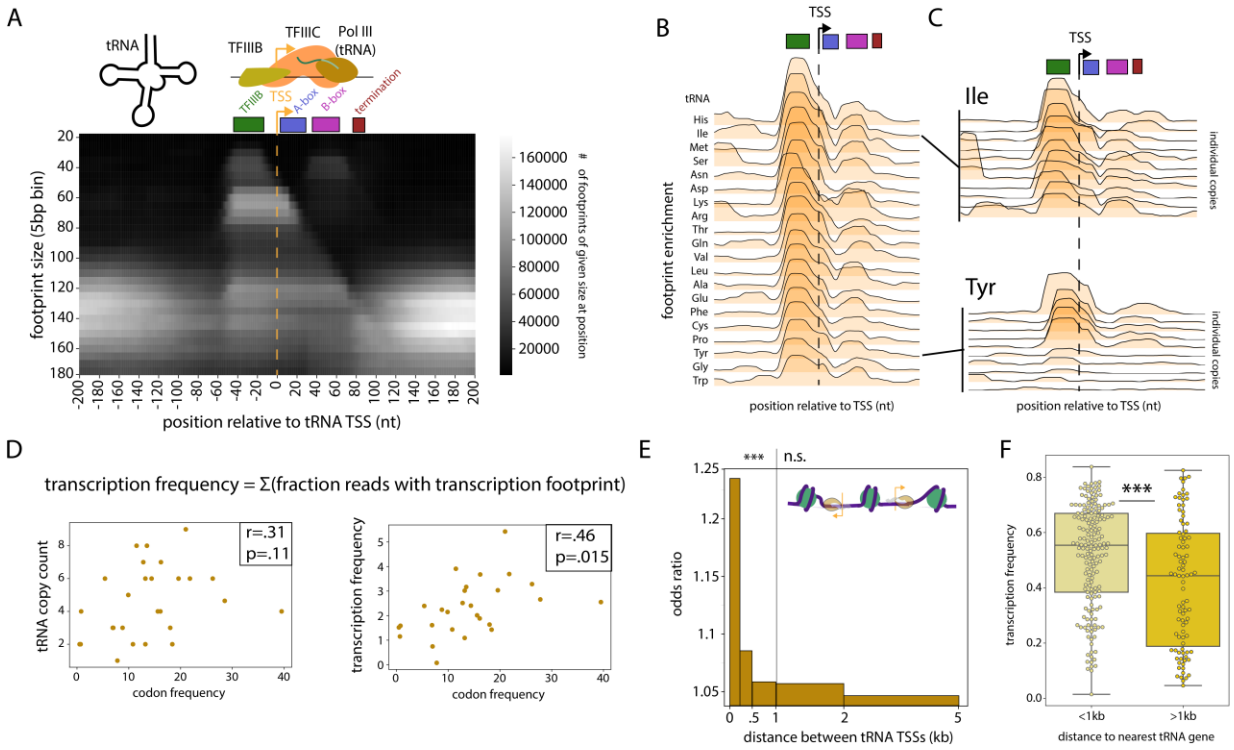


Fig. 2.6: Coordination in transcription activity between nearby tRNA genes
(A) Heatmap depicting the enrichment of differently sized Fiber-seq footprints with respect to the nearest tRNA TSS. The expected position of TFIIIB and TFIIIC are schematized as well as internal and external regulatory sequences of the tRNA. There is a strong enrichment of footprints corresponding to these sites, with the B-box associated footprints found in combination with footprints overlapping the TFIIIB site. Based on this plot, we defined Pol III transcription associated footprints as starting -55 to -45bp relative to the TSS and being 30-140bp in size. **(B)** Enrichment of 30-90bp footprints relative to the TSS split by tRNA amino acid. A similar pattern of enrichment is seen across all groups of tRNAs. **(C)** Enrichment of 30-90bp footprints relative to the TSS for individual isoleucine (Top) or tyrosine (Bottom) tRNA genes. tRNA genes are sorted by maximum footprint enrichment. Isoleucine tRNAs have moderate variation from copy to copy, but tyrosine tRNAs have strong variation, with half of the genes showing minimal evidence of expression. **(D)** (Top) transcription frequency scores were calculated for each family of tRNAs with identical anticodons and with all members having coverage in the middle 95% of overall Fiber-seq sequencing coverage. (Left) Scatterplot depicting tRNA family gene copy number plotted against their corresponding codon frequency in the *Drosophila melanogaster* genome. (Right) Scatterplot depicting tRNA isodecoder transcription frequency plotted against corresponding codon frequency in the *Drosophila melanogaster* genome. The Pearson correlation is shown in the top left corner of each plot, with only the transcription frequency correlation being significant. **(E)** Bar plot depicting the Fisher's exact test odds ratio of Pol III transcription footprint cooccupancy between pairs of tRNAs binned by distance, horizontally scaled on distance. tRNA genes found within 1kb of each other show significant coordination (Fisher's exact test, *** signifies p-value < 0.001). **(F)** Box- and swarm- plot showing the distribution of transcription frequency scores for tRNA genes either <1kb separated, or >1kb separated. tRNA genes found within 1kb of another tRNA gene have significantly higher overall transcription frequency (T-test, n.s. signifies p>.05, *** signifies p-value < 0.001).

tRNA genes that produce sequence-identical transcripts **(Fig. 2.6C)**. In addition, we found that the sum of these scores for all tRNA genes that contribute to the same codon strongly correlated with frequency of that codon within the *Drosophila* genome, indicating that the transcriptional activity score of a tRNA gene accurately captured the demand for that tRNA within *Drosophila* **(Fig. 2.6D)**.

Clustered Pol III genes exhibit coordinated transcription activity

As tRNA genes are often spatially clustered along the genome, we next tested whether neighboring Pol III tRNA promoters showed a similar coordination of transcription initiation as observed above between Pol II promoters. Overall, we observed that chromatin fibers overlapping neighboring tRNA promoters were preferentially co-bound by initiating Pol III footprints significantly more frequently than expected. In addition, we observed a strong distance-dependent relationship in this effect, with paired tRNA promoters located within 500 bp of each other showing the strongest coordination in Pol III co-occupancy **(Fig. 2.6E)**. Notably, in addition to having coordinated Pol III co-occupancy, tRNA genes located within 1 kb of each other also have significantly higher transcriptional activity scores, consistent with a model in which tRNA gene clustering enables higher overall expression via the spatial clustering of Pol III machinery **(Fig. 2.6F).**

We next wanted to determine whether the Pol III coordination observed at tRNA genes was unique to tRNA genes or a general feature of Pol III-mediated transcription initiation. RNA polymerase III is also responsible for the transcription of 5S rRNA genes, which are 120 bp genes spatially clustered together as an array of ~100 tandem copies of the same 375 bp repeat (109). Of note, this tandem duplication is highly repetitive, making reference-based assessments of co-occupancy using short read sequencing approaches largely impossible. At 5S rRNA genes we found a similar pattern of footprints as at tRNA genes, corresponding to the selective occupancy at binding sites for TFIIIA, TFIIIB, and TFIIIC **(Fig. 2.7A)** (109). As individual Fiber-seq reads on

average span ~30 repeats, we next determined whether Pol III 5S rRNA promoters showed a similar coordination of transcription initiation as observed above between Pol III tRNA promoters and between Pol II promoters. Overall, we observed that chromatin fibers overlapping neighboring 5S rRNA promoters were preferentially co-bound by Pol III transcription-associated footprints significantly more frequently than expected. In addition, we observed a strong distance-dependent relationship in this effect. However this effect was longer range than that observed at neighboring tRNA promoters, with 5S rRNA genes showing coordinated occupancy by Pol III transcription associated footprints up to 10 copies away **(Fig. 2.7B).** These results demonstrate that clustered, coordinated occupancy by Pol III is a consistent effect across the major Pol III transcribed genes, and further indicates that the condensed spatial clustering of 5S rRNA genes likely potentiates this effect.

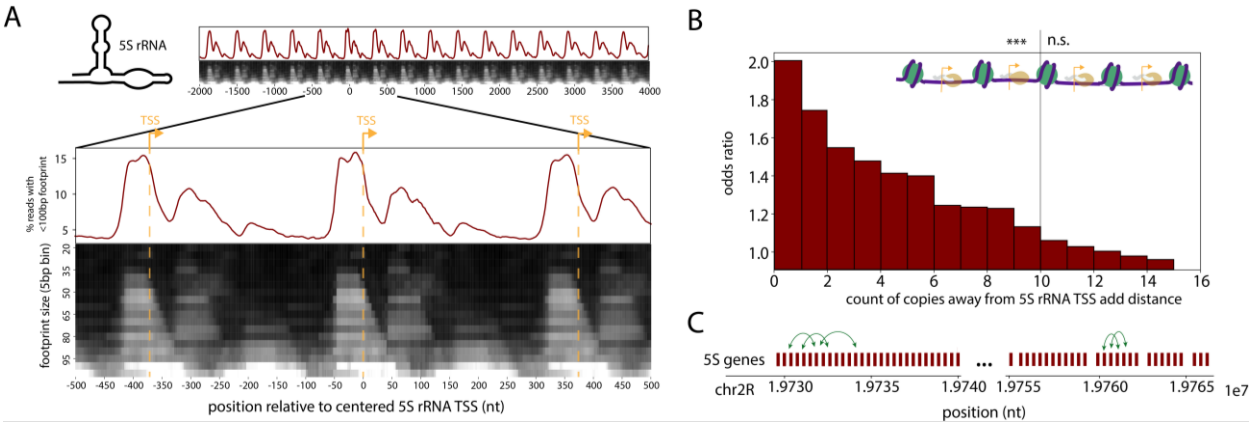


Fig. 2.7: Coordination in transcription activity between 5S rRNA genes
(A) (Overall top) wide view showing 20 5S rRNA copies. (Overall bottom) zoomed in view showing 3 5S rRNA genes. For both wide and zoomed plots: (Top) enrichment of 30-90bp footprints relative to 5S rRNA TSSs and (Bottom) heatmap depicting the enrichment of differently sized Fiber-seq footprints with respect to 5S rRNA TSSs. The overall pattern is similar to the pattern seen at tRNAs, with larger overall footprints due to TFIIIA binding. Pol III 5S rRNA transcription associated footprints were therefore defined similarly to Pol III tRNA transcription associated footprints as starting between -55 and -45nt relative to the TSS, with a size between 35 and 160 **(B)** Bar plot depicting the Fisher's exact test odds ratio of Pol III 5S rRNA transcription associated footprint cooccupancy between pairs of 5S rRNAs binned by count of 5S rRNA copies separating the pair. 5S rRNA genes found within 10 copies (~3750bp) of each other show significant coordination (Fisher's exact test, n.s. signifies p>.05, *** signifies p-value < 0.001). **(c)** 5S rRNA array found on chromosome 2R with cartoon illustrating the clustered, coordinated expression of 5S rRNA genes.

Transcription initiation is anti-coordinated between Pol II and Pol III genes

tRNA genes are often positioned adjacent pol II transcribed genes and the presence of these tRNA genes has been shown to disrupt the transcriptional output of these neighboring Pol II transcribed genes (66,111). However, the exact mechanism driving the feedback between Pol II and Pol III transcribed genes occurs is unknown, with current models suggesting that tRNA genes may act as boundary elements in a Pol III transcription dependent manner (66). To determine if transcription initiation at Pol III-transcribed tRNA genes influences Pol II-mediated transcription initiation at neighboring genes, we evaluated for transcription initiation footprints along chromatin fibers overlapping neighboring tRNA and Pol II-transcribed promoters. Overall, we observed that Pol III and Pol II genes showed the opposite effect as previously seen between pairs of Pol II promoters or pairs of Pol III genes. Specifically, we observed that PPP and PIC footprints were significantly depleted along chromatin fibers that contained Pol III transcription-associated footprints, and vice versa. This effect demonstrated a strong distance-dependent relationship, with anti-correlated Pol II and Pol III transcription initiation being observed for genes within 2.5 kb of each other **(Fig. 2.8A)**. Furthermore, as Pol III transcription-associated footprints at tRNA genes are significantly more prevalent than PPP and PIC footprints at pol II-transcribed

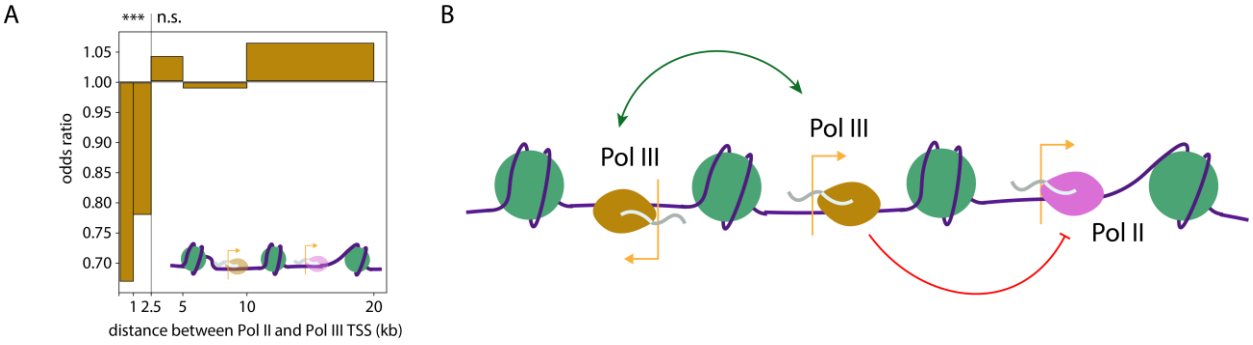


Fig. 2.8: Anti-coordination in transcription activity between Pol II and Pol III genes
(A) Bar plot depicting the Fisher's exact test odds ratio of Pol III tRNA transcription associated footprint cooccupancy with PPP/PIC footprints between pairs of tRNA and Pol II genes binned by distance, horizontally scaled on distance. tRNA genes found within 2.5kb of a Pol II gene show significant anti-coordination with the corresponding Pol II gene (Fisher's exact test, n.s. signifies p>.05, *** signifies p-value < 0.001), suggesting an insulator effect on Pol II expression. **(B)** Overall model of the coordination seen between Pol III and Pol II genes. Pol III genes are coordinated when in proximity but exhibit an anti-coordination effect on nearby Pol II genes.

genes, the outcome largely results in Pol III mediated silencing of transcription at Pol II genes. Overall, these results demonstrate that tRNA mediated silencing of Pol II-transcribed genes is largely occurring at the level of transcription initiation.

Discussion

Overall, we delineate at single-molecule and near single-nucleotide resolution the mechanisms by which transcription initiation mediates both competition and coordination along individual chromatin fibers. Specifically, we demonstrate both a mechanism for pause-inhibited initiation and pause-driven changes to the positioning and stability of nucleosome arrays that compete for the same genomic real estate. In addition, we demonstrate pervasive coordination of transcription initiation at nearby pairs of promoters and enhancer-promoter pairs, and anti-coordination between Pol II and Pol III transcribed genes. Together these findings illustrate transcription initiation as not merely a passenger occupying templated chromatin environments, but as an active participant in the regulation of gene expression and formation of chromatin architectures within these environments.

We find that Pol II pausing is directly responsible for a broad destabilization of the downstream nucleosome landscape up to 3 nucleosomes away from the TSS but has no effect on upstream nucleosomes. Furthermore, chromatin fibers maintain a short-term epigenetic memory of these destabilized downstream nucleosomes at genes with frequent pausing. Upon pause release, fibers with destabilized downstream nucleosomes appear to be actively transcribed at a higher frequency, suggesting a direct connection between pausing, altered downstream nucleosome architecture, and elongation activity. Further, the pause-associated changes to promoter structure likely permits a more efficient reestablishment of both initiation and pausing upon pause release, via maintaining both an accessible TSS and pause site, which would otherwise be occluded by nucleosomes. Consequently, this directional, pause-specific, and

expression-independent short-term epigenetic memory may serve to prime highly paused genes for rapid, regulated transcription, consistent with the patterns of expression seen at developmental and heat shock genes (28,34).

We also demonstrate the pervasive coordination of transcription initiation between spatially proximal Pol II transcribed promoters and enhancers. Coordination appears to be an effect shared across RNA polymerases and promoters and is largely limited to elements located within 1 kb of each other. Spatial proximity as a driver of synchronized, coregulated transcription provides a mechanistic basis for enhancer RNA expression, and the clustering of functionally similar genes within the *Drosophila* genome. In addition, spatial coordination helps explain the marked variability in expression we observe between otherwise identical Pol III transcribed tRNA genes, with higher or lower transcription activity based on their proximity to other tRNA genes. Together, these findings highlight the intricate dance that polymerases and the rest of the chromatin-bound proteome perform along individual chromatin fibers within the cell, and ground prior evolutionary-, microscopy-, and genomics-based observations within a single-molecule framework of competition and coordination of polymerase and the surrounding landscape of proteins on individual chromatin fibers

Chapter 3: Single-nucleoid architecture reveals heterogeneous packaging of mitochondrial DNA

<u>This chapter has been adapted from an accepted manuscript:</u> R. Stefan Isaac, Thomas W. Tullius, Katja G. Hansen, Danilo Dubocanin[2], Mary Couvillion, Andrew B. Stergachis, L. Stirling Churchman. **Single-nucleoid architecture reveals heterogeneous packaging of mitochondrial DNA.** ***Nature Structural and Molecular Biology. Accepted September 2023.***

Author Contributions

Conceptualization, R.S.I and L.S.C.; Methodology, R.S.I. (lead), K.G.H, T.W.T., A.B.S., and L.S.C.; Investigation, R.S.I. (lead) and K.G.H.; Formal Analysis, R.S.I, K.G.H., T.W.T., M.C., and D.D.; Writing - Original Draft, R.S.I.; Writing - Review & Editing, R.S.I, K.G.H., T.W.T., M.C., A.B.S, and L.S.C.; Funding Acquisition, R.S.I., K.G.H., A.B.S. and L.S.C.; Supervision, A.B.S. and L.S.C.

Abstract

Cellular metabolism relies on the regulation and maintenance of mitochondrial DNA (mtDNA). Hundreds to thousands of copies of mtDNA exist in each cell, yet because mitochondria lack histones or other machinery important for nuclear genome compaction, it remains unresolved how mtDNA is packaged into individual nucleoids. In this study, we used long-read single-molecule accessibility mapping to measure the compaction of individual full-length mtDNA molecules at nucleotide resolution. We found that, unlike the nuclear genome, human mtDNA largely undergoes all-or-none global compaction, with the majority of nucleoids existing in an inaccessible, inactive state. Highly accessible mitochondrial nucleoids are co-occupied by transcription and replication components and selectively form a triple-stranded D-loop structure. In addition, we showed that the primary nucleoid-associated protein TFAM directly modulates the fraction of inaccessible nucleoids both *in vivo* and *in vitro*, acting consistently with a nucleation-and-spreading mechanism to coat and compact mitochondrial nucleoids. Together, these findings reveal the primary architecture of mtDNA packaging and regulation in human cells.

Introduction

Originating from a eubacterial ancestor, mitochondria have retained a small (16.5 kb) circular genome that is present across the mitochondrial network in hundreds to thousands of copies per cell (**Error! Reference source not found.**). Human mitochondrial DNA (mtDNA) encodes thirteen core subunits of the oxidative phosphorylation (OXPHOS) complexes that are assembled with nuclear-encoded subunits at the mitochondrial inner membrane. The majority of the mitochondrial genome is coding, with two polycistronic transcripts originating from transcription start sites (TSS) that direct transcription along either the heavy or light strand; the names of the two strands reflect their distinct sedimentation properties **(Fig. 1a)**. The primary non-coding region (NCR) of mtDNA contains an origin of replication (O_H) and a displacement loop (D-loop) with unknown function. The D-loop is present in a fraction of human mtDNA molecules and contains a 400–650 nucleotide third strand of DNA, the 7S DNA, whose synthesis depends on the transcription and replication machinery (46,73,**Error! Reference source not found.**,**Error! Reference source not found.**,**Error! Reference source not found.**,**Error! Reference source not found.**,**Error! Reference source not found.**,**Error! Reference source not found.**). Mutations in mtDNA and its associated proteins cause numerous inherited and acquired diseases, and misregulation of mtDNA expression is implicated in neurodegenerative disorders, cancers, and aging-related illnesses (**Error! Reference source not found.**,**Error! Reference source not found.**,**Error! Reference source not found.**,123,**Error! Reference source not found.**,**Error! Reference source not found.**,**Error! Reference source not found.**).

Although mtDNA has a contour length of over 5 µm, super-resolution imaging studies have shown that it is compacted into 100-nanometer nucleoprotein complexes called nucleoids (112,113). However, the principles balancing mitochondrial nucleoid compaction, replication, and transcription remain poorly understood. Unlike nuclear chromosomes, mtDNA is not packaged with histones and their associated machinery. The mitochondrial transcription factor TFAM is the

primary constituent of mitochondrial nucleoids (**Error! Reference source not found.**,**Error! Reference source not found.**). High levels of TFAM compact DNA *in vitro* (70,130,130), yet the architecture formed by TFAM packaging and the extent to which other proteins contribute in cells remain unknown (**Error! Reference source not found.**,**Error! Reference source not found.**,**Error! Reference source not found.**,**Error! Reference source not found.**,**Error! Reference source not found.**,137).

To address these questions, we measured DNA accessibility of individual nucleoids at near single-nucleotide and single-molecule resolution, generating full-length accessibility profiles of individual mtDNA molecules. The results showed that the majority of nucleoids are in an inaccessible state, with the remainder exhibiting heterogeneity in packaging. Thus, to first order, mtDNA accessibility is all-or-none, in contrast to nuclear genome organization, which is characterized by local differences in chromatin accessibility. We found that TFAM levels directly modulate the fraction of accessible nucleoids in cells. We identified footprints from transcription and replication factors across the accessible nucleoids, resolving their co-occupancy across single molecules. Comparing footprinting patterns between nucleoids reconstituted *in vitro* and nucleoids in cells demonstrated that TFAM alone explains the majority of mtDNA packaging. These patterns were consistent with a nucleation-and-spreading model whereby TFAM binds first to high-affinity sites and subsequently binds cooperatively to coat and compact the genome.

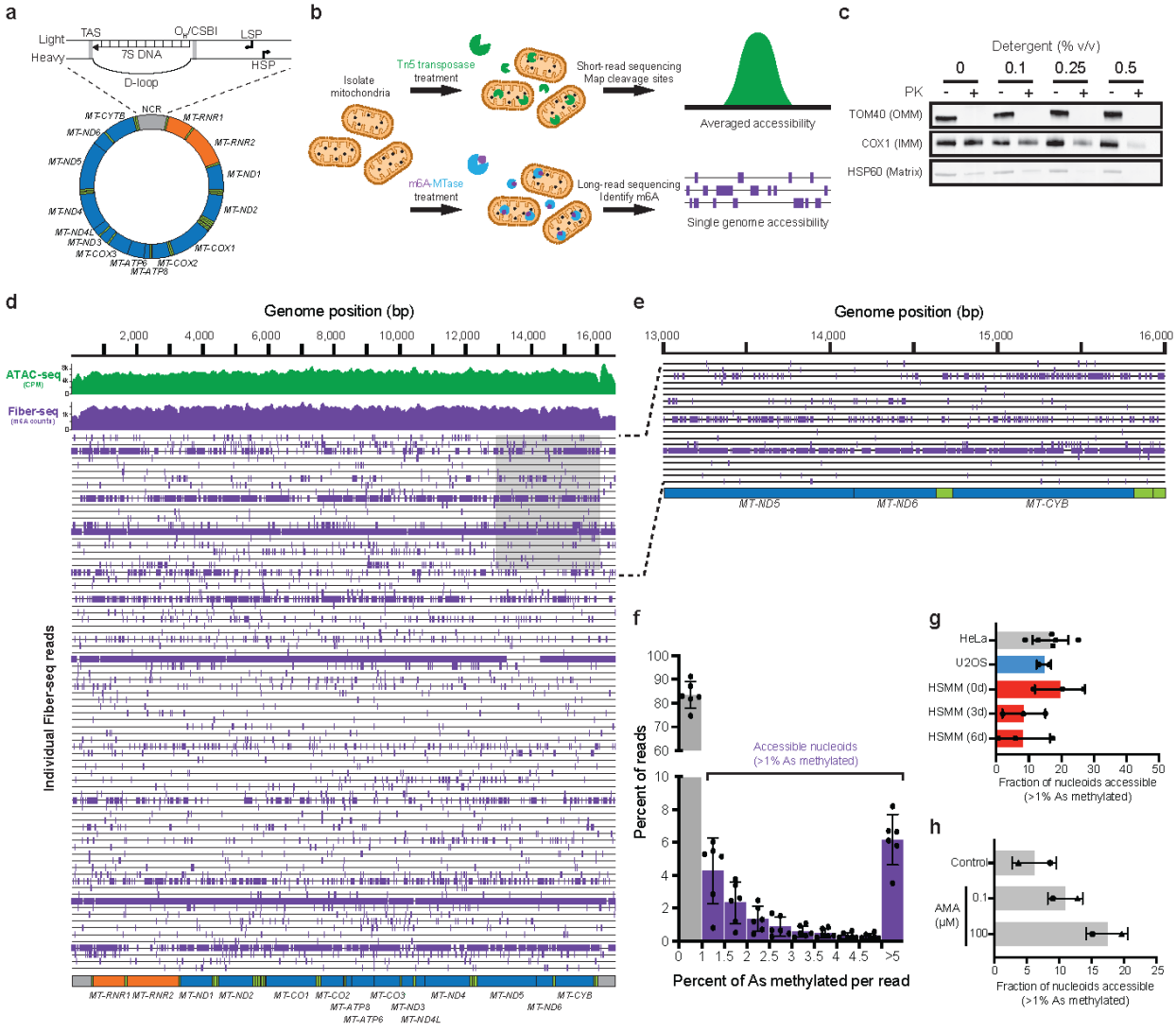
Results

mtFiber-seq probes individual nucleoid accessibility

To investigate mtDNA accessibility patterns, we first subjected isolated mitochondria to Assay for Transposase-Accessible Chromatin (ATAC-seq) **(Fig. 3.1b)**, which relies on the selective cleavage of accessible DNA by the hyperactive Tn5 transposase (**Error! Reference source not found.**). Similar to other short-read sequencing-based analyses of mtDNA

accessibility (139,**Error! Reference source not found.**,**Error! Reference source not found.**), ATAC-seq revealed near-uniform accessibility across the mitochondrial genome **(Fig. 3.1d)**, irrespective of the NCR and coding regions. Such homogeneity could arise from bulk averaging of accessibility across individual mitochondrial nucleoids. Since hundreds of copies of mtDNA are present in each cell, even single-cell based accessibility methods would result in bulk averaging, and the dynamic, interconnected mitochondrial network prevents a 'single-mitochondrion' approach. To overcome these barriers, we sought to leverage recent single-molecule chromatin fiber sequencing

Fig. 3.1. The majority of human mitochondrial nucleoids are inaccessible
(a) Schematic of the human mitochondrial genome. (Top) The non-coding region (NCR) of the mitochondrial genome, highlighting the Termination Association Sequence (TAS), the D-loop and 7S DNA, the heavy strand origin of replication (O_H) and Conserved Sequence Box I (CSBI), and the light and heavy strand promoters (LSP and HSP, respectively). (Bottom) The 16.5-kb circular human mitochondrial genome encodes 13 polypeptides (blue), 22 tRNAs (green), and 2 rRNAs (orange). **(b)** Schematic depicting the experimental design. Human cells were subjected to cellular fractionation to isolate mitochondria, which were then permeabilized and treated with Tn5 Transposase for ATAC-seq or the Hia5 m6A-MTase for mtFiber-seq. The resultant ATAC-seq libraries were then subjected to Illumina short-read sequencing. For mtFiber-seq, mtDNA was first linearized using a restriction enzyme with a single cut site before library preparation and PacBio sequencing. **(c)** Permeabilization of mitochondria with increasing concentrations of Tween-20 and NP-40 was assessed by Proteinase K digestion of the outer mitochondrial membrane (OMM) protein TOM40, the inner mitochondrial membrane (IMM) protein COX1, and the matrix-soluble protein HSP60, followed by western blot analysis. **(d)** Mitochondrial genome comparing the signal between ATAC-seq, aggregated methylation from mtFiber-seq, and 80 randomly sampled individual mtFiber-seq reads. Individual PacBio reads are indicated by horizontal black lines, and m6A-modified bases are marked by purple vertical dashes. **(e)** Zoom-in of positions 13,000–16,000 in the mitochondrial genome showing 20 sampled mtFiber-seq


a
Light
Heavy
TAS
OH/CSBI
LSP
7S DNA
HSP
D-loop
NCR
MT-CYTB
MT-RNR1
MT-ND6
MT-RNR2
MT-ND5
MT-ND1
MT-ND4
MT-ND2
MT-ND4L
MT-ND3
MT-COX3
MT-COX1
MT-ATP6
MT-COX2
MT-ATP8
b
Isolate mitochondria
Tn5 transposase treatment
Short-read sequencing
Map cleavage sites
Averaged accessibility
m6A-MTase treatment
Long-read sequencing
Identify m6A
Single genome accessibility
c
Detergent (% v/v)
0
0.1
0.25
0.5
PK
- + - + - + - +
TOM40 (OMM)
COX1 (IMM)
HSP60 (Matrix)
d
Genome position (bp)
2,000
4,000
6,000
8,000
10,000
12,000
14,000
16,000
ATAC-seq
(CPM)
Fiber-seq
(m6A counts)
Individual Fiber-seq reads
MT-RNR1 MT-RNR2 MT-ND1 MT-ND2 MT-CO1 MT-CO2 MT-ATP8 MT-ATP6 MT-CO3 MT-ND3 MT-ND4L MT-ND4 MT-ND5 MT-ND6 MT-CYB
e
Genome position (bp)
13,000
14,000
15,000
16,000
MT-ND5
MT-ND6
MT-CYB
f
Percent of reads
Accessible nucleoids
(>1% As methylated)
Percent of As methylated per read
0 1 1.5 2 2.5 3 3.5 4 4.5 >5
g
HeLa
U2OS
HSMM (0d)
HSMM (3d)
HSMM (6d)
Fraction of nucleoids accessible
(>1% As methylated)
h
Control
AMA (μM)
0.1
100
Fraction of nucleoids accessible
(>1% As methylated)

approaches (22,23,24,91,92). Specifically, we adapted the Fiber-seq method (23) for mtDNA, because it was capable of evaluating the accessibility along both the heavy and light strands from the same 16.5 kb mtDNA molecule. Fiber-seq uses a nonspecific adenine methyltransferase (MTase) to mark accessible DNA and application of Fiber-seq to the nuclear genome mirrors DNase I cleavage in bulk (24). The use of an adenine MTase offers superior resolution of the mitochondrial genome relative to an approach based on GC dinucleotide methylation (**Error! Reference source not found.**,**Error! Reference source not found.**) due to the frequency of adenines and distances between them **(Fig. S3.1a, Figure S3.1b)**. Fiber-seq 'stencils' the architecture of chromatin onto underlying DNA fragments via the selective modification of accessible A/T base pairs with N^6-methyl-deoxyadenosine (m6A). The m6A residues are then identified using highly accurate long-read sequencing of individual double-stranded 15-20 kb DNA fragments. We adapted this protocol to investigate the packaging of individual mitochondrial nucleoids **(Fig. 3.1b)**. For mitochondrial Fiber-seq (mtFiber-seq), isolated mitochondria were permeabilized **(Fig. 3.1c)** and then treated with the m6A-MTase Hia5. After DNA isolation, mtDNA was linearized and subjected to PacBio HiFi single-molecule sequencing, enabling identification of modified adenines along each sequenced molecule of DNA.

Application of mtFiber-seq to human HeLa cells yielded reads spanning the full 16.5-kb mitochondrial genome **(Fig. 3.1d, Figure 3.1e)**, with aggregated methylation profiles mirroring our observations by ATAC-seq **(Fig. 3.1d)**. However, methylation patterns along individual reads revealed substantial intra- and inter-read heterogeneity in the accessibility of mitochondrial nucleoids that had previously been obscured by bulk-averaged accessibility measurements. In the vast majority of nucleoids (~80%), less than 1% of their total adenines were methylated, a methylation level approaching that of mtDNA untreated with MTase **(Fig. S3.1c)**, indicating that these nucleoids are compacted in a predominately inaccessible state **(Fig. 3.1f)**. In the remaining population of nucleoids, the accessibility varied, with a maximum methylation level of ~80% of adenines.

To eliminate the possibility that the observed phenomenon is due to subsaturating reaction conditions, we varied the concentration of MTase and treatment times used in the mtFiber-seq reactions. These conditions did not substantially alter the proportion of nucleoids within the inaccessible state, demonstrating that this minimally methylated population is not due to insufficient enzyme or time **(Fig. S3.1d, Figure S3.1e)**. To determine whether this phenomenon is unique to HeLa S3 cells, we performed mtFiber-seq in U2-OS cells and in human skeletal muscle myoblasts (HSMM) at several stages of differentiation, a process associated with both changes in mitochondrial biogenesis and proliferation (**Error! Reference source not found.**). In all of these cell types, the accessible population of nucleoids remained in the minority **(Fig. 3.1g).** To assess whether this proportion of nucleoids is static, we treated HeLa cells with antimycin A, a Complex III inhibitor which stimulates oxidative stress. After treatment for 24 hours, we saw an increase in the accessible nucleoid population compared to the vehicle control (Fig. **3.1h**). Together, these results reveal that mitochondrial nucleoids undergo heterogeneous compaction, with the majority of nucleoids existing in a minimally accessible, compacted state, and that this population can shift in response to perturbed mitochondrial function.

To orthogonally validate this finding and resolve whether the variability in mtDNA accessibility is driven by inter- or intracellular differences in mtDNA compaction, we adapted and performed ATAC-see for mitochondria. ATAC-see uses Tn5 transposase to ligate fluorescent oligos into accessible DNA, which are visualized alongside mtDNA by fluorescence microscopy (**Error! Reference source not found.**) **(Fig. S3.3a, Figure S3.3b, Figure S3.3c)**. In bulk, ATAC-see signal at mtDNA puncta was distributed broadly but skewed towards lower intensities, with the majority of puncta exhibiting a low signal, mirroring the compacted state we observed with mtFiber-seq **(Fig. S3.3d)**. At the single-cell level, the ATAC-see and DNA signal distributions were similar to those in our population-level analysis, revealing that nucleoid accessibility is also highly variable across

individual nucleoids within a given cell **(Fig. S3.3e)**. Together, these results demonstrate that the majority of nucleoids are inaccessible and indicate that this heterogeneity is present in each cell.

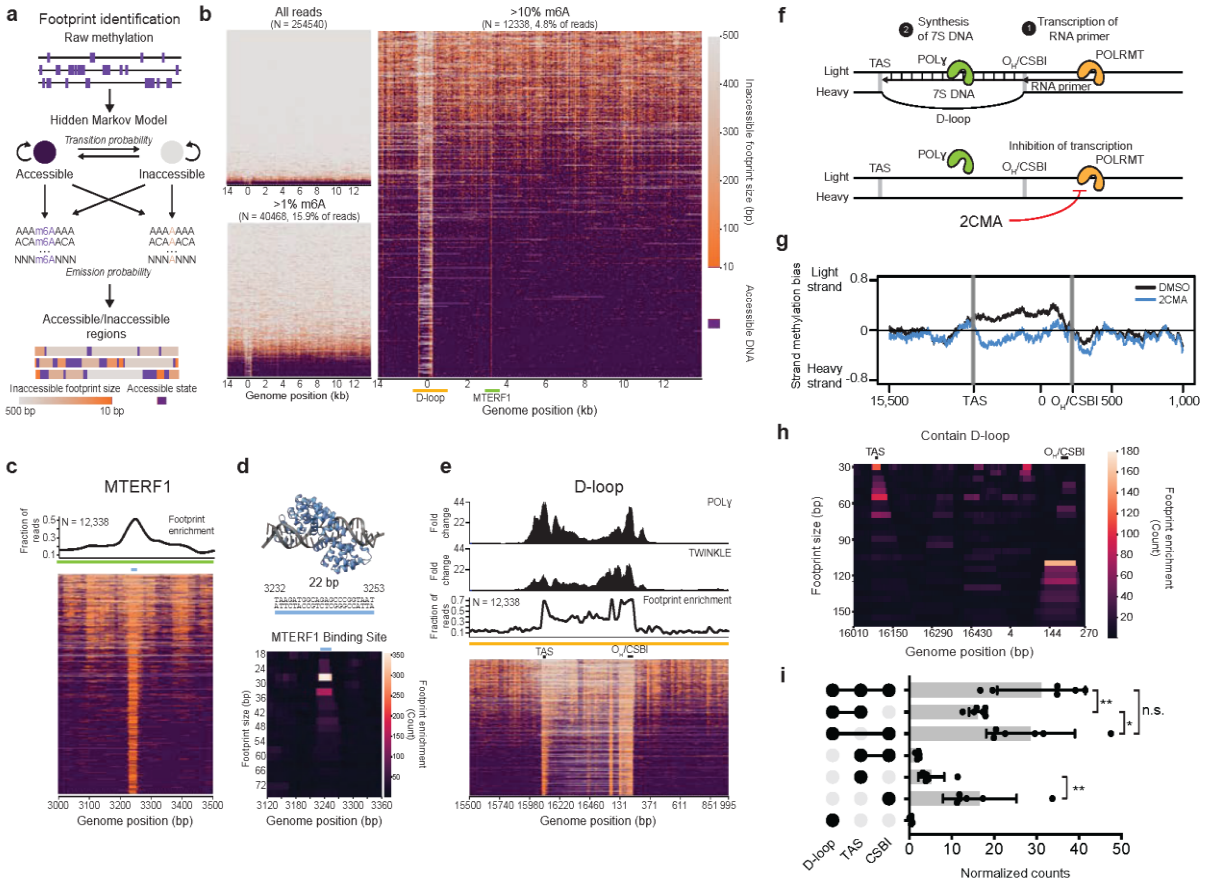
Single-molecule protein occupancy of accessible nucleoids

We next sought to determine whether accessible nucleoids exhibited features of mitochondrial transcription and replication, processes that are regulated by multiple DNA-binding proteins. To identify protein occupancy along each nucleoid, we developed a hidden Markov model (HMM)-based footprinting approach that controls for local methylation propensities **(Fig. 3.2a)**. Application of this HMM to accessible nucleoids from six HeLa datasets revealed diverse single-molecule patterns of protected footprints **(Fig. 3.2b)**. For example, within the mt-tRNA Leu(UUR) gene, 50% of highly accessible nucleoids contained a ~30 bp footprint overlapping the known binding site of the transcription termination factor MTERF1 (78) **(Fig. 3.2c)**, consistent with occupancy by a single MTERF1 molecule **(Fig. 3.2d)**. These results demonstrate that accessible nucleoids are preferentially bound by MTERF1 and highlight the power of mtFiber-seq to capture multiple molecular features on the same DNA molecules.

Single-molecule architecture of the mitochondrial D-loop

Seventy percent of highly accessible nucleoids exhibited protection of the D-loop region **(Fig. 3.2e)**, with single-molecule footprints localizing to two conserved sequence elements of the D-loop: the Termination Associated Sequence (TAS) and the Conserved Sequence Box I (CSBI). Notably, these single-molecule footprints mirrored ChIP-seq profiles from the DNA polymerase Polɣ and the replicative helicase TWINKLE (**Error! Reference source not found.**) **(Fig. 3.2e, top)**, indicating that the footprint patterns were formed, in part, by the residence of these factors at the TAS and CSBI. Notably, most reads with a TAS or CSBI footprint also contained an MTERF1 footprint, signifying that these factors were preferentially co-occupying the same nucleoids **(Fig. S3.5a)**.

Fig. 3.2. Accessibility patterns reveal the mtDNA architecture
(a) Schematic of HMM used to identify accessible and inaccessible regions of the genome using probabilities of methylation in hexamer sequence contexts from methylation of bare DNA. **(b)** Heatmaps showing the identified accessible and inaccessible regions of the genome for all reads, reads containing >1% m6A, and reads containing >10% m6A from six biological replicates from HeLa cells. Each row represents an individual read. Accessible regions are colored purple, and protected regions are colored according to size. **(c)** (Top) Metaplot of footprint enrichment showing the fraction of reads with >10% m6A protected at each position. (Bottom) Zoom-in showing the region surrounding the MTERF1 binding site within the tRNA-Leu(UUR) gene for reads containing >10% m6A. The 22 bp binding site is indicated by a blue horizontal line. **(d)** (Top) The 22 bp sequence recognized by MTERF1 and the crystal structure of MTERF1 bound to this sequence (PDB code 3MVA[65]). (Bottom) Heatmap of the footprint size enrichment at the MTERF1 binding site. Each row represents a footprint size, and each column shows a position in the genome. The 22 bp binding site is indicated by a blue horizontal line. **(e)** (Top) ChIP-seq tracks showing the fold change of signal over the respective control antibodies for Polɣ and TWINKLE[42]. Metaplot of the footprint enrichment showing the fraction of reads with >10% m6A protected at each position. (Bottom) Zoom-in showing the region surrounding the D-loop. Heatmap shows reads containing >10% m6A. **(f)** Cartoon depicting D-loop formation. POLRMT transcribes the RNA primers for POLɣ to synthesize the 7S DNA, which hybridizes with the light strand forming a D-loop. 2'-C-methyladenosine (2CMA) inhibits POLRMT, resulting in no 7S DNA synthesis or D-loop formation. **(g)** mtFiber-seq methylation strand bias at the NCR in HeLa cells treated with DMSO or the transcription inhibitor 2CMA. Methylation bias is calculated as the number of methylations on the light strand and heavy strand, averaged over a 150-nt sliding window and normalized against the region's AT content. Each window was required to have at least 2,250 methylations across all reads combined. **(h)** Heatmap of the footprint size enrichment at the D-loop region from HeLa cells in reads containing a D-loop. Each row represents a footprint size, and each column shows a position in the genome. Presence of a D-loop was calculated using a GMM with a threshold of 3.01 from the natural log distribution of the ratio of light strand and heavy strand methylation levels. **(i)** UpSet plot showing the normalized frequency of nucleoids containing combinations of footprints at TAS and CSBI and containing a D-loop. Maximum footprint sizes of 60 bp and 140 bp were used for TAS and CSBI, respectively. Groups were compared by paired Student's t-test, n.s. Signifies p-value > 0.05, * signifies p-value < 0.05, ** signifies p-value < 0.01. Individual dots represent six biological replicates. The categories in the plot represent 2.10%, 2.31%, 1.89%, 1.01%, 2.81%, and 2.67% of the total molecule population for replicates 1-6, respectively.


a
Footprint identification
Raw methylation
Hidden Markov Model
Transition probability
Accessible
Inaccessible
Emission probability
Accessible/Inaccessible regions
Inaccessible footprint size
Accessible state
500 bp
10 bp
b
All reads
(N = 254540)
>10% m6A
(N = 12338, 4.8% of reads)
>1% m6A
(N = 40468, 15.9% of reads)
Genome position (kb)
D-loop
MTERF1
Inaccessible footprint size (bp)
Accessible DNA
c
MTERF1
N = 12,338
Footprint enrichment
Fraction of reads
Genome position (bp)
d
22 bp
MTERF1 Binding Site
Footprint size (bp)
Footprint enrichment (Count)
Genome position (bp)
e
D-loop
POLγ
TWINKLE
Fold change
N = 12,338
Footprint enrichment
TAS
O_H/CSBI
Genome position (bp)
f
Synthesis of 7S DNA
Transcription of RNA primer
Light
Heavy
TAS
POLγ
O_H/CSBI
POLRMT
7S DNA
RNA primer
D-loop
Inhibition of transcription
2CMA
g
Strand methylation bias
Light strand
Heavy strand
DMSO
2CMA
TAS
O_H/CSBI
h
Contain D-loop
Footprint size (bp)
Footprint enrichment (Count)
Genome position (bp)
i
n.s.
D-loop
TAS
CSBI
Normalized counts

In some mtDNA molecules, the D-loop structure forms through 7S DNA hybridization to the light strand of the D-loop locus, displacing the heavy strand **(Fig. 3.2f, top)**. The structure remains enigmatic as it has been challenging to isolate the D-loop containing nucleoid population that is typically in the vast minority. To quantify D-loop formation among individual nucleoids, we leveraged two important features unique to mtFiber-seq. First, the Hia5 MTase is >15-fold less active on single-stranded DNA than on double-stranded DNA **(Fig. S3.5b)**, so the methylation patterns on the light and heavy strands across the D-loop region are predicted to differ when the 7S DNA is present. Second, PacBio sequencing detects methylation on both strands, so strand-specific methylation within a region can be calculated by averaging the number of m6As identified on each strand, normalized against the local AT content. Using this approach, we found that outside of the D-loop both the light and heavy strands were methylated to a similar extent **(Fig. S3.5c,e)**. However, specifically between the TAS and CSBI of the D-loop region, we observed more methylation on the light strand **(Fig. 3.2g, Fig. S3.5d,e)**. If the methylation strand bias is due to the presence of a 7S DNA molecule forming a D-loop, then this strand bias should disappear when replication is inhibited **(Fig. 3.2f, bottom)**. To test this, we treated cells with 2'-C-methyladenosine (2CMA), a nucleoside analogue that preferentially inhibits mitochondrial transcription **(Fig. S3.5f)** and thus blocks mtDNA replication by depleting the essential RNA primers (76,**Error! Reference source not found.**,**Error! Reference source not found.**,**Error! Reference source not found.**). Treatment of HeLa cells with 2CMA for 24 hours led to a loss of strand bias within the D-loop region, suggesting that the bias was indeed caused by the D-loop structure **(Fig. 3.2g, Fig. S3.5c-e)**. In total, an average of 4% of nucleoids from HeLa cells contained a 7S DNA **(Fig. S3.5g)**, consistent with previously measured 7S DNA content in this cell type (**Error! Reference source not found.**).

The single-molecule resolution of mtFiber-seq allows for the identification of regulatory features specifically associated with nucleoids containing a D-loop. Individual nucleoids with strand-specific methylation at the D-loop preferentially featured footprints that precisely

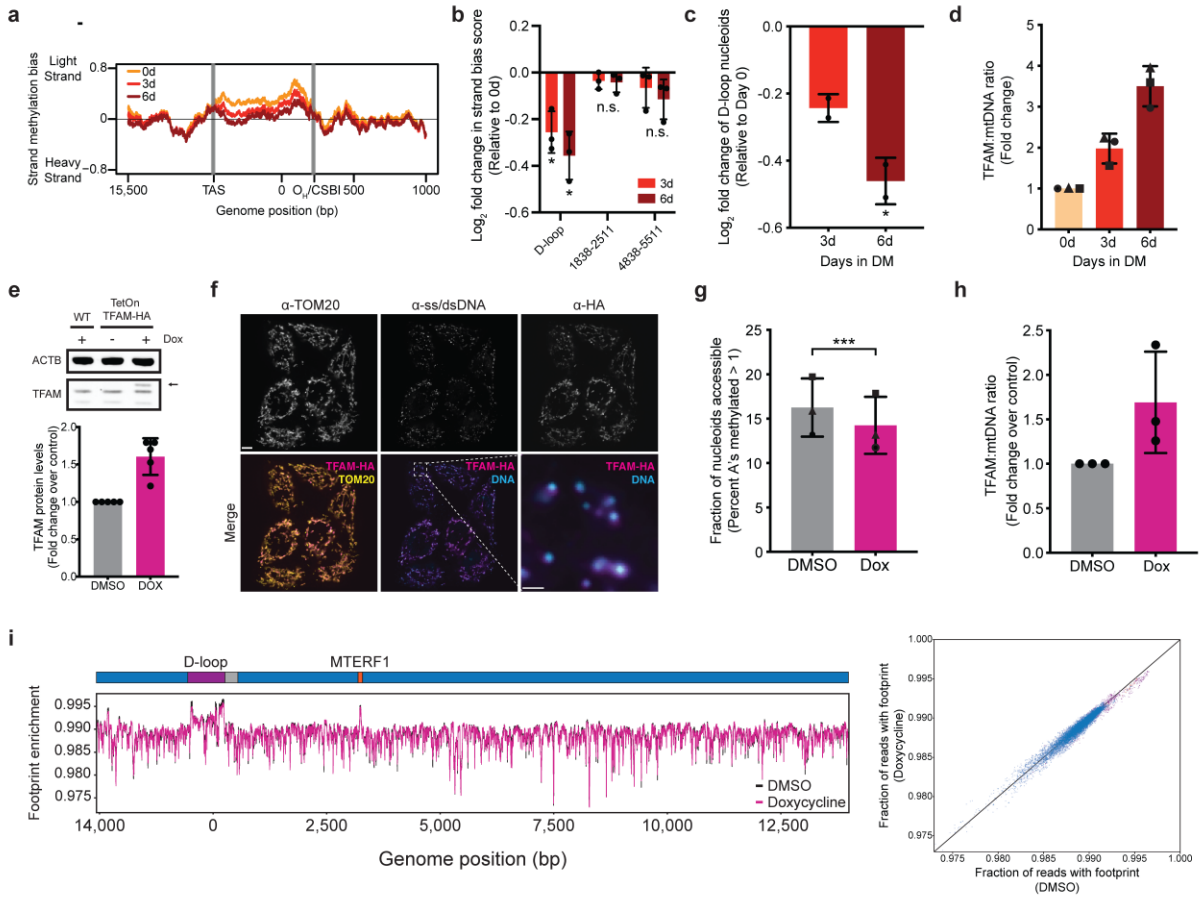
overlapped the TAS and/or CSBI, indicating that mtDNA molecules with 7S DNA are likely actively engaged by replication machinery **(Fig. 3.2h,i, S3.6b,c)**. The pronounced footprint at the TAS is consistent with a paused or terminating Polɣ, congruent with the proposed origin of the 7S DNA and frequently abortive nature of mtDNA replication (**Error! Reference source not found.**,**Error! Reference source not found.**). Interestingly, we observed nucleoids lacking a D-loop that contained one, but not both, replication machinery-associated footprints. Separating these nucleoids into those protected at the TAS and CSBI revealed that the majority of these were protected only at CSBI, suggesting a poised, pre-replication state (**Fig. 3.2i**). Together, these results demonstrate that nucleoids containing a 7S DNA are co-occupied with replication machinery at one or both ends of the D-loop. This suggests that D-loop containing nucleoids are continuously cycling through replication initiation and premature termination, consistent with the high turnover of the 7S DNA and indicating that the D-loop structure is the natural by-product of these cycles (46,**Error! Reference source not found.**).

mtDNA accessibility decreases during myoblast differentiation

To understand how nucleoid accessibility changes in a biological context, we analyzed mtDNA packaging during human skeletal muscle myoblasts (HSMM) differentiation, in which cells induce mitochondrial biogenesis **(Fig. S3.7a,b)**, exit the cell cycle, and fuse together to form multinucleated myotubes (**Error! Reference source not found.**,**Error! Reference source not found.**). We observed a decrease in the accessible nucleoid population during the first three days of differentiation that was maintained at the 6 days timepoint **(Fig. 3.1g)**. Closer inspection of these nucleoids revealed changes to the D-loop region that indicate a reduction in mtDNA replication. We found a significant decrease in the methylation strand bias within the D-loop at 3 and 6 days, but not at other loci **(Fig. 3.3a,b)**, similar to what we observed with 2CMA treatment (**Fig. 3.2g, Fig. S3.5d,e**), indicating lower levels of 7S DNA synthesis and mtDNA replication. We

also subsampled the datasets to account for changes in the overall methylation distribution and classified each read as containing a D-loop or not. We found that the proportion of reads with D-loops decreased across differentiation **(Fig. 3.3c)**. In

Fig. 3.3: Altered TFAM levels shift the population of accessible nucleoids
(a) mtFiber-seq methylation strand bias at the NCR in human skeletal muscle myoblasts (HSMM) throughout differentiation. Methylation bias is calculated as the number of methylations on the light and heavy strands, averaged over a 150 nt sliding window and normalized against the region's AT content. Each window was required to have at least 2,250 methylations across all reads combined **(b)** Log_2 fold change in the total strand bias score after 3 and 6 days in differentiation media (DM) relative to undifferentiated myoblasts across three genomic regions. Samples were compared with a Student's t-test, * signifies p-value < 0.05. Results from 3 biological replicates shown. **(c)** Log_2 fold change of D-loop containing nucleoids from HSMMs after 3 and 6 days in DM relative to undifferentiated myoblasts. Datasets were subsampled to each other to match methylation distributions. Samples were compared by Fisher's exact tests, * signifies p-value < 0.05. Result from 2 biological replicates shown. **(d)** Bar plot showing the ratio of TFAM protein levels to relative mtDNA levels as measured by western blot and qPCR from HSMMs after 0, 3, and 6 days in DM. Results from 3 biological replicates shown. **(e)** Overexpression of TFAM in HeLa cells. WT HeLa cells and TetOn-TFAM-HA HeLa cells were treated with DMSO or 100 ng/mL doxycycline for 48 hours. (Top) TFAM levels were assessed by western blot using an anti-TFAM antibody, and ACTB was used as a loading control. Treatment of TetOn-TFAM-HA cells with doxycycline results in the appearance of an upper TFAM band corresponding to the HA-tagged construct. (Bottom) Quantification of TFAM levels by western blot. TFAM bands were quantified and normalized against ACTB. Shown are the TFAM levels relative to each respective control. Results from 5 biological replicates are shown. **(f)** Confocal microscopy showing the TFAM-HA construct localized to mitochondria and to nucleoids. TetOn-TFAM-HA HeLa S3 cells were treated with doxycycline for 48 hours and labeled for TOM20, ss/dsDNA, and HA (Scale bars: 5 µm and 1 µm for the zoom-in. **(g)** Bar plot showing the fraction of nucleoids that are accessible, defined by having >1% m6A. Increasing TFAM levels decreases the accessible population. Individual replicates were compared with a chi-squared test, *** signifies p-value < 0.001. Results from 3 biological replicates are shown. **(h)** Bar plot showing the ratio of TFAM protein levels to relative mtDNA levels as measured by western blot and qPCR after treatment with DMSO or 100 ng/mL doxycycline for 48 hours. Results from 3 biological replicates shown. **(i)** (Left) Line plot and (Right) scatterplot showing the mtFiber-seq footprint enrichment across the genome in cells overexpressing TFAM relative to the control (Pearson's r = 0.97). Datasets were subsampled to each other to match methylation distributions.


a
Strand methylation bias
Light Strand
Heavy Strand
0d
3d
6d
15,500
TAS
0
O_H/CSBI
500
1000
Genome position (bp)
b
Log2 fold change in strand bias score (Relative to 0d)
n.s.
n.s.
3d
6d
D-loop
1838-2511
4838-5511
c
Log2 fold change of D-loop nucleoids (Relative to Day 0)
3d
6d
Days in DM
d
TFAM:mtDNA ratio (Fold change)
0d
3d
6d
Days in DM
e
WT
TetOn TFAM-HA
Dox
ACTB
TFAM
TFAM protein levels (Fold change over control)
DMSO
DOX
f
α-TOM20
α-ss/dsDNA
α-HA
Merge
TFAM-HA
TOM20
TFAM-HA
DNA
TFAM-HA
DNA
g
Fraction of nucleoids accessible (Percent A's methylated > 1)
***
DMSO
Dox
h
TFAM:mtDNA ratio (Fold change over control)
DMSO
Dox
i
D-loop
MTERF1
Footprint enrichment
DMSO
Doxycycline
14,000
0
2,500
5,000
7,500
10,000
12,500
Genome position (bp)
Fraction of reads with footprint (Doxycycline)
Fraction of reads with footprint (DMSO)

sum, myoblast nucleoids decrease in accessibility and show signs of reduced replication as myoblasts withdraw from the cell cycle towards terminal differentiation, suggesting that a higher accessible nucleoid population supports the needs of active proliferation.

TFAM levels modulate mtDNA accessibility in cells

As TFAM is the primary constituent of mitochondrial nucleoids and compacts DNA *in vitro* (70,130,**Error! Reference source not found.**), we hypothesized that this shift in the accessible population could be explained by an increase in the TFAM:mtDNA ratio during differentiation. Consistent with previous reports, we observed a ~1.3-fold increase in mtDNA copy number by 6 days differentiation **(Fig. S3.7e)**. However, we saw an even larger increase in TFAM protein levels **(Fig. S3.7f)**, leading to a 2-fold increase in the TFAM:mtDNA ratio at day 3 and a 3.5-fold increase by day 6 **(Fig. 3.3d)**. Together, these results support our hypothesis that the observed shift in the nucleoid population is due to a change in the TFAM:mtDNA ratio during myogenesis.

To determine whether an increase in TFAM levels alone is sufficient to shift nucleoid accessibility, we induced the expression of HA-tagged TFAM that localized to mitochondria in HeLa S3 cells, yielding a 50% increase in TFAM levels **(Fig. 3.3e,f, Fig. 3.8a)**. We found that TFAM overexpression shifted the population towards more inaccessible nucleoids compared to DMSO treated controls cells **(Fig. 3.3g)**. In some systems, TFAM overexpression leads to increased mtDNA levels, but mtDNA levels were unchanged in our cells **(Fig. 3.8d)**, consistent with previous TFAM overexpression in HeLa cells (**Error! Reference source not found.**). Therefore, the increase in the TFAM:mtDNA ratio **(Fig. 3.3h)** explains the shift in the accessible population. Interestingly, although the global fraction of accessible nucleoids decreased, the footprint occupancy patterns remained constant with higher levels of TFAM **(Fig. 3.3i, Fig. S3.8e,f)**. These findings indicate that TFAM levels predominantly impact mitochondrial nucleoid

packaging by modulating the global fraction of accessible nucleoids without altering their architecture.

TFAM levels modulate mtDNA accessibility in vitro

To investigate the mechanism by which TFAM mediates mitochondrial nucleoid compaction, we performed *in vitro* MTase reactions on full-length mtDNA equilibrated with recombinant TFAM **(Fig. 3.4a, Fig. S3.9a,b)**. TFAM alone was sufficient to block the ability of Hia5 to methylate DNA *in vitro* in a concentration-dependent manner **(Fig. 3.4b)**, with 30 μM TFAM virtually abolishing mtDNA methylation **(Fig. 3.4b, Fig. S3.9c)**. We then performed PacBio sequencing on these MTase-treated samples to measure single-molecule accessibility and TFAM occupancy patterns along each *in vitro* reconstituted nucleoid **(Fig. 3.4c)**. Congruent with our bulk measurements **(Fig. 3.4b)**, increasing TFAM concentrations was associated with decreasing levels of per-read methylation **(Fig. 3.4c,d)**.

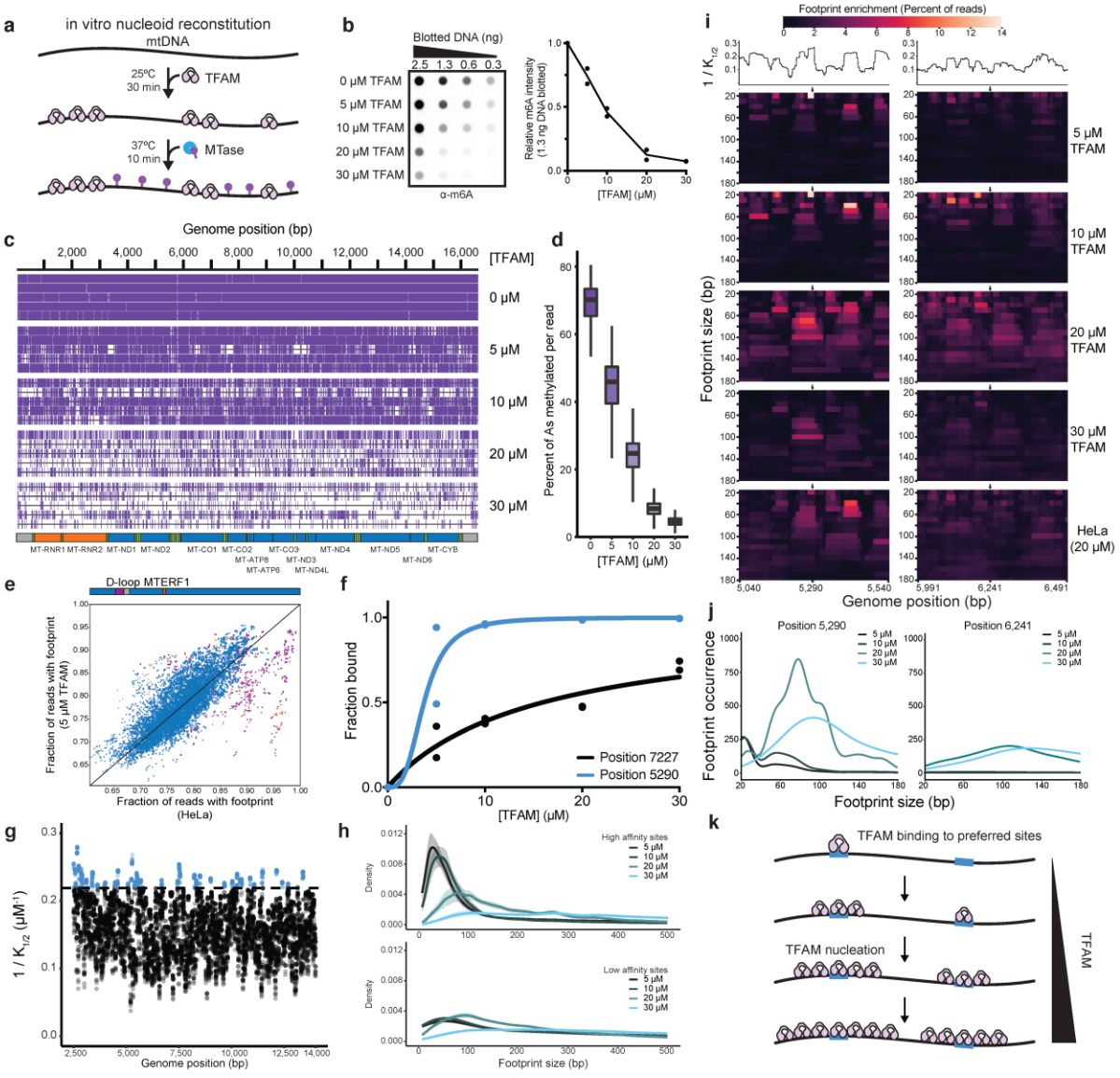
The high resolution of mtFiber-seq permits a deeper analysis of mtDNA packaging over bulk measures of compaction. We extracted the footprints across the reconstituted nucleoids using our HMM and compared the *in vitro* architecture to our measurements in cells. Notably, the *in vitro* reconstituted nucleoids lacked localized footprint densities at the D-loop and MTERF1 binding site, consistent with non-TFAM factors driving the formation of these features in cells **(Fig. 3.4e)**. However, across the rest of the mitochondrial genome, the footprint patterns along *in vitro* reconstituted nucleoids mirrored those derived from HeLa cells **(Fig. 3.4e)**. These results demonstrate that the protection patterns observed in cells are dominated by TFAM binding and indicate that the population of accessible nucleoids is driven by TFAM protein levels *in vivo*.

TFAM drives nucleoid packaging via a ‘nucleation-and-spreading’ mechanism

We next used the single-molecule footprint patterns on *in vitro* reconstituted nucleoids to elucidate the mechanism by which TFAM packages DNA. First, we calculated the relative binding

Fig. 3.4: *In vitro* reconstituted nucleoids reveal preferential TFAM binding and nucleation sites throughout the genome

(a) Schematic of the experimental design. Full-length linear mtDNA was generated by long-range PCR and equilibrated with increasing concentrations of recombinant TFAM. Samples were then methylated with 200 U Hia5 m6A-MTase. **(b)** (Left) Dot blot assessing methylation of mtDNA with increasing concentrations of TFAM. A range of amounts of DNA from (A) were adsorbed onto a nitrocellulose membrane, crosslinked, and detected with an antibody against m6A. (Right) Quantification of the dot blot, the 1.3 ng DNA sample. Intensities were quantified using ImageJ and normalized against the 0 µM sample. Values from two replicates are shown. **(c)** Genomic track of mtDNA. Five randomly sampled mtFiber-seq reads for each concentration of TFAM are shown. Individual reads are indicated by horizontal black lines and m6A bases are marked by vertical purple dashes. **(d)** Box plot showing the percent of As methylated per read for each concentration of TFAM. **(e)** Scatterplot showing the mtFiber-seq footprint enrichment across the genome in HeLa cells and *in vitro* with 5 µM TFAM for reads subsampled to match methylation distributions (Pearson's r = 0.665). **(f)** Example binding curve from *in vitro* mtFiber-seq data. Individual reads were classified as bound at a particular site if a footprint overlapped the position that was at least 20 bp long. Data were fit to a four-parameter logistic regression to determine $K_{1/2}$ constants. Results from two replicates are shown. **(g)** Affinities were determined from genomic position 2,500 to 14,000 based on the fraction of reads protected with a footprint at least 20 bp long. Reciprocals of $K_{1/2}$ constants are shown. Results are from the average of two replicates. **(h)** Meta density plots of footprint sizes from (top) 26 high affinity sites and (bottom) 64 low affinity sites at each TFAM concentration with the 95% confidence intervals shaded. Distributions between high and low affinity sites are significantly different at 5 µM TFAM (KS test, $p << 0.05$, $D = 0.32$) and 10 µM TFAM (KS test, $p << 0.05$, $D = 0.29$). (i) Heatmaps of footprint size enrichment from (left) position 5,040 to 5,540 and (right) position 5,991 to 6,491 from *in vitro* reconstituted nucleoids with 5, 10, 20, and 30 µM TFAM (top) and for HeLa nucleoids, subsampled to match the methylation distribution of the 20 µM TFAM dataset. Each row represents a footprint size, and each column shows a position in the genome. Line plots indicating the $1/K_{1/2}$ across this locus are shown above each heatmap **(j)** Density showing the footprint size distribution between 20 and 180 bp at positions 5,290 and 6,241 as a function of TFAM concentration. **(k)** Cartoon depicting TFAM binding and nucleation. TFAM binds preferred sites (blue) located throughout the genome. Due to TFAM's cooperative binding behavior, higher TFAM concentrations result in spreading from these specific nucleation sites.


a
in vitro nucleoid reconstitution
mtDNA
25ºC
30 min
TFAM
37ºC
10 min
MTase
b
Blotted DNA (ng)
2.5 1.3 0.6 0.3
0 µM TFAM
5 µM TFAM
10 µM TFAM
20 µM TFAM
30 µM TFAM
α-m6A
Relative m6A intensity
(1.3 ng DNA blotted)
[TFAM] (µM)
c
Genome position (bp)
[TFAM]
0 µM
5 µM
10 µM
20 µM
30 µM
MT-RNR1 MT-RNR2 MT-ND1 MT-ND2 MT-CO1 MT-CO2 MT-ATP8 MT-ATP6 MT-CO3 MT-ND3 MT-ND4L MT-ND4 MT-ND5 MT-ND6 MT-CYB
d
Percent of As methylated per read
[TFAM] (µM)
e
D-loop MTERF1
Fraction of reads with footprint
(5 µM TFAM)
Fraction of reads with footprint
(HeLa)
f
Fraction bound
[TFAM] (µM)
Position 7227
Position 5290
g
1 / K1/2 (µM-1)
Genome position (bp)
h
High affinity sites
Low affinity sites
5 µM
10 µM
20 µM
30 µM
Density
Footprint size (bp)
i
Footprint enrichment (Percent of reads)
1 / K1/2
Footprint size (bp)
5 µM TFAM
10 µM TFAM
20 µM TFAM
30 µM TFAM
HeLa (20 µM)
Genome position (bp)
j
Position 5,290
Position 6,241
Footprint occurrence
Footprint size (bp)
k
TFAM binding to preferred sites
TFAM nucleation
TFAM

affinity ($K_{1/2}$ constants) of TFAM for each base in the mitochondrial genome to ask whether certain sites are preferentially bound by TFAM. To compute these per-base $K_{1/2}$ constants, we calculated the fraction of reads with a footprint at each genomic position with each TFAM concentration and used a four parameter logistics regression to determine the concentration of TFAM at which 50% of mtDNA molecules are protected at that position **(Fig. 3.4f,g)**. While binding assays with short fragments of DNA show similar K_Ds irrespective of sequence[19], here we see a range of binding affinities, revealing preferential binding sites situated throughout the genome. This approach revealed 31 unique high-affinity TFAM binding sites along the mitochondrial genome that were preferentially bound even at low concentrations of TFAM. We were unable to identify strong motifs or common features that defined these sites. We next calculated the Hill coefficients of TFAM binding across the genome to determine the degree of cooperative TFAM binding. We found Hill coefficients exceeded 2 at the large majority of mtDNA loci **(Fig. S3.9f),** consistent with previous reports of cooperative TFAM binding to nonspecific DNA *in vitro* (**Error! Reference source not found.**,**Error! Reference source not found.**). Our results indicate that TFAM preferentially binds certain DNA sequences and has a strong propensity for cooperative binding genome-wide.

We next hypothesized that these properties of TFAM binding would lead to nucleoid compaction via a cooperative occupancy of TFAM adjacent to high-affinity TFAM-bound nucleation sites. Indeed, at lower concentrations of TFAM, high-affinity sites had footprint sizes that mirrored those expected from the occupancy of a single TFAM molecule (~30 bp) (**Fig. 3.4h-j, Fig. S3.9g, Fig. S3.10a-d)**. At increased TFAM concentrations, these footprints increased in size while remaining centered at the high-affinity site, consistent with cooperative binding of additional TFAM proteins rather than the appearance of additional single TFAM binding events. Low-affinity sites were substantially more accessible at lower TFAM concentrations and became protected by large footprints at higher concentrations, indicative of TFAM association aided by the cooperative extension from a proximal high affinity site **(Fig. 3.4h-j, Fig. S3.10e-h)**. Moreover,

the *in vivo* data mirrored the *in vitro* results; we observed nearly identical footprint patterns for our *in vivo* HeLa S3 data when we subsampled the reads to match the methylation densities of the *in vitro* reads **(Fig. 3.4i, Fig. S3.10a,b,e,f)**. These TFAM association patterns are consistent with computational modeling of proteins binding cooperatively to DNA (**Error! Reference source not found.**). Taken together, these results imply a model for mtDNA compaction in cells in which TFAM binds higher-affinity sites throughout the genome that nucleate the compaction of mtDNA through TFAM spreading out from these sites **(Fig. 3.4k)**.

Discussion

Our single-molecule analysis of the mitochondrial genome revealed substantial heterogeneity in the packaging of individual mitochondrial nucleoids, a feature previously obscured by the bulk averaging inherent to cleavage-based accessibility methods. These observations demonstrated that most mtDNA molecules are inaccessible, uncovering an unappreciated layer of mtDNA regulation. The remaining population shows heterogeneity in the extent and pattern of accessibility which could reflect multiple genomic processes including progression of the RNA and DNA polymerases and stochastic dissociation/reassociation of TFAM. We found that TFAM levels strongly correlate with methylation levels detected by mtFiber-seq. Our data are consistent with a model whereby TFAM nucleates from preferred binding sites throughout the genome and then spreads to compact mtDNA through cooperative binding. Thus, TFAM levels directly control DNA accessibility in cells, which would impact the ability of the transcription and replication machinery to bind and initiate their respective reactions. Consistent with this, high TFAM levels inhibit transcription and replication *in vitro* and *in vivo* (**Error! Reference source not found.**,**Error! Reference source not found.**,**Error! Reference source not found.**,**Error! Reference source not found.**), inactive nucleoids show higher TFAM signal by STED microscopy[18], and only a fraction of nucleoids undergo transcription or replication

(**Error! Reference source not found.**,**Error! Reference source not found.**). Moreover, we found that when mtDNA accessibility decreased, fewer hallmarks of active replication were observed within the reads **(Fig. 3.3a-c, Fig. S3.8g)**. Although accessibility and activity may not be directly coupled, we propose that the two are correlated, with inaccessible nucleoids having a higher likelihood of being inactive, and, conversely, accessible nucleoids having a higher potential to be active, as is indicated by the co-occupancy of these accessible nucleoids by transcription and replication components that are defining features of active mitochondrial nucleoids.

Why does the cell maintain a pool of inaccessible nucleoids? Two hypotheses could explain the phenomenon. First, inactive molecules could serve as a genetic reservoir. The maintenance of a healthy pool of mtDNA is critical, especially in long-lived cells such as neurons. mtDNA repair is inefficient, and mtDNA replication introduces mutations and deletions due to the intrinsic error rate of Polɣ and from replication-fork stalling (**Error! Reference source not found.**,**Error! Reference source not found.**,**Error! Reference source not found.**,**Error! Reference source not found.**,**Error! Reference source not found.**). Thus, damaged molecules could be selectively degraded and replaced by undamaged copies that were inactive in the reservoir. Second, an inactive population could serve as excess capacity, enabling transcription and replication throughout the mitochondrial network precisely where they are needed. As levels of nuclear-encoded TFAM vary across tissues and disease states (169,**Error! Reference source not found.**), the fraction of accessible and active nucleoids is likely to be dynamic as well, as we observed during differentiation (**Fig. 1g**), and may represent a crucial nuclear-controlled parameter in the balancing of OXPHOS production (**Error! Reference source not found.**). Our current data lack time resolution, so we are unable to comment on the dynamics and transitions between inaccessible and accessible states. Thus, addressing these hypotheses will require a time-resolved mtFiber-seq strategy, representing a priority for future studies.

Chapter 4: Discussion and outstanding questions

Discussion

The studies outlined in Chapter 2 and 3 encompass two very distinct systems which presented unique challenges and advantages for the application and analysis of Fiber-seq data. The starkest difference was the scale of the data we could produce for the system—in the case of *Drosophila* nuclear transcription, wide scale and richly annotated but of lower depth, in the case of the human mitochondrial genome, extreme depth but minimally annotated and covering only ~17kb in range. As such, the questions and insights we were able to explore using the same basic toolset of footprinted Fiber-seq data were wildly distinct. In the case of transcription, we uncovered principles of transcriptional coordination based on single-molecule observations pooled across thousands of genes: pause-driven steric inhibition of initiation, pause-driven, stable changes to nucleosome organization, and coordination of transcription activity between genes. In contrast, our work on the mitochondrial genome focused more on the mitochondrial genome itself as a single molecule, capturing a huge dynamic range of accessibility driven by TFAM, a single architectural protein, as well as the footprints of specific factors involved in transcription and replication. In the remainder of Chapter 4 I will lay out the open questions and future directions that arose from each of these studies.

Outstanding questions about pausing and nucleosomes

What is the mechanism of pause-driven nucleosome disruption?

Our results demonstrating a stable, pause-driven disruption of downstream nucleosomes represent a direct connection between pausing and the nearby chromatin landscape. While we were able to show a definitive causal relationship between pausing and these shifts, the exact mechanism behind them remains unclear. They do not appear to be driven by elongation, as has been previously reported (41), as we found that the shifted nucleosomes are associated with

pause index, but not with expression of a given gene. This leaves two likely explanations, neither of which is necessarily mutually exclusive. The first possibility is that chromatin remodelers are associated with the paused polymerase complex and drive shifts to nucleosomes during the duration of the promoter proximal pause. The second is a pause-driven shift, caused by positive supercoiling from the short distance transcribed prior to the pause (182). Supercoiling could explain the strong directionality of the shift, as well as the specificity to the PPP, which has transcribed 20-50 bases of RNA, over the PIC. Possible approaches to clarify this mechanism could be based on a targeted overexpression or depletion of candidate chromatin remodelers to identify any which are responsible, and a targeted overexpression or depletion of topoisomerases, which could modulate supercoiling and any connected effects. Alternatively, a supercoiling detection approach like psora-seq or GapR-seq (184,185) combined with the Fiber-seq methodology or perturbations to pausing could directly identify if there is an association between pausing, supercoiling, and the disruption of downstream nucleosomes.

What is the timescale of pause-driven nucleosome disruption?

A related question to the mechanism of the disruption of downstream nucleosomes is the duration of this disruption after Pol II has been released from the pause state. Is the disruption a brief change that must be reestablished with frequent pausing, or is it more stable? We can infer generally only that it is stable on the timescale that can be captured by Fiber-seq, but any increased temporal specificity would require additional experimental validation. This would likely require advances to the Fiber-seq method to add a temporal element to the data, perhaps involving multiple cycles of distinct accessibility-marking prior to the final methylation step. Alternatively, using orthogonal short-read accessibility-mapping approaches combined with a pulse-chase of pause-release inhibition with a drug like DRB could potentially capture more detailed time-points, assuming we could detect the changes to nucleosomes in short-read data.

What is the function of pause-driven nucleosome disruption?

Ultimately the most important remaining question about the pause-driven nucleosome disruption is what, if any, function the disruption serves. While some of our results hinted at a possible function, they did not provide definitive evidence. Of particular note was the observation that the frequency of elongating Pol II footprints was enriched specifically on the reads exhibiting a stable disruption of the +1 nucleosome after pause release. That these fibers contain active elongating polymerases could imply that there is a connection between the disruption of downstream nucleosomes and increased overall transcription elongation or even transcriptional bursting. However, we could not establish a causal relationship, as without a mechanism for the disruption we could not alter the nucleosome disruption without also altering pausing. Further, while we were confident that the putative elongating Pol II footprints we identified were representative of elongation activity in aggregate, we were not sufficiently confident in the individual identities of the footprints to attempt to categorize them as transcriptional bursting or more sporadic elongation. Future work delving into specific changes to transcription elongation associated with and driven by the pause-disrupted downstream nucleosomes could serve to connect pausing directly to subsequent changes in transcription downstream and provide an additional purpose to the promoter proximal pause in gene regulation.

Is it possible to perturb pause-driven nucleosome disruption?

A unifying dilemma raised by the above questions is how one can perturb the nucleosome disruption associated with pausing without altering pausing itself. The difficulty arises from the fact that the shifted nucleosomes seem sterically necessary to allow for pausing at most genes, making disentangling effects that alter the nucleosome shift and those that alter the possibility for pausing to actually occur will be paramount. A general strategy, we think, will be to use a similar

approach to the one employed to validate the mechanism of pause-inhibited initiation. In brief, that strategy is to take advantage of the variability in pause and +1 position across genes to identify a subset which do not require the shift for pausing, which could be used as the basis for future work attempting to probe the disruption.

Outstanding questions about transcriptional coordination

What genomic features affect transcriptional coordination, and to what extent?

The second major result from Chapter 2 was the observation of pervasive coordination in both accessibility and transcription activity at pairs of promoters and transcribed enhancers based on proximity. More so than the relationship between pausing and nucleosomes, coordination is a system that should be readily manipulable. A major next step towards understanding coordination will therefore be to rigorously quantify the effects of different features on coordination, through genome manipulation and targeted perturbations. Of particular interest would be the effect of insertion of additional genomic sequences between pairs of loci, to determine in a more fine-grained manner the distance dependence on coordination. Beyond distance alone, inserting specific sequences like tRNA genes, insulators, enhancers, or other genomic features could be an effective way to measure their effect on coordination in a more targeted and quantitative way. Building further on insulators, identifying the specific footprints associated with proteins like CP190, SuHW and CTCF, would allow us to test their effect on coordination in a single-molecule, direct manner. This would be a step beyond the associative approach we used in Chapter 2 based on ChIP-seq peaks, allowing us to characterize the degree to which these proteins have a direct role in blocking coordination.

Expanding the distances to identify transcriptional coordination

Our work demonstrated a role for local chromatin structure in modulating coordination on an up to 20kb scale, with a predominant effect found within 5kb. The main limitation in distance, though, was the length of Fiber-seq reads, which needed to capture both promoters in a single read. We expect that similar coordination may be seen in specific cases at ranges beyond those that we could observe given the extensive evidence in the literature of long range promoter-enhancer interactions and promoter-promoter interactions at huge scales. These interactions are generally mediated by 3D chromatin architecture, which seemed to have an effect even on our smaller-scale coordination, with genes in the same TAD having higher levels of coordination (45,46,47,48,49,53,54,55).

Currently, Fiber-seq cannot probe far beyond a 20kb distance on a single read. However, there are feasible adjustments to the Fiber-seq protocol that could capture longer range interactions, such as combining Fiber-seq with proximity crosslinking. In the longer term, the length of PacBio, and Fiber-seq reads in turn, have steadily increased over time with improvements to the technology, with the average read length doubling between the earliest datasets used in Chapter 2 to the most recent ones. As such, it is likely that more of these interactions, especially those at the 100 kb scale, will be visible in a single read without additional effort. Identifying long-range coordinated enhancer-promoter pairs would be particularly valuable, providing a direct method to categorize these otherwise difficult to identify regulatory relationships.

What other factors exhibit coordination?

A major question, outside the scope of Chapter 2, is to what degree coordination involves and/or is mediated by other factors, in particular transcription factors. Our work was limited to

PPP, PIC, and nucleosome footprints, but we found that most promoters and especially enhancers had one or more small footprints in the vicinity of the promoter that we did not identify. These footprints often exhibited strong and distinctive binding patterns in relation to polymerase footprints, suggesting a direct regulatory relationship. Identifying, and then characterizing the relationship between these transcription factor footprints, nucleosome footprints, and polymerase footprints at individual promoters could clarify the function of these factors in regulating gene expression. Are they coordinated with the PPP or PIC? Anti-coordinated? Beyond individual promoters, the degree to which these currently unidentified footprints are coordinated between different pairs of genes could help identify and characterize the factors that mediate coordination across the genome.

Does RNA Polymerase I exhibit coordination?

In addition, we expect that given the strong coordination shown by other polymerases, RNA Polymerase I (Pol I) will likely show a similar effect. Pol I is responsible for the production of most rRNAs, which are highly abundant, and found in variable arrays of hundreds of tandem-repeated copies on the X and Y chromosomes in *Drosophila* (186). The repeating unit consists of the 18S, 5.8S and 28S genes, with a length of 12-17kb (186,187). This length, ~40x longer than the 5S rRNA, limits the ability of a current Fiber-seq read to capture multiple sets of rRNA genes in a single read or even map these reads accurately with respect to each other. As a result, we decided to focus on Pol III as the additional RNA polymerase due to the smaller size of its transcripts and the frequent proximity of tRNA genes to Pol II genes. These advantages facilitated our coordination analysis and allowed for a direct connection to our work on Pol II, whereas analysis of Pol I transcripts would have been largely limited to footprint identification. Especially as Fiber-seq reads grow longer, though, future work should certainly be able to identify Pol I

transcription associated footprints and delve into the interplay and coordination between nearby Pol I genes.

Outstanding questions about mitochondrial genome organization

Is mtDNA accessibility tied to transcription activity?

Given the high variability in accessibility across mitochondrial genomes, and variability in the distribution of accessibility seen across treatment conditions, it appears that mitochondrial genome accessibility could serve as a general regulatory mechanism for globally modulating mitochondrial gene expression. TFAM concentration in general is also known to affect mitochondrial transcription, with a particular level necessary to initiate transcription while not fully blocking accessibility. However, we could not test for a direct connection between levels of accessibility of a given genome and transcription activity because we were not able to identify footprints of the mitochondrial transcriptional machinery within our analysis. That is not to say that those footprints are not present, of course, but a more targeted experimental approach to identifying those footprints would likely be necessary given the similar sizes of PolRMT and TFAM. One effective strategy that could facilitate identification of PolRMT footprints could be to take advantage of small molecule inhibitors of mitochondrial transcription such as 2CMA or IMT1 (**Error! Reference source not found.**,188). These inhibitors lead to polymerase stalling, which would likely deplete PolRMT footprints in comparison to TFAM footprints, similar to the role played by triptolide in the work outlined in Chapter 2.

Is mtDNA accessibility tied to genome replication?

Based on the strong enrichment of single-stranded D-loops and consistent footprints of Polγ and TWINKLE found at either end of the NCR in reads from accessible genomes, we expect that there is a connection between mitochondrial genome replication and TFAM occupancy on a given genome. While we were able to identify abundant footprints associated with abortive and primed mtDNA replication, we were not able to identify active replication machinery. Ideally we would be able to directly identify footprints of the replication machinery, but as with transcription, discerning the difference between these footprints and those of TFAM outside of their most stable binding sites could prove difficult. Alternatively, to identify actively replicating genomes, we could take advantage of a similar strategy to the one used to identify the D-loop, that is to identify the single-stranded DNA associated with active replication via methylation strand bias. As a complementary approach we could carry out metabolic labeling of newly synthesized mitochondrial DNA via incorporation and direct detection of modified nucleotides like BrdU in Fiber-seq (190), which would allow for identification of recently copied genomes and their accessibility. Overall, these strategies could allow us to draw a direct association between TFAM occupancy and mitochondrial genome replication, assigning a direct functional role to the variation in accessibility observed.

Is mtDNA accessibility tied to genome stability?

Mitochondrial genomes exist in a harsh environment relative to the rest of the cell with poor error correction and polymerases prone to errors (**Error! Reference source not found.**,**Error! Reference source not found.**,**Error! Reference source not found.**,**Error! Reference source not found.**,**Error! Reference source not found.**). One possible explanation for the large frequency of fully inaccessible mitochondrial genomes would be maintenance of a reservoir of intact genomes, sequestered from the harsh environment by nearly

saturated coverage by TFAM. In addition, TFAM itself has been linked to DNA damage recognition, although the functionality of this link is unclear (191). Future work detecting DNA damage levels of sequenced mitochondrial genomes, ideally directly within Fiber-seq reads, and connecting these levels with TFAM binding and compaction could clarify the role of mtDNA accessibility and TFAM binding specifically in protecting the integrity of the mitochondrial genome both proactively and reactively. Are more accessible genomes more error-prone, or are they similar to compacted ones?

How is mtDNA accessibility regulated?

Whatever the precise connection between mitochondrial genome accessibility and regulation of transcription, replication, and genome stability, understanding how such strong variability in mitochondrial genome accessibility arises will be essential. We demonstrated that this variation in accessibility can be almost entirely explained by the relative concentration of TFAM, with *in vitro* levels on synthetic mitochondrial genomes closely replicating the patterns seen *in vivo.* We also observed changes to the distribution of accessibility states across mitochondrial genomes in response to perturbations and development. The question, though, is how TFAM levels change on a mechanistic level– is the change a result of expression, import, stability, or another mechanism? Further, is it even TFAM that changes? mtDNA copy number could also play a major role, given its variability (155,178).

Beyond understanding how TFAM levels and mitochondrial compaction in general are regulated, another question is how variable accessibility is over time. Adding a temporal component to Fiber-seq to allow for characterization of the dynamics of mtDNA accessibility in steady-state and through perturbations could establish if mitochondrial genome accessibility is stable or frequently altered and actively regulated. Further, given the single-genome nature of the

data, such experiments could establish if there is a subset of genomes that are stable and a subset that are mutable, and connect that status to overall genome accessibility.

Appendix 1: Supplemental Figures for Chapter 2

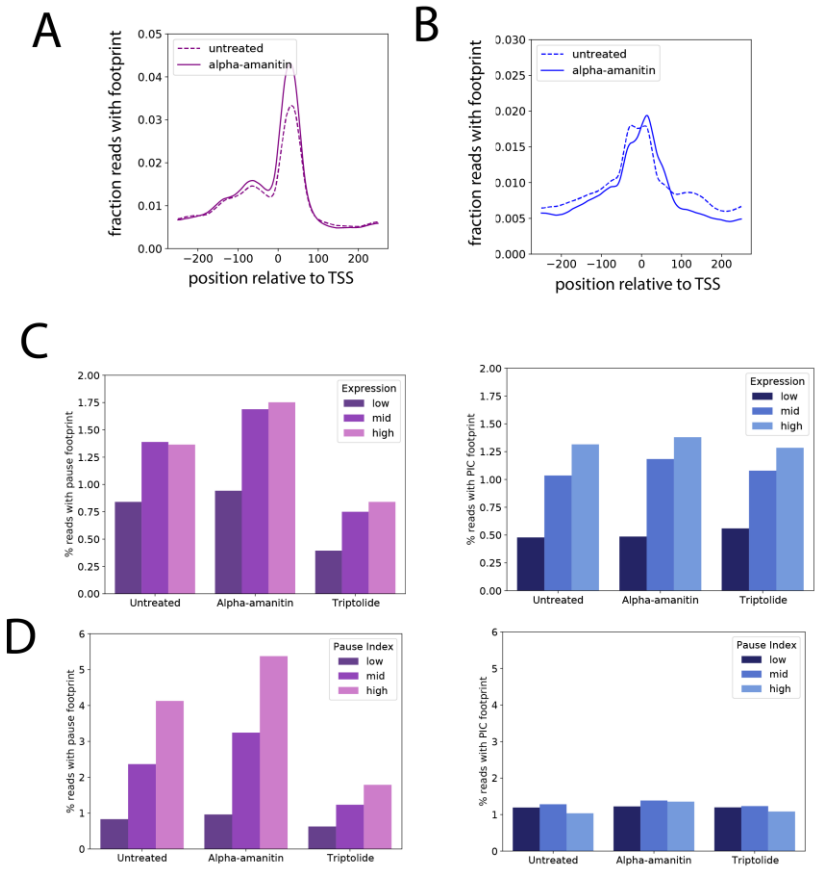
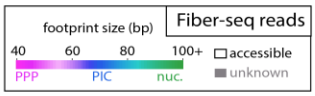
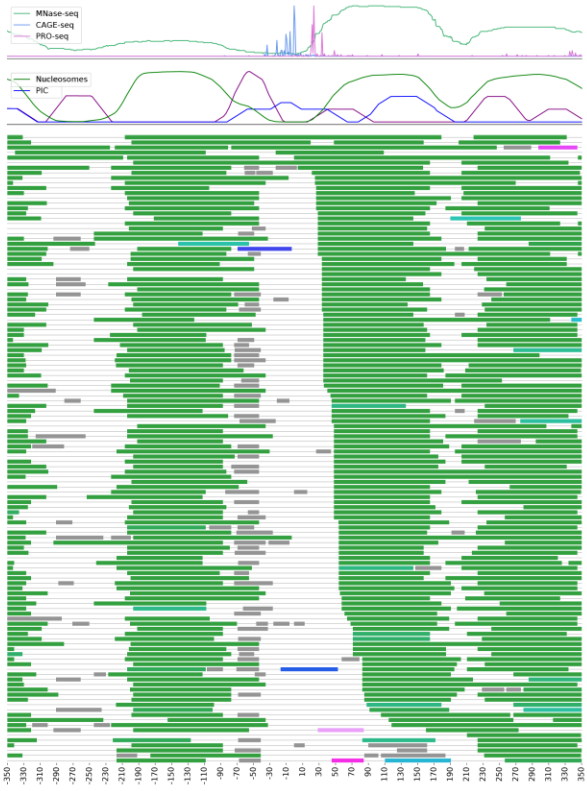


Figure S2.1 (related to Figure 2.1)
(A) Plot showing enrichment of PPP footprints with (solid) or without (dashed) alpha-amanitin treatment showing an increase in overall footprints. (B) Plot showing enrichment of PIC footprints with (solid) or without (dashed) alpha-amanitin treatment showing a similar level of overall footprints. (C) Enrichment of PPP (Left) or PIC (Right) footprints at different expression levels as calculated from PRO-seq signal in the gene body in untreated, triptolide-treated, or alpha-amanitin-treated conditions. PIC footprints show a steady enrichment with increased expression, PPP footprints see a lesser effect except at low expression. (D) Enrichment of PPP (Left) or PIC (Right) footprints at different pause index levels in untreated, triptolide-treated, or alpha-amanitin-treated conditions. PIC footprints show no effect with increased pause index, PPP footprints see a strong, steady increase with increased pause index. Across both C and D triptolide exhibits a depletion of PPP footprints and alpha-amanitin shows an increase. Neither have a strong effect on PIC footprints.

Figure S2.2 (related to Figure 2.1)
Fiber-seq reads at example protein coding loci. Each plot contains a track with MNAse-seq, PRO-seq, and CAGE-seq, as well as a track showing enrichment of PPP, PIC, and nucleosome footprints for comparison. Note that Hsp70Bbb is lacking PRO-seq signal—this is due to difficulties aligning PRO-seq to the 6 sequence-identical copies of Hsp70. Below are a series of lines showing Fiber-seq footprints at each locus. Footprints are colored based on predicted identity (PPP = pink, PIC = blue, nucleosome = green, unknown = grey). Note the variety in PPP and PIC occupancy, the numerous unknown (likely transcription factor) footprints, and the variety in nucleosome positioning.

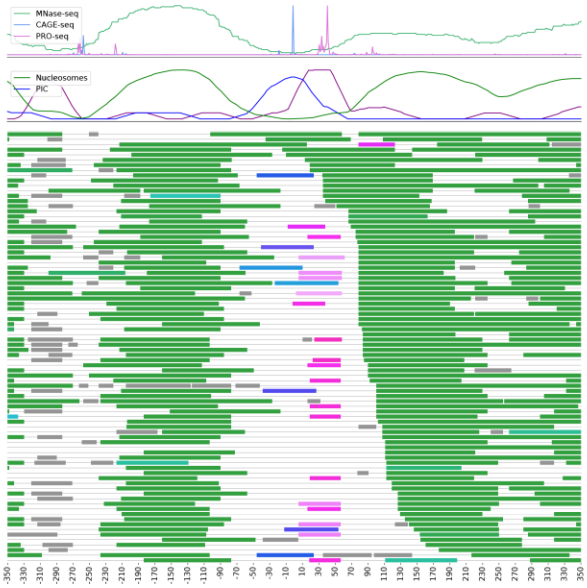


FBgn0003278 (Polr1B)



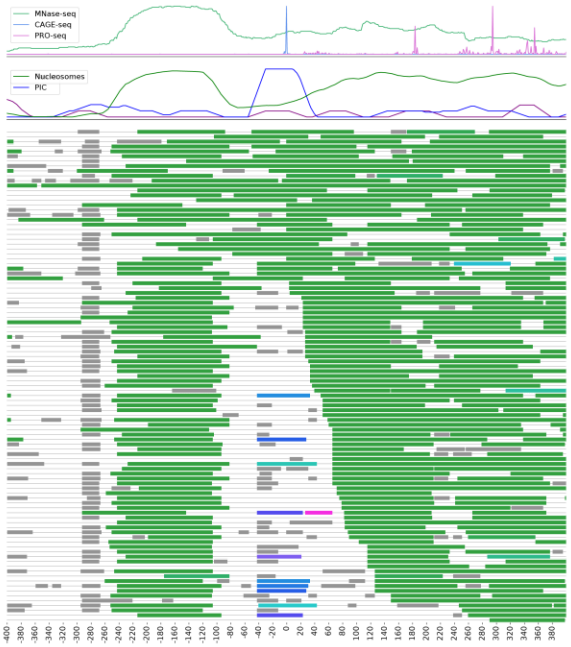
position relative to TSS (bp)

FBgn0051955 (CG31855)



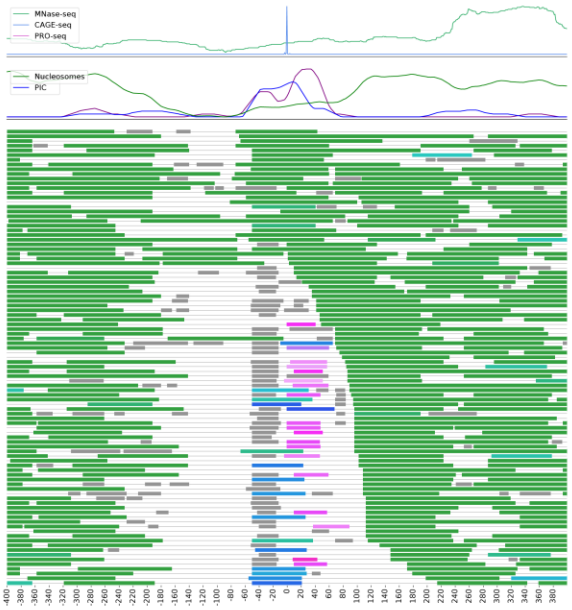
position relative to TSS (bp)

FBgn0001223 (Hsp22)



position relative to TSS (bp)

FBgn0051354 (Hsp70Bbb)



position relative to TSS (bp)

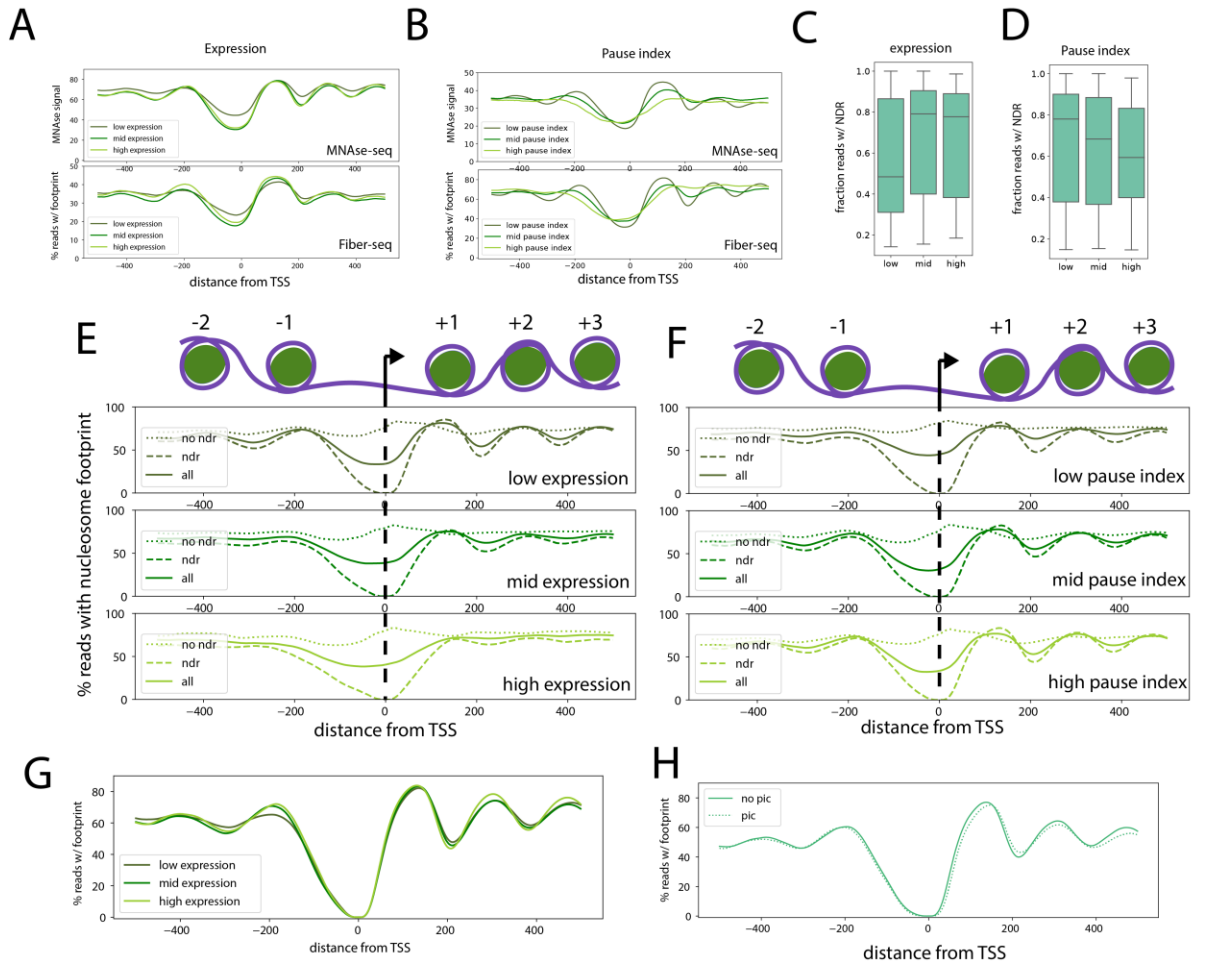


Figure S2.3 (related to Figure 2.3)
(A) Comparison between (Top) MNAse-seq signal and (Bottom) Fiber-seq nucleosome enrichment at different expression levels based on normalized gene-body PRO-seq coverage. (B) Comparison between (Top) MNAse-seq signal and (Bottom) Fiber-seq nucleosome enrichment at different pause index levels based on normalized gene-body PRO-seq coverage. (C) Boxplot showing the fraction of Fiber-seq reads with an accessible promoter at different expression levels, with increasing accessibility with higher expression. (D) Boxplot showing the fraction of Fiber-seq reads with an accessible promoter at different pause index levels, with decreasing accessibility with higher pasue index. (E) Comparison between Fiber-seq nucleosome enrichment at different expression levels with (dashed) or without (dotted) an accessible promoter, in comparison to the nucleosome enrichment across all genes, demonstrating the effect of inaccessible promoters on that signal. (F) Comparison between Fiber-seq nucleosome enrichment at different pause index levels with (dashed) or without (dotted) an accessible promoter, in comparison to the nucleosome enrichment across all genes, demonstrating the effect of inaccessible promoters on that signal. (G) Enrichment of nucleosome footprints in Fiber-seq reads at different levels of expression, only including reads with an accessible promoter. The distribution of nucleosomes is nearly identical once accessibility is controlled for. (H) Comparison of nucleosome enrichment in Fiber-seq reads with (dotted) or without (solid) a PIC footprint, only including reads with an accessible promoter and sampled to capture an equal amount of PIC and no PIC reads from each gene. There is no significant difference in nucleosome positioning associated with the PIC footprint.

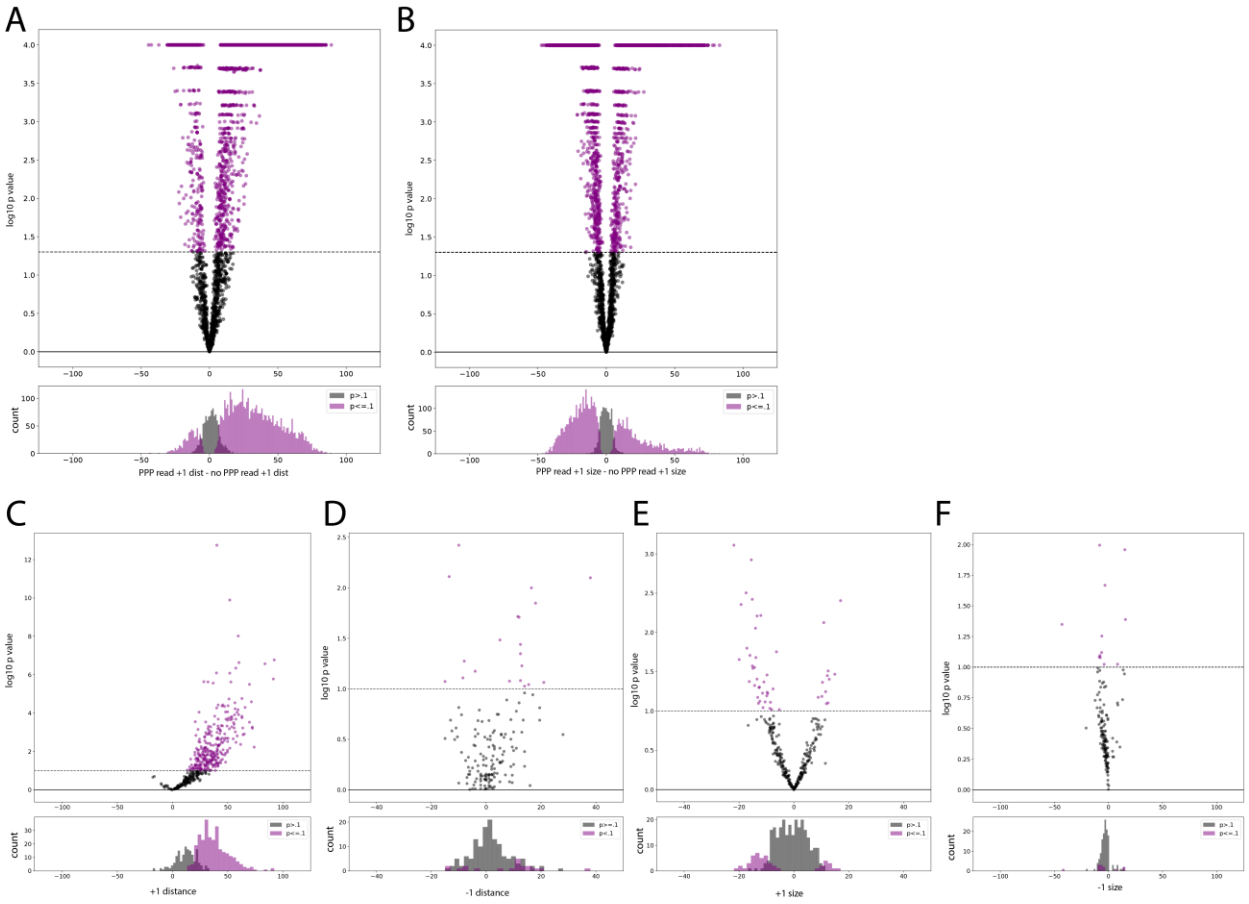


Figure S2.4 (related to Figure 2.3)
(A and B) Each individual read with a PPP footprint was compared to the bootstrapped mean distance (A) and size (B) of the +1 nucleosome. (Top) Plot of difference in +1 distance or size for each PPP read compared to the bootstrapped mean +1 distance or size vs the associated -log10(p-value), generated by comparing the value to the bootstrapped distribution at the origin locus. There is a strong relative shift in downstream in +1 distance and smaller in +1 size visible on an individual read basis. (Bottom) Histogram of differences compared to bootstrapped mean for all reads with PPP, shaded based on p-value. (C-F) Comparing reads with or without a PPP footprint at each locus with at least 5 PPP reads. Graphs in order from left to right are +1 distance, -1 distance, +1 size, -1 size. (Top) Plot of the mean distance or size of the +1 or -1 nucleosome for reads with a PPP footprint at each locus, vs the -log10(p-value) generated by the Wilcoxon rank sums test. (Bottom) Histogram of the mean distance or size of the +1 or -1 nucleosome for reads with a PPP footprint at each locus, shaded based on the p-value of the Wilcoxon rank sums test. Reads with a PPP at each locus are enriched for a shifted +1 nucleosome and to a small degree a smaller +1 nucleosome. There is minimal change associated with the -1 nucleosome.

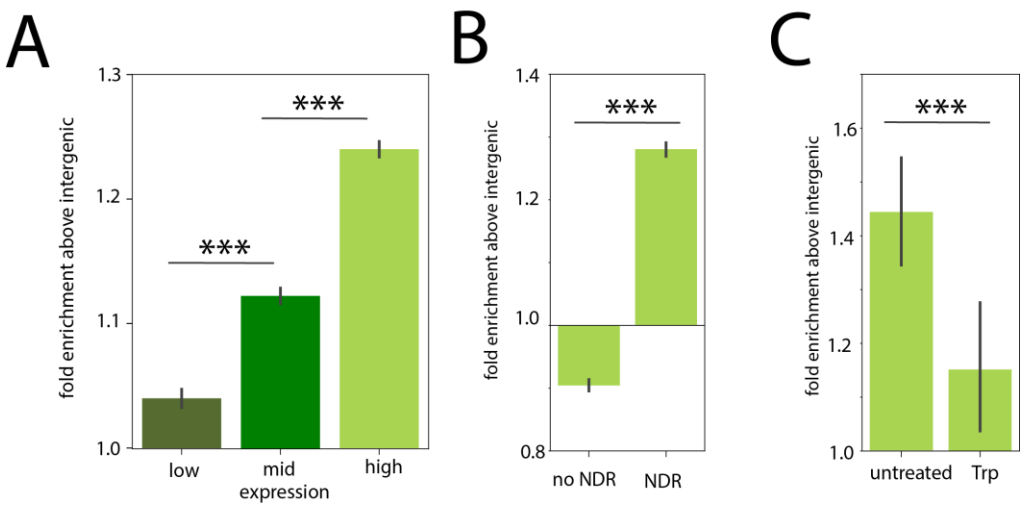


Figure S2.5 (related to Figure 2.4)
(A) Bar plot quantifying the fold enrichment of putative elongating Pol II footprints within gene bodies of genes split by expression based on PRO-seq signal within gene bodies. There is a significant increase in the fold enrichment going from low to mid and mid to high expression (T-test, *** signifies p-value < 0.001). (B) Bar plot quantifying the fold enrichment of putative elongating Pol II footprints within gene bodies at reads with an inaccessible (no NDR) or accessible (NDR) promoter corresponding to that gene. Reads with an accessible promoter have a significantly higher enrichment of footprints (T-test, *** signifies p-value < 0.001). (C) Bar plot quantifying the fold enrichment of putative elongating Pol II footprints within gene bodies at reads with a PIC footprint at that gene. A comparison is made between these reads from the untreated dataset and the triptolide-treated dataset, wherein reads with a PIC footprint are expected to represent reads actively inhibited by triptolide. There is a significantly elevated frequenecy of elongating Pol II footprints in the untreated dataset compared to the triptolide dataset, wherein the level of these footprints minimally higher than the background level (T-test, *** signifies p-value < 0.001).

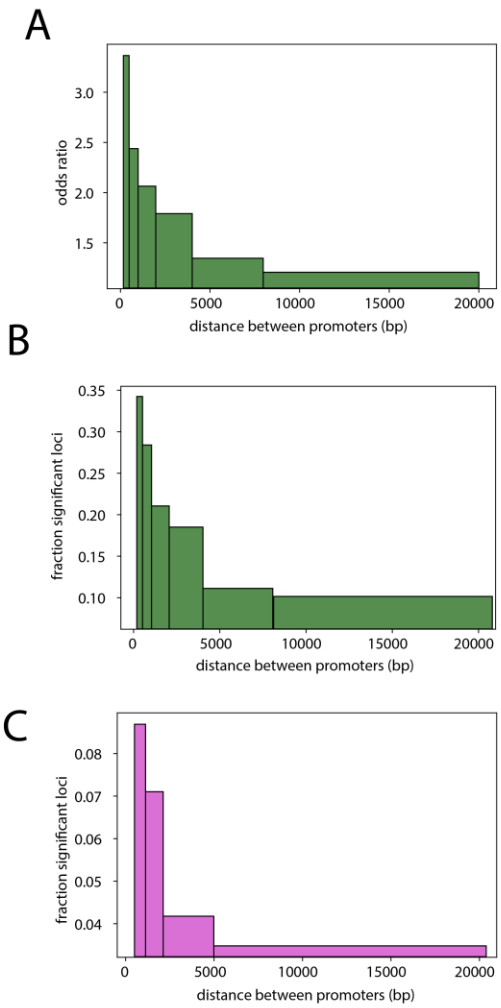


Figure S2.6 (related to Figure 2.5)
(A) Coordination of promoter accessibility based on distance via pooled Fisher's exact test based on distance. (B) Coordination of promoter accessibility based on the fraction of pairs of loci that are significantly enriched for co-accessibility within each distance bin. (C) Coordination of PPP/PIC occupancy based on the fraction of pairs of loci that are significantly enriched for cooccupation within each distance bin.

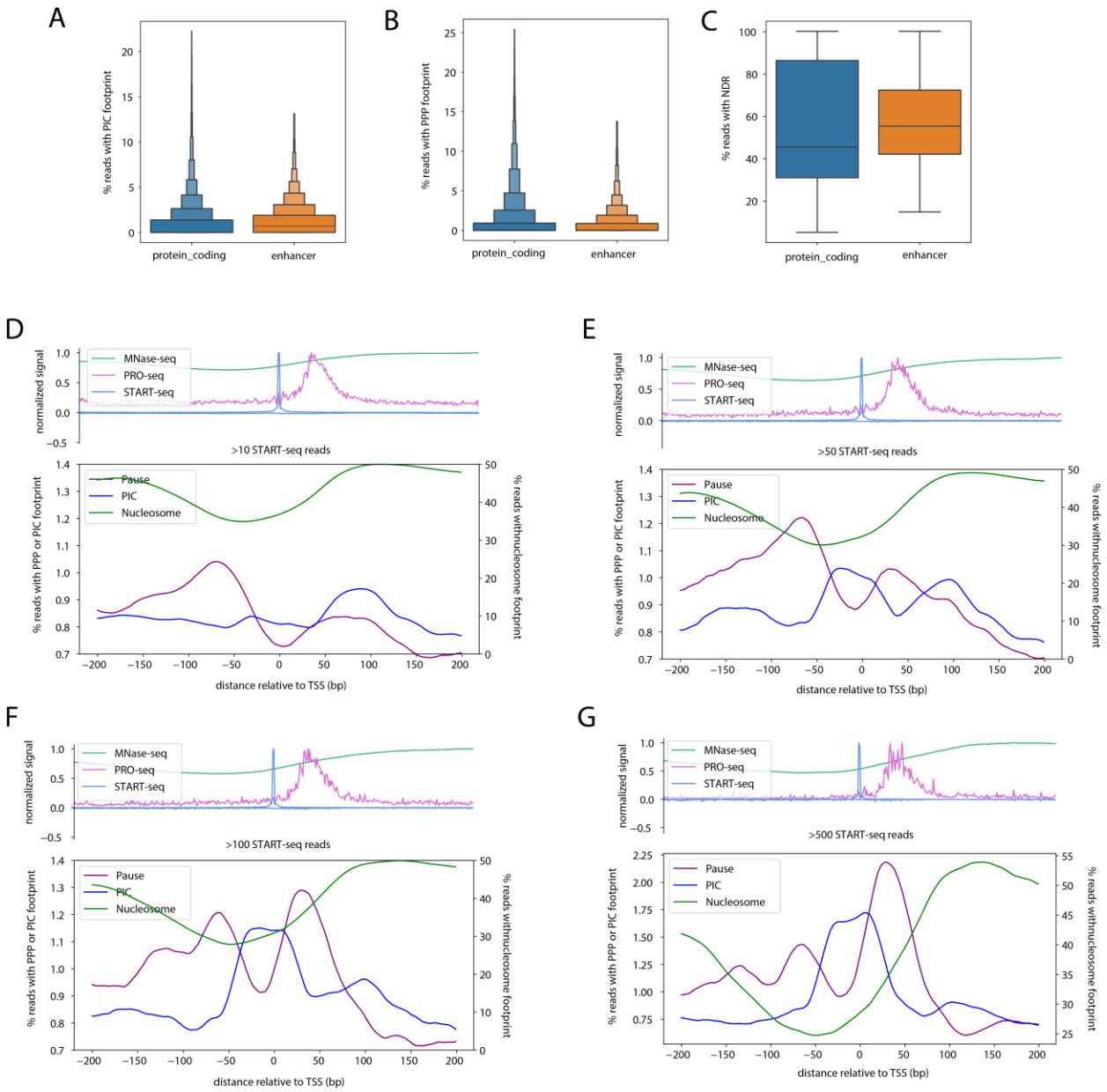
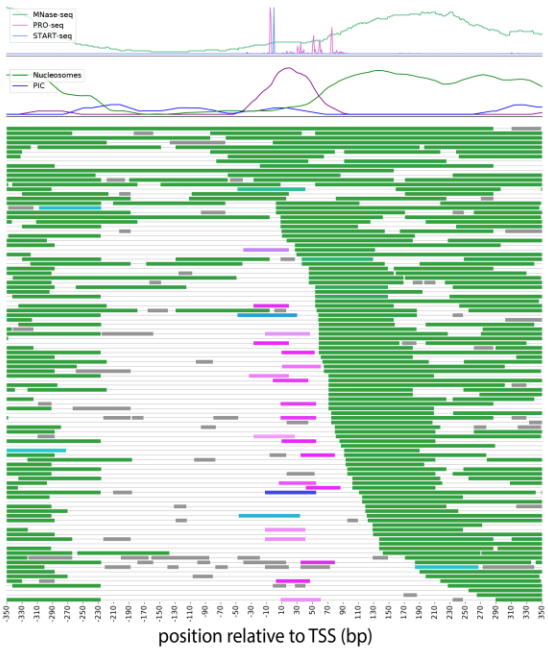
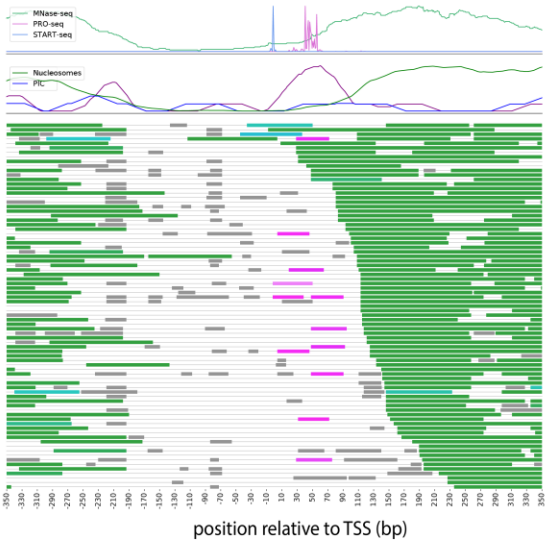
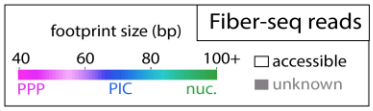


Figure S2.7 (related to Figure 2.5)
(A) Boxenplot showing distribution of percent reads with PIC footprints at protein-coding genes and transcribed enhancers. Transcribed enhancers have a higher average count of PIC footprints, but a lower maximum count. (B) Boxenplot showing distribution of percent reads with PPP footprints at protein-coding genes and transcribed enhancers. Transcribed enhancers have fewer PPP footprints. (C) Boxplot showing fraction reads with an accessible promoter for protein coding genes and transcribed enhancers, showing that enhancers have overall higher accessibility. (D-G) Plots showing (Top) MNAse-seq, PRO-seq, and START-seq signal at transcribed enhancers compared to (Bottom) enrichment of PPP, PIC and nucleosome footprints in Fiber-seq reads. Plots D-G show increasing minimum START-seq reads at the genes included in the plot, demonstrating that increasing evidence of transcription is associated with increased PPP and PIC footprints, as well as the strong enrichment of smaller footprints upstream of the TSS.

Figure S2.8 (related to Figure 2.5)
Fiber-seq reads at example enhancer loci. Footprints are colored based on predicted identity (PPP = pink, PIC = blue, nucleosome = green, unknown = grey). Note the numerous unknown footprints, likely transcription factors, abundant PPP and PIC footprints, and variety in nucleosome positioning.


MNase-seq
PRO-seq
START-seq
Nucleosomes
PIC
position relative to TSS (bp)


MNase-seq
PRO-seq
START-seq
Nucleosomes
PIC
position relative to TSS (bp)


footprint size (bp)
Fiber-seq reads
40
60
80
100+
accessible
unknown
PPP
PIC
nuc.


MNase-seq
PRO-seq
START-seq
Nucleosomes
PIC
position relative to TSS (bp)

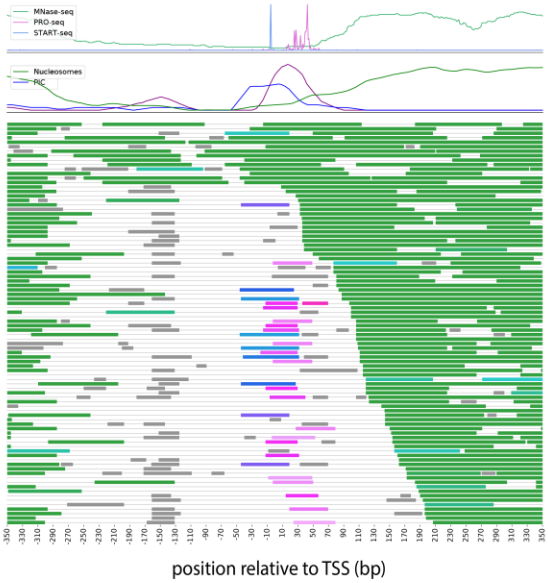

MNase-seq
PRO-seq
START-seq
Nucleosomes
PIC
position relative to TSS (bp)

Figure S2.9 (related to Figure 2.6)
(A) Footprinted Fiber-seq reads at 3 example tRNA loci demonstrating the variety of Pol III tRNA transcription associated footprint enrichment. Footprints are colored based on the definitions set out previously for nucleosomes (green) or Pol III transcription associated footprints (yellow to orange with increasing size).

Appendix 2: Supplemental Figures for Chapter 3

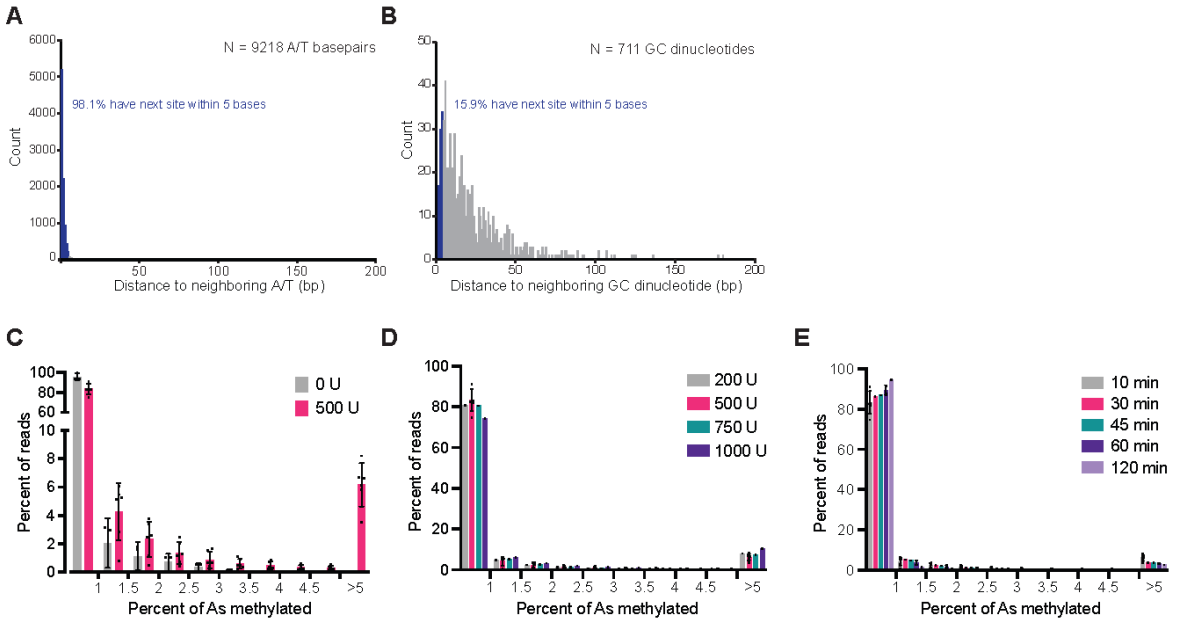


Figure S3.1 (Related to Figure 3.1)
(a) Histogram showing the distance to the next neighboring A/T nucleotide from the 9,218 A/T nucleotides present in the human mitochondrial genome. **(b)** Histogram showing the distance to the next GC dinucleotide from the 711 present in the human mitochondrial genome. **(c)** Bar plot showing the percent of reads binned by the number of m6A modifications per read comparing untreated samples and those treated with 500 U of the m6A-MTase Hia5. Individual dots represent three and six biological replicates for 0 U Hia5 and 500 U Hia5, respectively **(d)** Bar plot showing the percent of reads binned by the number of m6A modifications per read for samples treated with 200, 500, 750, and 1000 U of the m6A-MTase Hia5. **(e)** Bar plot showing the percent of reads binned by the number of m6A modifications per read for samples treated with 500 U of the m6A-MTase Hia5 for 10, 30, 45, 60, and 120 minutes. The 120 minute sample received an additional SAM spike-in after 60 minutes.

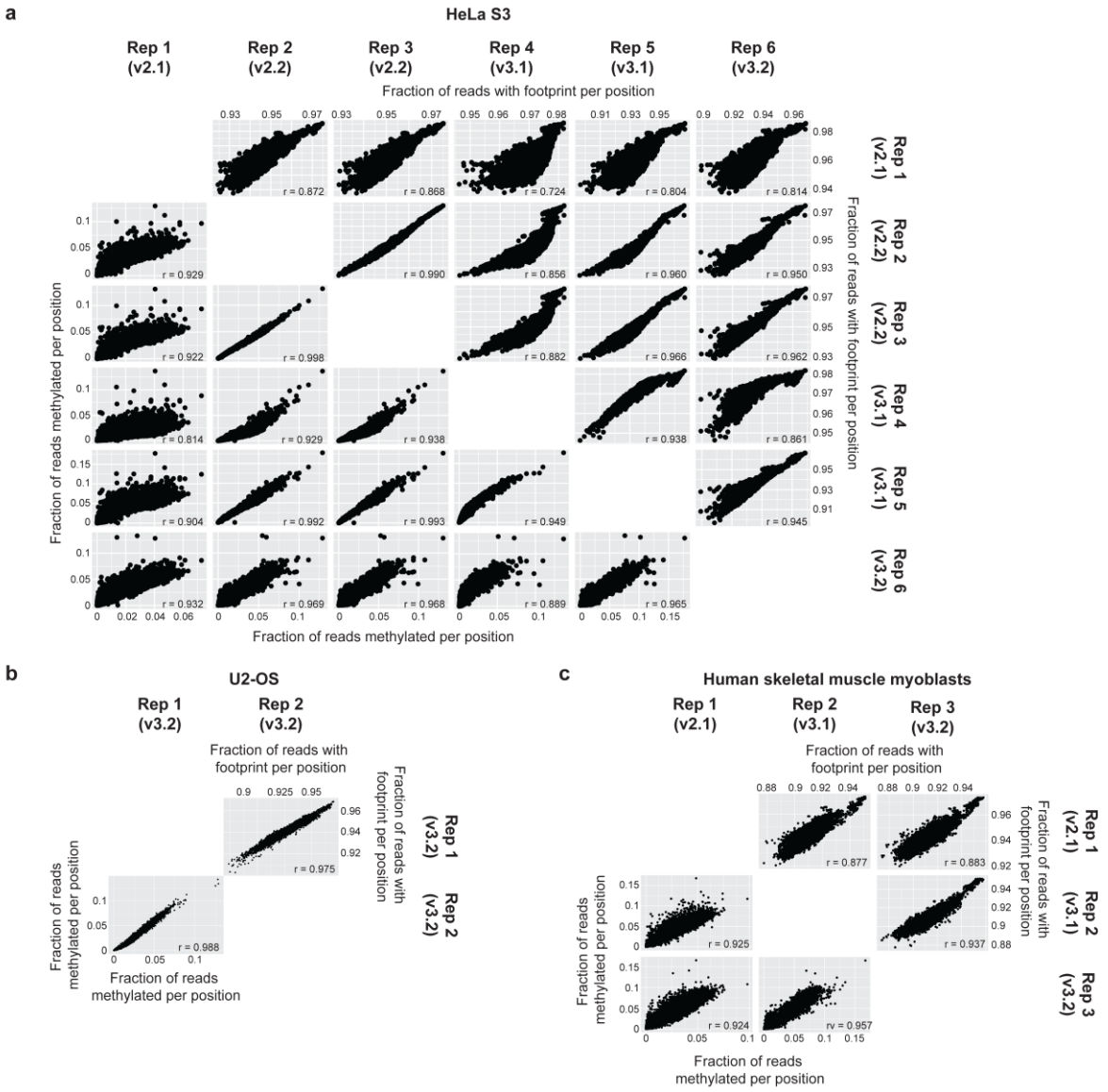


Figure S3.2 (Related to Figures 1, 2)
(a) Correlation scatter plots comparing six mtFiber-seq samples from HeLa S3 cells for (bottom left) fraction of reads methylated at each adenine and (top right) fraction of reads with a footprint at each genomic position. PacBio chemistry version is indicated for each replicate. Pearson's correlation coefficient is shown for each correlation. **(b)** Correlation scatter plots comparing two mtFiber-seq samples from U2-OS cells for (bottom left) fraction of reads methylated at each adenine and (top right) fraction of reads with a footprint at each genomic position. PacBio chemistry version is indicated for each replicate. Pearson's correlation coefficient is shown for each correlation. **(c)** Correlation scatter plots comparing three mtFiber-seq samples from undifferentiated human skeletal muscle myoblasts for (bottom left) fraction of reads methylated at each adenine and (top right) fraction of reads with a footprint at each genomic position. PacBio chemistry version is indicated for each replicate. Pearson's correlation coefficient is shown for each correlation.

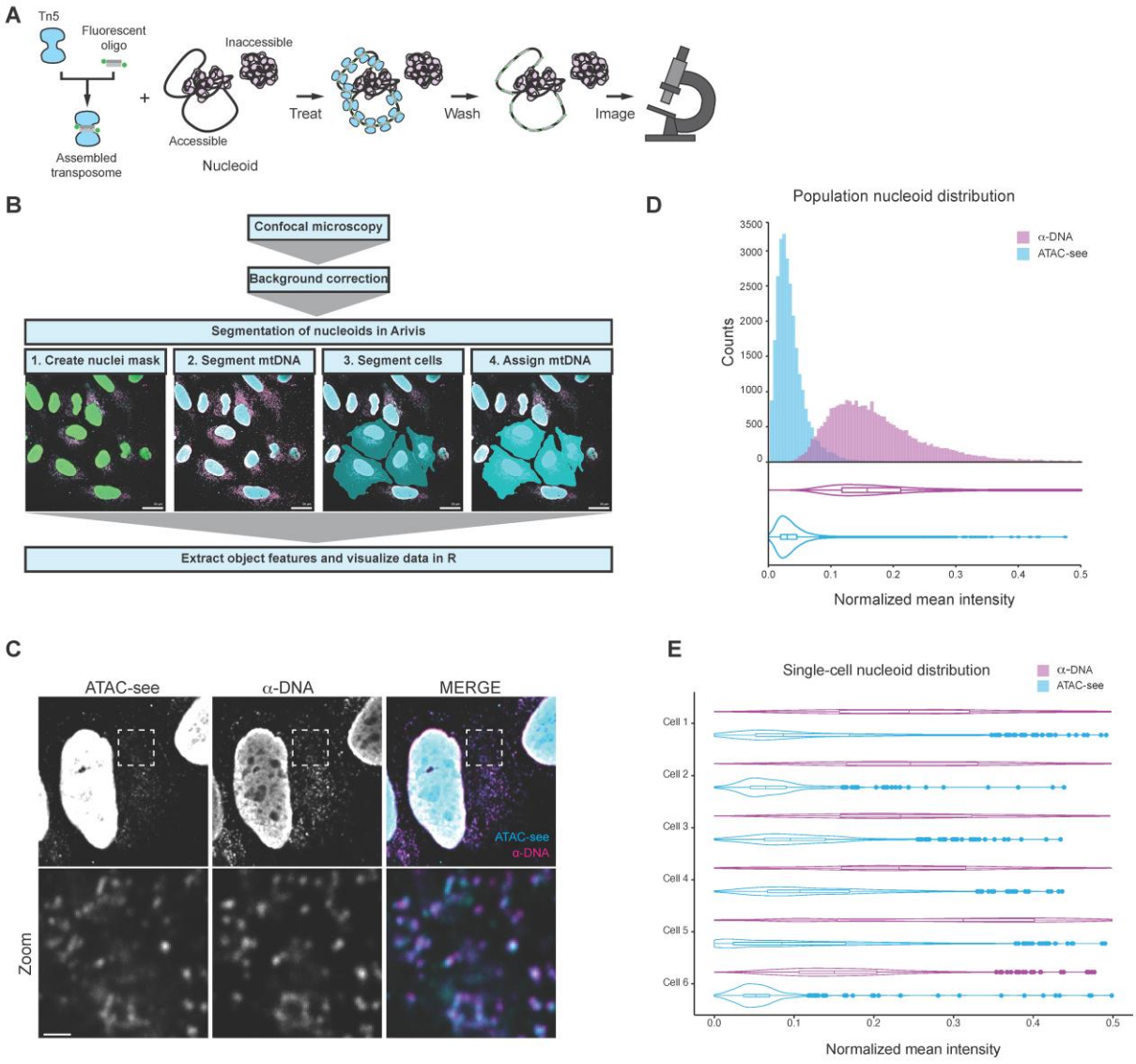


Figure S3.3
(a) Schematic depicting experimental design. Tn5 was loaded with ATTO488-labeled oligos to form active transposomes. U2-OS cells were treated with transposome and imaged by confocal fluorescence microscopy **(b)** Schematic depicting the ATAC-see segmentation and analysis pipeline. Background corrected images were masked and segmented in Arivis. Features of assigned objects were extracted and analyzed. **(c)** Representative image of a U2-OS cell showing ATAC-see and DNA signals. DNA was labeled with an anti-ss/dsDNA antibody that shows preferential labeling of mtDNA[66–68] (Scale bars, 5 µm for single cell, 1 µm for zoom) **(d)** Histogram and violin plot showing the distribution of ATAC-see and DNA signal. Shown are the min-max normalized mean intensities from 27,079 segmented objects **(e)** Distribution of ATAC-see and DNA signal from 6 individual U2-OS cells. Shown are the min-max normalized mean intensities from each segmented object.

Figure S3.4
(a) Tn5 *in vitro* activity in four different reaction buffers measured by DNA fragment analysis. Assembled transposome was mixed with 50 ng plasmid DNA for 30 minutes at 37ºC and DNA fragments were assessed by Agilent TapeStation D1000. Tn5 activity fragments the plasmid DNA, resulting in the appearance of smaller (<500 bp) bands. Four buffers were tested: B1 (50 mM Tris, pH 7.4, 10 mM potassium chloride, 75 µM disodium phosphate, 274 mM sodium chloride), B2 (33 mM Tris, pH 7.8, 66 mM potassium acetate, 11 mM magnesium acetate, 16% N,N-dimethylformamide), B3 (20 mM Tris, pH 7.6, 10 mM magnesium chloride, 20% N,N-dimethylformamide), and B4 (50 mM TAPS, pH 8.5, 25 mM magnesium chloride, 40% PEG8000). Buffer B2 was used for ATAC-see reactions performed in Extended Data Figures 2 and 3. **(b)** Z-projection of the max intensities of a background corrected field-of-view of U2-OS cells treated with Tn5 transposomes in the presence and absence of EDTA. DNA was labeled with an α-ss/dsDNA antibody. (Scale bar, 10 µm) **(c)** Representative images of U2-OS cells sowing ATAC-see signal after treatment with Tn5 over a 60 minute time course. Two intensity ranges are shown to highlight the nuclear and mitochondrial signals .(Scale bar, 10 µm) **(d)** Confocal fluorescence microscopy showing mtDNA labeling throughout the mitochondrial network. A single Z-plane (0.3 µm) is shown. The mitochondrial network is labeled with an α-TOM20 antibody, mtDNA with an α-ss/dsDNA antibody, and chromatin with DAPI (Scale bars, 10 µm, 2 µm for zoom).

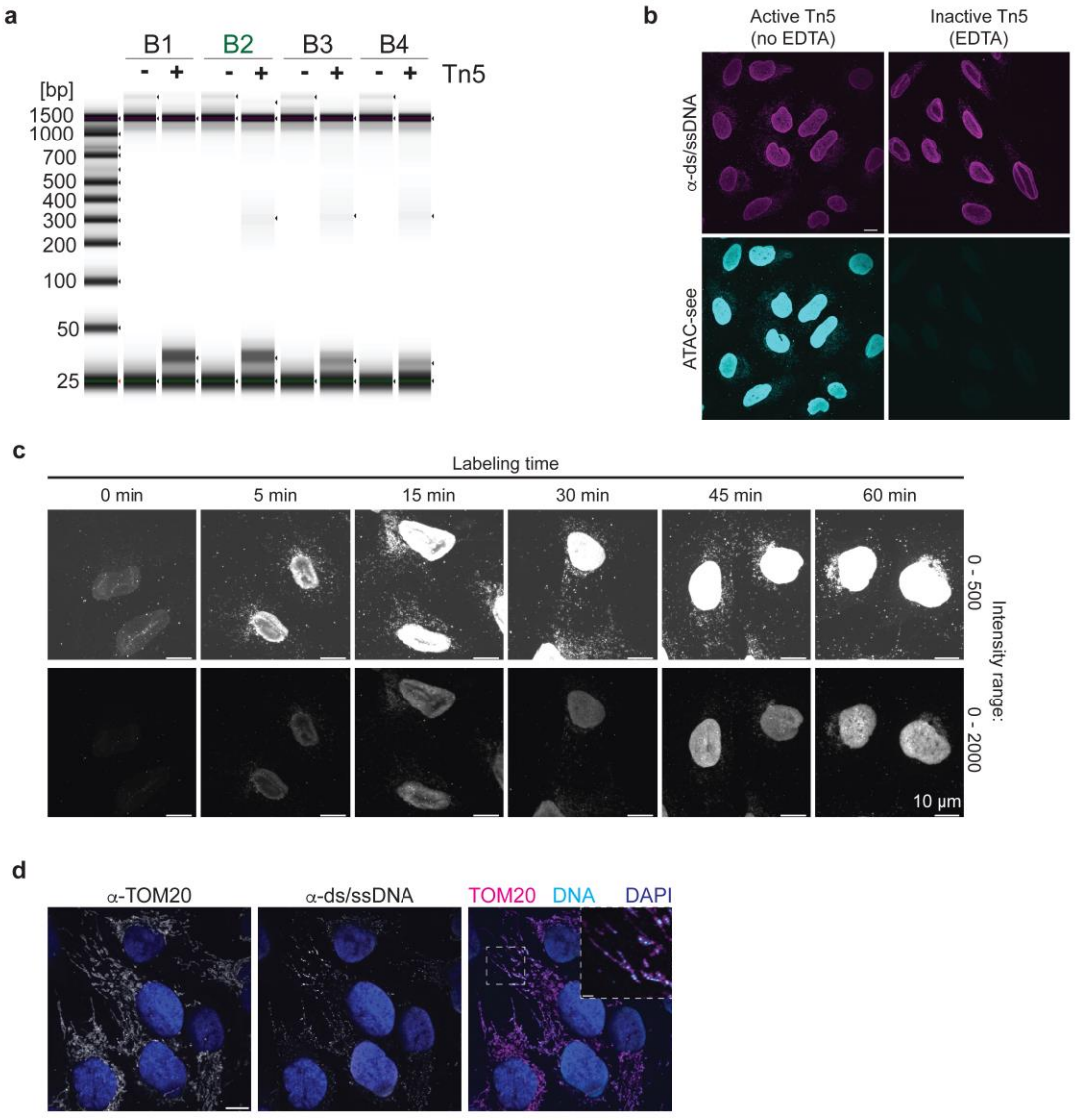
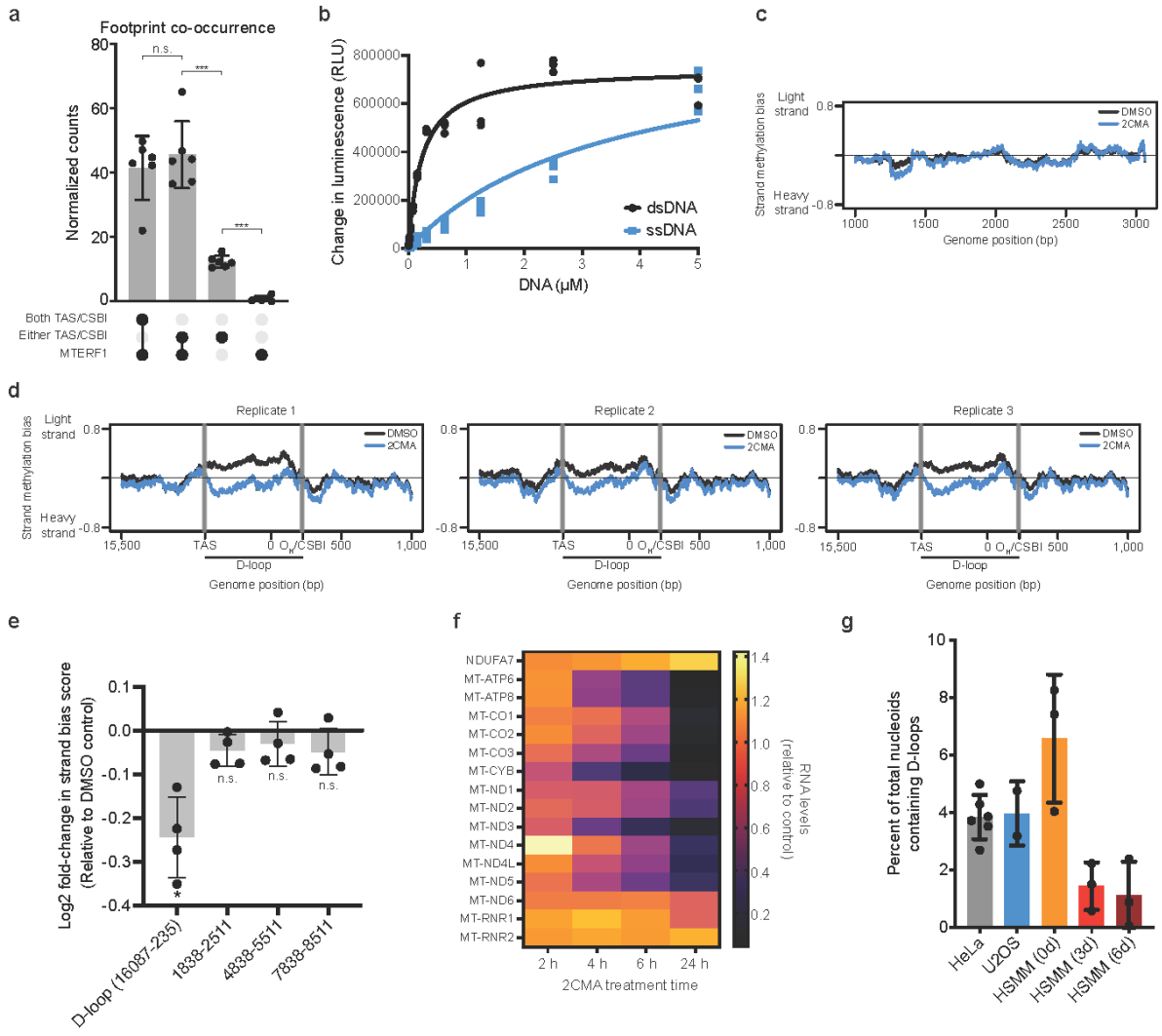

a
B1
B2
B3
B4
- + - + - + - +
Tn5
[bp]
1500
1000
700
500
400
300
200
100
50
25
b
Active Tn5
(no EDTA)
Inactive Tn5
(EDTA)
α-ds/ssDNA
ATAC-see
c
Labeling time
0 min
5 min
15 min
30 min
45 min
60 min
0 - 500
Intensity range:
0 - 2000
10 µm
d
α-TOM20
α-ds/ssDNA
TOM20 DNA DAPI

Figure S3.5 (Related to Figure 3.2)
(a) UpSet plot showing the co-occurrence of footprints on the same molecule at the Termination Associated Sequence (TAS), Conserved Sequence Box I (CSBI), and MTERF1 binding site. Maximum footprint sizes were set for each footprint based on the footprint size distribution at these loci: 60 bp for TAS, 140 bp for CSBI, and 35 bp for MTERF1. Reads with larger footprints at these loci were not considered. Paired Student t-tests were used to compare the frequency of footprint co-occurrence, n.s. signifies p-value > 0.05, *** signifies p-value < 0.001. The categories in the plot represent 1.74%, 1.94%, 3.75%, 1.66%, 4.05%, and 2.08% of the total molecule population for replicates 1-6, respectively. **(b)** Hia5 MTase activity on single-stranded (ssDNA) and double-stranded DNA (dsDNA) substrates. The K_m for dsDNA is 0.233 µM. A lower limit of 3.48 µM was set for the K_m for ssDNA as the reaction never reached saturation even at 5 µM substrate. Results shown are the mean with s.d. from three replicates. **(c)** mtFiber-seq methylation strand bias from genomic positions 1,000 to 3,000 in untreated HeLa cells and cells treated with 2CMA. Methylation bias is calculated as the number of methylations on the light strand and heavy strand, averaged using a 150 nt window and normalized against the region's AT content. Each window was required to have at least 2,250 methylations across all reads combined. **(d)** mtFiber-seq methylation strand bias at the NCR from three biological replicates of HeLa cells treated with DMSO or 2CMA. Methylation bias is calculated as the number of methylations on the light and heavy strands, averaged over a 150 nt sliding window and normalized against the region's AT content. Each window was required to have at least 2,250 methylations across all reads combined **(e)** Log_2 fold-change in the methylation strand bias score between 2CMA treated and control samples at the D-loop and three alternate genomic loci. Individual dots represent four biological replicates. Samples were compared with a Student's t-test, * signifies p-value < 0.05, n.s. Signifies p-values > 0.05. Results from 4 biological replicates shown. **(f)** Heatmap showing RNA levels in HeLa cells after treatment with 2CMA as measured by NanoString. RNA counts were internally normalized to GAPDH. Shown are levels for different treatment times with 2CMA relative to the DMSO control. All RNAs shown are mitochondrially encoded except for NDUFA7, which is nuclear encoded. Shown is the mean from three biological replicates. **(g)** Bar plot showing the percent of total nucleoids containing D-loops in five cell types. A nucleoid was defined as having a D-loop if it had at least 7 methylation events within the region and a Light:Heavy strand methylation ratio of 3.01 or greater. Individual dots represent 6 biological replicates for HeLa S3, two for U2-OS, and 3 for each HSMM differentiation time point.


a
Footprint co-occurrence
Normalized counts
n.s.
***
***
Both TAS/CSBI
Either TAS/CSBI
MTERF1
b
Change in luminescence (RLU)
DNA (μM)
dsDNA
ssDNA
c
Strand methylation bias
Light strand
Heavy strand
DMSO
2CMA
Genome position (bp)
d
Replicate 1
Replicate 2
Replicate 3
Strand methylation bias
Light strand
Heavy strand
DMSO
2CMA
TAS
CSBI
D-loop
Genome position (bp)
e
Log2 fold-change in strand bias score
(Relative to DMSO control)
n.s.
D-loop (16087-235)
1838-2511
4838-5511
7838-8511
f
NDUFA7
MT-ATP6
MT-ATP8
MT-CO1
MT-CO2
MT-CO3
MT-CYB
MT-ND1
MT-ND2
MT-ND3
MT-ND4
MT-ND4L
MT-ND5
MT-ND6
MT-RNR1
MT-RNR2
2 h
4 h
6 h
24 h
2CMA treatment time
RNA levels
(relative to control)
g
Percent of total nucleoids
containing D-loops
HeLa
U2OS
HSMM (0d)
HSMM (3d)
HSMM (6d)

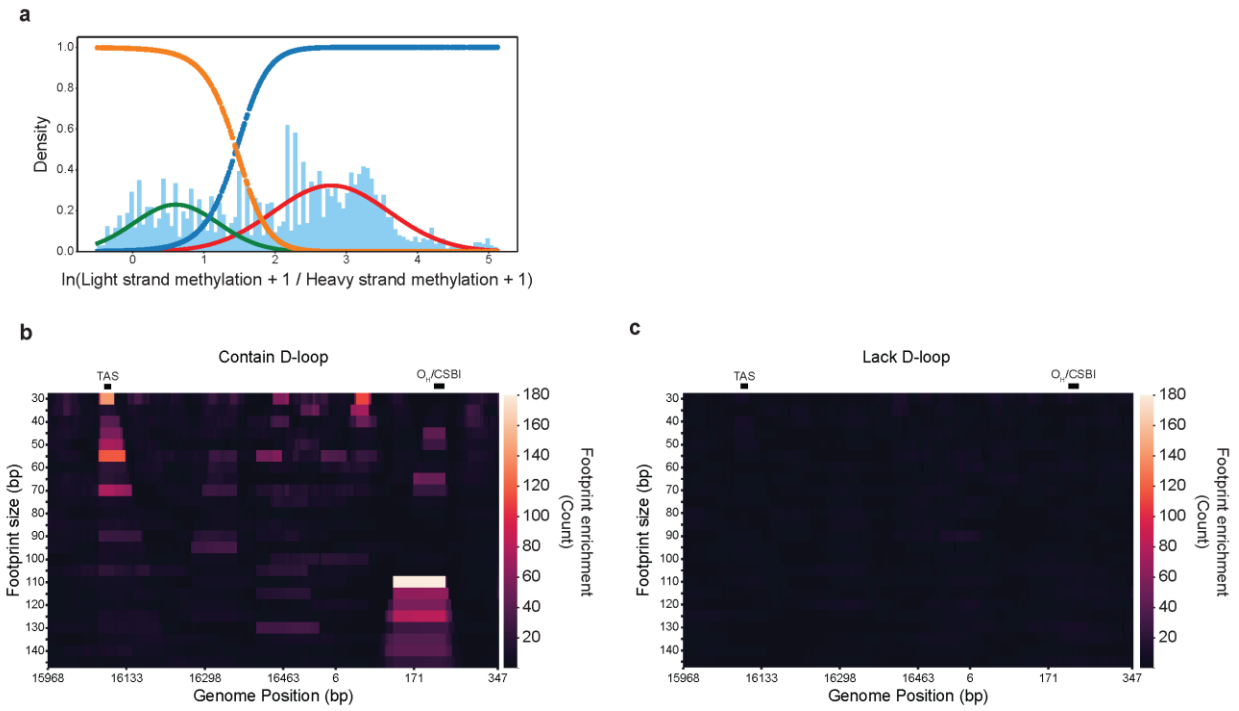


Figure S3.6 (Related to Figure 3.2)
(a) Histogram of the natural log of the ratio of light strand methylation to heavy strand methylation for reads with a minimum of 7 methylations in the D-loop. A Gaussian mixture model was applied and a threshold was identified based on a GMM posterior probability of 0.99. Red and green lines indicate each Gaussian fit. The blue and orange lines indicate the posterior probability of each population. A threshold of 3.01 was determined and used to identify reads with D-loop as shown in main Figure 2h,i and Extended Data Figure 6b,c **(b,c)** Heatmap of the footprint size enrichment at the D-loop region in (b) reads containing a D-loop and in (c) reads lacking a D-loop in HeLa cells. Each row represents a footprint size, and each column shows a position in the genome. Presence of a D-loop was calculated using a GMM with a threshold of 3.01 from the log distribution of the ratio of Light Strand and Heavy Strand methylation.

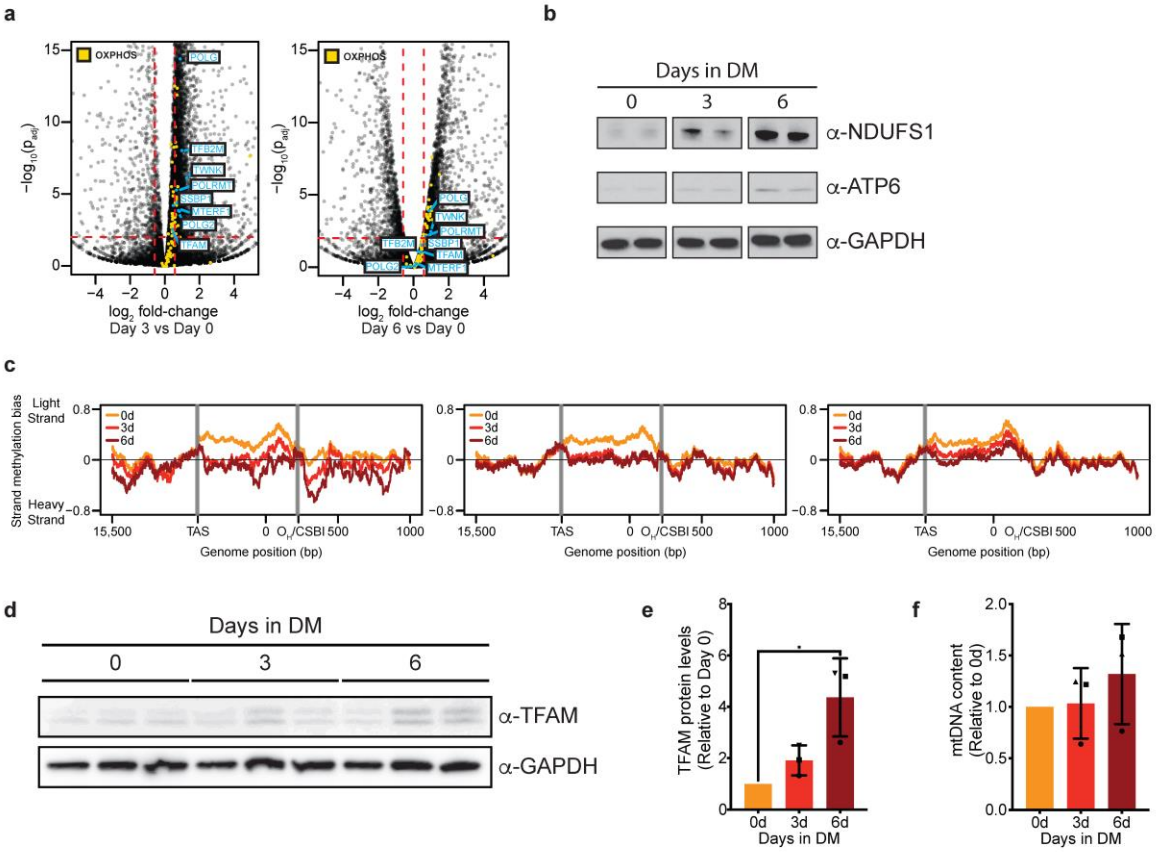


Figure S3.7 (Related to Figure 3.3)
(a) Volcano plot showing differential expression analysis of human skeletal muscle myoblasts in differentiation media for 3 days compared to 0 days (left) and 6 days compared to 0 days (right). Red dotted lines are shown for a p_{adj} value of 0.01 and a fold-change value of 1.5. OXPHOS genes are shown in yellow and key nucleoid-associated proteins are labeled in light blue. **(b)** Western blot of nuclear-encoded (NDUFS1) and mitochondrial-encoded (ATP6) OXPHOS subunits from human skeletal muscle myoblasts in differentiation media for 0, 3, and 6 days. GAPDH was used as a loading control. Shown are two biological replicates. **(c)** mtFiber-seq methylation strand bias at the NCR from three biological replicates of human skeletal muscle myoblasts after 0, 3, and 6 days in differentiation media. Methylation bias is calculated as the number of methylations on the light and heavy strands, averaged over a 150 nt sliding window and normalized against the region's AT content. Each window was required to have at least 2,250 methylations across all reads combined. **(d)** Western blot of TFAM levels from human skeletal muscle myoblasts in differentiation media for 0, 3, and 6 days. GAPDH was used as a loading control. Shown are three biological replicates. **(e)** Quantification of relative mtDNA levels by qPCR from human skeletal muscle myoblasts in differentiation media for 0, 3, and 6 days. Shown are mtDNA levels relative to day 0 from three biological replicates. Separate replicates are indicated by circle, triangle, and square shapes. **(f)** Quantification of TFAM levels by western blot. TFAM bands were quantified and normalized against GAPDH. Shown are the TFAM levels relative to day 0. Shown are three biological replicates. Separate replicates are indicated by circle, triangle, and square shapes.

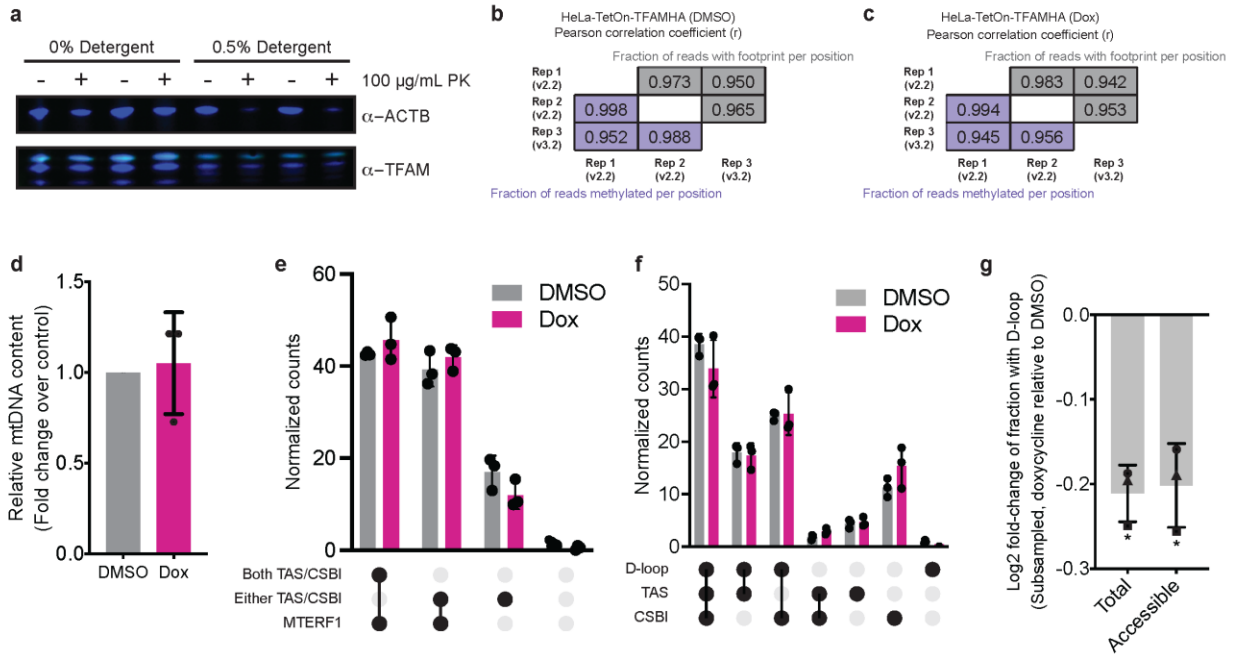
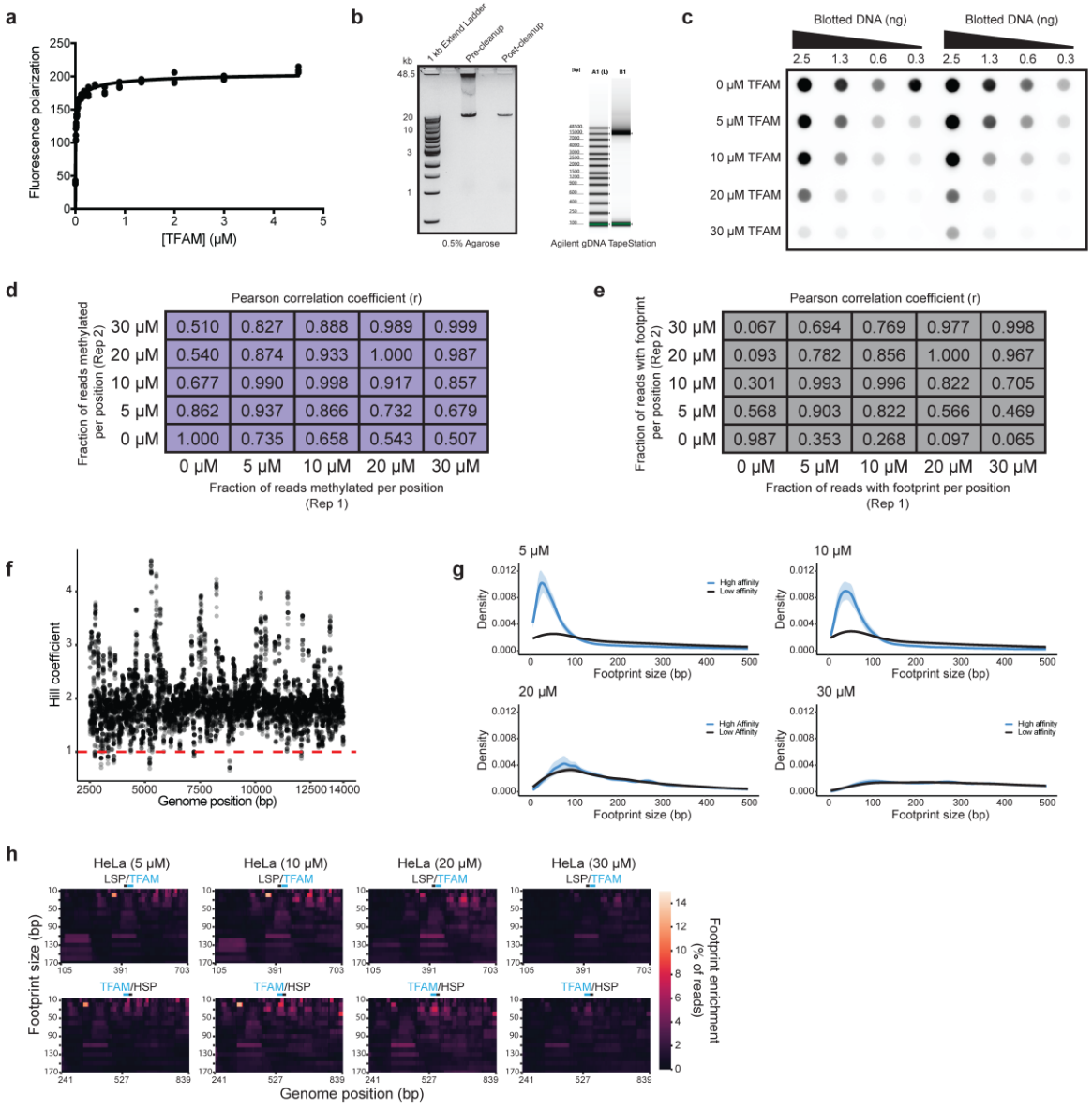


Figure S3.8 (Related to Figure 3.3)
(a) Western blot of Proteinase K protection assay in the presence or absence of 0.5% Tween-20 and NP-40/Igepal-630 to validate overexpressed TFAM-HA localization to mitochondria. TFAM-HA appears as a third upper band and was protected from digestion except in the presence of 0.5% detergent. **(b,c)** Tables of Pearson correlation coefficients from three mtFiber-seq replicates from HeLa S3 TetOn-TFAM-HA cells treated with (b) DMSO or (c) doxycycline for 48 hours comparing (bottom left) fraction of reads methylated at each adenine and (top right) fraction of reads with a footprint at each genomic position. PacBio chemistry version is indicated for each replicate. Pearson's correlation coefficient is shown for each correlation. **(d)** Quantification of relative mtDNA levels by qPCR from HeLa S3 TetOn-TFAM-HA cells treated with DMSO or doxycycline for 48 hours. Shown are mtDNA levels relative to the DMSO control from three biological replicates. **(e)** UpSet plot showing the co-occurrence of footprints on the same molecule at the Termination Associated Sequence (TAS) and Conserved Sequence Box I (CSBI) with footprints at the MTERF1 binding site from HeLa S3 TetOn-TFAM-HA cells treated with DMSO or doxycycline for 48 hours. Maximum footprint sizes were set for each footprint based on the footprint size distribution at these loci: 60 bp for TAS, 140 bp for CSBI, and 35 bp for MTERF1. A maximum footprint size was set at 170 bp. Reads with larger footprints at these loci were not considered. Individual dots represent three biological replicates. The categories in the plot represent 0.59%, 0.98%, and 1.67% of the total molecule population for DMSO replicates 1-3, respectively, and 0.44%, 0.54% and 1.31% of the total molecule population for doxycycline replicates 1-3, respectively. **(f)** UpSet plot showing the co-occurrence of footprints on the same molecule at the TAS and CSBI with containing a D-loop in HeLa S3 TetOn-TFAM-HA cells treated with DMSO or doxycycline for 48 hours. Maximum footprint sizes of 60 bp and 140 bp were used for TAS and CSBI, respectively. Individual dots represent three biological replicates. The categories in the plot represent 1.23%, 1.85%, and 3.37% of the total molecule population for DMSO replicates 1-3, respectively, and 1.00%, 1.14%, and 2.61% of the total molecule population for doxycycline replicates 1-3, respectively. **(g)** Log_2 fold-change in the fraction of reads containing a D-loop from all nucleoids and accessible nucleoids (>1% adenines methylated per read) in HeLa S3 TetOn-TFAM-HA cells treated with doxycycline for 48 hours relative to the DMSO control. Data were subsampled to each other to match methylation distributions.

Figure S3.9 (Related to Figure 3.4)
(a) TFAM binding to a 28mer corresponding to the HSP TFAM binding site measured by fluorescence polarization. The K_D was determined to be 6.2 nM. Results shown are the mean with s.d. from three replicates. **(b)** (Left) 0.5% agarose gel showing mtDNA LR-PCR product before and after column cleanup. (Right) Genomic DNA ScreenTape analysis showing mtDNA LR-PCR product. **(c)** Two replicates of a dot blot assessing methylation of mtDNA with increasing concentrations of TFAM. A dilution series of DNA amounts were adsorbed onto a nitrocellulose membrane, crosslinked, and detected with an anti-m6A antibody. **(d,e)** Tables of Pearson correlation coefficients between replicates and TFAM concentrations for (d) the fraction of reads methylated at each adenine and (e) the fraction of reads with a footprint at each genomic position from two replicates with each TFAM concentration. **(f)** Hill coefficients calculated from genomic position 2,500 to 14,000 from a four parameter logistics regression based on the fraction of reads protected with a footprint at least 20 bp long. Results are from the average of two replicates. **(g)** Meta density plots of footprint sizes from (top) 26 high affinity sites and (bottom) 64 low affinity sites at each TFAM concentration with the 95% confidence intervals shaded. Distributions between high and low affinity sites are significantly different at 5 μM TFAM (KS test, $p << 0.05$, $D = 0.32$) and 10 μM TFAM (KS test, $p << 0.05$, $D = 0.29$). **(h)** Heatmaps of footprint size enrichment at the light strand promoter (LSP) (top) and heavy strand promoter (HSP) (bottom) from HeLa cells subsampled to each TFAM concentration. Each heatmap row represents a footprint size, and each column shows a position in the genome.


a
Fluorescence polarization
[TFAM] (μM)
b
1 kb Extend Ladder
Pre-cleanup
Post-cleanup
kb
48.5
20
10
3
1
0.5% Agarose
Agilent gDNA TapeStation
c
Blotted DNA (ng)
Blotted DNA (ng)
2.5 1.3 0.6 0.3 2.5 1.3 0.6 0.3
0 μM TFAM
5 μM TFAM
10 μM TFAM
20 μM TFAM
30 μM TFAM
d
Pearson correlation coefficient (r)
Fraction of reads methylated per position (Rep 2)
30 μM 0.510 0.827 0.888 0.989 0.999
20 μM 0.540 0.874 0.933 1.000 0.987
10 μM 0.677 0.990 0.998 0.917 0.857
5 μM 0.862 0.937 0.866 0.732 0.679
0 μM 1.000 0.735 0.658 0.543 0.507
0 μM 5 μM 10 μM 20 μM 30 μM
Fraction of reads methylated per position (Rep 1)
e
Pearson correlation coefficient (r)
Fraction of reads with footprint per position (Rep 2)
30 μM 0.067 0.694 0.769 0.977 0.998
20 μM 0.093 0.782 0.856 1.000 0.967
10 μM 0.301 0.993 0.996 0.822 0.705
5 μM 0.568 0.903 0.822 0.566 0.469
0 μM 0.987 0.353 0.268 0.097 0.065
0 μM 5 μM 10 μM 20 μM 30 μM
Fraction of reads with footprint per position (Rep 1)
f
Hill coefficient
Genome position (bp)
g
5 μM
10 μM
20 μM
30 μM
Density
Footprint size (bp)
High affinity
Low affinity
h
HeLa (5 μM)
HeLa (10 μM)
HeLa (20 μM)
HeLa (30 μM)
LSP/TFAM
TFAM/HSP
Footprint size (bp)
Genome position (bp)
Footprint enrichment (% of reads)

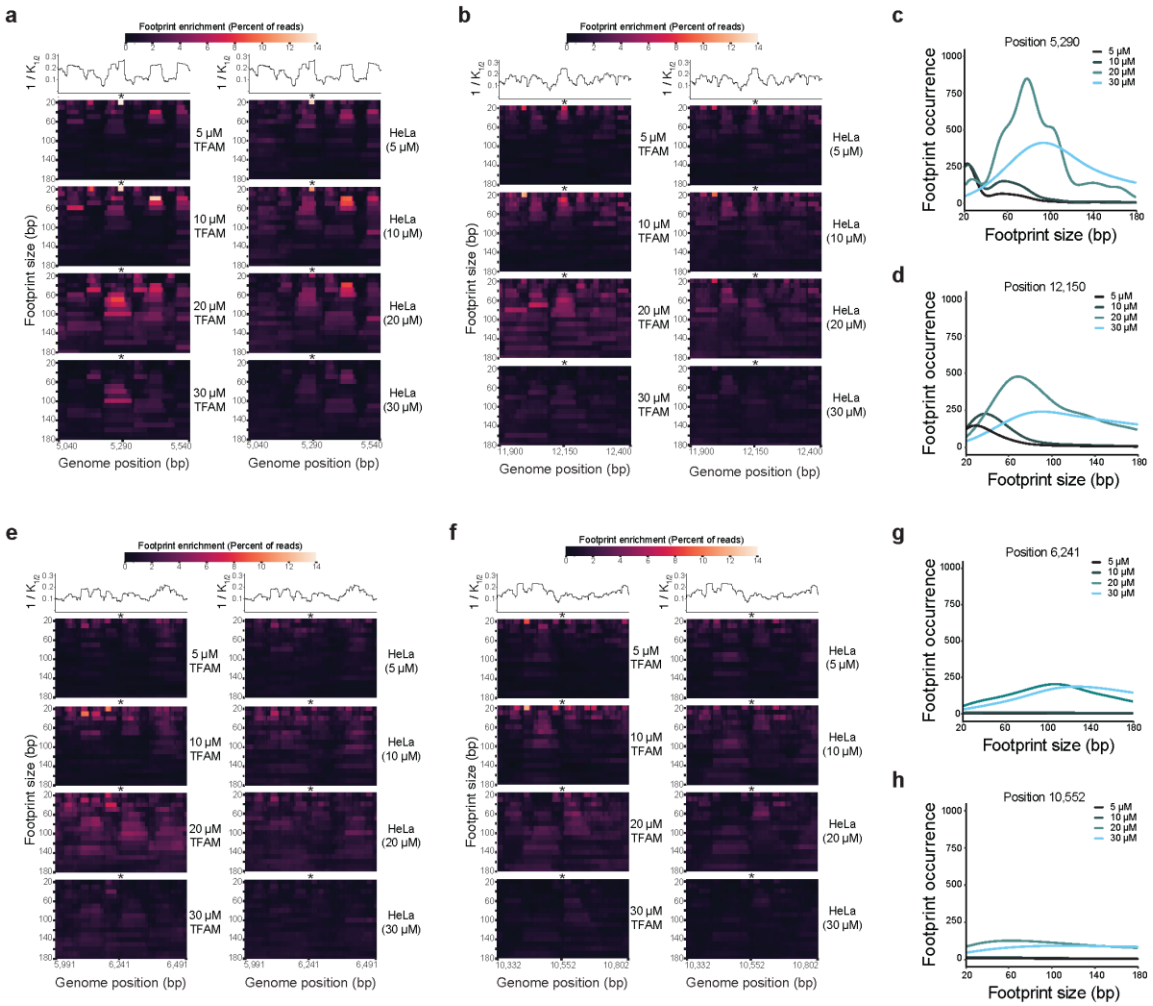


Figure S3.10 (Related to Figure 3.4)
(a,b) Heatmaps of the footprint size enrichment from two high affinity sites for each concentration of TFAM and from HeLa cells subsampled to each concentration: (a) from position 5,040 to 5,540 and (b) from position 11,900 to 12,400. Each heatmap row represents a footprint size, and each column shows a position in the genome. Line plots indicating the $1/K_{1/2}$ across these loci are shown above the heatmaps. **(c,d)** Density plots showing the footprint size distribution between 20 and 180 bp at (c) position 5,290 and (d) position 12,150 as a function of TFAM concentration. **(e,f)** Heatmaps of the footprint size enrich from two low affinity sites for each concentration of TFAM and from HeLa cells subsampled to each concentration: (a) from position 5,991 to 6,491 and (b) from position 10,332 to 10,802. Each heatmap row represents a footprint size, and each column shows a position in the genome. Line plots indicating the $1/K_{1/2}$ across these loci are shown above the heatmaps. **(g,h)** Density plots showing the footprint size distribution between 20 and 180 bp at (g) position 6,241 and (h) position 10,552 as a function of TFAM concentration.

Appendix 3: Methods for Chapter 2

Defining TSSs, CAGE-seq peaks and PRO-seq peaks

To call CAGE- and PRO-seq peaks near TSSs, we used a sliding 20bp window from -50 to +50nt relative to the TSS (CAGE-seq) or 0 to 100nt relative to the TSS (PRO-seq) around the TSS, assigning maxima within each window as peaks. A minimum of 10 reads was required to call a peak, with the strongest peak above that threshold being assigned as the primary PRO-seq or CAGE-seq peak. In the case of the primary CAGE-seq peak, we adjusted the TSS used in subsequent analyses from the reference annotation to that peak.

Calculating pause index and expression levels

Pause index and expression were both calculated using PRO-seq data. Pause index was calculated as the ratio of PRO-seq coverage in the promoter region (-100 to 300nt relative to the TSS) to the body of the gene, normalized by gene length. Expression was based on PRO-seq coverage in the gene body normalized by gene length.

Fiber-seq

We carried out Fiber-seq as described in Stergachis *et al.* 2020 on *Drosophila melanogaster* S2 cells, 8 million cells per sample. For samples treated with triptolide, a final concentration of 10uM of triptolide was added to the cells 30 minutes prior to harvesting, with an identical molarity maintained through the Fiber-seq protocol. For samples treated with alpha-amanitin, alpha-amanitin was added from the nuclear isolation step onwards at a final concentration of 25uM. The control Fiber-seq datasets were carried out identically, except that the -Hia5 sample had no added Hia5 and that the dechromatinized sample used purified genomic DNA.

Fiber-seq processing

Fiber-seq datasets were assigned ccs values using the CCS tool from Pacific Biosciences. They were then mapped to the dm6 genome using the pbmm2 package. Methylations were called using the fibertools package with the SCNN mode and converted to a bed file using a threshold of 250 for the m6A score.

FiberHMM

Footprints were called on Fiber-seq reads using FiberHMM. FiberHMM is based on a hidden Markov model (HMM) with two hidden states-- accessible and inaccessible. To account for sequence-related biases from Hia5 or the methylation caller, at each position the model takes into account both the base and its +/-3nt sequence context. The emission probabilities used in the model are the probabilities of methylation of a given base with its +/-3nt sequence context in an accessible or inaccessible state based on experimental-derived methylation rates from control datasets. The probability of methylation given an accessible state was based on the methylation frequency in a dataset generated from dechromatinized *Drosophila* S2 cell genomic DNA. The probability of methylation given an inaccessible state was based on the methylation frequency in a dataset generated from *Drosophila* S2 cells untreated with Hia5. Transition and starting probabilities for the HMM were trained 20 times on 1000 reads sampled in equal proportions from all untreated datasets, with initial probabilities picked from the Dirichlet distribution with all parameters set to 1. The best model was chosen and then used for all subsequent footprint calling.

Identifying size ranges for Nucleosome footprints

As the vast majority of footprints over 90 bp are likely nucleosomes based on phasing and the expected footprint size of a nucleosome, we defined nucleosomes in general as footprints greater than 90bp. Most nucleosome footprints are single nucleosomes, which range from 90-200bp, which is the definition used for analyses involving single nucleosomes. Two nucleosomes

combined into a single footprint also occur, with a size of 200-350bp. Finally, further combinations of 3, 4, and more nucleosomes are found, but only represent a tiny fraction of the population of nucleosome footprints.

Identifying size ranges for PPP and PIC footprints

To identify the size ranges of PPP and PIC footprints, we first identified segments of Fiber-seq reads overlapping promoters with a pause index greater than 10. We then aligned these read segments around the primary TSS of each promoter, plotting the enrichment of different footprint sizes at positions around the TSS in the form of a heatmap (Fig. 1C). This heatmap allowed us to identify size ranges of two enriched footprint populations around the TSS, corresponding to putative PPP and PIC footprints at 40-60bp and 60-80bp respectively.

Defining PPP and PIC footprints

For our analyses we defined PPP and PIC footprints using two requirements. The first requirement was that the footprint must be within the expected size range of 40-60bp or 60-80bp respectively. The second requirement was that the footprint must have a PRO-seq or CAGE-seq peak overlapping the footprint within the middle 80% of the footprint, and that the PRO-seq peak or CAGE-seq peak within a range of 0 to +100nt or +/- 50nt around the TSS respectively. Due to higher false positive rates of PPP calls at genes with low pausing which were treated with triptolide, we used a slightly stricter definition of the PPP footprint for the analysis in Figure 2.4, requiring that the PRO-seq peak overlap the middle 50% of the putative PPP footprint.

Comparing to existing structures

To compare footprint sizes to structures, structures from the referenced studies were accessed and downloaded from RCSB PDB and opened in PyMOL. The scaffold DNA bases

found within the structure and obscured by the protein structure were counted manually to estimate the predicted footprint size associated with the structure.

Pause-inhibited initiation quantification

To quantify the inhibition of simultaneous PPP and PIC occupancy, we first identified all genes which had at least one PPP and PIC footprint across all corresponding Fiber-seq reads. We then binned genes based on the distance between the primary CAGE-seq and PRO-seq peak and created a contingency table for the count of individual, simultaneous, and absent PPP and PIC footprints at these binned loci. We then used Fisher's exact test to determine the odds ratio and p-value of having a simultaneous PPP and PIC footprint within each distance bin.

Identifying reads with an accessible promoter

Reads with an accessible promoter for a given gene were defined as those reads lacking a nucleosome footprint overlapping within +/-10nt around the primary TSS of that gene.

Sampling reads based on PPP and PIC footprints

To allow for direct comparison between features of reads with or without a PPP or PIC footprint on a global scale we employed a sampling approach. At each gene we identified all reads with an NDR. These reads were then split based on if they contained a PPP footprint, a PIC footprint, or neither. Reads at each gene were then sampled to match the count of reads with a PPP footprint to those with neither a PPP or PIC, or to match the count of reads with a PIC footprints to the count of those with neither a PPP or PIC. All analyses involving sampling were repeated with a different randomization seed at least 50 times to make sure that any conclusions were stable and not a result of sampling biases. This was found to be universally true, as the count of PPP and PIC footprints was sufficient to capture a stable sample of reads.

Sampling reads based on distance

For comparison of coordination of pairs of genes based on the existence of TAD boundaries or ChIP-seq peaks, we carried out a similar sampling method as described above. We randomly sampled a selection of pairs of genes based on their distance to capture an equal distribution of distances for each condition (with or without a TAD boundary or ChIP-seq peak). This sampling was repeated 10,000x to capture the sampling variance.

Global nucleosome feature analyses

To identify global changes to nucleosome size and positioning, we first sampled as previously outlined our reads to match the proportion of reads with or without a PPP or PIC footprint at each gene. We then calculated a feature-- the distance to the edge of a nucleosome footprint, the distance to the center of a nucleosome, or the size of a nucleosome footprint-- or plotted the overall distribution of nucleosome footprints in each subset of reads.

Segmenting downstream and upstream nucleosomes

To identify the most likely position of nucleosomes upstream and downstream of the -1 and +1 nucleosome respectively, we used a gaussian mixture model (GMM) trained on the distribution of centers of nucleosome footprints to identify 95% confidence intervals for each nucleosome within the range. The number of states was identified by eye depending on the range used.

+1 and -1 nucleosome size and distance on a per-gene basis

To compare nucleosome footprint distance and size on a per-gene basis, we calculated the size and distance across all reads at each individual locus and carried out a Wilcoxon rank sums test, comparing the reads with a PPP footprint to those without a PPP footprint.

+1 nucleosome size and distance on a per-read basis

To compare nucleosome footprint distance and size on a per-read basis, we calculated the size and distance of the +1 nucleosome across all reads at each individual gene. We then used bootstrapping at each individual gene across all reads with an NDR to calculate a mean size or distance of the +1 nucleosome as well as a confidence interval. Each read with a PPP footprint was then tested against the mean and confidence interval to identify the difference and significance of the +1 distance or size of that read compared to the reads at the origin gene.

Defining elongating Pol II footprints

Elongating Pol II footprints were defined based on matching the size range of a PPP footprint, 40-60bp, and existing in gene bodies outside of a range of +/-300bp around an annotated TSS or predicted eRNA TSS in the genome. The frequency of these footprints was defined based on the baseline footprints of that size per kilobase in intergenic regions outside of a range of +/-300bp from an annotated TSS or predicted eRNA TSS in the genome. The mean intergenic footprints per kilobase for a given dataset was used as a normalization factor to determine relative enrichment of putative elongating Pol II footprints for reads originating from that dataset.

Identifying transcribed enhancers

Transcribed enhancers were called by identifying TSSs within a 1kb window of enhancers from STARR-seq using START-seq data. Enhancer TSSs could not be within 500bp of any annotated TSS in the genome, including pseudogenes, noncoding RNAs, and other transcribed enhancers with lower peak START-seq signal.

Defining tRNA Pol III transcription-associated footprints

We used a similar strategy as for PPP and PIC footprints to identify Pol III transcription associated footprints. First, we generated a heatmap to identify regions and sizes of enriched footprints aligned around all tRNA TSSs. We identified several unique populations of footprints overlapping known binding sites of TFIIIB and TFIIIC within and upstream of tRNA genes, with sizes ranging from 30-140bp. Of note, while there were several distinct populations of fibers with different patterns of these footprints, we found that all fibers with any of the footprints had a footprint starting at the TBP binding site, -55 to -45 bp upstream of the TSS. As a result, we defined fibers with Pol III tRNA transcription-associated footprints as those with a footprint between 30 and 140bp starting 45-55bp upstream of the TSS.

Defining 5S rRNA Pol III transcription-associated footprints

5S rRNA genes showed a similar pattern of footprints to tRNA genes, but with slightly larger footprints overall. As such, Pol III 5S rRNA transcription associated footprints were defined identically as at tRNA genes, except that the maximum size was set to 160bp.

Correlation of tRNA expression score to codon frequency

The tRNA expression score was calculated as the fraction of reads containing a Pol III transcription-associated footprint compared to those without. We then identified all tRNA genes with coverage within the middle 95% of the overall distribution of coverage across the genome. We then identified all families of tRNA genes sharing a given codon sequence where all known members had an acceptable level of coverage. We then compared the Pearson correlation of the overall count of copies of tRNA genes in each family with their codon frequency in the *Drosophila* genome to the correlation of codon frequency with the sum of expression scores for each family.

Global coordination analysis

The following describes the methodology used for the PPP/PIC coding promoter coordination analysis-- however this approach holds true for all global coordination analyses, substituting the feature (PPP/PIC, promoter accessibility, Pol III transcription associated footprints) and location (Pol II promoter, enhancer, tRNA, 5S rRNA). First, we found all pairs of genes with a distance between their TSSs of less than 20kb. We then identified all reads overlapping both members of each pair of genes, and counted the frequency of individual, simultaneous, and absent PPP/PIC footprints to generate a contingency table. Genes were then binned by distance in such a manner as to allow for a roughly equivalent amount of simultaneous PPP and PIC footprints and a minimum of 10 reads. The contingency tables for all pairs of genes in each bin were then merged. We then used Fisher's exact test to calculate an odds ratio and p-value associated with simultaneous PPP/PIC occupancy in each set of reads binned by distance.

Global coordination, alternative randomization test

We also carried out an alternative statistical test to test coordination on a global scale but based on randomization at a local scale, again identically across different features and locations. First, we found all pairs of genes with a distance between their TSSs of less than 20kb. We then identified all reads overlapping both members of each pair of genes and counted the frequency of simultaneous PPP/PIC footprints. We then randomized the read order of one of the genes in each pair 100,000 times, counting the number of simultaneous PPP/PIC footprints found at each pair each time. We then binned the genes by distance in such a manner as to allow for a roughly equivalent count of simultaneous PPP/PIC footprints in each bin. We then summed the true simultaneous count of PPP/PIC footprints as well as the count in each randomized trial for each bin, and then calculated an empirical odds ratio and p-value for the true count of simultaneous footprints based on the distribution of the randomized trials. These tests gave a concurring result in terms of distance dependence and significance to the previously outlined tests.

Per-locus coordination analysis

Coordination on a per-locus level was again quantified identically across features and locations. First, we identified all pairs of genes with a distance between their TSSs of less than 20kb. We then identified all reads overlapping both members of each pair of genes, and counted the frequency of individual, simultaneous, and absent PPP/PIC footprints to generate a contingency table for each pair. We then carried out Fisher's exact test to identify an associated p-value and odds ratio at each individual locus. Pairs of genes were then assigned as significantly coordinated if the odds ratio was greater than 1 and the p-value was less than .05. Gene pairs were then binned based on distance and the fraction of coordinated pairs was calculated for each bin.

Appendix 4: Methods for Chapter 3

Cell Lines and culture

HeLa S3 cells (ATCC CCL-2.2) were grown in DMEM containing glucose and pyruvate (Thermo Fisher Scientific 11995073) supplemented with 10% Fetal Bovine Serum (Thermo Fisher Scientific 10437028). U2-OS cells (ATCC HTB-96) were grown in McCoy's 5A Media (ATCC 30-2007) supplemented with 10% Fetal Bovine Serum. Human Skeletal Muscle Myoblasts were from anonymous healthy control samples kindly provided by Dr. Brendan Battersby (Institute of Biotechnology, University of Helsinki). Myoblasts were grown in Human Skeletal Muscle Cell Basal Media containing growth supplement (Cell Applications, INC. 151K-500). To differentiate into myocytes, myoblasts at 80% confluency were switched to DMEM containing glucose and pyruvate (Thermo Fisher Scientific 11995073) supplemented with 2% Heat Inactivated Horse Serum (Thermo Fisher Scientific 26050-070) and 0.4 µg/mL dexamethasone. Media was then replaced every 24 hours, and cells were harvested at either 3 days or 6 days post induction. Dermal human fibroblasts were grown in DMEM containing glucose and pyruvate supplemented with 10% Fetal Bovine Serum. Cell lines were tested for mycoplasma contamination and confirmed negative by PCR using a Universal Mycoplasma Detection Kit (ATCC 30-1012K). For 2'-C-methyladenosine (2CMA) treatment, HeLa S3 cells at roughly 60% confluency were treated with either 100 µM 2CMA in DMSO or DMSO alone and harvested after 2, 4, 6, or 24 hours.

Constructs and Cloning

For recombinant TFAM, cDNA corresponding to TFAM lacking the mitochondrial targeting sequence (amino acids 50-246) was cloned into pET30a using the BamHI and NotI cut sites. The construct contains an N-terminal 6xHis tag and a TEV cleavage site just upstream of the TFAM coding region to allow for the generation of untagged protein. For the TFAM-HA overexpression HeLa cell line, full-length TFAM with a C-terminal HA-tag connected by an SGGS linker was

cloned into pCW57.1 using the NheI and BamHI cut sites. Lentivirus was generated using HEK293 cells using pCW57.1_TFAMHA, pRSV-REV, pMDLg, and pMD2.G with Lipofectamine 3000 (Thermo Fisher Scientific L3000008) according to manufacturer's instructions. Collected virus was added to HeLa S3 cells grown to 50% confluency in 6-well tissue culture dishes. 1 mL virus was added along with 1 mL DMEM + 10% FBS and polybrene to a final concentration of 8 µg/mL. After 24 hours, cells were selected with 2 µg/mL puromycin until all negative control cells had died.

Immunoblotting and antibodies

Lysates used for immunoblotting were prepared in RIPA buffer in the presence of protease inhibitors. Lysates were normalized using Qubit Protein Assay (Invitrogen Q33211), and standard SDS-PAGE and blotting protocols were used. The following antibodies were used in this study: N6-methyladenosine (Active Motif, 61995), TFAM (Santa Cruz, sc-376672), TFAM (Proteintech, 22586-1-AP), TOMM40 (Proteintech, 18409-1-AP), TOM20 (Santa Cruz, sc-17764), HSP60 (Cell Signaling, D307), MT-CO1/COX1 (Abcam, ab14705), ACTB (Cell Signaling, 3700), ds/ssDNA (PROGEN, 61014), Sodium Potassium ATPase (Abcam, ab76020), NDUFS1 (Abcam, ab169540), ATP6 (Proteintech, 55313-1-AP), GAPDH (Invitrogen, AM4300), HA (Cell Signaling 3724), HRP-conjugated anti-rabbit IgG (Cell Signaling, 7074S), HRP-conjugated anti-mouse IgG (Cell Signaling, 70765), Goat anti-rabbit IgG Alexa Fluor 647 (Thermo Fisher, A-21245), Goat anti-mouse IgG2a Alexa Fluor 488 (Thermo Fisher, A-21131), Goat anti-mouse IgG2a Alexa Fluor 555 (Thermo Fisher, A-21137), Goat anti-mouse IgM Alexa Fluor 647 (Abcam, ab150123), Goat anti-mouse IgM Alexa Fluor 555 (Thermo Fisher, A-21426), Goat anti-rabbit IgG Alexa Fluor 546 (Invitrogen A-11010).

Hia5 expression and purification

pHia5ET was expressed and purified as previously described[32]. pHia5ET was transformed into T7 Express lysY/Iq *Escherichia coli* cells (NEB C3013I). Overnight cultures were added to two 1 L cultures of LB medium supplemented with 50 µg/mL kanamycin and grown with shaking at 37ºC to an OD_{600} of 0.8-1.0. Isopropyl β-D-1-thiogalactopyranoside (IPTG) was added to a final concentration of 1 mM and protein was expressed for 4 hours at 20ºC with shaking. Cells were pelleted at 5,000 rpm for 10 minutes at 4ºC. The pellet was resuspended in 35 mL lysis buffer (50 mM HEPES, pH 7.5; 300 mM NaCl; 10% glycerol; 0.5% Triton X-100; 10 mM β-mercaptoethanol) supplemented with 2X Complete, EDTA-free Protease Inhibitor Cocktail (Millipore Sigma 11873580001). Cells were lysed by probe sonication (Qsonica Q125) for 10 minutes on ice at 50% amplitude, 30 seconds on/off. Lysate was clarified by centrifuging for 1 hour at 40,000 x g. Ni-NTA Agarose (Qiagen 30210) was prepared by washing 5 mL slurry with 30 mL equilibration buffer (50 mM HEPES, pH 7.5; 300 mM NaCl; 20 mM Imidazole) and centrifuged at 500 x g for 3 minutes, repeating once. The clarified lysate and Ni-NTA agarose were combined and rotated at 4ºC for 1 hour. The lysate mixture was poured over a disposable gravity flow column (Bio-Rad 7321010) and washed with 20 mL Buffer 1 (50 mM HEPES, pH 7.5; 300 mM NaCl; 50 mM imidazole) and 15 mL Buffer 2 (50 mM HEPES, pH 7.5; 300 mM NaCl; 70 mM imidazole). Protein was eluted with 15 mL Elution Buffer (50 mM HEPES, pH 7.5; 300 mM NaCl; 250 mM imidazole). 6-8K MWCO SnakeSkin dialysis tubing (Spectrum Laboratories 132650) was pre-wet in dialysis buffer (50 mM Tris-HCl, pH 8; 100 mM NaCl; 1 mM DTT) and the eluate was added to the tubing and dialyzed against 2 L dialysis buffer overnight at 4ºC. Dialyzed sample was concentrated using a 10K Amicon Ultra-15 spin concentrator (Millipore Sigma UFC901008) and centrifuged at 3,220 x g to concentrate to a volume of less than 1 mL. The concentrated sample was injected onto Tandem HiTrap Q HP (Cytiva 17115301) and HiTrap SP HP (Cytiva 17115101) columns equilibrated with FPLC Buffer A (50 mM Tris-HCl, pH 8; 100 mM NaCl; 1 mM DTT). Columns were washed with 5 column volumes of FPLC Buffer A. The Q column was removed and the sample eluted from the SP column using a linear gradient over 20 column volumes of 0 to 100%

FPLC Buffer B (50 mM Tris-HCl, pH 8; 1 M NaCl; 1 mM DTT). Peak fractions were collected and concentrated to 250 µL using a 10K Amicon Ultra-15 spin concentrator. Protein was supplemented with glycerol to a final concentration of 10%, frozen, and stored at -80ºC. Protein purity was assessed by SDS-PAGE, and its activity measured by an *in vitro* methylation assay.

In vitro MTase activity assay

Hia5 activity was quantified as previously described[32] with minor modifications. Substrate DNA was prepared by PCR of the pHia5ET using T7 Forward and Reverse primers. The PCR product was purified using Monarch PCR & Cleanup Kit (NEB T1030S) according to manufacturer's instructions. A series of eleven 60 µL MTase reactions were prepared with 1 µg substrate DNA with alternating two-fold and five-fold enzyme dilutions (10, 5, 1, 0.5, 0.1, 0.05, 0.01, 0.005, 0.001, 0.0005, and 0.0001 µL MTase) in MTase buffer (15 mM Tris, pH 8.0; 15 mM NaCl; 60 mM KCl; 1 mM EDTA, pH 8.0; 0.5 mM EGTA, pH 8.0; 0.5 mM Spermidine) supplemented with 0.8 mM S-adenosyl-methionine (NEB B9003S). A negative control was prepared without MTase. The reactions were mixed by gentle flicking before a 1 h incubation at 37ºC. Reactions were quenched with Monarch PCR & DNA Cleanup Kit (NEB T1030S**)** and the purified DNA eluted in 20 µL EB buffer. Twelve restriction enzyme digests were prepared by combining 15 µL of each purified DNA sample with 1 µL DpnI (NEB R0176S) and 4 µL 10X CutSmart Buffer (NEB) in a 40 µL reaction. The reactions were mixed by flicking and incubated at 37ºC for 1.5 hours. 1 µL of each reaction was combined with 2 µL of 6X Purple Gel Loading Dye (NEB B7024S) and 11 µL H2O and run on a 1.2% agarose gel containing 1X GelGreen Nucleic Acid Stain (Biotium 41005) at 130 V for 1.5 hours. The gel was imaged on an Azure C200 Gel Imager. MTase activity was determined by the highest MTase dilution that methylates 1 µg of DNA substrate leading to no fully intact DNA molecules after DpnI digestion.

Mitochondrial isolation

HeLa and U2-OS cells were grown in 150 mm plates to a confluency of 70%, and myoblasts to a confluency of 80-90%. Cells were pelleted by spinning at 150 x g for 5 minutes at 4ºC. Media was removed and cells were resuspended in 4 mL Cell Lysis Buffer (10 mM Tris, pH 7.5; 10 mM NaCl; 1.5 mM $MgCl_2$). Cells were allowed to swell on ice for 7.5 minutes. Cells were then lysed by douncing in a 7 mL dounce with 20 strokes. 2 M Sucrose T10E20 Buffer (10 mM Tris, pH 7.6; 1 mM EDTA, pH 8.0; 2 M Sucrose) was added to bring the final sucrose concentration to 250 mM. Cell debris was pelleted by centrifuging lysates at 1,300 x g for 3 minutes at 4ºC. Supernatant was transferred to fresh tubes and spun again at 1,300 x g for 3 minutes at 4ºC. Supernatant was transferred to fresh tubes, and mitochondria were pelleted by spinning at 18,000 x g for 15 minutes at 4ºC.

Proteinase K protection assay

HeLa S3 cells were grown in 150 mm plates to a confluency of 70%. Mitochondria were isolated as described above. The mitochondria pellet was resuspended in 100 µL permeabilization buffer (20 mM Tris, pH 7.4; 70 mM Potassium Acetate; 250 mM Sucrose; 0% / 0.1% / 0.25% / 0.5% Tween-20; 0% / 0.1% / 0.25% / 0.5% NP-40/Igepal-630). Mitochondria were permeabilized on ice for 10 minutes and then pelleted by spinning at 18,000 x g for 15 minutes at 4ºC. Mitochondrial pellets were resuspended in 80 µL PK Reaction Buffer (20 mM Tris, pH 7.5; 70 mM Potassium Acetate; 250 mM Sucrose). Samples were split in two and either 5 µL buffer or 5 µL 1 µg/µL Proteinase K (Millipore Sigma 3115887001) was added. Samples were mixed by gently flicking and incubated for 20 minutes at 37ºC. Reactions were quenched by adding 5 µL 100 mM PMSF and heat inactivated for 10 minutes at 95ºC. Proteins were assessed by western analysis using anti-TOMM40 (Proteintech 18409-1-AP), anti-COX1 (Abcam 14705), and anti-HSP60 (Cell Signaling D307).

Mitochondrial ATAC-seq

ATAC-seq was performed as previously described[28] using TDE1 (Illumina 20034197) with modifications. Mitochondria were first isolated as described above. Mitochondria were resuspended in 50 µL cold lysis buffer (10 mM Tris-HCl, pH 7.4, 10 mM NaCl, 3 mM $MgCl_2$, 0.1% NP-40, 0.1% Tween-20) and immediately centrifuged for 10 minutes at 4ºC at 18,000 x g. Mitochondrial pellets were resuspended in 50 µL transposition reaction mix (25 µL TD, 2.5 µL TDE1, 22.5 µL H_2O) and incubated at 37ºC for 30 minutes. Following transposition, samples were immediately purified using the Qiagen MinElute PCR Purification kit (Qiagen 28004). DNA was minimally amplified with 1 cycle of 5 minutes at 72ºC and 30 seconds at 98ºC followed by 5 cycles of 10 seconds at 98ºC, 30 seconds at 63ºC, and 1 minute at 72ºC. 5 µL of partially amplified DNA was used in qPCR reactions, and the number of additional cycles to amplify was determined as the number of cycles corresponding to one third of the maximum fluorescence intensity signal by qPCR. The samples were then amplified with 1 cycle of 30 seconds at 98ºC followed by the determined number of cycles for 10 seconds at 98ºC, 30 seconds at 63ºC, and 1 minute at 72ºC. Following amplification, samples were assessed on an Agilent BioAnalyzer and subjected to paired-end Illumina sequencing.

ATAC-seq alignment and processing

Paired-end reads were aligned using bowtie2 (version 2.2.9) either to Hg38 or to just chrM of Hg38. The sam file output was converted to a bam file using samtools (version 1.3.1) and sorted. Duplicate reads were filtered using PICARD (version 2.8.0). Aligned reads were then filtered for those having a mapping quality score greater than 30 using samtools (-q 30). Read coverage was determined using the bamCoverage command of deeptools (version 3.0.2) with a

binSize of 1, normalized using counts per million (CPM), and extending the reads by 100. Data were visualized using IGV (version 2.4.9).

Mitochondrial Fiber-seq (mtFiber-seq)

Mitochondria were first isolated as described above. Mitochondrial pellets were resuspended in Permeabilization Buffer containing 0.25% Tween-20 and 0.25% NP-40/Igepal-630 unless otherwise indicated. After permeabilization, mitochondria were pelleted by centrifugation at 18,000 x g for 15 minutes at 4ºC. Supernatant was removed and pellets were resuspended in 56 µL mtFiber-seq Reaction Buffer (20 mM Tris, pH 7.4; 70 mM Potassium Acetate; 250 mM Sucrose). 1.5 µL 32 mM S-adenosylmethionine was added to a final concentration of 0.8 mM. Tubes were transferred to a thermocycler prewarmed to 37ºC. Reactions were started by adding 500 U Hia5, unless otherwise indicated, and mixing. Methylation reaction was allowed to proceed for 10 minutes at 37ºC, unless otherwise indicated. For the 120 minute sample, and additional 1.5 µL of 32 mM SAM was spiked in after 60 minutes. Reactions were quenched by adding 3 µL 20% SDS and mixing. Sample volume was then increased to 200 µL by adding 7 µL 20% SDS and buffer. Protein was degraded by adding 2 µL 18.2 mg/mL Proteinase K (Millipore Sigma 3115887001) and incubating at 55ºC for 1 hour. Phenol:chloroform:isoamyl alcohol extractions were performed to extract DNA, and DNA was precipitated by standard ethanol precipitation protocols using 1/10 volumes Sodium Acetate and 2.5 volumes ethanol. DNA was pelleted and washed with 70% ethanol. DNA was air dried for 10 minutes at room temperature before resuspending in 81 µL. 10 µL of NEB CutSmart Buffer was added along with 3 µL RNase A (Thermo Fisher Scientific AM2270) and restriction enzymes. For all samples, 3 µL XmaI (NEB R0180S) was added to fragment chromatin. An additional 3 µL of BamHI-HF (NEB R3136L) was added to linearize mtDNA from all cell lines with the exception of human skeletal muscle myoblasts, for which 3 µL of EagI-HF (NEB R3505) was added due to the

mitochondrial genome in this cell line having a SNP resulting in a second BamHI cut site. Reactions were incubated at 37ºC for 1 hour. A phenol:chloroform:isoamyl extraction was performed followed by a chloroform:isoamyl extraction to remove any remaining phenol. DNA was precipitated as before. DNA was pelleted and washed with 70% ethanol. After air drying, DNA was quantified by Qubit hsDNA (Thermo Fisher Scientific Q32851). Samples were then subjected to PacBio library preparation protocols according to manufacturer's instructions.

Mapping single-molecule mtFiber-seq reads and m6A identification

Using the raw subread bam files and modified chrM reference genomes from Hg38, we ran a custom pipeline combined with Fibertools (172) on a SLURM managed cluster (see data and code availability). Modified chrM genomes were used as the reference depending on how the particular library was constructed: 1) linearized with BamHI (hg38_chrM_BamHI.fa) 2) linearized with EagI (hg38_chrM_EagI.fa) or 3) synthesized by long-range PCR (hg38_chrM_lrpcr.fa). This pipeline takes the raw PacBio subread bam files and converts them into a 12 column BED file containing mapped m6A positions.

Specifically, subread bam files are first split into 60 chunks using the split_subreads.sh and bamseive_batch.slurm scripts. CCS reads were then generated for each chunk using *ccs* (v6.4.0) with the –hifi-kinetics flag to include averaged kinetic information. Next, chunks were recombined into a single bam file using merge_bam.sh. Fibertools (v0.1.3) predict-m6a was then run on the combined bam file with the -s flag. Reads were then aligned to the appropriate reference genome with *pbmm2* (v1.9.0) using the –preset CCS, –sort, –sort-memory 1G, –unmapped, –log-level INFO flags. Following this, Fibertools extract was run using the -r and –min-ml-score. To account for differing PacBio chemistries, a range of –min-ml-score thresholds from 232 to 255 were used to generate bed files at each cutoff.

Myoblast differentiation

Myoblasts were grown in Human Skeletal Muscle Cell Basal Media containing growth supplement (Cell Applications, INC. 151K-500). To differentiate, at 80% confluency they were switched to DMEM containing glucose and pyruvate (Thermo Fisher Scientific 11995073) supplemented with 2% Heat Inactivated Horse Serum (Thermo Fisher Scientific 26050-070) and 0.4 µg/mL dexamethasone. Media was then replaced every 24 hours, and cells were harvested at either 3 days or 6 days post differentiation.

Tn5 expression and purification

Tn5 was expressed as previously described (173) with modifications. pTXB1-Tn5 (Addgene #60240) was transformed into T7 Express LysY/Iq (NEB C3013). Two 1 L of LB supplemented with 100 µg/mL ampicillin were inoculated with 10 mL of overnight culture. Cells were grown at 37ºC with shaking until the OD600 reached 0.55. Protein expression was induced with 0.2 mM IPTG, temperature was reduced to 18ºC, and cells were allowed to express for 18 hours. Cells were spun down at 4,000 rpm at 4ºC for 20 minutes. The cell pellet was resuspended in 80 mL of Tn5 Lysis/Wash buffer (20 mM HEPES, pH 7.2; 0.8 M NaCl; 1 mM EDTA; 10% glycerol; 0.2% Triton X-100) supplemented with 1X Complete, EDTA-free Protease Inhibitor Cocktail (Millipore Sigma 11873580001). Cells were lysed by probe sonication (Qsonica Q125) for 10 minutes on ice at 50% amplitude, 30 seconds on/off for a total of 20 minutes. The lysate was clarified by centrifugation at 16,000 rpm for 30 minutes at 4ºC. To the supernatant, 2.1 mL 10% polyethyleneimine was added dropwise on a magnetic stirrer. The precipitate was removed by centrifugation at 12,000 rpm for 10 minutes at 4ºC. Clarified lysate was added to 10 mL Chitin Resin (NEB S6651S) pre-equilibrated with Tn5 Lysis/Wash buffer and rocked for 1 hour at 4ºC. Resin was washed with 300 mL Tn5 lysis/wash buffer. Protein was eluted by adding Tn5 Chitin Elution Buffer (20 mM HEPES, pH 7.2; 0.8 M NaCl; 1 mM EDTA; 10% Glycerol; 0.2% Triton X-

100; 100 mM DTT). The first 11 mL were collected and discarded. The resin was kept at 4ºC for 48 hours to allow the intein fusion to cleave and elute the protein from the Chitin resin. After 48 hours, 1 mL fractions were collected and the fractions containing protein were determined using Detergent Compatible Bradford assay (Thermo Scientific 23246). Fractions were dialyzed against 1 L of Tn5 Size Exclusion Chromatography (SEC) Buffer (50 mM Tris, pH 7.5; 800 mM NaCl; 0.2 mM EDTA; 2 mM DTT; 10% glycerol) overnight at 4ºC. Sample was injected and run on ENrich SEC 650 10 x 300 24 mL size exclusion column (Bio-Rad 7801650) using Tn5 SEC Buffer. Peak fractions were collected and dialyzed into 2X Tn5 Dialysis Buffer (100 mM HEPES, pH 7.2; 0.2 M NaCl; 0.2 mM EDTA; 2 mM DTT; 0.2% Triton X-100; 20% glycerol). Protein was concentrated using 10K Spin Concentrator (Millipore Sigma UFC801008) and the concentration was determined using Detergent Compatible Bradford. Glycerol was added to a final concentration of 55% and protein was stored at -80ºC. Protein purity was confirmed by SDS-PAGE.

Tn5 transposome assembly

The transposome assembly was performed as previously described[40] with modifications. In brief, adaptor oligos were ordered from IDT, the reverse oligo Tn5MErev (5′-[phos]CTGTCTCTTATACACATCT-3′), Tn5ME-A-ATTO488 (5′-/ATTO488N/TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG-3′) and Tn5ME-B-ATTO488 (5′-/ATTO488N/GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG-3′). 100 µM Tn5MErev was mixed in equal amounts either with 100 µM Tn5ME-A-ATTO488 or 100 µM Tn5ME-B-ATTO488. The oligo mixtures were denatured for 5 minutes at 95ºC in a thermocycler. The cycler was shut off and the oligos left inside the thermomixer until they reached room temperature. The transposome was assembled in a mixture of mixed oligos (Tn5MErev/Tn5ME-A-ATTO488 and Tn5MErev/Tn5ME-B-ATTO488) (final concentration 12.5 µM for each mix), 31.75% glycerol (final concentration reached 47.9 %), 0.24x Tn5 Dialysis Buffer and 5 µM Tn5. The mixture was mixed

by carefully pipetting up and down and incubated at room temperature for 60 minutes. The assembled transposome was stored at -20ºC.

Tn5 in vitro activity assay

To determine optimal buffer conditions for the assembled transposome, an *in vitro* tagmentation of a plasmid (pET30a-6xHis-ΔN-TFAM) was performed. 50 ng plasmid were incubated with or without 100 nM preassembled Tn5. Four reaction buffers were tested; Buffer 1 (2xTD-Buffer (10 mM KCl, 75 µM $Na_2HPO_4 \cdot 7H_2O$, 274 mM NaCl, 50 mM Tris, pH 7.4) (174), Buffer 2 (2xTD-Buffer(33 mM Tris, pH 7.8, 66 mM potassium acetate, 11 mM magnesium acetate, and 16% N,N-dimethylformamide (175)), Buffer 3 (2xTD-Buffer (20 mM Tris-HCl, pH 7.6, 10 mM $MgCl_2$, 20% N,N-dimethylformamide (176)) and Buffer 4 (5xTAPS-PEG (50 mM TAPS-NaOH, pH 8.5, 25 mM $MgCl_2$, 40% PEG 8000 (173)). Reactions were performed in 1x concentration of the respective buffer. The reaction mixtures were incubated for 30 minutes at 37ºC. Reactions were quenched by the addition of 2x concentrated Washing Buffer (100 mM EDTA, 0.02% SDS, 2x PBS) and incubated for 15 minutes at 55ºC. Samples were brought to a volume of 200 µl, and DNA was precipitated by the addition of 2 mM $MgCl_2$, 60 mM Sodium Acetate pH 5.5, and 80% v/v ethanol. The samples were incubated for 2 hours on ice and the DNA was pelleted for 15 minutes at full speed in a tabletop centrifuge at 4ºC. The pellet was washed with 70% ethanol and dried at room temperature. The DNA was resuspended in water and the concentration measured by Nanodrop. To analyze the tagmentation process, the samples were analyzed using the Agilent TapeStation D1000.

Mitochondrial ATAC-see

ATAC-see was performed as previously described[40,72] with modifications**.** U2-OS cells were grown in µ-Slide 8 Well Glass Bottom chambers (#1.5H glass bottom, ibidi, Cat.No:80827)

until 60-70% confluent. Cells were washed three times with 1X PBS and fixed for 10 minutes with 1% methanol-free formaldehyde (Thermo Fisher 28906) in 1X PBS at room temperature. After three washes with 1X PBS, cells were permeabilized with 1X PBS, 0.25% TX-100 for 20 minutes. Cells were washed three times with 1X PBS and the chamber slide was incubated in a hybridization oven (Boekel Scientific, RapidFISH) prewarmed to 37ºC. 100 nM Tn5 was used in tagmentation buffer (16.5 mM Tris, pH 7.8, 33 mM Potassium Acetate, 5.5 mM Magnesium Acetate and 8% N,N-dimethylformamide). For the inactive Tn5 control, the mixture included 50 mM EDTA. After mixing by pipetting, the transposase was centrifuged for either 10 (EFD 3) or 20 min (EFD 4C) at max speed at 4ºC in a tabletop centrifuge to reduce aggregates in the sample. The cells were incubated with the transposase for 60 minutes at 37ºC in the hybridization oven in the dark. The transposase was inactivated and washed away in three washes for 15 minutes at 37ºC with prewarmed washing buffer (1X PBS, 50 mM EDTA, 0.1% TX-100). Afterwards the cells were rinsed three times with 1X PBS at room temperature.

Immunofluorescence and DAPI staining after mitochondrial ATAC-see

For immunofluorescence, ATAC-see samples were incubated for 1 hour at room temperature protected from light with blocking buffer (1X PBS, 0.1% TX-100 and 5% normal goat serum (Vector Laboratories, S-1000-20). ds/ssDNA mouse monoclonal IgM antibody clone AC-30-10 (PROGEN, Cat. no. 61014) was used in a 1:100 dilution in blocking buffer. To be able to segment single cells, the rabbit monoclonal anti-Sodium Potassium ATPase antibody [EP1845Y] (Abcam, ab76020) was used in a 1:250 dilution in blocking buffer. The antibody mix was added to the cells, and they were incubated at 4ºC protected from light overnight or for 1 hour at room temperature. The samples were washed three times for 5 minutes each at room temperature with 1X PBS and 0.1% TX-100 followed by an incubation for 1 hour at room temperature protected from light with the secondary antibodies. The goat anti-mouse IgM mu chain (Alexa Fluor® 647)

(Abcam, ab150123) and the goat anti-rabbit IgG (H+L) cross-adsorbed secondary antibody (Alexa Fluor™ 546) (Invitrogen, A-11010) were used in a 1:1000 dilution in blocking buffer. After the 1 hour incubation, the samples were washed three times for 5 min at room temperature protected from light. After three rinses with 1X PBS, 5 µg/ml DAPI (4′,6-Diamidin-2-phenylindol) in 1X PBS was added. The samples were incubated for 10 minutes protected from light and washed again three times with 1X PBS. Glycerol mounting media (80% glycerol, 1X PBS, 20 mM Tris, pH 8, 2.5 g/ml n-Propyl-Gallate) was added to all wells.

Immunofluorescence and DAPI staining

To test the detection of mtDNA, U2-OS cells were grown in µ-Slide 8 Well Glass Bottom chambers until 80% confluent. Cells were washed once with 1X PBS and fixed using 4% formaldehyde in 1X PBS for 1 hour at room temperature. After three washes with 1X PBS, cells were permeabilized for 20 minutes with 1X PBS and 0.1% TX-100 followed by three washes using 1X PBS. For the TFAM overexpression, HeLa S3 cells were grown in µ-Slide 8 Well Glass Bottom chambers with high walls (Cat. #80807) and TFAM-HA expression was induced by addition of 100 ng/ml doxycycline for 48 hours. DMSO (diluted 1:1000) served as a control. TFAM-HA was expressed for 48 hours. Cells were rinsed once with 1X PBS and fixed for 20 minutes with 4% formaldehyde in 1X PBS. Washing and permeabilization were performed as described before. The immunofluorescence was performed the same for both experiments.

After permeabilization and washing, the buffer was replaced by blocking buffer and the samples were incubated for 1 hour at room temperature. The primary antibodies were diluted in blocking buffer, added to the samples, and incubated overnight at 4ºC. The antibodies were used in the following dilutions: for the DNA detection in the mitochondrial network, the ds/ssDNA mouse monoclonal IgM antibody was used in a 1:250 dilution and in the TFAM-HA overexpression experiment in a 1:162 dilution. The TOM20 (F-10) (Santa Cruz, sc-17764) and the HA-Tag

(C29F4) Rabbit mAb (Cell Signaling, Cat. No.: 3724) were used in a 1:500 dilution. On the next day samples were washed 3 times for 5 minutes using 1X PBS and 0.1% TX-100. The secondary antibodies were diluted 1:1000 in blocking buffer. For Fig. 3B, Extended Data Fig. 2B, Extended Data Fig. 3B,C mtDNA anti-mouse IgM Alexa Fluor 647 and Goat anti-Mouse IgG2a Cross-Adsorbed Secondary Antibody, Alexa Fluor™ 555 (Thermo Fisher, Cat. No.: A-21137) were used. For Fig. 3B TFAM overexpression the following antibodies were used: Goat anti-Mouse IgM (Heavy chain) Cross-Adsorbed Secondary Antibody, Alexa Fluor™ 555 (Thermo Fisher, Cat. No.: A-21426), Goat anti-Rabbit IgG (H+L) Highly Cross-Adsorbed Secondary Antibody, Alexa Fluor™ 647 (Thermo Fisher, Cat. No.: A-21245), Goat anti-Mouse IgG2a Cross-Adsorbed Secondary Antibody, Alexa Fluor™ 488 (Thermo Fisher, Cat. No.: A-21131). The samples were incubated for 1 hour at room temperature and in the dark. Afterwards, the cells were washed again 3 times for 5 minutes with 1X PBS and 0.1% TX-100. After three rinses with 1X PBS, 5 µg/ml DAPI in 1X PBS was added. The samples were incubated for 10 minutes in the dark and washed again three times with 1X PBS. No DAPI stain was used for the TFAM-HA overexpression experiment. Glycerol mounting media (80% glycerol, 1X PBS, 20 mM Tris, pH 8, 2.5 g/ml n-Propyl-Gallate) was added to all wells.

Microscopy

As controls for imaging, one well was empty and was used to image the background for flatfield correction. To be able to perform collar correction with the silicon oil objective and to test the point spread function as well as chromatic aberrations, fluorescent blue/green/orange/dark red TetraSpec Microspheres size 0.2 µm (Thermo Fisher Scientific, T7280) were used. The spheres were first vortexed to allow for well resuspension. 1 µl spheres were mixed with 9 µl 100% ethanol and vortexed. The mix was spread onto the well with a nipped tip and let dry for 30

minutes. Glycerol mounting media was added. The spheres were always prepared on the same day as the samples to allow the addition of the mounting media at the same time.

For imaging a Nikon Ti2 inverted microscope with a W1 Yokogawa Spinning disk with a 50 µm pinhole disk, laser lines 405, 488, 561, 640 (Nikon Instruments, Laser Unit model Lun-F), emission filters (455/50, 480/40, 525/36, 605/52, 630/75, 705/72), a Nikon motorized stage, a Physik Instrument Piezo Z motor, Andor Zyla 4.2 Plus sCMOS monochrome VSC-04833 camera, and a Plan Apo λS SR HP 100xC/1.45 Silicon DIC objective were used. The Nikon Elements Acquisition Software AR 5.02 was used. Slides were cleaned first with Sparkle, water, and finally with 100% ethanol before imaging to remove any salt and oil residues. A silicon immersion oil 30cc (Nikon, Cat. No.: MXA22179) was used in combination with the silicon oil objective. The laser power was adjusted using noEDTA samples. The same exposure times (green channel: 700 ms, all others 500 ms) were used for the biological replicates. No binning was used. The 16-bit dual gain ¼ setting was used. The final image had a pixel size of 0.065 µm x 0.065 µm. 30 field of views and the 27 z-planes were imaged on each session of an empty well to perform averaging of the field of views and use for flatfield correction. To assess the noise detected by the detector, 100 frames with lasers off were taken and averaged for background subtraction of samples and the flatfield image. Z-stacks of 29 or 27 steps of 200 nm were imaged using the NIDAQ piezo stage. The shutter was off during image acquisition and all fluorescence channels one after the other (starting with green, red, far red, and finally blue) were taken for each z-plane before moving to the next plane. All images were acquired at room temperature and saved as .ND2 files. The immunofluorescence of DNA and TOM20 was imaged with the same microscope as described above. Again, the silicon oil objective was used. A z-stack of 19 planes with each plane sized 300 nm was imaged. The 16-bit dual gain ¼ setting was used. The final image had a pixel size of 0.065 µm x 0.065 µm. The exposure time for each channel was 500 ms and again all channels were first imaged before moving to the next z-plane. For Fig. 3B the Plan Apo λ 100x/1.45 Oil DIC objective in combination with immersion oil (Cargille, non-drying immersion oil

Type 37, Cat. No.: 16237) was used. Again, no binning and 16-bit dual gain ¼ setting was used. A z-stack of 38 planes with a plane size of 300 nm was imaged. The final image had a pixel size of 0.065 µm x 0.065 µm. Again the exposure times were 500 ms.

Image Analysis

For image analysis, Fiji, Arivis, and R were used. Images were exported as ome.tiff files for analysis. First, sample images were flat field corrected. The fields of view (FoVs) of the empty well were averaged using a custom macro in Fiji. Next, the camera noise was subtracted from the averaged flatfield correction image. Since the background subtraction of the camera noise leads to negative pixel values, an offset of 200 was added to the sample images as well as the corrected flatfield image. Next, the camera noise was subtracted from the sample image and the result of this was divided by the flatfield corrected image.

((Sample + 200) - Camera noise) / ((Flatfield - Camera noise) + 200)

For replicate 1, the first 5 images had 29 z-planes. Two z planes were deleted to gain the same number as the other images. Z-planes without signals were chosen to be removed. The corrected images could be used then for segmentation in Arivis. First, nuclei were segmented based on the DAPI stain and masked from the image. Therefore, the denoising module was used with the following settings: bilateral, diameter was set to 1 pixel, sensitivity was set to 24.3. Next, an intensity threshold segmenter was used with the settings, “simple”, “bright” and the “visible range” of 0.789 - 25. The threshold was set to 1. The segment morphology module was used to erode objects for 11 pixels to reduce detected mtDNA. The erode function was performed plane wise. In order to detect the full nuclei size the objects were dilated again planewise using the segment morphology filter by 11 pixels. The resulting objects were used to mask the nuclei in the next segmentation step. mtDNA was segmented based on the ds/ssDNA-antibody signal. A morphology filter was used to preserve bright objects with a size of 80 pixels. The option “spheres”

was chosen. Afterwards, the nuclei were masked using the before created objects. The blob finder function was used to segment the nucleoids with the following settings: the probability threshold was set to 5% and the split sensitivity to 60%. The diameter was set to 1 pixel. The blob finder function segments objects in 3D space creating. Objects smaller than 5 voxels were filtered out, since they are considered artifacts. The object features were extracted. Single cells were segmented by hand based on the signal of the anti-Sodium Potassium ATPase antibody. These masks were used to assign the before segmented nucleoids to single cells using the compartment function of Arivis. 30 cells were segmented for noEDTA (7 field of views) and EDTA (5 field of views) for replicate 1 and 29 cells for noEDTA (7 field of views) and 30 cells (9 field of views) for EDTA for replicate 2. The object features of those objects were extracted and used for analysis in R. In case of cells where the nuclei mask didn't cover the full signal and the mtDNA segmentation pipeline detected objects, those objects were deleted by hand to avoid artifacts. The extracted intensities were either analyzed in bulk, only considering objects assigned to single cells, or as single cells. For figures, brightness and contrast were adapted in Fiji. For Fig. 3B of the TFAM overexpression cell line, a median filter of pixel size 2 was used. For supplemental Extended Data Fig. 3D screenshots of the different segmentation steps were taken in Arivis. Before, brightness and contrast were adjusted to represent the signals.

mtFiber-seq footprint identification

To identify footprints, a hidden Markov model (HMM) was applied with two hidden states: accessible and inaccessible. To account for any sequence biases, the HMM observations were considered to be a methylation (m6A) or a lack of methylation(A) observed at every possible hexamer sequence context (-3nt, +3nt). For example, AAAm6AAAA was treated as a separate observation from TAAm6AAA. The emission probabilities for these hexamer sequence context observations were calculated based on experimental data. For the probability of methylation given

an accessible state, data from mtFiber-seq performed on mtDNA generated from Long-Range PCR (LRPCR) was used. For the probability of methylation given an inaccessible state and to account for biases in the methylation caller or low levels of endogenous methylation, data from mtFiber-seq performed on isolated mitochondria without Hia5 (MTase) was used. Observations at a position with a C or G were given a methylation emission probability of 0. The transition and starting probabilities were trained using 1% of HeLA mtFiber-seq data 20 times, with the reads shuffled each time and different starting parameters used. The starting parameters for each training were generated by sampling the Dirichlet distribution with all parameters set to 1. Each training converged to almost identical starting and transition probabilities. The model was applied to both experimental data and to a sampled portion of the LR-PCR product data. A minimum mean posterior probability of .95 was identified as a threshold that provided a 1% false discovery rate for footprints in the accessible data. Footprint enrichment was calculated as the fraction of reads, either total or with a defined level of methylation, protected at a particular base. Although developed independently, a similar approach to calling footprints has been successful in other contexts[36].

Strand specificity of methylation

To calculate methylation strand bias, A (for light strand methylation) and T (for heavy strand methylation) were represented as the values 1 and 0, respectively. Using a 150 nt sliding window, methylations (1 or 0) across all reads were averaged. Each window was required to have at least 2250 methylations across all reads combined or was assigned a value of 'NA'. Values > 0.5 thus represent a light strand methylation bias, and vice versa. The reference sequence A/T content was then evaluated using an identical approach and methylation A/T values were normalized by the reference A/T values and $\log_2$-transformed. D-loop boundaries are shown at the coordinates 16080-16100 and 210-230. To calculate an overall strand bias score, all

methylations across the region (e.g. 16100 - 210) for all reads were recorded as 1 or 0, then the mean was calculated.

7S DNA D-loop identification.

Reads with a high ratio of light strand methylations to heavy strand methylations in the D-loop region were determined and called as likely containing a D-Loop. This ratio was defined as $(mA_{light}+1)/(mA_{heavy}+1)$. A minimum count of 7 total methylations in the region was used to call a D-loop and a Gaussian mixture model (GMM) was applied to the distribution of the log of the ratios. A ratio threshold of 3.01 was identified to call a D-loop, based on a GMM posterior probability of .99.

Footprint co-occurrence

Footprint co-occurrence on a read was tallied using the following filters: minimum size for all, 20 nt; maximum size across the left D-loop boundary (16080:16100), 60 nt; maximum size across the right D-loop boundary (210:230), 140 nt; maximum size across the MTERF binding site (3232:3253), 35 nt. Minimum overlap of footprint with site: 6 nt. D-loop presence in a read was determined by the ratio of light to heavy strand methylations as described above. Counts were normalized by the total number of reads in the categories considered and shown as a percentage.

In vitro MTase substrate specificity assay

Hia5 methyltransferase substrate specificity was measured using the MTase-Glo Methyltransferase Assay (Promega) according to manufacturer's instructions. 27 nt oligos were used as substrate: A) 5'-TGACATGAACACAGGTGCTCAGATAGC-3' and B) 5'-

GCTATCTGAGCACCTGTGTTCATGTCA-3'. Oligos were resuspended in Duplex Buffer (100 mM Potassium Acetate; 30 mM HEPES, pH 7.5) to a final concentration of 100 µM. Oligos A and B were mixed at a ratio of 1:1, heated to 94ºC for 2 minutes, and slowly cooled to room temperature to anneal the dsDNA substrate. Oligo A was used as the ssDNA substrate. Optimal enzyme concentration was determined by varying Hia5 with 1 µM dsDNA. Reactions were allowed to proceed for 30 minutes at room temperature before quenching, developing, and measuring luminescence using a Tecan Infinite 200 FPlex plate reader. To measure substrate specificity, 0.02 U/µL Hia5 was used with increasing concentrations of dsDNA or ssDNA. Reactions were allowed to proceed for 30 minutes before quenching with 0.5% trifluoroacetic acid (TFA), developing, and measuring luminescence.

mtRNA NanoString quantification

HeLa S3 cells were treated with either DMSO or 2'-C-methyladenosine dissolved in DMSO (100 µM final concentration) for 2, 4, 6, or 24 hours. For each time point, cells were harvested and counted using a Countess II Cell Counter (Thermo Fisher Scientific). Cells were diluted to a concentration of 2000 cells/µL and pelleted by spinning at 300 x g for 5 minutes at 4ºC. Cells were resuspended in 1 mL RLT Buffer (Qiagen 79216) and vortexed for 1 minute. Cells were then diluted in RLT Buffer to make 200 µL aliquots at 1000 cells/µL and 100 µL aliquots at 500 cells/µL. Cells were flash frozen in liquid nitrogen and stored at -80ºC. To hybridize NanoString probes (modified from MitoString profiling probe set[74]), frozen cells were thawed on ice. 4 µL Probe Stocks A and B were each diluted with 29 µL TE Tween (TE + 0.1% Tween-20). 70 µL hybridization buffer was added to the Probe Set. 7 µL of Diluted Probe Stock A was added to the diluted Probe Set, mixed by inverting, and spun down for 2-3 seconds at 400 rpm. 7 µL of Diluted Probe Stock B and 84 µL H_2O were added, mixed by inverting, and spun down for 2-3 seconds at 400 rpm. 14 µL of this master mix was transferred to each required PCR tube. 1 µL of thawed

lysate was added to each tube, sample was mixed by flicking and briefly spun down using a table-top centrifuge. Samples were then hybridized by incubating at 67ºC for 16 hours.

Relative mtDNA content quantification

Relative mtDNA content was quantified as previously described (178). Total DNA was isolated from cell culture using the Qiagen QIAamp DNA mini kit. DNA concentrations were determined by Nanodrop. PCR reactions were assembled by combining 10 µL SsoFast EvaGreen Supermix (Bio-Rad 1725201), 0.8 µL of 10 µM forward primer, 0.8 µL of 10 µM reverse primer, 3 µL of 20 ng/µL template DNA, and 5.4 µL H_2O. Samples were cycled as follows: 50ºC for 2 minutes, 95ºC for 10 minutes, 40 cycles of 95ºC for 15 seconds and 62ºC for 60 seconds. Samples were performed in technical replicate for each PCR target. C_q values for the three technical replicates were averaged. The relative mtDNA content was calculated as $2*2^{\Delta CT}$ where ΔCT is defined as $Nuclear_{CT}$ - $Mitochondrial_{CT}$. For nuclear DNA, primers were used targeting the µglobulin gene (Forward: TGCTGTCTCCATGTTTGATGTATCT, Reverse: TCTCTGCTCCCCACCTCTAAGT), and for mtDNA primers were used targeting the Leu tRNA gene (Forward: CACCCAAGAACAGGGTTTGT, Reverse: TGGCCATGGGTATGTTGTTA).

RNA-seq

RNA-seq libraries from differentiating human skeletal muscle myoblasts were prepared using the SMARTer Stranded Total RNA HI Mammalian Kit (Takara 634873) in biological duplicates (Days 0, 3, and 6). Libraries were sequenced on an Illumina NovaSeq with paired-end 100 nt reads at the Harvard Medical School's Biopolymers Facility. Reads were trimmed with cutadapt v1.14 and aligned using STAR's default parameters to the reference hg38 genome. Rsubread's featureCounts was used to count uniquely-mapping reads across coding sequences and differential expression analysis was performed using DESeq2 (179). The genes GAPDH,

HPRT1, and RPS12 were used in the controlGenes parameter of the DESeq2 estimateSizeFactors function (180).

TFAM overexpression

For TFAM overexpression, HeLa TFAM-HA TetOn HeLa S3 cells were grown on 150 mm plates to a confluency of 50%. Cells were treated with either DMSO or 100 ng/mL doxycycline for 48 hours. Media containing either DMSO or 100 ng/mL doxycycline was replaced after 24 hours. Following treatment, mtFiber-seq was performed according to the above protocol using equal cell numbers, and TFAM overexpression levels were confirmed by western blot analysis using an anti-TFAM antibody (Santa Cruz sc-376672) and normalized with an anti-β-Actin antibody (Cell Signaling 3700).

Subsampling of datasets.

To directly compare the footprinting patterns of datasets with often widely varying levels of methylation, a subsampling approach to match their overall methylation distributions was employed using *sample_bed.py*. The overall number of methylations were counted either across the entire read for the overexpression datasets, or in the coding region (Genome positions 4,000 - 14,000) to avoid the NCR and mTERF binding site, for the *in vitro* datasets. The distribution of methylation counts were then equalized in both datasets by downsampling the set of reads with a given methylation count in a dataset to match the count in the other dataset. These samplings were repeated 10 times with different random seeds to make sure that any conclusions derived from them were stable.

TEV protease expression and purification

TEV protease was purified as previously described (181) with modifications. MBP-TEV plasmid was transformed into T7 Express LysY/Iq cells. Overnight cultures were grown and four 1 L cultures of 2X LB supplemented with 100 µg/mL ampicillin were inoculated with 10 mL overnight culture. Cultures were grown at 37ºC with shaking until the OD_{600} reached 0.4-0.5. Cultures were transferred to 18ºC and grown until the OD_{600} reached 0.6-0.8. Expression was induced by adding 400 µL of 1 M IPTG per 1 L of culture. Protein was expressed for 18 hours at 18ºC with shaking. Cells were spun down at 5,000 rpm for 25 minutes. Pellets were resuspended in a total of 60 mL TEV Lysis/Wash Buffer (1X PBS, pH 7.5; 300 mM KCl; 10% glycerol; 7.5 mM imidazole). Cells were lysed by probe sonication (Qsonica Q125) for 10 minutes on ice at 50% amplitude, 30 seconds on/off. Lysate was clarified by centrifuging for 20 minutes at 25,000 x g at 4ºC. Supernatant was transferred to new tubes and spun at 40,000 x g for 25 minutes at 4ºC. Clarified lysate was added to 6 mL Ni-NTA agarose resin pre-equilibrated in TEV Lysis/Wash buffed. Lysate mixture was incubated for 1 hour at 4ºC with rocking. Resin was spun down in 50 mL conicals at 1,000 x g for 3 min at 4ºC. Resin was washed with 150 mL TEV Lysis/Wash buffer. Protein was eluted with 20 mL TEV Elution buffer (25 mM HEPES, pH 7.5; 100 mM KCl; 500 mM Imidazole). Protein was concentrated using a 3 kDa cutoff spin concentrator (Millipore Sigma UFC9003) to a volume of 2 mL. Protein was injected and run on a HiLoad 16/60 S200 column (Cytiva 28989335) pre-equilibrated in TEV Storage Buffer (25 mM HEPES, pH 7.5; 300 mM KCl; 10% glycerol; 2 mM DTT). Fractions were analyzed by SDS-PAGE, and pure fractions were pooled. Concentration was determined by A_{280} using an extinction coefficient of 36,130 M^{-1} cm^{-1}. TEV was stored at 3 mg/mL at -80ºC.

TFAM expression and purification

TFAM lacking the N-terminal mitochondrial targeting sequence (ΔN-TFAM) corresponding to amino acids 50-246 was cloned into pET30a using BamHI and NotI. The construct contains an

N-terminal 6xHis tag and a TEV cleavage just upstream of the TFAM coding region to allow for the generation of untagged protein. The plasmid was transformed into C43 (DE3) *E. coli* (BioSearch Technologies 60345-1), and overnight cultures were grown in LB supplemented with 25 µg/mL kanamycin. Four 1 L cultures of LB supplemented with 25 µg/mL kanamycin were inoculated with 15 mL each of overnight culture. Cultures were grown at 37ºC with shaking until the OD_{600} reached 0.6. Protein expression was induced with 1 mM IPTG and cells were induced for 24 hours at 25ºC with shaking. Cells were harvested by spinning at 5,000 rpm for 20 minutes. Pellets were resuspended in 80 mL TFAM Lysis Buffer (50 mM Tris, pH 7.4; 300 mM NaCl; 5 mM Imidazole; 1 mM β-mercaptoethanol) supplemented with 1X Complete, EDTA-free Protease Inhibitor Cocktail. Cells were lysed by probe sonication (Qsonica Q125) for 10 minutes on ice at 50% amplitude, 30 seconds on/off. Lysate was supplemented with 5 µL benzonase to reduce *E. coli* genomic DNA, and lysate was clarified by centrifugation in Oak Ridge tubes at 16,000 rpm for 30 minutes. Lysate was combined with Cobalt TALON resin pre-equilibrated with TFAM Lysis Buffer and incubated for 1 hour at 4ºC with rocking. Resin was washed with 120 mL TFAM lysis buffer, and protein was eluted with 10 mL TFAM Elution Buffer (50 mM Tris, pH 7.4; 300 mM NaCl; 500 mM Imidazole; 1 mM βME). Elution was concentrated to 2 mL using a 10 kDa cutoff spin concentrator (Millipore Sigma UFC9010) and supplemented with 100 µL 3 mg/mL TEV protease. Sample was dialyzed overnight against Heparin Load/Wash Buffer (25 mM Tris, pH 7.4; 100 mM NaCl; 5 mM DTT; 10% glycerol). Sample was injected on Heparin HP column and eluted with a 0 to 100% gradient with Heparin Load/Wash Buffer and Heparin Elution Buffer (25 mM Tris, pH 7.4; 1 M NaCl; 5 mM DTT; 10% glycerol) over 20 column volumes. Peak fractions were concentrated to 1 mL and injected onto a HiLoad 16/60 S200 column pre-equilibrated with TFAM Size Exclusion Buffer (25 mM Tris, pH 7.4; 150 mM NaCl; 5 mM DTT; 10% glycerol). Peak fractions were pooled and concentrated, the purity determined by SDS-PAGE, and the concentration determined by NanoDrop using the absorbance at 280 nm with the 35,410 M^{-1} cm^{-1}.

TFAM in vitro DNA binding

TFAM binding activity was confirmed using a fluorescence polarization-based method. DNA oligos corresponding to the HSP promoter were annealed by heating to 95ºC and slowly cooling to room temperature. The forward oligo (GGTTGGTTCGGGGTATGGGGTTAGCAGC) contained a 5' FAM fluorophore. 100 nM DNA was mixed with protein at a volume of 20 µL in TFAM Binding Buffer (25 mM Tris, pH 7.4; 150 mM NaCl; 1 mM DTT; 0.01% NP-40) in a black Corning 384 well plate (Corning 3575). Samples were equilibrated for 30 min at room temperature. Polarization was measured using a Tecan Infinite 200 Pro FPlex plate reader using 485 nm/20 nm bandwidth excitation filters and 530 nm/25 nm bandwidth emission filters fitted with polarizers. Data were fit using FPobs = [Protein]*FPmax + KD*FPmin / [Protein] + Kd.

mtDNA long-range PCR amplification

Long-range PCR was performed to generate linear full-length mtDNA. mtDNA was amplified from HeLa genomic DNA that was isolated from HeLa S3 cells using the Qiagen QiAmp DNA Mini Kit according to the manufacturer's instructions. Takara LA HS Polymerase (Takara RR042) was used with 16426 Forward (5'-CCGCACAAGAGTGCTACTCTCCTCGCTC-3') and 16425 Reverse (5'-GATATTGATTTCACGGAGGATGGTGGTCAAGGGACC-3') primers using 100 ng template in 50 µL reactions. DNA was amplified using a two step amplification protocol: 95ºC for 2 minutes, 30 cycles of 95ºC for 20 seconds and 68ºC for 13 minutes, and a final extension at 68ºC for 18 minutes. DNA was purified using a Zymo Genomic DNA Clean & Concentrator 10 kit (Zymo Research D4011), pooling five 50 µL reactions per column. DNA was eluted with 50 µL Elution Buffer pre-warmed to 50ºC by applying to the column, incubating for 5 minutes, and spinning for 1 minute at 16,000 x g. Elution was reapplied to the column, incubated for 5 minutes, and spun for 1 minute at 16,000 x g. Multiple elutions were pooled and concentrated

by SpeedVac. DNA concentration was determined using Qubit hsDNA kit (Thermo Fisher Scientific Q32851) according to the manufacturer's instructions. DNA length and purity was confirmed using a 0.5% agarose gel as well as by TapeStation using a gDNA tape.

In vitro mtFiber-seq

Methylation of mtDNA generated by long-range PCR was performed using 1 µg template DNA in a final volume of 60 µL with a range of recombinant TFAM concentrations. DNA was mixed with TFAM in TFAM binding buffer (25 mM Tris, pH 7.4; 150 mM NaCl; 1 mM DTT) and equilibrated for 30 minutes at room temperature. Samples were supplemented with 0.8 mM S-adenosylmethionine and reactions were started by adding 200 U Hia5 and mixing. Samples were incubated for 10 min at 37ºC before quenching using a Zymo Genomic DNA Clean & Concentrator 10 (Zymo Research D4011) kit according to manufacturer's instructions. Samples were eluted with 50 µL elution buffer prewarmed to 50ºC. Elution buffer was incubated on the column membrane for 5 minutes before spinning. Elution was reapplied to the membrane, incubated for 5 minutes, and then centrifuged. DNA concentration was determined by Qubit hsDNA assay (Thermo Fisher Scientific Q32851). Integrity of the DNA was confirmed by TapeStation using a gDNA tape, and methylation was measured by DNA dot blot. DNA was subjected to PacBio library construction and sequencing.

DNA methylation dot blot

DNA was diluted in 20X SSC (3M NaCl; 300 mM trisodium citrate) buffer in a 96 well plate and denatured at 95ºC for 10 minutes. Nitrocellulose membrane (Bio-Rad 1620112) was wetted in 20X SSC buffer and secured in a Bio-Dot Microfiltration Apparatus (Bio-Rad 1706545) according to the manufacturer's instructions. Membrane was washed with 100 µL 20X SSC per well and the vacuum applied until all liquid was pulled through. 150 µL sample was applied to

each well (150 µL 20X SSC was applied to unused wells) and pulled through by vacuum. Membrane was placed face up on dry Whatman filter paper and crosslinked with 125 mJoule in a GS Gene Linker UV Chamber (Bio-Rad) using the C-L setting. Membrane was washed briefly with 1X TBST (10 mM Tris, pH 7.5; 0.25 mM EDTA; 150 mM NaCl; 0.1% TWEEN-20) and blocked in 1X TBST + 5% non-fat dry milk for 1 hour at room temperature. Rabbit polyclonal anti-N6-methyladenosine antibody (Active Motif 61995) was diluted to 2 µg/mL in 1X TBST + 5% non-fat dry milk and incubated overnight at 4ºC. The blot was washed 3 times in 1X TBST for 15 minutes each. The anti-rabbit IgG, HRP-linked secondary (Cell Signaling 7074S) was diluted 1:5000 in 1X TBST + 5% non-fat dry milk and incubated with the blot for 1 hour at room temperature. Three washes were repeated and the blot was developed with ECL substrate (Cytiva RPN2106) and imaged using a Bio-Rad ChemiDoc MP system.

Genome-wide TFAM $K_{1/2}$ determination and $GN_{10}G$ enrichment

Footprints for *in vitro* mtFiber-seq datasets were identified as described above. For each read, each genome position was determined as being bound if it overlapped with a footprint at least 20 bp in size. The fraction bound was determined at each position as the number of reads protected at a site with a footprint at least 20 bp in size divided by the total number of reads. The fraction bound was performed at each genomic position for each TFAM concentration: 0, 5, 10, 20, and 30 µM. The $K_{1/2}$ was determined using a four parameter logistics regression, with the minimum and maximum forced to be 0 and 1, respectively. The strongest affinity sites throughout the genome were identified as those with the top 5% of $1/K_{1/2}$ values. The center of each stretch of high affinity position was determined, resulting in the identification of 35 highest affinity sites. A site was determined to have a $GN_{10}G$ motif if one was present within a +/- 15 bp window of the high affinity site.

References

1. Slatko BE, Gardner AF, Ausubel FM. Overview of Next-Generation Sequencing Technologies. Curr Protoc Mol Biol. 2018 Apr;122(1):e59. doi: 10.1002/cpmb.59. PMID: 29851291; PMCID: PMC6020069

2. Akbari V, Garant JM, O'Neill K, Pandoh P, Moore R, Marra MA, Hirst M, Jones SJM. Megabase-scale methylation phasing using nanopore long reads and NanoMethPhase. Genome Biol. 2021 Feb 22;22(1):68. doi: 10.1186/s13059-021-02283-5. PMID: 33618748; PMCID: PMC7898412.

3. Wang Y, Zhao Y, Bollas A, Wang Y, Au KF. Nanopore sequencing technology, bioinformatics and applications. Nat Biotechnol. 2021 Nov;39(11):1348-1365. doi: 10.1038/s41587-021-01108-x. Epub 2021 Nov 8. PMID: 34750572; PMCID: PMC8988251.

4. Korlach J, Bjornson KP, Chaudhuri BP, Cicero RL, Flusberg BA, Gray JJ, Holden D, Saxena R, Wegener J, Turner SW. Real-time DNA sequencing from single polymerase molecules. Methods Enzymol. 2010;472:431-55. doi: 10.1016/S0076-6879(10)72001-2. PMID: 20580975.

5. Wang Y, Zhao Y, Bollas A, Wang Y, Au KF. Nanopore sequencing technology, bioinformatics and applications. Nat Biotechnol. 2021 Nov;39(11):1348-1365. doi: 10.1038/s41587-021-01108-x. Epub 2021 Nov 8. PMID: 34750572; PMCID: PMC8988251.

6. Leger A, Amaral PP, Pandolfini L, Capitanchik C, Capraro F, Miano V, Migliori V, Toolan-Kerr P, Sideri T, Enright AJ, Tzelepis K, van Werven FJ, Luscombe NM, Barbieri I, Ule J, Fitzgerald T, Birney E, Leonardi T, Kouzarides T. RNA modifications detection by comparative Nanopore direct RNA sequencing. Nat Commun. 2021 Dec 10;12(1):7198. doi: 10.1038/s41467-021-27393-3. PMID: 34893601; PMCID: PMC8664944.

7. Xu L, Seki M. Recent advances in the detection of base modifications using the Nanopore sequencer. J Hum Genet. 2020 Jan;65(1):25-33. doi: 10.1038/s10038-019-0679-0. Epub 2019 Oct 11. PMID: 31602005; PMCID: PMC7087776.

8. Stephenson W, Razaghi R, Busan S, Weeks KM, Timp W, Smibert P. Direct detection of RNA modifications and structure using single-molecule nanopore sequencing. Cell Genom. 2022 Feb 9;2(2):100097. doi: 10.1016/j.xgen.2022.100097. PMID: 35252946; PMCID: PMC8896822.

9. Martin-Baniandres P, Lan WH, Board S, Romero-Ruiz M, Garcia-Manyes S, Qing Y, Bayley H. Enzyme-less nanopore detection of post-translational modifications within long polypeptides. Nat Nanotechnol. 2023 Jul 27. doi: 10.1038/s41565-023-01462-8. Epub ahead of print. PMID: 37500774.

10. Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, Peluso P, Rank D, Baybayan P, Bettman B, Bibillo A, Bjornson K, Chaudhuri B, Christians F, Cicero R, Clark S, Dalal R, Dewinter A, Dixon J, Foquet M, Gaertner A, Hardenbol P, Heiner C, Hester K, Holden D, Kearns G, Kong X, Kuse R, Lacroix Y, Lin S, Lundquist P, Ma C, Marks P, Maxham M, Murphy D, Park I, Pham T, Phillips M, Roy J, Sebra R, Shen G, Sorenson J, Tomaney A, Travers K, Trulson M, Vieceli J, Wegener J, Wu D, Yang A, Zaccarin D, Zhao P, Zhong F, Korlach J, Turner S. Real-time DNA sequencing from single polymerase molecules. Science. 2009 Jan 2;323(5910):133-8. doi: 10.1126/science.1162986. Epub 2008 Nov 20. PMID: 19023044.

11. Rhoads A, Au KF. PacBio Sequencing and Its Applications. Genomics Proteomics Bioinformatics. 2015 Oct;13(5):278-89. doi: 10.1016/j.gpb.2015.08.002. Epub 2015 Nov 2. PMID: 26542840; PMCID: PMC4678779.

12. Nurk S, Koren S, Rhie A, Rautiainen M, Bzikadze AV, Mikheenko A, Vollger MR, Altemose N, Uralsky L, Gershman A, Aganezov S, Hoyt SJ, Diekhans M, Logsdon GA, Alonge M, Antonarakis SE, Borchers M, Bouffard GG, Brooks SY, Caldas GV, Chen NC, Cheng H, Chin CS, Chow W, de Lima LG, Dishuck PC, Durbin R, Dvorkina T, Fiddes IT, Formenti G, Fulton RS, Fungtammasan A, Garrison E, Grady

PGS, Graves-Lindsay TA, Hall IM, Hansen NF, Hartley GA, Haukness M, Howe K, Hunkapiller MW, Jain C, Jain M, Jarvis ED, Kerpedjiev P, Kirsche M, Kolmogorov M, Korlach J, Kremitzki M, Li H, Maduro VV, Marschall T, McCartney AM, McDaniel J, Miller DE, Mullikin JC, Myers EW, Olson ND, Paten B, Peluso P, Pevzner PA, Porubsky D, Potapova T, Rogaev EI, Rosenfeld JA, Salzberg SL, Schneider VA, Sedlazeck FJ, Shafin K, Shew CJ, Shumate A, Sims Y, Smit AFA, Soto DC, Sović I, Storer JM, Streets A, Sullivan BA, Thibaud-Nissen F, Torrance J, Wagner J, Walenz BP, Wenger A, Wood JMD, Xiao C, Yan SM, Young AC, Zarate S, Surti U, McCoy RC, Dennis MY, Alexandrov IA, Gerton JL, O'Neill RJ, Timp W, Zook JM, Schatz MC, Eichler EE, Miga KH, Phillippy AM. The complete sequence of a human genome. Science. 2022 Apr;376(6588):44-53. doi: 10.1126/science.abj6987. Epub 2022 Mar 31. PMID: 35357919; PMCID: PMC9186530.

13. Drexler HL, Choquet K, Merens HE, Tang PS, Simpson JT, Churchman LS. Revealing nascent RNA processing dynamics with nano-COP. Nat Protoc. 2021 Mar;16(3):1343-1375. doi: 10.1038/s41596-020-00469-y. Epub 2021 Jan 29. PMID: 33514943; PMCID: PMC8713461.

14. Schlesinger F, Smith AD, Gingeras TR, Hannon GJ, Hodges E. De novo DNA demethylation and noncoding transcription define active intergenic regulatory elements. Genome Res. 2013 Oct;23(10):1601-14. doi: 10.1101/gr.157271.113. Epub 2013 Jun 28. PMID: 23811145; PMCID: PMC3787258.

15. Boyle AP, Davis S, Shulha HP, Meltzer P, Margulies EH, Weng Z, Furey TS, Crawford GE. High-resolution mapping and characterization of open chromatin across the genome. Cell. 2008 Jan 25;132(2):311-22. doi: 10.1016/j.cell.2007.12.014. PMID: 18243105; PMCID: PMC2669738.

16. Frommer M, McDonald LE, Millar DS, Collis CM, Watt F, Grigg GW, Molloy PL, Paul CL. A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. Proc Natl Acad Sci U S A. 1992 Mar 1;89(5):1827-31. doi: 10.1073/pnas.89.5.1827. PMID: 1542678; PMCID: PMC48546.

17. Zubradt M, Gupta P, Persad S, Lambowitz AM, Weissman JS, Rouskin S. DMS-MaPseq for genome-wide or targeted RNA structure probing in vivo. Nat Methods. 2017 Jan;14(1):75-82. doi: 10.1038/nmeth.4057. Epub 2016 Nov 7. PMID: 27819661; PMCID: PMC5508988.

18. Johnson DS, Mortazavi A, Myers RM, Wold B. Genome-wide mapping of in vivo protein-DNA interactions. Science. 2007 Jun 8;316(5830):1497-502. doi: 10.1126/science.1141319. Epub 2007 May 31. PMID: 17540862.

19. Zhao J, Ohsumi TK, Kung JT, Ogawa Y, Grau DJ, Sarma K, Song JJ, Kingston RE, Borowsky M, Lee JT. Genome-wide identification of polycomb-associated RNAs by RIP-seq. Mol Cell. 2010 Dec 22;40(6):939-53. doi: 10.1016/j.molcel.2010.12.011. PMID: 21172659; PMCID: PMC3021903.

20. Garibaldi A, Carranza F, Hertel KJ. Isolation of Newly Transcribed RNA Using the Metabolic Label 4-Thiouridine. Methods Mol Biol. 2017;1648:169-176. doi: 10.1007/978-1-4939-7204-3_13. PMID: 28766297; PMCID: PMC5783291.

21. Carlile TM, Rojas-Duran MF, Gilbert WV. Pseudo-Seq: Genome-Wide Detection of Pseudouridine Modifications in RNA. Methods Enzymol. 2015;560:219-45. doi: 10.1016/bs.mie.2015.03.011. Epub 2015 May 28. PMID: 26253973; PMCID: PMC7945874.

22. Krebs AR, Imanci D, Hoerner L, Gaidatzis D, Burger L, Schübeler D. Genome-wide Single-Molecule Footprinting Reveals High RNA Polymerase II Turnover at Paused Promoters. Mol Cell. 2017 Aug 3;67(3):411-422.e4. doi: 10.1016/j.molcel.2017.06.027. Epub 2017 Jul 20. PMID: 28735898; PMCID: PMC5548954.

23. Stergachis AB, Debo BM, Haugen E, Churchman LS, Stamatoyannopoulos JA. Single-molecule regulatory architectures captured by chromatin fiber sequencing. Science. 2020 Jun 26;368(6498):1449-1454. doi: 10.1126/science.aaz1646. PMID: 32587015.

24. Abdulhay NJ, McNally CP, Hsieh LJ, Kasinathan S, Keith A, Estes LS, Karimzadeh M, Underwood JG, Goodarzi H, Narlikar GJ, Ramani V. Massively multiplex single-molecule oligonucleosome footprinting. Elife. 2020 Dec 2;9:e59404. doi: 10.7554/eLife.59404. PMID: 33263279; PMCID: PMC7735760.

25. Jha A, Bohaczuk SC, Mao Y, Ranchalis J, Mallory BJ, Min AT, Hamm MO, Swanson E, Finkbeiner C, Li T, Whittington D, Noble WS, Stergachis AB, Vollger MR. Fibertools: fast and accurate DNA-m6A calling using single-molecule long-read sequencing. bioRxiv [Preprint]. 2023 Jul 6:2023.04.20.537673. doi: 10.1101/2023.04.20.537673. PMID: 37131601; PMCID: PMC10153250.

26. Spector BM, Parida M, Li M, Ball CB, Meier JL, Luse DS, Price DH. Differences in RNA polymerase II complexes and their interactions with surrounding chromatin on human and cytomegalovirus genomes. Nat Commun. 2022 Apr 14;13(1):2006. PMCID: PMC901040

27. Judd J, Duarte FM, Lis JT. Pioneer-like factor GAF cooperates with PBAP (SWI/SNF) and NURF (ISWI) to regulate transcription. Genes Dev. 2021 Jan 1;35(1-2):147-156. doi: 10.1101/gad.341768.120. Epub 2020 Dec 10. PMID: 33303640; PMCID: PMC7778264.

28. Core L, Adelman K. Promoter-proximal pausing of RNA polymerase II: a nexus of gene regulation. Genes Dev. 2019 Aug 1;33(15-16):960-982. doi: 10.1101/gad.325142.119. Epub 2019 May 23. PMID: 31123063; PMCID: PMC6672056.

29. Muse GW, Gilchrist DA, Nechaev S, Shah R, Parker JS, Grissom SF, Zeitlinger J, Adelman K. RNA polymerase is poised for activation across the genome. Nat Genet. 2007 Dec;39(12):1507-11. doi: 10.1038/ng.2007.21. Epub 2007 Nov 11. PMID: 17994021; PMCID: PMC2365887.

30. Gilmour DS, Lis JT. RNA polymerase II interacts with the promoter region of the noninduced hsp70 gene in Drosophila melanogaster cells. Mol Cell Biol. 1986 Nov;6(11):3984-9. doi: 10.1128/mcb.6.11.3984-3989.1986. PMID: 3099167; PMCID: PMC367162.

31. Rougvie AE, Lis JT. The RNA polymerase II molecule at the 5' end of the uninduced hsp70 gene of D. melanogaster is transcriptionally engaged. Cell. 1988 Sep 9;54(6):795-804. doi: 10.1016/s0092-8674(88)91087-2. PMID: 3136931.

32. Martell DJ, Merens HE, Caulier A, Fiorini C, Ulirsch JC, Ietswaart R, Choquet K, Graziadei G, Brancaleoni V, Cappellini MD, Scott C, Roberts N, Proven M, Roy NBA, Babbs C, Higgs DR, Sankaran VG, Churchman LS. RNA polymerase II pausing temporally coordinates cell cycle progression and erythroid differentiation. Dev Cell. 2023 Aug 8:S1534-5807(23)00365-9. doi: 10.1016/j.devcel.2023.07.018. Epub ahead of print. PMID: 37586368.

33. Zeitlinger J, Stark A, Kellis M, Hong JW, Nechaev S, Adelman K, Levine M, Young RA. RNA polymerase stalling at developmental control genes in the Drosophila melanogaster embryo. Nat Genet. 2007 Dec;39(12):1512-6. doi: 10.1038/ng.2007.26. Epub 2007 Nov 11. PMID: 17994019; PMCID: PMC2824921.

34. Gaertner B, Zeitlinger J. RNA polymerase II pausing during development. Development. 2014 Mar;141(6):1179-83. doi: 10.1242/dev.088492. PMID: 24595285; PMCID: PMC3943177.

35. Shao W, Zeitlinger J. Paused RNA polymerase II inhibits new transcriptional initiation. Nat Genet. 2017 Jul;49(7):1045-1051. doi: 10.1038/ng.3867. Epub 2017 May 15. PMID: 28504701.

36. Gressel S, Schwalb B, Cramer P. The pause-initiation limit restricts transcription activation in human cells. Nat Commun. 2019 Aug 9;10(1):3603. doi: 10.1038/s41467-019-11536-8. PMID: 31399571; PMCID: PMC6689055.

37. Gressel S, Schwalb B, Decker TM, Qin W, Leonhardt H, Eick D, Cramer P. CDK9-dependent RNA polymerase II pausing controls transcription initiation. Elife. 2017 Oct 10;6:e29736. doi: 10.7554/eLife.29736. PMID: 28994650; PMCID: PMC5669633.

38. Gilchrist DA, Dos Santos G, Fargo DC, Xie B, Gao Y, Li L, Adelman K. Pausing of RNA polymerase II disrupts DNA-specified nucleosome organization to enable precise gene regulation. Cell. 2010 Nov 12;143(4):540-51. doi: 10.1016/j.cell.2010.10.004. PMID: 21074046; PMCID: PMC2991113.

39. Ardehali MB, Lis JT. Tracking rates of transcription and splicing in vivo. Nat Struct Mol Biol. 2009 Nov;16(11):1123-4. doi: 10.1038/nsmb1109-1123. PMID: 19888309.

40. Petesch SJ, Lis JT. Overcoming the nucleosome barrier during transcript elongation. Trends Genet. 2012 Jun;28(6):285-94. doi: 10.1016/j.tig.2012.02.005. Epub 2012 Mar 31. PMID: 22465610; PMCID: PMC3466053.

41. Ramachandran S, Ahmad K, Henikoff S. Transcription and Remodeling Produce Asymmetrically Unwrapped Nucleosomal Intermediates. Mol Cell. 2017 Dec 21;68(6):1038-1053.e4. doi: 10.1016/j.molcel.2017.11.015. Epub 2017 Dec 7. PMID: 29225036; PMCID: PMC6421108.

42. Farnung L, Ochmann M, Garg G, Vos SM, Cramer P. Structure of a backtracked hexasomal intermediate of nucleosome transcription. Mol Cell. 2022 Sep 1;82(17):3126-3134.e7. doi: 10.1016/j.molcel.2022.06.027. Epub 2022 Jul 19. PMID: 35858621.

43. Teves SS, Henikoff S. Transcription-generated torsional stress destabilizes nucleosomes. Nat Struct Mol Biol. 2014 Jan;21(1):88-94. doi: 10.1038/nsmb.2723. Epub 2013 Dec 8. PMID: 24317489; PMCID: PMC3947361.

44. Petesch SJ, Lis JT. Rapid, transcription-independent loss of nucleosomes over a large chromatin domain at Hsp70 loci. Cell. 2008 Jul 11;134(1):74-84. doi: 10.1016/j.cell.2008.05.029. PMID: 18614012; PMCID: PMC2527511.

45. Zheng H, Xie W. The role of 3D genome organization in development and cell differentiation. Nat Rev Mol Cell Biol. 2019 Sep;20(9):535-550. doi: 10.1038/s41580-019-0132-4. PMID: 31197269.

46. Hafner A, Boettiger A. The spatial organization of transcriptional control. Nat Rev Genet. 2023 Jan;24(1):53-68. doi: 10.1038/s41576-022-00526-0. Epub 2022 Sep 14. PMID: 36104547.

47. Sutherland H, Bickmore WA. Transcription factories: gene expression in unions? Nat Rev Genet. 2009 Jul;10(7):457-66. doi: 10.1038/nrg2592. PMID: 19506577.

48. Kimura H, Sato Y. Imaging transcription elongation dynamics by new technologies unveils the organization of initiation and elongation in transcription factories. Curr Opin Cell Biol. 2022 Feb;74:71-79. doi: 10.1016/j.ceb.2022.01.002. Epub 2022 Feb 17. PMID: 35183895.

49. Patange S, Ball DA, Karpova TS, Larson DR. Towards a 'Spot On' Understanding of Transcription in the Nucleus. J Mol Biol. 2021 Jul 9;433(14):167016. doi: 10.1016/j.jmb.2021.167016. Epub 2021 May 2. PMID: 33951451; PMCID: PMC8184600.

50. Yang L, Yu J. A comparative analysis of divergently-paired genes (DPGs) among Drosophila and vertebrate genomes. BMC Evol Biol. 2009 Mar 11;9:55. doi: 10.1186/1471-2148-9-55. PMID: 19284596; PMCID: PMC2670823.

51. Lawrence JG. Shared strategies in gene organization among prokaryotes and eukaryotes. Cell. 2002 Aug 23;110(4):407-13. doi: 10.1016/s0092-8674(02)00900-5. PMID: 12202031.

52. Herr DR, Harris GL. Close head-to-head juxtaposition of genes favors their coordinate regulation in Drosophila melanogaster. FEBS Lett. 2004 Aug 13;572(1-3):147-53. doi: 10.1016/j.febslet.2004.07.026. PMID: 15304339.

53. Small S, Arnosti DN. Transcriptional Enhancers in *Drosophila*. Genetics. 2020 Sep;216(1):1-26. doi: 10.1534/genetics.120.301370. PMID: 32878914; PMCID: PMC7463283.

54. Arnold CD, Gerlach D, Stelzer C, Boryń ŁM, Rath M, Stark A. Genome-wide quantitative enhancer activity maps identified by STARR-seq. Science. 2013 Mar 1;339(6123):1074-7. doi: 10.1126/science.1232542. Epub 2013 Jan 17. PMID: 23328393.

55. Cubeñas-Potts C, Rowley MJ, Lyu X, Li G, Lei EP, Corces VG. Different enhancer classes in Drosophila bind distinct architectural proteins and mediate unique chromatin interactions and 3D architecture. Nucleic Acids Res. 2017 Feb 28;45(4):1714-1730. doi: 10.1093/nar/gkw1114. PMID: 27899590; PMCID: PMC5389536.

56. Henriques T, Scruggs BS, Inouye MO, Muse GW, Williams LH, Burkholder AB, Lavender CA, Fargo DC, Adelman K. Widespread transcriptional pausing and elongation control at enhancers. Genes Dev. 2018 Jan 1;32(1):26-41. doi: 10.1101/gad.309351.117. Epub 2018 Jan 29. PMID: 29378787; PMCID: PMC5828392.

57. Sartorelli V, Lauberth SM. Enhancer RNAs are an important regulatory layer of the epigenome. Nat Struct Mol Biol. 2020 Jun;27(6):521-528. doi: 10.1038/s41594-020-0446-0. Epub 2020 Jun 8. PMID: 32514177; PMCID: PMC7343394.

58. Mikhaylichenko O, Bondarenko V, Harnett D, Schor IE, Males M, Viales RR, Furlong EEM. The degree of enhancer or promoter activity is reflected by the levels and directionality of eRNA transcription. Genes Dev. 2018 Jan 1;32(1):42-57. doi: 10.1101/gad.308619.117. Epub 2018 Jan 29. PMID: 29378788; PMCID: PMC5828394.

59. Wang Q, Sun Q, Czajkowsky DM, Shao Z. Sub-kb Hi-C in D. melanogaster reveals conserved characteristics of TADs between insect and mammalian cells. Nat Commun. 2018 Jan 15;9(1):188. doi: 10.1038/s41467-017-02526-9. PMID: 29335463; PMCID: PMC5768742.

60. Beagan JA, Phillips-Cremins JE. On the existence and functionality of topologically associating domains. Nat Genet. 2020 Jan;52(1):8-16. doi: 10.1038/s41588-019-0561-1. Epub 2020 Jan 10. PMID: 31925403; PMCID: PMC7567612.

61. Schwartz YB, Cavalli G. Three-Dimensional Genome Organization and Function in Drosophila. Genetics. 2017 Jan;205(1):5-24. doi: 10.1534/genetics.115.185132. PMID: 28049701; PMCID: PMC5223523.

62. Kahn TG, Savitsky M, Kuong C, Jacquier C, Cavalli G, Chang JM, Schwartz YB. Topological screen identifies hundreds of Cp190- and CTCF-dependent *Drosophila* chromatin insulator elements. Sci Adv. 2023 Feb 3;9(5):eade0090. doi: 10.1126/sciadv.ade0090. Epub 2023 Feb 3. PMID: 36735780; PMCID: PMC9897668.

63. Batut PJ, Bing XY, Sisco Z, Raimundo J, Levo M, Levine MS. Genome organization controls transcriptional dynamics during development. Science. 2022 Feb 4;375(6580):566-570. doi: 10.1126/science.abi7178. Epub 2022 Feb 3. PMID: 35113722; PMCID: PMC10368186.

64. Turowski TW, Tollervey D. Transcription by RNA polymerase III: insights into mechanism and regulation. Biochem Soc Trans. 2016 Oct 15;44(5):1367-1375. doi: 10.1042/BST20160062. PMID: 27911719; PMCID: PMC5095917.

65. Dieci G, Sentenac A. Facilitated recycling pathway for RNA polymerase III. Cell. 1996 Jan 26;84(2):245-52. doi: 10.1016/s0092-8674(00)80979-4. PMID: 8565070.

66. Raab JR, Chiu J, Zhu J, Katzman S, Kurukuti S, Wade PA, Haussler D, Kamakaka RT. Human tRNA genes function as chromatin insulators. EMBO J. 2012 Jan 18;31(2):330-50. doi: 10.1038/emboj.2011.406. Epub 2011 Nov 15. PMID: 22085927; PMCID: PMC3261562.

67. Roger AJ, Muñoz-Gómez SA, Kamikawa R. The Origin and Diversification of Mitochondria. Curr Biol. 2017 Nov 6;27(21):R1177-R1192. doi: 10.1016/j.cub.2017.09.015. PMID: 29112874.

68. Couvillion MT, Soto IC, Shipkovenska G, Churchman LS. Synchronized mitochondrial and cytosolic translation programs. Nature. 2016 May 26;533(7604):499-503. doi: 10.1038/nature18015. Epub 2016 May 11. PMID: 27225121; PMCID: PMC4964289.

69. Isaac RS, McShane E, Churchman LS. The Multiple Levels of Mitonuclear Coregulation. Annu Rev Genet. 2018 Nov 23;52:511-533. doi: 10.1146/annurev-genet-120417-031709. Epub 2018 Sep 19. PMID: 30230928.

70. Kukat C, Davies KM, Wurm CA, Spåhr H, Bonekamp NA, Kühl I, Joos F, Polosa PL, Park CB, Posse V, Falkenberg M, Jakobs S, Kühlbrandt W, Larsson NG. Cross-strand binding of TFAM to a single mtDNA molecule forms the mitochondrial nucleoid. Proc Natl Acad Sci U S A. 2015 Sep 8;112(36):11288-93. doi: 10.1073/pnas.1512131112. Epub 2015 Aug 24. PMID: 26305956; PMCID: PMC4568684.

71. Ngo HB, Lovely GA, Phillips R, Chan DC. Distinct structural features of TFAM drive mitochondrial DNA packaging versus transcriptional activation. Nat Commun. 2014;5:3077. doi: 10.1038/ncomms4077. PMID: 24435062; PMCID: PMC3936014.

72. Choi WS, Garcia-Diaz M. A minimal motif for sequence recognition by mitochondrial transcription factor A (TFAM). Nucleic Acids Res. 2022 Jan 11;50(1):322-332. doi: 10.1093/nar/gkab1230. PMID: 34928349; PMCID: PMC8754647.

73. Milenkovic D, Matic S, Kühl I, Ruzzenente B, Freyer C, Jemt E, Park CB, Falkenberg M, Larsson NG. TWINKLE is an essential mitochondrial helicase required for synthesis of nascent D-loop strands and complete mtDNA replication. Hum Mol Genet. 2013 May 15;22(10):1983-93. doi: 10.1093/hmg/ddt051. Epub 2013 Feb 7. PMID: 23393161; PMCID: PMC3633371.

74. Falkenberg M, Larsson NG, Gustafsson CM. DNA replication and transcription in mammalian mitochondria. Annu Rev Biochem. 2007;76:679-99. doi: 10.1146/annurev.biochem.76.060305.152028. PMID: 17408359.

75. Falkenberg M. Mitochondrial DNA replication in mammalian cells: overview of the pathway. Essays Biochem. 2018 Jul 20;62(3):287-296. doi: 10.1042/EBC20170100. PMID: 29880722; PMCID: PMC6056714.

76. Chang DD, Clayton DA. Priming of human mitochondrial DNA replication occurs at the light-strand promoter. Proc Natl Acad Sci U S A. 1985 Jan;82(2):351-5. doi: 10.1073/pnas.82.2.351. PMID: 2982153; PMCID: PMC397036.

77. Barshad G, Marom S, Cohen T, Mishmar D. Mitochondrial DNA Transcription and Its Regulation: An Evolutionary Perspective. Trends Genet. 2018 Sep;34(9):682-692. doi: 10.1016/j.tig.2018.05.009. Epub 2018 Jun 23. PMID: 29945721.

78. Terzioglu M, Ruzzenente B, Harmel J, Mourier A, Jemt E, López MD, Kukat C, Stewart JB, Wibom R, Meharg C, Habermann B, Falkenberg M, Gustafsson CM, Park CB, Larsson NG. MTERF1 binds mtDNA to prevent transcriptional interference at the light-strand promoter but is dispensable for rRNA gene transcription regulation. Cell Metab. 2013 Apr 2;17(4):618-26. doi: 10.1016/j.cmet.2013.03.006. PMID: 23562081.

79. Roberti M, Polosa PL, Bruni F, Manzari C, Deceglie S, Gadaleta MN, Cantatore P. The MTERF family proteins: mitochondrial transcription regulators and beyond. Biochim Biophys Acta. 2009 May;1787(5):303-11. doi: 10.1016/j.bbabio.2009.01.013. Epub 2009 Jan 30. PMID: 19366610.

80. Fisher RP, Clayton DA. A transcription factor required for promoter recognition by human mitochondrial RNA polymerase. Accurate initiation at the heavy- and light-strand promoters dissected and reconstituted in vitro. J Biol Chem. 1985 Sep 15;260(20):11330-8. PMID: 4030791.

81. Ramachandran A, Basu U, Sultana S, Nandakumar D, Patel SS. Human mitochondrial transcription factors TFAM and TFB2M work synergistically in promoter melting during transcription initiation. Nucleic Acids Res. 2017 Jan 25;45(2):861-874. doi: 10.1093/nar/gkw1157. Epub 2016 Nov 29. PMID: 27903899; PMCID: PMC5314767.

82. Inatomi T, Matsuda S, Ishiuchi T, Do Y, Nakayama M, Abe S, Kasho K, Wanrooij S, Nakada K, Ichiyanagi K, Sasaki H, Yasukawa T, Kang D. TFB2M and POLRMT are essential for mammalian mitochondrial DNA replication. Biochim Biophys Acta Mol Cell Res. 2022 Jan;1869(1):119167. doi: 10.1016/j.bbamcr.2021.119167. Epub 2021 Oct 30. PMID: 34744028.

83. Mayer A, Landry HM, Churchman LS. Pause & go: from the discovery of RNA polymerase pausing to its functional implications. Curr Opin Cell Biol. 2017 Jun;46:72-80. doi: 10.1016/j.ceb.2017.03.002. Epub 2017 Mar 28. PMID: 28363125; PMCID: PMC5505790.

84. Churchman LS, Weissman JS. Native elongating transcript sequencing (NET-seq). Curr Protoc Mol Biol. 2012 Apr;Chapter 4:Unit 4.14.1-17. doi: 10.1002/0471142727.mb0414s98. PMID: 22470065.

85. Farnung L, Vos SM, Cramer P. Structure of transcribing RNA polymerase II-nucleosome complex. Nat Commun. 2018 Dec 21;9(1):5432. doi: 10.1038/s41467-018-07870-y. PMID: 30575770; PMCID: PMC6303367.

86. Wang H, Schilbach S, Ninov M, Urlaub H, Cramer P. Structures of transcription preinitiation complex engaged with the +1 nucleosome. Nat Struct Mol Biol. 2023 Feb;30(2):226-232. doi: 10.1038/s41594-022-00865-w. Epub 2022 Nov 21. PMID: 36411341; PMCID: PMC9935396.

87. Zhijie Chen Ronen Gabizon Aidan I Brown Antony Lee Aixin Song César Díaz-Celis Craig D Kaplan Elena F Koslover Tingting Yao Carlos Bustamante (2019) High-resolution and high-accuracy topographic and transcriptional maps of the nucleosome barrier eLife 8:e48281.

88. Hodges C, Bintu L, Lubkowska L, Kashlev M, Bustamante C. Nucleosomal fluctuations govern the transcription dynamics of RNA polymerase II. Science. 2009 Jul 31;325(5940):626-8. doi: 10.1126/science.1172926. PMID: 19644123; PMCID: PMC2775800.

89. Cisse II, Izeddin I, Causse SZ, Boudarene L, Senecal A, Muresan L, Dugast-Darzacq C, Hajj B, Dahan M, Darzacq X. Real-time dynamics of RNA polymerase II clustering in live human cells. Science. 2013 Aug 9;341(6146):664-7. doi: 10.1126/science.1239053. Epub 2013 Jul 4. PMID: 23828889.

90. Cho WK, Spille JH, Hecht M, Lee C, Li C, Grube V, Cisse II. Mediator and RNA polymerase II clusters associate in transcription-dependent condensates. Science. 2018 Jul 27;361(6400):412-415. doi: 10.1126/science.aar4199. Epub 2018 Jun 21. PMID: 29930094; PMCID: PMC6543815.

91. Lee I, Razaghi R, Gilpatrick T, Molnar M, Gershman A, Sadowski N, Sedlazeck FJ, Hansen KD, Simpson JT, Timp W. Simultaneous profiling of chromatin accessibility and methylation on human cell lines with nanopore sequencing. Nat Methods. 2020 Dec;17(12):1191-1199. doi: 10.1038/s41592-020-01000-7. Epub 2020 Nov 23. PMID: 33230324; PMCID: PMC7704922.

92. Shipony Z, Marinov GK, Swaffer MP, Sinnott-Armstrong NA, Skotheim JM, Kundaje A, Greenleaf WJ. Long-range single-molecule mapping of chromatin accessibility in eukaryotes. Nat Methods. 2020 Mar;17(3):319-327. doi: 10.1038/s41592-019-0730-2. Epub 2020 Feb 10. PMID: 32042188; PMCID: PMC7968351.

93. Vos SM, Farnung L, Urlaub H, Cramer P. Structure of paused transcription complex Pol II-DSIF-NELF. Nature. 2018 Aug;560(7720):601-606. doi: 10.1038/s41586-018-0442-2. Epub 2018 Aug 22. PMID: 30135580; PMCID: PMC6245578.

94. Kwak H, Fuda NJ, Core LJ, Lis JT. Precise maps of RNA polymerase reveal how promoters direct initiation and pausing. Science. 2013 Feb 22;339(6122):950-3. doi: 10.1126/science.1229386. PMID: 23430654; PMCID: PMC3974810.

95. ENCODE

96. Titov DV, Gilman B, He QL, Bhat S, Low WK, Dang Y, Smeaton M, Demain AL, Miller PS, Kugel JF, Goodrich JA, Liu JO. XPB, a subunit of TFIIH, is a target of the natural product triptolide. Nat Chem Biol. 2011 Mar;7(3):182-8. doi: 10.1038/nchembio.522. Epub 2011 Jan 30. PMID: 21278739; PMCID: PMC3622543.

97. Bushnell DA, Cramer P, Kornberg RD. Structural basis of transcription: alpha-amanitin-RNA polymerase II cocrystal at 2.8 A resolution. Proc Natl Acad Sci U S A. 2002 Feb 5;99(3):1218-22. doi: 10.1073/pnas.251664698. Epub 2002 Jan 22. PMID: 11805306; PMCID: PMC122170.

98. Chereji RV, Bryson TD, Henikoff S. Quantitative MNase-seq accurately maps nucleosome occupancy levels. Genome Biol. 2019 Sep 13;20(1):198. doi: 10.1186/s13059-019-1815-z. PMID: 31519205; PMCID: PMC6743174.

99. Zabidi MA, Arnold CD, Schernhuber K, Pagani M, Rath M, Frank O, Stark A. Enhancer-core-promoter specificity separates developmental and housekeeping gene regulation. Nature. 2015 Feb 26;518(7540):556-9. doi: 10.1038/nature13994. Epub 2014 Dec 15. PMID: 25517091; PMCID: PMC6795551.

100. Ulianov SV, Khrameeva EE, Gavrilov AA, Flyamer IM, Kos P, Mikhaleva EA, Penin AA, Logacheva MD, Imakaev MV, Chertovich A, Gelfand MS, Shevelyov YY, Razin SV. Active chromatin and transcription play a key role in chromosome partitioning into topologically associating domains. Genome Res. 2016 Jan;26(1):70-84. doi: 10.1101/gr.196006.115. Epub 2015 Oct 30. PMID: 26518482; PMCID: PMC4691752.

101. Yang J, Ramos E, Corces VG. The BEAF-32 insulator coordinates genome organization and function during the evolution of Drosophila species. Genome Res. 2012 Nov;22(11):2199-207. doi: 10.1101/gr.142125.112. Epub 2012 Aug 15. PMID: 22895281; PMCID: PMC3483549.

102. Ong CT, Van Bortle K, Ramos E, Corces VG. Poly(ADP-ribosyl)ation regulates insulator function and intrachromosomal interactions in Drosophila. Cell. 2013 Sep 26;155(1):148-59. doi: 10.1016/j.cell.2013.08.052. Epub 2013 Sep 19. PMID: 24055367; PMCID: PMC3816015.

103. Kaushal A, Mohana G, Dorier J, Özdemir I, Omer A, Cousin P, Semenova A, Taschner M, Dergai O, Marzetta F, Iseli C, Eliaz Y, Weisz D, Shamim MS, Guex N, Lieberman Aiden E, Gambetta MC. CTCF loss has limited effects on global genome architecture in Drosophila despite critical regulatory functions. Nat Commun. 2021 Feb 12;12(1):1011. doi: 10.1038/s41467-021-21366-2. PMID: 33579945; PMCID: PMC7880997.

104. Graczyk D, Cieśla M, Boguta M. Regulation of tRNA synthesis by the general transcription factors of RNA polymerase III - TFIIIB and TFIIIC, and by the MAF1 protein. Biochim Biophys Acta Gene Regul Mech. 2018 Apr;1861(4):320-329. doi: 10.1016/j.bbagrm.2018.01.011. Epub 2018 Feb 6. PMID: 29378333.

105. Male G, von Appen A, Glatt S, Taylor NM, Cristovao M, Groetsch H, Beck M, Müller CW. Architecture of TFIIIC and its role in RNA polymerase III pre-initiation complex assembly. Nat Commun. 2015 Jun 10;6:7387. doi: 10.1038/ncomms8387. PMID: 26060179; PMCID: PMC4490372.

106. Geslain, R. & Pan, T. Functional analysis of human tRNA isodecoders. J. Mol. Biol. 396, 821–831 (2010).

107. Marygold SJ, Chan PP, Lowe TM. Systematic identification of tRNA genes in Drosophila melanogaster. MicroPubl Biol. 2022 May 1;2022:10.17912/micropub.biology.000560. doi: 10.17912/micropub.biology.000560. PMID: 35789696; PMCID: PMC9249942.

108. Behrens A., Rodschinka G., Nedialkova D.D. High-resolution quantitative profiling of tRNA abundance and modification status in eukaryotes by mim-tRNAseq. Mol. Cell. 2021; 81:1802–1815.

109. Lucas MC, Pryszcz LP, Medina R, Milenkovic I, Camacho N, Marchand V, Motorin Y, Ribas de Pouplana L, Novoa EM. Quantitative analysis of tRNA abundance and modifications by nanopore RNA sequencing. Nat Biotechnol. 2023 Apr 6. doi: 10.1038/s41587-023-01743-6. Epub ahead of print. PMID: 37024678.

110. Procunier JD, Tartof KD. Genetic analysis of the 5S RNA genes in Drosophila melanogaster. Genetics. 1975 Nov;81(3):515-23. doi: 10.1093/genetics/81.3.515. PMID: 812776; PMCID: PMC1213417.

111. Hull MW, Erickson J, Johnston M, Engelke DR. tRNA genes as transcriptional repressor elements. Mol Cell Biol. 1994 Feb;14(2):1266-77. doi: 10.1128/mcb.14.2.1266-1277.1994. PMID: 8289806; PMCID: PMC358482

112. Farge G, Falkenberg M. Organization of DNA in Mammalian Mitochondria. Int J Mol Sci. 2019 Jun 5;20(11):2770. doi: 10.3390/ijms20112770. PMID: 31195723; PMCID: PMC6600607.

113. Ojala D, Crews S, Montoya J, Gelfand R, Attardi G. A small polyadenylated RNA (7 S RNA), containing a putative ribosome attachment site, maps near the origin of human mitochondrial DNA replication. J Mol Biol. 1981 Aug 5;150(2):303-14. doi: 10.1016/0022-2836(81)90454-x. PMID: 6172590.

114. Xu B, Clayton DA. RNA-DNA hybrid formation at the human mitochondrial heavy-strand origin ceases at replication start sites: an implication for RNA-DNA hybrids serving as primers. EMBO J. 1996 Jun 17;15(12):3135-43. PMID: 8670814; PMCID: PMC450256.

115. Di Re M, Sembongi H, He J, Reyes A, Yasukawa T, Martinsson P, Bailey LJ, Goffart S, Boyd-Kirkup JD, Wong TS, Fersht AR, Spelbrink JN, Holt IJ. The accessory subunit of mitochondrial DNA polymerase gamma determines the DNA content of mitochondrial nucleoids in human cultured cells. Nucleic Acids Res. 2009 Sep;37(17):5701-13. doi: 10.1093/nar/gkp614. Epub 2009 Jul 22. PMID: 19625489; PMCID: PMC2761280.

116. Clayton DA. Replication of animal mitochondrial DNA. Cell. 1982 Apr;28(4):693-705. doi: 10.1016/0092-8674(82)90049-6. PMID: 6178513.

117. Vos SM, Farnung L, Boehning M, Wigge C, Linden A, Urlaub H, Cramer P. Structure of activated transcription complex Pol II-DSIF-PAF-SPT6. Nature. 2018 Aug;560(7720):607-612. doi: 10.1038/s41586-018-0440-4. Epub 2018 Aug 22. PMID: 30135578.

118. Robberson DL, Clayton DA. Pulse-labeled components in the replication of mitochondrial deoxyribonucleic acid. J Biol Chem. 1973 Jun 25;248(12):4512-4. PMID: 4736430.

119. Wang K, Zhang S, Zhou X, Yang X, Li X, Wang Y, Fan P, Xiao Y, Sun W, Zhang P, Li W, Huang S. Unambiguous discrimination of all 20 proteinogenic amino acids and their modifications by nanopore. Nat Methods. 2023 Sep 25. doi: 10.1038/s41592-023-02021-8. Epub ahead of print. PMID: 37749214.

120. Nicholls TJ, Minczuk M. In D-loop: 40 years of mitochondrial 7S DNA. Exp Gerontol. 2014 Aug;56:175-81. doi: 10.1016/j.exger.2014.03.027. Epub 2014 Apr 4. PMID: 24709344.

121. Suomalainen A, Battersby BJ. Mitochondrial diseases: the contribution of organelle stress responses to pathology. Nat Rev Mol Cell Biol. 2018 Feb;19(2):77-92. doi: 10.1038/nrm.2017.66. Epub 2017 Aug 9. PMID: 28792006.

122. Schon EA, DiMauro S, Hirano M. Human mitochondrial DNA: roles of inherited and somatic mutations. Nat Rev Genet. 2012 Dec;13(12):878-90. doi: 10.1038/nrg3275. PMID: 23154810; PMCID: PMC3959762.

123. Nunnari J, Suomalainen A. Mitochondria: in sickness and in health. Cell. 2012 Mar 16;148(6):1145-59. doi: 10.1016/j.cell.2012.02.035. PMID: 22424226; PMCID: PMC5381524.

124. Kim M, Mahmood M, Reznik E, Gammage PA. Mitochondrial DNA is a major source of driver mutations in cancer. Trends Cancer. 2022 Dec;8(12):1046-1059. doi: 10.1016/j.trecan.2022.08.001. Epub 2022 Aug 27. PMID: 36041967; PMCID: PMC9671861.

125. Gorman GS, Schaefer AM, Ng Y, Gomez N, Blakely EL, Alston CL, Feeney C, Horvath R, Yu-Wai-Man P, Chinnery PF, Taylor RW, Turnbull DM, McFarland R. Prevalence of nuclear and mitochondrial DNA mutations related to adult mitochondrial disease. Ann Neurol. 2015 May;77(5):753-9. doi: 10.1002/ana.24362. Epub 2015 Mar 28. PMID: 25652200; PMCID: PMC4737121.

126. Keogh MJ, Chinnery PF. Mitochondrial DNA mutations in neurodegeneration. Biochim Biophys Acta. 2015 Nov;1847(11):1401-11. doi: 10.1016/j.bbabio.2015.05.015. Epub 2015 May 23. PMID: 26014345.

127. Sanchez-Contreras M, Kennedy SR. The Complicated Nature of Somatic mtDNA Mutations in Aging. Front Aging. 2022;2:805126. doi: 10.3389/fragi.2021.805126. Epub 2022 Jan 10. PMID: 35252966; PMCID: PMC8896747.

128. Kukat C, Wurm CA, Spåhr H, Falkenberg M, Larsson NG, Jakobs S. Super-resolution microscopy reveals that mammalian mitochondrial nucleoids have a uniform size and frequently contain a single copy of mtDNA. Proc Natl Acad Sci U S A. 2011 Aug 16;108(33):13534-9. doi: 10.1073/pnas.1109263108. Epub 2011 Aug 1. PMID: 21808029; PMCID: PMC3158146.

129. Brüser C, Keller-Findeisen J, Jakobs S. The TFAM-to-mtDNA ratio defines inner-cellular nucleoid populations with distinct activity levels. Cell Rep. 2021 Nov 23;37(8):110000. doi: 10.1016/j.celrep.2021.110000. PMID: 34818548.

130. Ngo HB, Lovely GA, Phillips R, Chan DC. Distinct structural features of TFAM drive mitochondrial DNA packaging versus transcriptional activation. Nat Commun. 2014;5:3077. doi: 10.1038/ncomms4077. PMID: 24435062; PMCID: PMC3936014.

131. Kaufman BA, Durisic N, Mativetsky JM, Costantino S, Hancock MA, Grutter P, Shoubridge EA. The mitochondrial transcription factor TFAM coordinates the assembly of multiple DNA molecules into nucleoid-like structures. Mol Biol Cell. 2007 Sep;18(9):3225-36. doi: 10.1091/mbc.e07-05-0404. Epub 2007 Jun 20. PMID: 17581862; PMCID: PMC1951767.

132. Altemose N, Maslan A, Smith OK, Sundararajan K, Brown RR, Mishra R, Detweiler AM, Neff N, Miga KH, Straight AF, Streets A. DiMeLo-seq: a long-read, single-molecule method for mapping protein-DNA interactions genome wide. Nat Methods. 2022 Jun;19(6):711-723. doi: 10.1038/s41592-022-01475-6. Epub 2022 Apr 8. PMID: 35396487; PMCID: PMC9189060.

133. Lu B, Lee J, Nie X, Li M, Morozov YI, Venkatesh S, Bogenhagen DF, Temiakov D, Suzuki CK. Phosphorylation of human TFAM in mitochondria impairs DNA binding and promotes degradation by the AAA+ Lon protease. Mol Cell. 2013 Jan 10;49(1):121-32. doi: 10.1016/j.molcel.2012.10.023. Epub 2012 Nov 29. PMID: 23201127; PMCID: PMC3586414.

134. Cuppari A, Fernández-Millán P, Battistini F, Tarrés-Solé A, Lyonnais S, Iruela G, Ruiz-López E, Enciso Y, Rubio-Cosials A, Prohens R, Pons M, Alfonso C, Tóth K, Rivas G, Orozco M, Solà M. DNA specificities modulate the binding of human transcription factor A to mitochondrial DNA control region. Nucleic Acids Res. 2019 Jul 9;47(12):6519-6537. doi: 10.1093/nar/gkz406. PMID: 31114891; PMCID: PMC6614842.

135. Farge G, Laurens N, Broekmans OD, van den Wildenberg SM, Dekker LC, Gaspari M, Gustafsson CM, Peterman EJ, Falkenberg M, Wuite GJ. Protein sliding and DNA denaturation are essential for DNA organization by human mitochondrial transcription factor A. Nat Commun. 2012;3:1013. doi: 10.1038/ncomms2001. PMID: 22910359.

136. King GA, Hashemi Shabestari M, Taris KH, Pandey AK, Venkatesh S, Thilagavathi J, Singh K, Krishna Koppisetti R, Temiakov D, Roos WH, Suzuki CK, Wuite GJL. Acetylation and phosphorylation of human TFAM regulate TFAM-DNA interactions via contrasting mechanisms. Nucleic Acids Res. 2018 Apr 20;46(7):3633-3642. doi: 10.1093/nar/gky204. PMID: 29897602; PMCID: PMC5909435.

137. Ngo HB, Kaiser JT, Chan DC. The mitochondrial transcription and packaging factor Tfam imposes a U-turn on mitochondrial DNA. Nat Struct Mol Biol. 2011 Oct 30;18(11):1290-6. doi: 10.1038/nsmb.2159. PMID: 22037171; PMCID: PMC3210390.

138. Rubio-Cosials A, Sidow JF, Jiménez-Menéndez N, Fernández-Millán P, Montoya J, Jacobs HT, Coll M, Bernadó P, Solà M. Human mitochondrial transcription factor A induces a U-turn structure in

the light strand promoter. Nat Struct Mol Biol. 2011 Oct 30;18(11):1281-9. doi: 10.1038/nsmb.2160. Erratum in: Nat Struct Mol Biol. 2012 Mar;19(3):364. PMID: 22037172.

139. Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nat Methods. 2013 Dec;10(12):1213-8. doi: 10.1038/nmeth.2688. Epub 2013 Oct 6. PMID: 24097267; PMCID: PMC3959825.

140. Blumberg A, Danko CG, Kundaje A, Mishmar D. A common pattern of DNase I footprinting throughout the human mtDNA unveils clues for a chromatin-like organization. Genome Res. 2018 Aug;28(8):1158-1168. doi: 10.1101/gr.230409.117. Epub 2018 Jul 12. PMID: 30002158; PMCID: PMC6071632.

141. Mercer TR, Neph S, Dinger ME, Crawford J, Smith MA, Shearwood AM, Haugen E, Bracken CP, Rackham O, Stamatoyannopoulos JA, Filipovska A, Mattick JS. The human mitochondrial transcriptome. Cell. 2011 Aug 19;146(4):645-58. doi: 10.1016/j.cell.2011.06.051. PMID: 21854988; PMCID: PMC3160626.

142. Lareau CA, Liu V, Muus C, Praktiknjo SD, Nitsch L, Kautz P, Sandor K, Yin Y, Gutierrez JC, Pelka K, Satpathy AT, Regev A, Sankaran VG, Ludwig LS. Mitochondrial single-cell ATAC-seq for high-throughput multi-omic detection of mitochondrial genotypes and chromatin accessibility. Nat Protoc. 2023 May;18(5):1416-1440. doi: 10.1038/s41596-022-00795-3. Epub 2023 Feb 15. PMID: 36792778; PMCID: PMC10317201.

143. Hensen F, Potter A, van Esveld SL, Tarrés-Solé A, Chakraborty A, Solà M, Spelbrink JN. Mitochondrial RNA granules are critically dependent on mtDNA replication factors Twinkle and mtSSB. Nucleic Acids Res. 2019 Apr 23;47(7):3680-3698. doi: 10.1093/nar/gkz047. PMID: 30715486; PMCID: PMC6468249.

144. Sakamoto Y, Zaha S, Nagasawa S, Miyake S, Kojima Y, Suzuki A, Suzuki Y, Seki M. Long-read whole-genome methylation patterning using enzymatic base conversion and nanopore sequencing. Nucleic Acids Res. 2021 Aug 20;49(14):e81. doi: 10.1093/nar/gkab397. PMID: 34019650; PMCID: PMC8373077.

145. Chen J, Cheng J, Chen X, Inoue M, Liu Y, Song CX. Whole-genome long-read TAPS deciphers DNA methylation patterns at base resolution using PacBio SMRT sequencing technology. Nucleic Acids Res. 2022 Oct 14;50(18):e104. doi: 10.1093/nar/gkac612. PMID: 35849350; PMCID: PMC9561279.

146. Frangini M, Franzolin E, Chemello F, Laveder P, Romualdi C, Bianchi V, Rampazzo C. Synthesis of mitochondrial DNA precursors during myogenesis, an analysis in purified C2C12 myotubes. J Biol Chem. 2013 Feb 22;288(8):5624-35. doi: 10.1074/jbc.M112.441147. Epub 2013 Jan 7. PMID: 23297407; PMCID: PMC3581417.

147. Chen X, Shen Y, Draper W, Buenrostro JD, Litzenburger U, Cho SW, Satpathy AT, Carter AC, Ghosh RP, East-Seletsky A, Doudna JA, Greenleaf WJ, Liphardt JT, Chang HY. ATAC-see reveals the accessible genome by transposase-mediated imaging and sequencing. Nat Methods. 2016 Dec;13(12):1013-1020. doi: 10.1038/nmeth.4031. Epub 2016 Oct 17. PMID: 27749837; PMCID: PMC5509561.

148. Qin J, Guo Y, Xue B, Shi P, Chen Y, Su QP, Hao H, Zhao S, Wu C, Yu L, Li D, Sun Y. ER-mitochondria contacts promote mtDNA nucleoids active transportation via mitochondrial dynamic tubulation. Nat Commun. 2020 Sep 8;11(1):4471. doi: 10.1038/s41467-020-18202-4. PMID: 32901010; PMCID: PMC7478960.

149. Jemt E, Persson Ö, Shi Y, Mehmedovic M, Uhler JP, Dávila López M, Freyer C, Gustafsson CM, Samuelsson T, Falkenberg M. Regulation of DNA replication at the end of the mitochondrial D-loop involves the helicase TWINKLE and a conserved sequence element. Nucleic Acids Res. 2015 Oct 30;43(19):9262-75. doi: 10.1093/nar/gkv804. Epub 2015 Aug 7. PMID: 26253742; PMCID: PMC4627069.

150. Arnold JJ, Sharma SD, Feng JY, Ray AS, Smidansky ED, Kireeva ML, Cho A, Perry J, Vela JE, Park Y, Xu Y, Tian Y, Babusis D, Barauskus O, Peterson BR, Gnatt A, Kashlev M, Zhong W, Cameron CE. Sensitivity of mitochondrial transcription and resistance of RNA polymerase II dependent nuclear transcription to antiviral ribonucleosides. PLoS Pathog. 2012;8(11):e1003030. doi: 10.1371/journal.ppat.1003030. Epub 2012 Nov 15. PMID: 23166498; PMCID: PMC3499576.

151. Pham XH, Farge G, Shi Y, Gaspari M, Gustafsson CM, Falkenberg M. Conserved sequence box II directs transcription termination and primer formation in mitochondria. J Biol Chem. 2006 Aug 25;281(34):24647-52. doi: 10.1074/jbc.M602429200. Epub 2006 Jun 21. PMID: 16790426.

152. Yakubovskaya E, Mejia E, Byrnes J, Hambardjieva E, Garcia-Diaz M. Helix unwinding and base flipping enable human MTERF1 to terminate mitochondrial transcription. Cell. 2010 Jun 11;141(6):982-93. doi: 10.1016/j.cell.2010.05.018. PMID: 20550934; PMCID: PMC2887341.

153. Gillum AM, Clayton DA. Mechanism of mitochondrial DNA replication in mouse L-cells: RNA priming during the initiation of heavy-strand synthesis. J Mol Biol. 1979 Dec 5;135(2):353-68. doi: 10.1016/0022-2836(79)90441-8. PMID: 537082.

154. Brown WM, Shine J, Goodman HM. Human mitochondrial DNA: analysis of 7S DNA from the origin of replication. Proc Natl Acad Sci U S A. 1978 Feb;75(2):735-9. doi: 10.1073/pnas.75.2.735. PMID: 273237; PMCID: PMC411331.

155. Brown TA, Clayton DA. Release of replication termination controls mitochondrial DNA copy number after depletion with 2',3'-dideoxycytidine. Nucleic Acids Res. 2002 May 1;30(9):2004-10. doi: 10.1093/nar/30.9.2004. PMID: 11972339; PMCID: PMC113833.

156. Shen X, Collier JM, Hlaing M, Zhang L, Delshad EH, Bristow J, Bernstein HS. Genome-wide examination of myoblast cell cycle withdrawal during differentiation. Dev Dyn. 2003 Jan;226(1):128-38. doi: 10.1002/dvdy.10200. PMID: 12508234.

157. Kozhukhar N, Alexeyev MF. Limited predictive value of TFAM in mitochondrial biogenesis. Mitochondrion. 2019 Nov;49:156-165. doi: 10.1016/j.mito.2019.08.001. Epub 2019 Aug 13. PMID: 31419493; PMCID: PMC6885536.

158. van der Heijden T, Dekker C. Monte carlo simulations of protein assembly, disassembly, and linear motion on DNA. Biophys J. 2008 Nov 15;95(10):4560-9. doi: 10.1529/biophysj.108.135061. Epub 2008 Jul 25. PMID: 18658217; PMCID: PMC2576399.

159. Uchida A, Murugesapillai D, Kastner M, Wang Y, Lodeiro MF, Prabhakar S, Oliver GV, Arnold JJ, Maher LJ, Williams MC, Cameron CE. Unexpected sequences and structures of mtDNA required for efficient transcription from the first heavy-strand promoter. Elife. 2017 Jul 26;6:e27283. doi: 10.7554/eLife.27283. PMID: 28745586; PMCID: PMC5552277.

160. Uchida A, Murugesapillai D, Kastner M, Wang Y, Lodeiro MF, Prabhakar S, Oliver GV, Arnold JJ, Maher LJ, Williams MC, Cameron CE. Unexpected sequences and structures of mtDNA required for efficient transcription from the first heavy-strand promoter. Elife. 2017 Jul 26;6:e27283. doi: 10.7554/eLife.27283. PMID: 28745586; PMCID: PMC5552277.

161. Bonekamp NA, Jiang M, Motori E, Garcia Villegas R, Koolmeister C, Atanassov I, Mesaros A, Park CB, Larsson NG. High levels of TFAM repress mammalian mitochondrial DNA transcription in vivo. Life Sci Alliance. 2021 Aug 30;4(11):e202101034. doi: 10.26508/lsa.202101034. PMID: 34462320; PMCID: PMC8408345.

162. Shutt TE, Lodeiro MF, Cotney J, Cameron CE, Shadel GS. Core human mitochondrial transcription apparatus is a regulated two-component system in vitro. Proc Natl Acad Sci U S A. 2010 Jul 6;107(27):12133-8. doi: 10.1073/pnas.0910581107. Epub 2010 Jun 18. PMID: 20562347; PMCID: PMC2901451.

163. Lewis SC, Uchiyama LF, Nunnari J. ER-mitochondria contacts couple mtDNA synthesis with mitochondrial division in human cells. Science. 2016 Jul 15;353(6296):aaf5549. doi: 10.1126/science.aaf5549. PMID: 27418514; PMCID: PMC5554545.

164. Phillips AF, Millet AR, Tigano M, Dubois SM, Crimmins H, Babin L, Charpentier M, Piganeau M, Brunet E, Sfeir A. Single-Molecule Analysis of mtDNA Replication Uncovers the Basis of the Common Deletion. Mol Cell. 2017 Feb 2;65(3):527-538.e6. doi: 10.1016/j.molcel.2016.12.014. Epub 2017 Jan 19. PMID: 28111015.

165. Larsson NG. Somatic mitochondrial DNA mutations in mammalian aging. Annu Rev Biochem. 2010;79:683-706. doi: 10.1146/annurev-biochem-060408-093701. PMID: 20350166.

166. Falkenberg M, Gustafsson CM. Mammalian mitochondrial DNA replication and mechanisms of deletion formation. Crit Rev Biochem Mol Biol. 2020 Dec;55(6):509-524. doi: 10.1080/10409238.2020.1818684. Epub 2020 Sep 24. PMID: 32972254.

167. Fontana GA, Gahlon HL. Mechanisms of replication and repair in mitochondrial DNA deletion formation. Nucleic Acids Res. 2020 Nov 18;48(20):11244-11258. doi: 10.1093/nar/gkaa804. PMID: 33021629; PMCID: PMC7672454.

168. Lujan SA, Longley MJ, Humble MH, Lavender CA, Burkholder A, Blakely EL, Alston CL, Gorman GS, Turnbull DM, McFarland R, Taylor RW, Kunkel TA, Copeland WC. Ultrasensitive deletion detection links mitochondrial DNA replication, disease, and aging. Genome Biol. 2020 Sep 17;21(1):248. doi: 10.1186/s13059-020-02138-5. PMID: 32943091; PMCID: PMC7500033.

169. Kang I, Chu CT, Kaufman BA. The mitochondrial transcription factor TFAM in neurodegeneration: emerging evidence and mechanisms. FEBS Lett. 2018 Mar;592(5):793-811. doi: 10.1002/1873-3468.12989. Epub 2018 Feb 15. PMID: 29364506; PMCID: PMC5851836.

170. Hsieh YT, Tu HF, Yang MH, Chen YF, Lan XY, Huang CL, Chen HM, Li WC. Mitochondrial genome and its regulator TFAM modulates head and neck tumourigenesis through intracellular metabolic reprogramming and activation of oncogenic effectors. Cell Death Dis. 2021 Oct 18;12(11):961. doi: 10.1038/s41419-021-04255-w. PMID: 34663785; PMCID: PMC8523524.

171. Soto I, Couvillion M, Hansen KG, McShane E, Moran JC, Barrientos A, Churchman LS. Balanced mitochondrial and cytosolic translatomes underlie the biogenesis of human respiratory complexes. Genome Biol. 2022 Aug 9;23(1):170. doi: 10.1186/s13059-022-02732-9. PMID: 35945592; PMCID: PMC9361522.

172. Jha, A. *et al.* Fibertools: fast and accurate DNA-m6A calling using single-molecule long-read sequencing. *bioRxiv* 2023.04.20.537673 (2023) doi:10.1101/2023.04.20.537673.

173. Picelli S, Björklund AK, Reinius B, Sagasser S, Winberg G, Sandberg R. Tn5 transposase and tagmentation procedures for massively scaled sequencing projects. Genome Res. 2014

Dec;24(12):2033-40. doi: 10.1101/gr.177881.114. Epub 2014 Jul 30. PMID: 25079858; PMCID: PMC4248319.

174. TD buffer. *Cold Spring Harb. Protoc.* **2010**, db.rec12138 (2010).

175. Yan K, Rousseau J, Machol K, Cross LA, Agre KE, Gibson CF, Goverde A, Engleman KL, Verdin H, De Baere E, Potocki L, Zhou D, Cadieux-Dion M, Bellus GA, Wagner MD, Hale RJ, Esber N, Riley AF, Solomon BD, Cho MT, McWalter K, Eyal R, Hainlen MK, Mendelsohn BA, Porter HM, Lanpher BC, Lewis AM, Savatt J, Thiffault I, Callewaert B, Campeau PM, Yang XJ. Deficient histone H3 propionylation by BRPF1-KAT6 complexes in neurodevelopmental disorders and cancer. Sci Adv. 2020 Jan 22;6(4):eaax0021. doi: 10.1126/sciadv.aax0021. PMID: 32010779; PMCID: PMC6976298.

176. Corces MR, Trevino AE, Hamilton EG, Greenside PG, Sinnott-Armstrong NA, Vesuna S, Satpathy AT, Rubin AJ, Montine KS, Wu B, Kathiria A, Cho SW, Mumbach MR, Carter AC, Kasowski M, Orloff LA, Risca VI, Kundaje A, Khavari PA, Montine TJ, Greenleaf WJ, Chang HY. An improved ATAC-seq protocol reduces background and enables interrogation of frozen tissues. Nat Methods. 2017 Oct;14(10):959-962. doi: 10.1038/nmeth.4396. Epub 2017 Aug 28. PMID: 28846090; PMCID: PMC5623106.

177. Wolf AR, Mootha VK. Functional genomic analysis of human mitochondrial RNA processing. Cell Rep. 2014 May 8;7(3):918-31. doi: 10.1016/j.celrep.2014.03.035. Epub 2014 Apr 18. PMID: 24746820; PMCID: PMC4289146.

178. Rooney JP, Ryde IT, Sanders LH, Howlett EH, Colton MD, Germ KE, Mayer GD, Greenamyre JT, Meyer JN. PCR based determination of mitochondrial DNA copy number in multiple species. Methods Mol Biol. 2015;1241:23-38. doi: 10.1007/978-1-4939-1875-1_3. PMID: 25308485; PMCID: PMC4312664.

179. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15(12):550. doi: 10.1186/s13059-014-0550-8. PMID: 25516281; PMCID: PMC4302049.

180. Masilamani TJ, Loiselle JJ, Sutherland LC. Assessment of reference genes for real-time quantitative PCR gene expression normalization during C2C12 and H9c2 skeletal muscle differentiation. Mol Biotechnol. 2014 Apr;56(4):329-39. doi: 10.1007/s12033-013-9712-2. PMID: 24146429.

181. Shih YP, Wu HC, Hu SM, Wang TF, Wang AH. Self-cleavage of fusion protein in vivo using TEV protease to yield native protein. Protein Sci. 2005 Apr;14(4):936-41. doi: 10.1110/ps.041129605. Epub 2005 Mar 1. PMID: 15741334; PMCID: PMC2253439.

182. Liu LF, Wang JC. Supercoiling of the DNA template during transcription. Proc Natl Acad Sci U S A. 1987 Oct;84(20):7024-7. doi: 10.1073/pnas.84.20.7024. PMID: 2823250; PMCID: PMC299221.

183. Ma J, Wang MD. DNA supercoiling during transcription. Biophys Rev. 2016 Nov;8(Suppl 1):75-87. doi: 10.1007/s12551-016-0215-9. Epub 2016 Jul 13. PMID: 28275417; PMCID: PMC5338639

184. Visser BJ, Sharma S, Chen PJ, McMullin AB, Bates ML, Bates D. Psoralen mapping reveals a bacterial genome supercoiling landscape dominated by transcription. Nucleic Acids Res. 2022 May 6;50(8):4436-4449. doi: 10.1093/nar/gkac244. PMID: 35420137; PMCID: PMC9071471

185. Guo MS, Kawamura R, Littlehale ML, Marko JF, Laub MT. High-resolution, genome-wide mapping of positive supercoiling in chromosomes. Elife. 2021 Jul 19;10:e67236. doi: 10.7554/eLife.67236. PMID: 34279217; PMCID: PMC8360656.

186. Roy V, Monti-Dedieu L, Chaminade N, Siljak-Yakovlev S, Aulard S, Lemeunier F, Montchamp-Moreau C. Evolution of the chromosomal location of rDNA genes in two Drosophila species subgroups: ananassae and melanogaster. Heredity (Edinb). 2005 Apr;94(4):388-95. doi: 10.1038/sj.hdy.6800612. PMID: 15726113.

187. Indik ZK, Tartof KD. Long spacers among ribosomal genes of Drosophila melanogaster. Nature. 1980 Apr 3;284(5755):477-9. doi: 10.1038/284477a0. PMID: 6244507.

188. Lu KL, Nelson JO, Watase GJ, Warsinger-Pepe N, Yamashita YM. Transgenerational dynamics of rDNA copy number in *Drosophila* male germline stem cells. Elife. 2018 Feb 13;7:e32421. doi: 10.7554/eLife.32421. PMID: 29436367; PMCID: PMC5811208.

189. Li SP, Ou L, Zhang Y, Shen FR, Chen YG. A first-in-class POLRMT specific inhibitor IMT1 suppresses endometrial carcinoma cell growth. Cell Death Dis. 2023 Feb 23;14(2):152. doi: 10.1038/s41419-023-05682-7. PMID: 36823110; PMCID: PMC9950144.

190. Prokop A, Technau GM. BrdU incorporation reveals DNA replication in non dividing glial cells in the larval abdominal CNS of Drosophila. Rouxs Arch Dev Biol. 1994 Oct;204(1):54-61. doi: 10.1007/BF00189068. PMID: 28305806.

191. Xu W, Tang J, Zhao L. DNA-protein cross-links between abasic DNA damage and mitochondrial transcription factor A (TFAM). Nucleic Acids Res. 2023 Jan 11;51(1):41-53. doi: 10.1093/nar/gkac1214. PMID: 36583367; PMCID: PMC9841407.