Notes to paperes

Some subtitle for my thesis

Johan Ulstrup

August 19, 2025

Table of contents

1 index

2 DNA language models are powerful predictors of genome-wide variant effects

The idea for this project was to evaluate the Genome-Pretrained Network (GPN) introduced in (Benegas, Batra, and Song 2024) and determine whether it could achieve greater accuracy than traditional methods in predicting genome-wide variant effects.

The model is designed as a convolutional neural network (CNN) and takes input sequences with a window size of 512. During training, 15% of the positions within each window are masked to enable the model to learn meaningful representations.

The architecture consists of 25 layers, each structured as follows: a dilated convolution layer, followed by an add-and-norm layer with a skip connection from before the dilated convolution. This is followed by a feedforward layer, another add-and-norm layer, and additional skip connections.

A feedforward layer is a fundamental component of neural networks, where inputs pass through one or more fully connected layers with activation functions, transforming data without looping back. This structure helps the model learn complex representations by applying weighted transformations and non-linearities.

A dilated convolution expands the receptive field of a convolutional layer without increasing the number of parameters or reducing resolution. By spacing kernel elements apart, it captures long-range dependencies in sequences, making it particularly useful in genomic data analysis. When combined, dilated convolutions and feedforward layers enhance a model's ability to recognize patterns across different scales efficiently.

After passing through the 25 layers, the model produces a contextual embedding with a dimension of 512 (D=512), followed by classification layers. The final layer outputs the probabilities of the four nucleotides at each masked position.

The GPN variant effect prediction score is calculated as the log-likelihood ratio between the alternate (ALT) and reference (REF) alleles. Here, L represents the window length in base pairs, and D denotes the embedding dimension.

2.1 evolution on X

The production of gametes in sexually reproducing organisms is a complex, tightly regulated process involving numerous cellular and molecular mechanisms. In males, spermatogenesis—the process by which sperm are formed—progresses through four main stages of differentiation: from germ cells to spermatogonia, then to pachytene spermatocytes, round spermatids, and finally to mature spermatozoa **?**(Wang et al. 2019). This process is fundamental to male fertility. Meiosis, a specialized form of cell division, ensures the proper pairing, recombination, and segregation of homologous chromosomes, thereby maintaining genetic integrity and promoting variation. A deep understanding of these molecular events is essential for comprehending inheritance, speciation, population genetics, and the biological causes of infertility.

Sex chromosomes differ from autosomes in several key ways due to their distinct patterns of inheritance and copy number. The Y chromosome is found only in males and exists as a single copy. In contrast, the X chromosome follows a more complex inheritance pattern: females carry two copies, while males carry only one. Consequently, the X chromosome resides two-thirds of the time in females and one-third in males, a distribution that influences how selection acts upon it. In males, the hemizygous nature of the X chromosome means that any mutations or deleterious alleles are fully exposed, without a second copy to mask their effects.

This feature is central to Haldane's rule **?** (Haldane 1922), which posits that in hybrids between two species, the heterogametic sex (e.g., XY in mammals or ZW in birds) is more likely to be inviable or sterile. While widely observed, the precise causes of this pattern remain a subject of ongoing debate. Several hypotheses have been proposed to explain it. One suggests that incompatibilities involving the Y chromosome can arise if it fails to remain compatible with the X chromosome or autosomes. Another centers on dosage compensation, proposing that hybridization may disrupt the regulatory mechanisms that balance gene expression between sex chromosomes. The dominance hypothesis argues that recessive deleterious alleles may be unmasked in the heterogametic sex, leading to sterility. Additional ideas include the faster evolution of male-biased reproductive genes, which can result in functional mismatches, and the accelerated divergence of X-linked loci compared to autosomes, known as faster-X evolution. Lastly, meiotic drive—genetic conflicts between selfish elements on sex chromosomes—has also been implicated in hybrid sterility.

These hypotheses underscore the complex interplay between sex chromosomes, selection, and hybridization **?** (Cowell 2023). Although Haldane's rule is consistently observed across diverse taxa—and even across kingdoms—the underlying mechanisms are not universally agreed upon. In many cases, multiple processes may act together **?** (Lindholm et al. 2016).

The X chromosome, in particular, appears to be subject to strong and unique selective pressures, especially in primates. This topic is explored in greater depth in the following sections.

2.2 Chromatin structure

Regions under strong selection across primate species—including humans and baboons—are consistently found on the X chromosome, spanning megabase-scale genomic areas. As previously noted, factors such as chromosomal architecture and structural rearrangements may play crucial roles in enabling, regulating, or insulating these regions from selective pressures.

Understanding genome organization and variation is therefore essential for uncovering the mechanisms behind selection. Chromatin has long been recognized as central to gene regulation and cellular function (Lieberman-Aiden et al. 2009), in part because the 3D structure of chromosomes brings distant genomic elements into close contact. Disruption of this spatial organization can lead to developmental abnormalities **?** (Dixon et al. 2015). Chromatin is organized hierarchically, with multiple levels of structure nested within one another (see Figure 2). At the broadest level, chromatin compartments can span several megabases **?** (Lieberman-Aiden et al. 2009), while at finer scales, even 500 base pairs can contribute to subgenic structural organization. Structures such as topologically associating domains (TADs) and chromatin loops, which operate at sub-megabase scales, further help partition regulatory elements and maintain proper genomic function **??**(Ramírez et al. 2018; Zuo et al. 2021).

kigger på cromatin overgange er der noget der er anderledens i de områder konseveret er der noge specifikke motiver diversiteten

Chromatin, the complex of DNA wrapped around histone proteins and associated with various regulatory factors, plays a central role in the spatial and temporal control of gene expression, as well as in maintaining the structural integrity and stability of the genome. Far from being a static scaffold, chromatin is a highly dynamic entity, subject to extensive remodeling in response to cellular signals, developmental cues, and environmental stimuli. Its organization within the nucleus is hierarchical, ranging from nucleosomes to higher-order domains, and this architecture can differ substantially across cell types, developmental stages, and even between closely related species. One of the key features of chromatin organization is its division into functionally distinct states—broadly classified as euchromatin, which is generally open and transcriptionally active, and heterochromatin, which is compacted and repressive. Transitions between these states, whether developmentally programmed or evolutionarily emergent, are often tightly coupled to shifts in gene regulatory activity and cellular identity.

In this study, we focus on characterizing these chromatin state transitions, with a particular interest in understanding how they are distributed across the genome and whether they exhibit predictable patterns linked to specific sequence features or regulatory elements. We ask whether certain genomic regions are more prone to undergoing structural reorganization and whether such regions harbor conserved sequence motifs, epigenetic signatures, or binding sites for key regulatory proteins. Furthermore, we investigate whether particular chromatin configurations are disproportionately associated with genes involved in critical biological processes or with regions of the genome that show signatures of evolutionary constraint or adaptation.

By examining the diversity of chromatin states and transitions across individuals and lineages, we also aim to uncover broader principles governing chromatin architecture and its functional consequences. These insights may inform our understanding of how chromatin organization evolves, how it contributes to phenotypic diversity, and how its dysregulation may underlie disease states. Ultimately, this work contributes to a growing body of knowledge on the interplay between genome structure, function, and evolution.

2.3 A/B configurations

One widely studied aspect of chromatin architecture is its partitioning into A and B compartments, as revealed by Hi-C and other genome conformation capture techniques. These compartments represent large-scale topological domains of the genome that differ markedly in their regulatory environment and functional output. A compartments are typically associated with euchromatic regions—open, accessible chromatin that is rich in actively transcribed genes, regulatory elements like enhancers and promoters, and active histone marks such as H3K4me3 and H3K27ac. In contrast, B compartments correspond to heterochromatic, transcriptionally repressive regions characterized by reduced accessibility, fewer active genes, and enrichment of repressive epigenetic marks like H3K9me3 and DNA methylation.

This spatial segregation of the genome into A and B compartments plays a critical role in regulating which genes are expressed in a given cell or context. Genes located in A compartments are more likely to be in close physical proximity to other regulatory elements, transcriptional machinery, and nuclear compartments that promote gene activation, such as nuclear speckles. Conversely, genes embedded in B compartments are more isolated from such transcriptional hubs and may be sequestered near the nuclear lamina, where gene silencing is reinforced.

Importantly, the assignment of a genomic region to either an A or B compartment is not fixed; it can shift in response to developmental signals, environmental stress, or disease states. Such compartment switching is a powerful mechanism of gene regulation. For example, when a gene or regulatory locus moves from a B to an A compartment, it may become transcriptionally activated due to its new, more permissive chromatin environment. Likewise, genes that transition into B compartments often show downregulation or complete silencing. These dynamic shifts allow the cell to orchestrate complex gene expression programs with spatial precision, integrating structural and regulatory layers of genome function.

Moreover, comparative studies across species and populations suggest that A/B compartmentalization is both conserved and evolutionarily flexible. Certain compartments—particularly those associated with developmental and housekeeping genes—are stably maintained across evolutionary time, reflecting strong selective pressure to preserve core regulatory networks. Other regions, especially those involved in species-specific traits or environmental responses, exhibit greater compartmental plasticity. By studying the distribution and dynamics of A/B compartments, especially in non-model organisms and hybrid genomes, we can gain novel insights into how chromatin structure evolves and how it contributes to phenotypic diversity and adaptation.

2.4 GPN

Recent advances in artificial intelligence have brought powerful tools to genomics, particularly through the development of Genomic Pre-trained Networks (GPNs)—deep learning models trained on large-scale genomic sequence data using architectures inspired by natural language processing. GPNs apply self-supervised learning approaches, analogous to models like BERT or GPT, to the DNA sequence, learning to predict masked or missing bases in context. Through this training, GPNs develop an internal representation of the "language of the genome," capturing patterns of sequence composition, motif syntax, and long-range dependencies that underlie regulatory function.

A key advantage of GPNs lies in their use of transformer architectures, which differ significantly from traditional models like convolutional neural networks (CNNs). While CNNs are adept at recognizing short, fixed-length patterns—such as transcription factor binding motifs—they struggle with learning dependencies over long genomic distances, such as enhancer-promoter interactions or 3D chromatin looping. Transformers, in contrast, leverage self-attention mechanisms, which allow the model to compare and weigh relationships between all nucleotide positions in a sequence, regardless of distance. This makes them exceptionally well-suited for modeling the complex, multi-scale structure of regulatory DNA.

The utility of GPNs in functional genomics has been demonstrated in tasks such as predicting the effects of noncoding variants, chromatin accessibility, and transcription factor binding. For instance, the model introduced by Benegas and Batra ("DNA language models are powerful predictors of genome-wide variant effects") showed that transformer-based GPNs outperform prior models in predicting variant impact across a variety of cell types and epigenomic contexts, even without explicit annotations. These models can generate base-resolution functional predictions across the genome, enabling insight into the likely regulatory role of any given sequence segment.

In the context of chromatin structure, GPNs hold particular promise for identifying sequence features that influence or correlate with A/B compartment identity. By learning patterns that distinguish sequences in open (A) versus closed (B) chromatin, GPNs can be used to predict how shifts in compartmentalization may arise from sequence variation. Furthermore, these predictions can be extended across species or populations to assess functional conservation—that is, whether a sequence maintains regulatory potential or compartment association despite genetic divergence.

In this study, we leverage GPN-derived embeddings and predictions to investigate the sequence determinants of chromatin organization. We aim to identify genomic regions where the GPN model predicts strong, conserved regulatory activity, and determine whether these regions also show structural stability in chromatin compartments across individuals or hybrid genomes. This approach allows us to assess whether chromatin transitions are driven primarily by changes in DNA sequence, or whether other regulatory layers (e.g., epigenetic modifications, chromatin remodelers) are required to explain shifts between A and B compartments.

Ultimately, integrating GPN predictions with chromatin conformation data (e.g., from Hi-C) provides a powerful framework for exploring genome function at multiple scales—from base-level regulatory logic to large-scale nuclear organization. These models not only enhance our ability to interpret genomic variation and conservation but also open new avenues for studying the evolution of regulatory architecture in systems such as primates, where hybridization and population structure offer unique opportunities to dissect genome regulation in action.

2.5 cnn vs transformer

Convolutional Neural Networks (CNNs) have been the foundational architecture in computer vision for over a decade. Introduced by LeCun et al. (1998) and popularized with the success of AlexNet (Krizhevsky et al., 2012), CNNs exploit spatial hierarchies in image data through local receptive fields and shared weights. Architectures such as VGGNet (Simonyan & Zisserman, 2015), ResNet (He et al., 2016), and EfficientNet (Tan & Le, 2019) have progressively improved the depth, scalability, and performance of CNNs, enabling state-of-the-art results in tasks such as image classification, object detection, and semantic segmentation.

A simple Convolutional Neural Network (CNN) architecture typically consists of several key types of layers. The first are convolutional layers, which apply learnable filters (also known as kernels) to local regions of the input data. These filters are designed to capture spatial features such as edges, textures, and, as the network deepens, more complex patterns. Following the convolutional operations, activation functions are applied—most commonly the Rectified Linear Unit (ReLU)—to introduce non-linearity into the model, enabling it to learn more complex representations.

Pooling layers are another essential component of CNNs. These layers downsample the feature maps, typically using max pooling or average pooling, which reduces the spatial dimensions of the data while retaining the most salient features. This not only lowers computational requirements but also helps mitigate overfitting by providing a form of translation invariance. Toward the end of the architecture, fully connected layers are often used to perform high-level reasoning. These layers integrate the features extracted by earlier layers and are typically responsible for producing the final classification or regression outputs.

One of the main strengths of CNNs lies in their use of local receptive fields and weight sharing, which makes them highly effective at learning spatial hierarchies and translation-invariant features. Even relatively shallow CNNs can perform well on image-related tasks, particularly when training data is limited.

CNNs offer several advantages. They are parameter-efficient due to filter sharing, which significantly reduces the number of learnable weights compared to fully connected networks. Their strong inductive bias—emphasizing locality and translation invariance—makes them particularly well-suited to visual data. Additionally, CNNs benefit from computational efficiency, as they are compatible with highly optimized GPU-based implementations, and they often require less data than more recent architectures to achieve good performance.

However, CNNs also come with limitations. They tend to focus on local features, which makes capturing long-range dependencies challenging without additional architectural mechanisms. Their receptive field is fixed and limited, so incorporating larger contextual information often requires making the network deeper or using special techniques like dilated convolutions or skip connections. Furthermore, CNN performance can be sensitive to architectural choices such as filter sizes, layer depth, and stride configurations, which often require manual tuning.

These challenges have inspired researchers to explore alternative architectures, most notably Transformer models. Unlike CNNs, Transformers are designed to model global relationships more naturally, though often at the expense of requiring more data and computational resources.

3 Materials and Methods

Data Sources: Baboon genome datasets, known hybridization cases, population size data.

Model and Tools: GPN model architecture and usage.

Preprocessing: Alignment, feature extraction, and data cleaning.

Entropy Calculation: Methodology for genome-wide entropy measurement.

3.1 GPN arcitecture

The idea for this project was to evaluate the Genome-Pretrained Network (GPN) introduced in **?** and determine whether it could achieve greater accuracy than traditional methods in predicting genome-wide variant effects.

The model is designed as a convolutional neural network (CNN) and takes input sequences with a window size of 512. During training, 15% of the positions within each window are masked to enable the model to learn meaningful representations.

The architecture consists of 25 layers, each structured as follows: a dilated convolution layer, followed by an add-and-norm layer with a skip connection from before the dilated convolution. This is followed by a feedforward layer, another add-and-norm layer, and additional skip connections.

A feedforward layer is a fundamental component of neural networks, where inputs pass through one or more fully connected layers with activation functions, transforming data without looping back. This structure helps the model learn complex representations by applying weighted transformations and non-linearities.

A dilated convolution expands the receptive field of a convolutional layer without increasing the number of parameters or reducing resolution. By spacing kernel elements apart, it captures long-range dependencies in sequences, making it particularly useful in genomic data analysis. When combined, dilated convolutions and feedforward layers enhance a model's ability to recognize patterns across different scales efficiently.

After passing through the 25 layers, the model produces a contextual embedding with a dimension of 512 (D=512), followed by classification layers. The final layer outputs the probabilities of the four nucleotides at each masked position.

The GPN variant effect prediction score is calculated as the log-likelihood ratio between the alternate (ALT) and reference (REF) alleles. Here, L represents the window length in base pairs, and D denotes the embedding dimension.

3.2 Data Preparation and Pretraining Workflow

To evaluate the Genome-Pretrained Network (GPN), the initial step involved reproducing the pretraining workflow provided in the original publication. This workflow outlines the procedure for generating a dataset compatible with the model, specifically for masked language modeling over genomic sequences.

Initial attempts to apply the workflow to a baboon genome dataset encountered technical difficulties. These issues were not intrinsic to the baboon data itself but stemmed from the implementation of the workflow. To ensure compatibility with the model and maintain consistency with the original study, the Anatoptis genome—used as the reference in the published work—was selected as a pilot dataset.

After resolving the implementation issues, a dataset suitable for GPN pretraining was successfully generated. The resulting data consisted of masked genomic sequences, formatted according to the model's input requirements.

3.3 Model Execution Workflow and Hardware Constraints

Following the successful generation of a pretraining-compatible dataset, the next step involved attempting to run the Genome-Pretrained Network (GPN) using the available inference and training workflows. The documentation provided on the model's GitHub repository lacked detailed guidance, making it challenging to understand the complete execution process and the rationale behind certain steps.

To proceed, an alternative workflow was identified—one specifically designed for use with the Anatoptis genome. This workflow was adopted as a reference implementation for executing the GPN model.

However, significant hardware limitations emerged during this phase. The original study reported using four NVIDIA A100 GPUs, each equipped with 80 GB of RAM. In contrast, the available system utilized NVIDIA L43R GPUs with 45 GB of RAM, resulting in substantial differences in computational capacity.

As a result, attempts to execute the GPN model encountered frequent out-of-memory (OOM) errors, particularly during forward passes through the deeper layers of the network. These memory constraints posed a major obstacle and significantly hindered reproducibility under the available hardware setup.

Initial efforts to train or fine-tune the Genome-Pretrained Network (GPN) focused on applying the workflow to a baboon genome using a multi-GPU setup. The goal was to replicate the original training environment, which employed four NVIDIA A100 GPUs (80 GB RAM each). However, the available infrastructure utilized NVIDIA L43R GPUs, each with 45 GB of RAM, which introduced several compatibility and performance challenges.

Early testing began with the Anatoptis genome. Attempts to process the entire genome using the Snakemake workflow were unsuccessful due to GPU detection issues. These issues stemmed from the cluster configuration, which prevented the Snakemake environment from correctly interfacing with the available GPU hardware.

After resolving these initial configuration issues, the workflow was executed with GPU support. While this improved performance, the model began encountering memory errors after a few hours of training, even when utilizing two to four GPUs in parallel. In an effort to overcome these constraints, the number of GPUs was incrementally increased—from four to six, and finally to eight—providing an aggregate of 360 GB of GPU RAM.

Despite these efforts, the model continued to encounter out-of-memory (OOM) errors during execution. Even with eight GPUs, the memory demands of the GPN model remained prohibitive under the given cluster configuration. These persistent failures highlighted a significant limitation: the model's architecture and resource demands exceeded the practical limits of the available hardware.

Ultimately, these constraints prevented successful execution of the GPN model on the tested system.

Given the persistent memory limitations encountered during full-genome processing, an alternative strategy was implemented. Rather than inputting the entire Anatoptis genome, the data was partitioned at the chromosome level. This allowed individual chromosomes to be processed independently, while still allocating the full available GPU resources (eight NVIDIA GPUs with a combined 360 GB of RAM) to each run.

This approach proved successful. When processing a single chromosome at a time, the GPN model was able to execute without encountering memory-related failures. As a result, contextual embeddings were successfully generated for individual chromosomes of the Anatoptis genome.

However, due to time constraints near the conclusion of the project timeline, the extracted embeddings were not subjected to further downstream analysis. Although they were successfully retrieved and stored, no additional modeling, interpretation, or evaluation steps were completed within the project scope.

3.4 using HIC to extract transtion areas

3.5 distance between the edges

To evaluate the availability of data across genomic distances and mitigate potential bias in downstream analyses, we calculated reverse cumulative counts of compartment edge half-distances. First, edges were identified at compartment transitions between A and B (in either direction). For each consecutive pair of such edges, we measured the genomic distance between their start positions and divided it by two, yielding the half-distance between edges.

We then binned these half-distances and computed the reverse cumulative count for each bin, which represents the number of edges with half-distances greater than or equal to that bin. This approach effectively measures the number of edges capable of contributing data at a given spatial scale. Using this metric allows us to normalize other measurements by the available data density, reducing the risk that apparent trends are driven by varying numbers of contributing edges at different distances rather than true biological effects. Counts were calculated up to 2 Mb to encompass the maximum observed separation.

3.6 compare GC conent to transition in macaca

We compared GC content between A and B chromatin compartments for chromosome 8 and chromosome X across five Macaca cell types: fibroblasts, sperm, round spermatids, pachytene spermatocytes, and spermatogonia. Chromatin compartments were defined from A–B and B–A structural transitions, and GC content was measured separately for each compartment.

For chromosome 8, the A compartment consistently showed higher GC content than the B compartment in all cell types, which aligns with expected genomic organization. In chromosome X, however, the differences between compartments were smaller. In round spermatids, GC content was nearly identical between A and B compartments, while in pachytene spermatocytes, spermatogonia, and sperm, A and B compartments had closely overlapping GC values. Interestingly, in fibroblasts the pattern was reversed, with the B compartment displaying higher GC content than the A compartment see Figure 3.2.

3.6.1 CLAUDE NOTES

3.7 X chromosome compartments and GC content: An unexplored frontier

The well-established pattern of higher GC content in A compartments versus B compartments on autosomes has *not been quantitatively investigated for the X chromosome*, despite extensive

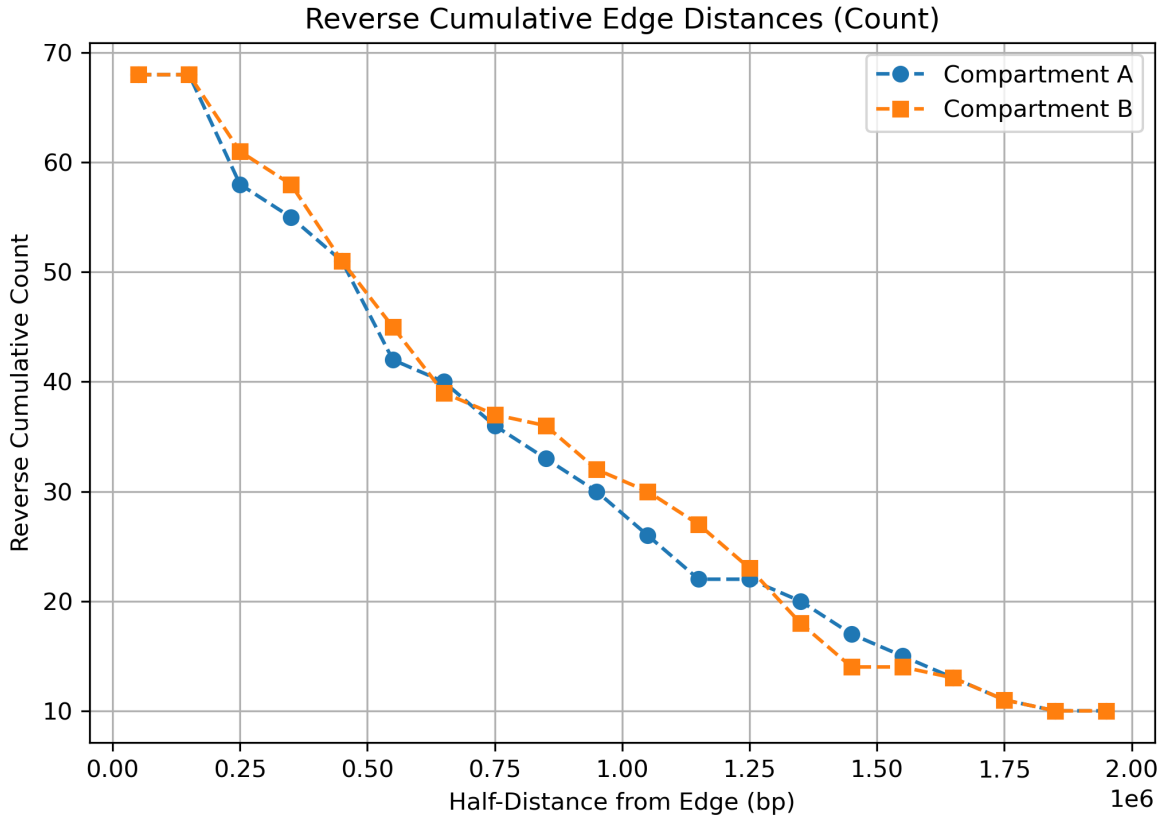

Figure 3.1: Reverse cumulative counts of compartment edge half-distances. Distances were calculated by first identifying sequential transitions between compartments A and B (or B and A). For each pair of transitions, the genomic distance between their start positions was measured and halved, producing the "half-distance" metric. The reverse cumulative count at a given bin shows the number of edges with half-distances greater than or equal to that bin's center, representing the number of edges that could contribute data to measurements at that spatial scale. Extending the plot to 2 Mb captures the full range of half-distances observed.

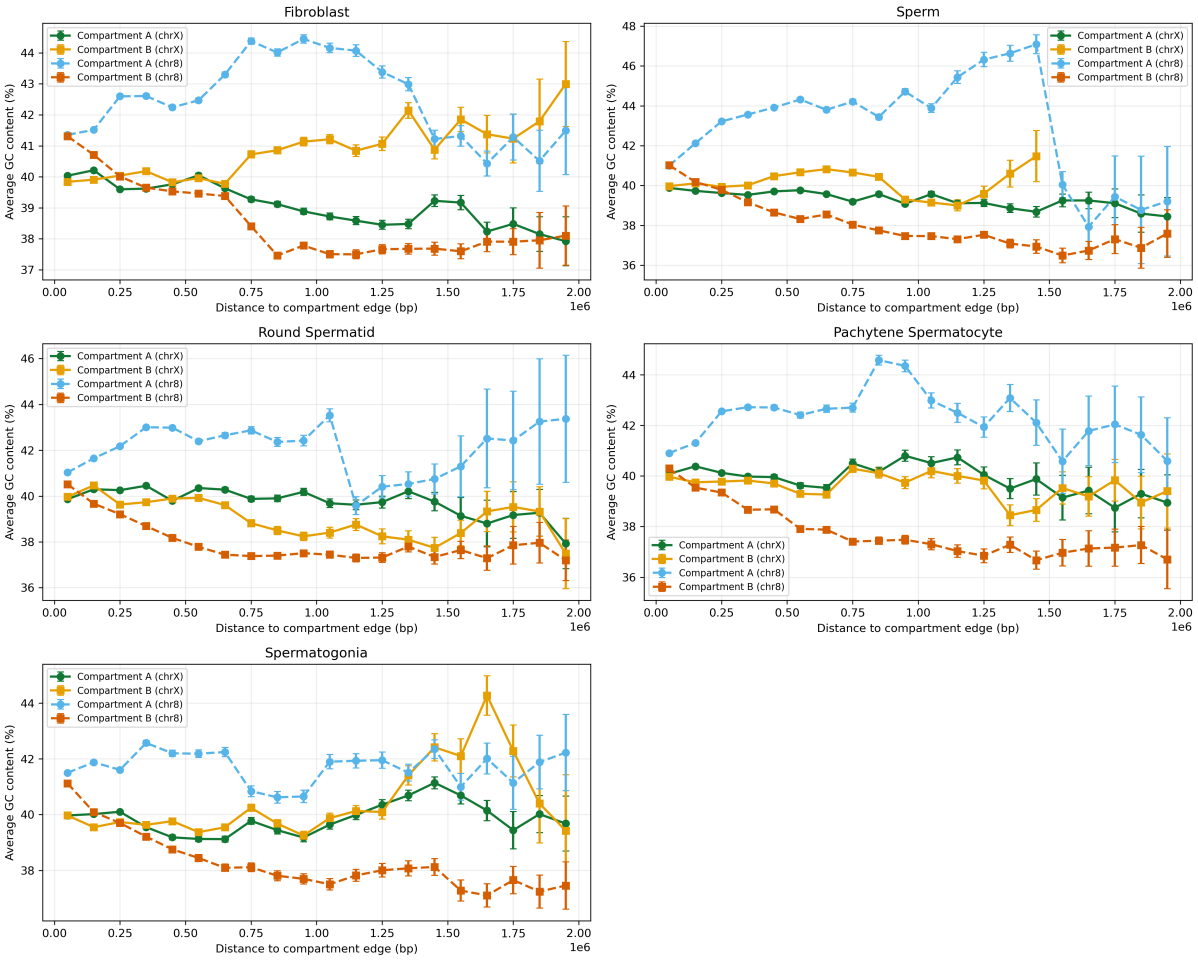

Figure 3.2: Compartment analysis for chromosomes X and 8 in five Macaca cell types—fibroblasts, sperm, round spermatids, pachytene spermatocytes, and spermatogonia. GC content was remapped to genomic locations where chromatin structure transitions between compartments A and B. The data were separated so that each compartment contains only A or B structure. For each genomic bin, the plot shows the mean GC content, with error bars representing the standard deviation of the mean, adjusted by the effective edge contribution (accounting for how many edges contribute at longer distances).For chromosome 8, the A compartment exhibits higher GC content than the B compartment. In tissues involved in spermatogenesis, chromosome X shows similar GC content in both A and B compartments. In fibroblasts, GC content is higher in the B compartment than in the A compartment.

research on X chromosome compartmentalization. This represents a significant and surprising gap in our understanding of 3D genome organization, particularly given recent discoveries about the unique compartmental architecture of the X chromosome. ## Studies confirm extensive X chromosome compartment research Research on X chromosome compartmentalization has advanced dramatically, particularly with the 2021 discovery that even the inactive X chromosome contains hidden A/B-like compartments. Studies have mapped approximately *75 A/B-like compartments on the inactive X* and *90 compartments on the active X* in neural precursor cells, revealing a complex organizational structure previously thought to be absent. The original compartment discovery by Lieberman-Aiden and colleagues in 2009 established that A compartments are gene-rich, GC-rich, and transcriptionally active, while B compartments are gene-poor, AT-rich, and heterochromatic. This principle has been consistently validated across autosomal chromosomes, with studies showing strong correlations ($r = 0.80$) between compartment identity and GC content. ## The missing comparative analysis Despite extensive searching through major scientific databases and journals, *no studies provide quantitative GC content percentages for A versus B compartments specifically on the X chromosome*, nor do any offer comparative statistics between X chromosomal and autosomal compartment GC content. A 2017 structural modeling study confirmed that A and B compartments "are found to be associated with GC rich and AT rich sequences, respectively," but provided no X chromosome-specific data. This gap is particularly striking given that researchers have characterized numerous other aspects of X chromosome compartmentalization, including epigenetic signatures, replication timing, and structural features. The absence of this fundamental comparison likely stems from several factors. X chromosome analysis presents unique technical challenges due to X-inactivation in females and hemizygosity in males, leading many genome-wide studies to exclude sex chromosomes from their analyses. Additionally, the discovery that inactive X chromosomes possess compartmental organization is very recent (2021), suggesting this area requires further investigation. ## X chromosome compartments display unique organizational features The X chromosome exhibits several distinctive compartmental characteristics that could affect GC content patterns differently than autosomes. During X-inactivation, standard A/B compartments undergo remodeling into specialized *S1/S2 compartments*, where S1 compartments are Xist-rich and correspond to gene-dense regions, while S2 compartments are Xist-poor and correspond to gene-poor regions. These then merge to create the characteristic "compartment-less" superstructure of the inactive X. The inactive X's A-like compartments differ fundamentally from classical A compartments – they're enriched in Xist RNA and H3K27me3 repressive marks rather than the active transcription marks (H3K4me3) typical of autosomal A compartments. Recent advances using micro-C and single-cell Hi-C technologies have revealed that the X chromosome maintains distinct *bipartite megadomain structures* separated by the Dxz4 boundary element, a configuration not seen in autosomes. These megadomains form stepwise during development through cohesin-dependent mechanisms, creating a unique architectural framework that could influence local GC content evolution patterns. Studies in 2024 demonstrated that X-megadomains appear transiently in extraembryonic lineages before establishing the final inactive X structure, representing a developmental trajectory unique to sex chromosomes. ## X-inactivation creates complex relationships with GC content The relationship between X-inactivation and compartment organization adds layers of complexity to potential GC content patterns. *Escape genes*, which

maintain expression from the inactive X, preferentially localize to specific subcompartments and can comprise up to 30% of genes in escape-rich clusters. These genes tend to reside in regions that retain some characteristics of A-type compartments, remaining at the periphery of or outside the Xist RNA domain. SINE elements are enriched near promoters of escape genes, potentially facilitating their resistance to inactivation and creating local environments with distinct sequence composition. The mechanistic pathway of X-inactivation – involving Xist recruitment of HNRNPK, PRC1, and SMCHD1 – drives compartment formation through protein self-association, likely via liquid-liquid phase separation. This process creates spatial segregation between different types of heterochromatin on the inactive X, with S1 compartments enriched in Polycomb marks and S2 compartments enriched in H3K9me2/3-associated proteins. This unique heterochromatin organization could create GC content distributions that differ from the standard active/inactive dichotomy seen in autosomal compartments. ## Comparative analyses reveal broader evolutionary patterns While direct X chromosome compartment GC content data remains absent, comparative genomic studies provide relevant context. Research has demonstrated that *X chromosomes show unique recombination suppression patterns* that affect GC-biased gene conversion differently than autosomes. The reduced recombination on sex chromosomes influences the efficacy of selection on GC content evolution, potentially creating different equilibrium GC levels in compartments. Analysis across placental mammals revealed that genomic architecture constrains X chromosome evolution, with unique 3D folding patterns conserved across species despite extensive sequence divergence. The X chromosome maintains distinct superloop formations involving escapee loci, which may influence local GC content patterns through altered mutational processes or selection pressures. Studies have shown that active domains on the X localize to territory boundaries where GC content patterns may be influenced by transcriptional accessibility and DNA repair mechanisms that differ from those in autosomal compartments. ## Conclusion The lack of quantitative data comparing GC content between X chromosomal and autosomal A/B compartments represents a significant gap in our understanding of 3D genome organization. While the general principle that A compartments are GC-rich and B compartments are AT-rich is well-established for autosomes, *whether this pattern holds equally on the X chromosome remains undemonstrated*. Given the X chromosome's unique architectural features – including specialized compartment types during inactivation, escape gene clustering, and distinctive evolutionary constraints – this comparison would provide crucial insights into the fundamental relationship between chromosome structure, sequence composition, and gene regulation. This unexplored frontier presents an important opportunity for future research to quantitatively characterize how sex chromosome compartmentalization relates to the GC content patterns so well documented in autosomes.

3.8 CPG islands

CpG islands (CGIs) are typically unmethylated DNA regions that are closely associated with gene promoters and play a crucial role in regulating chromatin accessibility and three-dimensional genome organization. Accumulating evidence demonstrates that changes in CGI methylation, histone modifications, or chromatin-binding proteins can drive or accompany transitions between

chromatin compartments, shifting from active A compartments to repressive B compartments or vice versa. Such transitions are frequently observed in developmental processes, cancer progression, X-chromosome inactivation, and transcriptional regulation. In many cases, CGIs act as anchors or boundaries that help define chromatin domains, and their epigenetic state can strongly influence large-scale compartmental reorganization.

Recent research on meiotic chromatin architecture provides insights that support this broader view of CGI function. A study by Wang *et al.* (2019) examined spermatogenesis in rhesus monkeys and mice using low-input Hi-C and found that canonical topologically associating domains (TADs) dissolve during the pachytene stage of meiosis, even while transcription continues. This observation suggests that TAD integrity is not strictly required for gene expression at this stage. Instead of the standard large-scale A/B compartmentalization, the study identified the emergence of smaller, refined A/B compartments that alternate between transcriptionally active and inactive domains. The formation of these fine-scale compartments was shown to depend on the synaptonemal complex (SC). In SC-defective mutants, canonical TADs reappeared, indicating that the SC suppresses TADs while enabling this unique compartmental structure. Notably, this reorganization of 3D chromatin architecture was conserved between monkeys and mice, suggesting a mammalian-wide regulatory principle (Wang *et al.*, 2019).

Although CpG islands were not the central focus of this study, the findings are consistent with the hypothesis that CGIs contribute to compartmental organization by functioning as boundary or anchoring elements. Because refined compartment transitions align with transcriptional activity, and CGIs are commonly associated with active transcription start sites, it is likely that they play a role in stabilizing the boundaries between alternating compartments. In this way, CGIs may provide local chromatin insulation while also influencing higher-order genome organization.

we investigatet how the CPG islands where locakedm compared to the compartment transition. this was done by remapping the compartment trasition to CPG islands location on the genome. the outgoing data where then filteret acording to the filter below

```
A_val = result[
    ((result['comp'] == 'A') & (result['start'] < 0)) |
    ((result['comp'] == 'B') & (result['start'] > 0))
].copy()
B_val = result[
    ((result['comp'] == 'A') & (result['start'] > 0)) |
    ((result['comp'] == 'B') & (result['start'] < 0))
].copy()
```

In this section, we restricted the analysis to chromosome X (chrX). We separated the data into A and B compartments and computed the absolute distance of each CpG island to the nearest compartment edge. The resulting distributions are shown in Figure Figure 4.2. The figure contains five panels—fibroblast, spermatogonia, pachytene spermatocyte, round spermatid, and

sperm. In fibroblast, compartments A and B are similar up to ~0.8–1.0 Mb, after which the B compartment shows higher CpG-island counts. In spermatogonia and pachytene spermatocyte, A and B remain similar across distances. In round spermatid, we observe a divergence between compartments with higher counts between ~0.3 and 1.2 Mb. In sperm, A and B are similar up to ~1 Mb; beyond this distance there is little to no additional signal in the B compartment.

Applying the same analysis to chromosome 8 (Figure 4.5), fibroblast shows elevated CpG-island counts in compartment A from ~0.6 to 1.4 Mb. In spermatogonia, A and B remain similar. In pachytene spermatocyte and round spermatid, counts are higher in compartment A from ~0 to 1.0 Mb compared with compartment B. The same pattern is observed in sperm, with substantially higher counts in compartment A than in compartment B.

#fig-CPG_normalized_chrx

4

#fig-CPG_normalized_chr_ab

4.1 notes for cpg islands

CpG islands, which are typically unmethylated and associated with gene promoters, play a pivotal role in chromatin accessibility and 3D genome organization. Multiple studies have demonstrated that chromatin compartment transitions (A to B or vice versa) can be driven or accompanied by changes in CpG island methylation, histone modifications, or chromatin-binding proteins. These changes are often observed in development, cancer progression, X-chromosome inactivation, and transcriptional regulation. CGIs frequently serve as anchors or boundaries for chromatin domains and their epigenetic status can correlate with large-scale compartment reorganization, particularly between transcriptionally active A compartments and repressive B compartments.

4.1.1 artikler

Reprogramming of Meiotic Chromatin Architecture during Spermatogenesis (6) Key Findings Summary This paper investigates how 3D chromatin architecture is reorganized during spermatogenesis—specifically in rhesus monkey and mouse models—using low-input Hi-C (a chromosome conformation capture method).

Major Contributions: Loss and Re-establishment of TADs (Topologically Associating Domains):

During meiosis, particularly in pachytene spermatocytes, most canonical TADs dissolve even though transcription continues. This suggests that TAD structure is not strictly required for active gene expression at this stage.

Emergence of Refined Local A/B Compartments:

- Instead of standard large-scale A/B compartmentalization, fine-grained compartmental structures emerge. These “refined-A/B compartments” alternate between transcriptionally active and inactive domains.
- hese compartments are smaller, and their emergence implies a compartmental organization decoupled from transcription activity.-

Synaptonemal Complex Dependency:

- These refined compartments require a functional synaptonemal complex (SC). In SC-defective mutants, canonical TADs are restored, suggesting that the SC actively suppresses TADs while enabling fine-scale compartmentalization.
- This is a novel role for the SC in 3D genome reorganization, beyond its known mechanical role in chromosome pairing and crossover formation.

Conservation Across Species:

- The unique meiotic chromatin organization found in rhesus monkeys is conserved in mice, suggesting this is a mammalian-wide regulatory principle during gametogenesis.

4.1.1.0.1 Implication for CpG Islands:

While CpG islands (CGIs) are not the primary focus of this paper, the study supports the broader hypothesis that CGIs may serve as boundary or anchoring elements in fine-scale compartment formation during meiosis:

The refined-A/B structure shows domain transitions that correlate with transcription activity, a feature often influenced by CGI presence at promoter regions.

Since CGIs are associated with active transcription start sites, they likely anchor boundaries between alternating compartments, effectively contributing to local chromatin insulation or regulation.

(**article?**){Wang2019, year = {2019}, title = {{Reprogramming of Meiotic Chromatin Architecture during Spermatogenesis}}, author = {Wang, Yu and Wang, Hui and Zhang, Yixiao and Du, Zheng and Si, Wei and Fan, Shuhan and Qin, Dong and Liu, Yuliang and Liu, Yuhan and Xu, Shiyuan and others}, journal = {Molecular Cell}, issn = {1097-4164}, doi = {10.1016/j.molcel.2018.11.019}, abstract = {{Chromatin organization undergoes drastic reconfiguration during gametogenesis. However, the molecular reprogramming of three-dimensional chromatin structure in this process remains poorly understood for mammals, including primates. Here, we examined three-dimensional chromatin architecture during spermatogenesis in rhesus monkey using low-input Hi-C. Interestingly, we found that topologically associating domains (TADs) undergo dissolution and reestablishment in spermatogenesis. Strikingly, pachytene spermatocytes, where synapsis occurs, are strongly depleted for TADs despite their active transcription state but uniquely show highly refined local compartments that alternate between transcribing and non-transcribing regions (refined-A/B). Importantly, such chromatin organization is conserved in mouse, where it remains largely intact upon transcription inhibition. Instead, it is attenuated in mutant spermatocytes, where the synaptonemal complex failed to be established. Intriguingly, this is accompanied by the restoration of TADs, suggesting that the synaptonemal complex may restrict TADs and promote local compartments. Thus, these data revealed extensive reprogramming of higher-order meiotic chromatin architecture during mammalian gametogenesis.}}, volume = {73}, number = {3}, pages = {547–561.e6}, pmid =

{30686580}, local-url = {https://www.cell.com/molecular-cell/fulltext/S1097-2765(18)30989-4}
}

4.1.1.0.2 maybe relevant articles:

Reconstructing A/B compartments as revealed by Hi-C using long-range correlations in epigenetic data (3)

Smchd1-Dependent and -Independent Pathways Determine Developmental Dynamics of CpG Island Methylation on the Inactive X Chromosome (4)

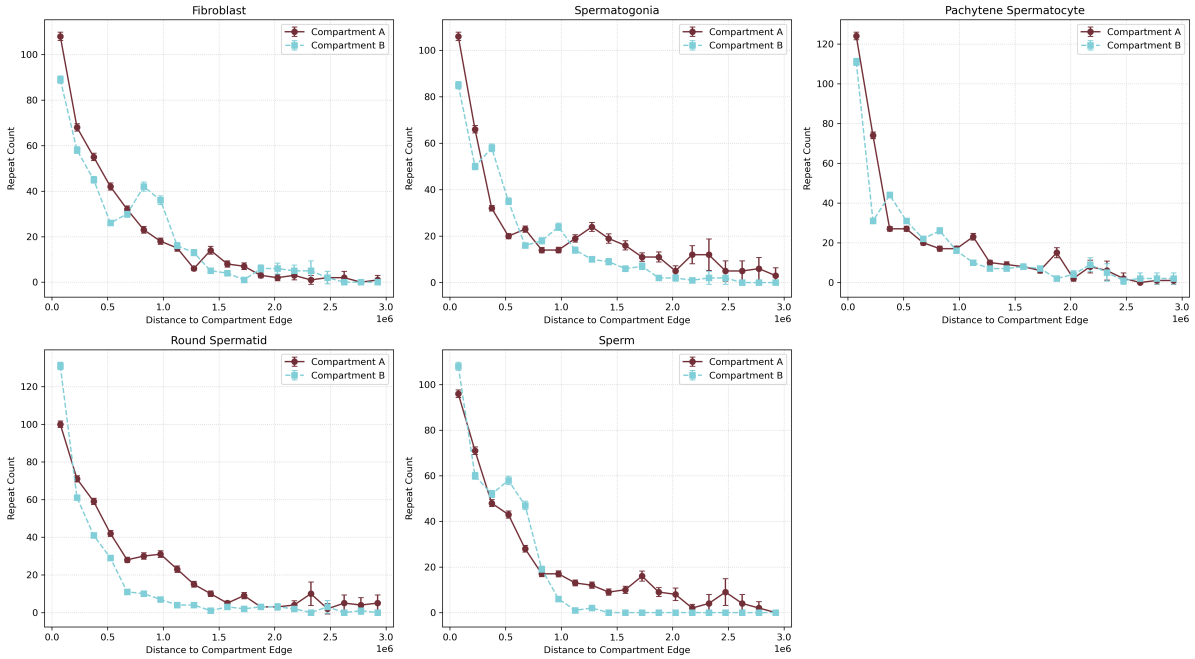

Figure 4.1: chrX

4.2 chr 8

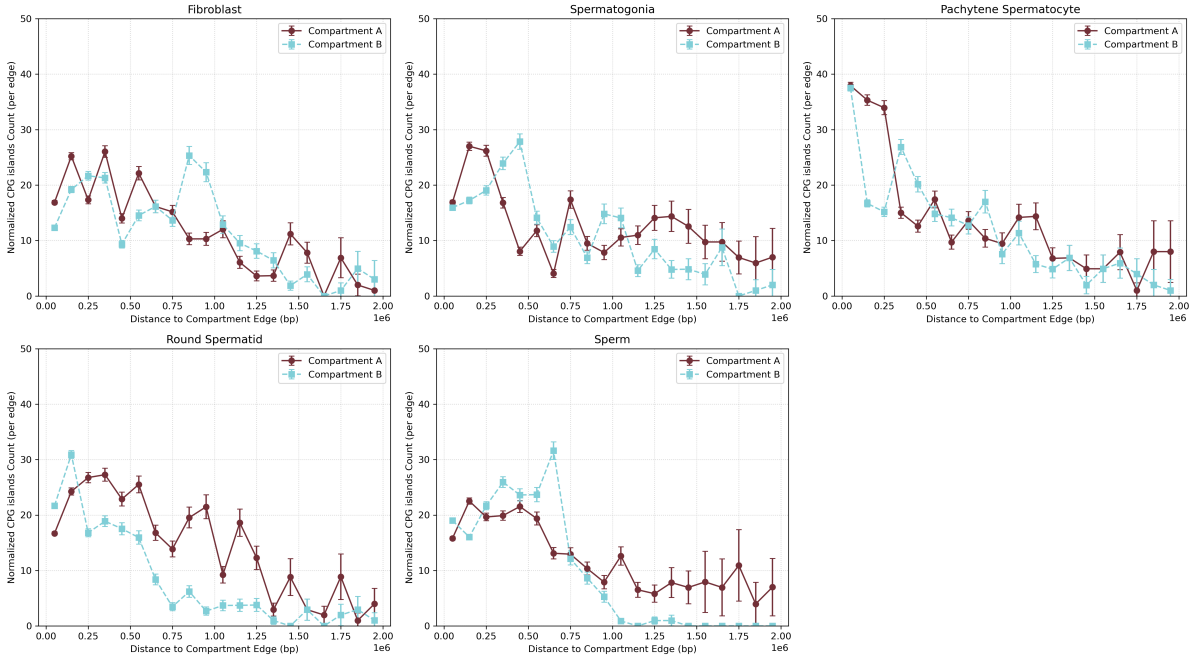

Figure 4.2: chrX

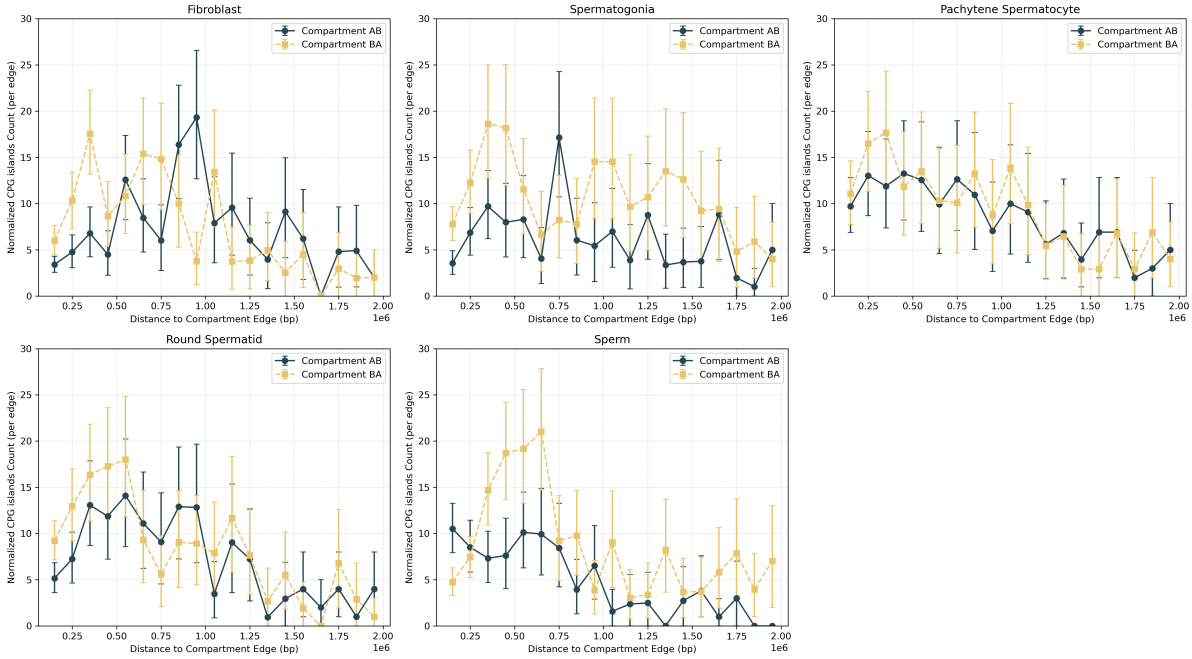

Figure 4.3: chrX

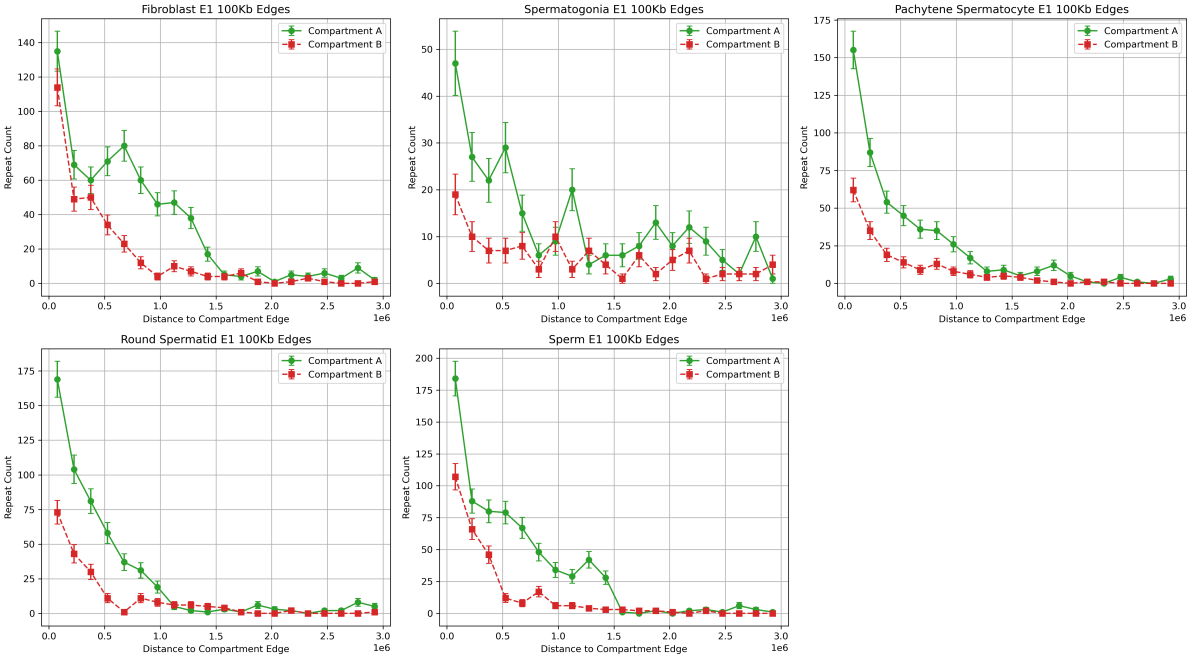

Figure 4.4: chr8

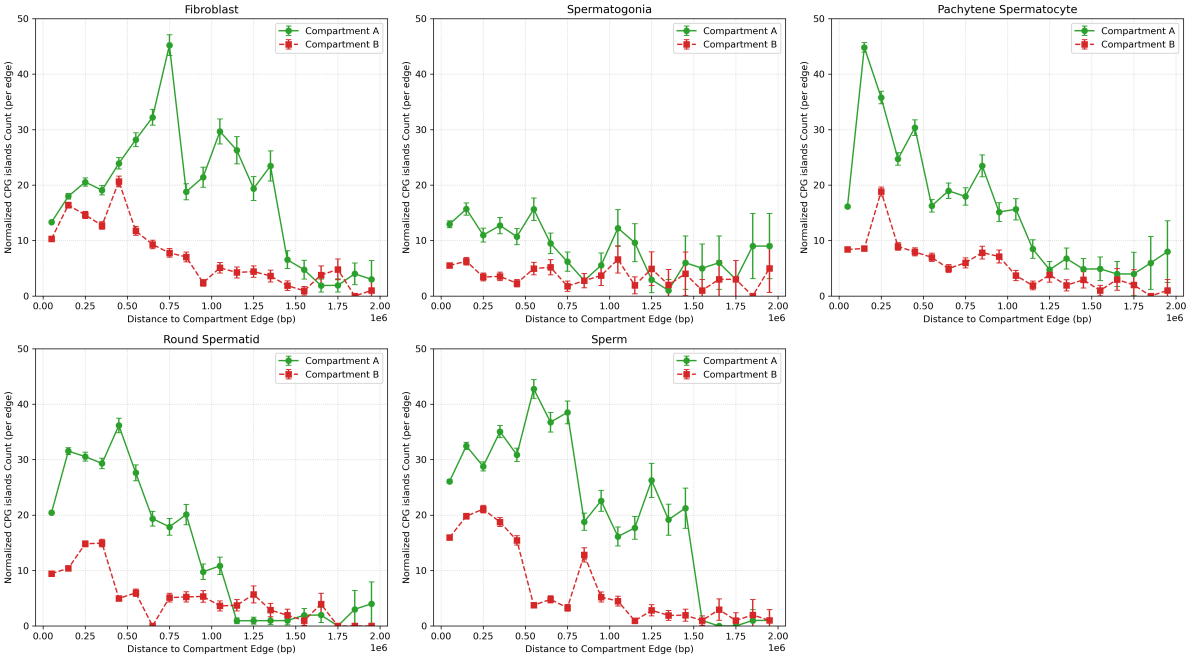

Figure 4.5: chr8

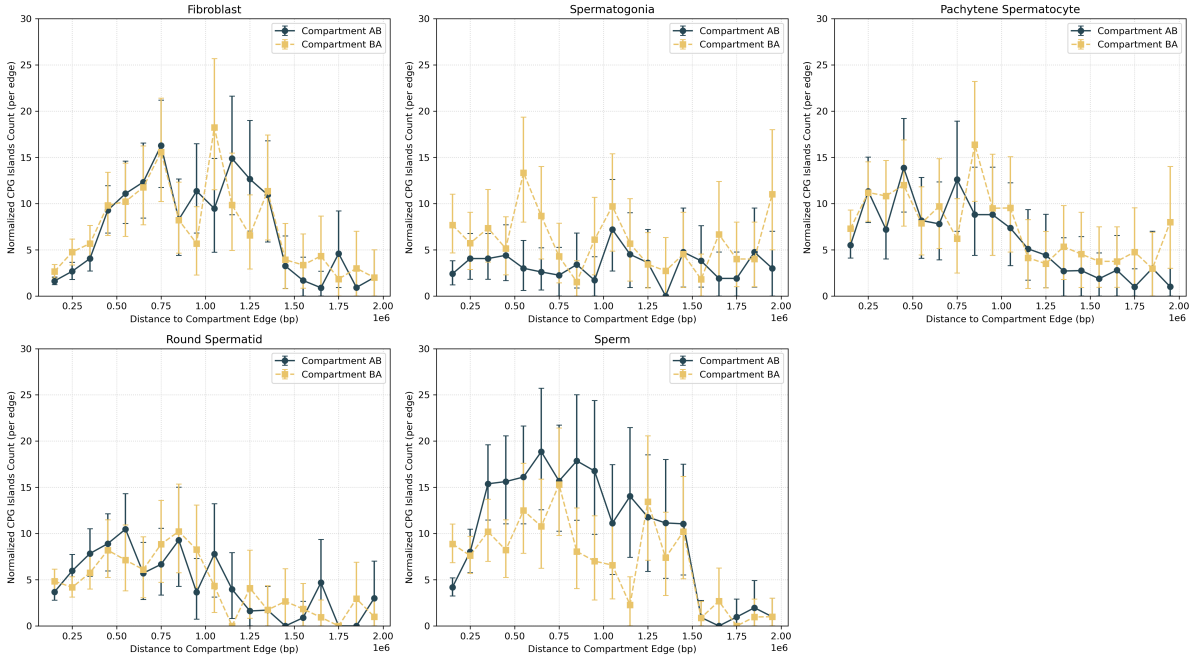

Figure 4.6: chr8

5 repeat

6 notes

7 cromatin structure and regulatiopn of genes

2. Regulation of Gene Expression The chromatin state influences gene accessibility:

Open chromatin is associated with active gene transcription.

Closed chromatin represses gene activity. Repetitive elements near regulatory genes can shift chromatin structure, acting as silencers or enhancers depending on the epigenetic landscape (Venkatesh & Workman, 2015).

Venkatesh & Workman (2015) – Histone exchange, chromatin structure and transcription regulation Nature Reviews Molecular Cell Biology Highlights the plasticity of chromatin around repeats and its impact on transcription.

8 repeats sorunding edges

1. Genome Stability and Repression of Transposable Elements Repetitive sequences, such as LINEs, SINEs, and satellite DNA, are often silenced via heterochromatin formation. This closed chromatin state is essential to suppress transposable elements, which can otherwise mobilize and disrupt genome integrity (Moazed, 2011).

Makova & Hardison (2015) – The effects of chromatin organization on variation in mutation rates Nature Reviews Genetics Discusses how heterochromatin protects repeats but also contributes to mutational heterogeneity.

4. Development and Cell Differentiation During development, regions with repetitive DNA undergo reconfiguration. The switch from euchromatin to heterochromatin is essential for cell lineage specification, and misregulation can cause developmental disorders (Gelato & Fischle, 2008).

Ghosh & Woodcock (2010) – Chromatin higher-order structure and dynamics Cold Spring Harbor Perspectives Explains how chromatin compaction is governed by nucleosome positioning and repeat regions.

8.0.1 Three-dimensional genome reorganization during mouse spermatogenesis

- Highlights how A/B compartment transitions coincide with major cell state transitions during spermatogenesis.
- Found TF motifs and simple repeats enriched in switching regions.
- Sheds light on regulatory elements linked to chromatin phase switching.

Core Findings This study investigates how the 3D organization of the genome changes across seven sequential stages of mouse spermatogenesis, using Hi-C technology. The work specifically focuses on topologically associating domains (TADs), chromatin loops, and A/B compartments, revealing highly dynamic reorganization events.

1. TAD and Loop Dynamics Present in Type A spermatogonia (early stage).

Disappear at pachytene stage (meiosis I prophase).

Reappear in mature spermatozoa.

This pattern indicates that TADs are not static during development and may not be essential for transcription at all stages.

2. CTCF Binding CTCF remains bound at TAD boundary regions even when TADs are absent (e.g., in pachytene spermatocytes).

This suggests CTCF binding is not sufficient by itself to maintain 3D structure.

3. Enhancers, Promoters, and Active Transcription Despite the absence of TADs at pachytene, enhancers and promoters retain open chromatin.

Transcription continues, implying that gene expression can occur independently of higher-order chromatin loops.

4. A/B Compartment Changes A/B compartments are largely conserved on autosomes throughout spermatogenesis.

However, specific A -> B switching events are correlated with changes in gene activity.

The X chromosome loses its A/B compartment structure during stages like pachytene spermatocytes (pacSC), round spermatids (rST), and elongating spermatids (eST).

Repeats and Chromatin Compartment Transitions TF motif enrichment was analyzed in differentially active chromatin regions, which also included repetitive DNA content.

During A/B compartment switching, particularly in regions transitioning from B (closed) to A (open), some simple and low-complexity repeats were enriched, suggesting a role in:

Regulatory rewiring during cell fate transition.

Chromatin accessibility and transcription activation.

The X chromosome, which undergoes massive reorganization and loss of A/B compartment structure during specific stages (pacSC, rST, eST), may exhibit a distinct repeat signature, though this was only briefly hinted at in the paper.

(**article?**){Luo2019, year = {2019}, title = {{Three-dimensional genome reorganization during mouse spermatogenesis}}, author = {Luo, Zhen and Wang, Xin and Wang, Rui and Chen, Jie and Chen, Yan and Xu, Qing and Cao, Jun}, journal = {bioRxiv}, doi = {10.1101/585281}, abstract = {{Dynamic changes in chromatin architecture accompany spermatogenesis. Using Hi-C and ATAC-seq, this study reveals A/B compartment transitions and transcription factor motif enrichments that align with germ cell stage-specific chromatin changes.}}, pages = {}, number = {}, volume = {}, pmid = {}, pmcid = {}, issn = {}, keywords = {}, local-url = {https://www.biorxiv.org/content/10.1101/585281v1} }

8 repeats sorunding edges

8.0.2 Single-cell long-read Hi-C, scNanoHi-C2

- Uses single-cell Hi-C to resolve chromatin changes in early germ cells.
- Finds TE subfamilies enriched differently in A/B compartments, suggesting a regulatory function of repeat elements.

Key Findings Summary Development of scNanoHi-C2

The authors introduce scNanoHi-C2, a novel single-cell long-read Hi-C technique.

It overcomes limitations of standard Hi-C by capturing long-range, multi-contact chromatin interactions in individual cells.

3D Genome Dynamics in Germ Cells

Applied to embryonic germline cells, the technique reveals:

Stage-specific topological reorganization of the genome.

Highly dynamic A/B compartment transitions, even at early embryonic stages.

Repeat Element Profiling

The study shows distinct enrichment patterns of transposable elements (TEs) in A and B compartments.

LINEs and SINEs were differentially represented, correlating with developmental chromatin remodeling.

These repeats may contribute to insulation, compartmental identity, or nuclear architecture.

Chromatin Rewiring during Primordial Germ Cell (PGC) Development

Early embryonic germ cells exhibit weaker TAD boundaries and less stable compartments compared to somatic cells.

As differentiation progresses, compartment strength increases, indicating a progressive maturation of 3D genome organization.

X Chromosome Inactivation Patterns

The paper provides high-resolution views of X chromosome organization, noting changes in compartmental asymmetry and pairing during early germ cell development.

(**article?**){Lu2025, year = {2025}, title = {{Single-cell long-read Hi-C, scNanoHi-C2, details 3D genome reorganization in embryonic-stage germ cells}}, author = {Lu, Jiawei and Li, Wen and Tang, Fuchou}, journal = {Nature Structural & Molecular Biology}, doi = {10.1038/s41594-025-01604-7}, issn = {1545-9985}, abstract = {{This study applies scNanoHi-C2 to reveal high-resolution single-cell 3D chromatin structure changes during germline development. Transposable elements and repeat families were differentially enriched in A/B compartments, impacting developmental regulation.}}, pages = {}, number = {}, volume = {}, pmid = {}, pmcid = {}, keywords = {}, local-url = {https://www.nature.com/articles/s41594-025-01604-7} }

8.0.3 3D genome remodeling and homologous pairing during meiotic prophase

- Simple and LINE repeats enriched in B compartments, involved in meiotic structural integrity.

Key Findings Summary This study integrates Hi-C chromatin conformation capture with super-resolution microscopy to uncover how 3D genome architecture is remodeled during meiotic prophase I in both male and female germlines (spermatogenesis and oogenesis) in mice.

Homologous Chromosome Pairing Chromosomes exhibit precise and progressive pairing during prophase I.

The study captures early pairing events in spermatocytes and oocytes, using high-resolution contact maps.

Contact frequency increases significantly in homologous regions during zygotene and pachytene stages.

2. Chromatin Compartment Remodeling A/B compartments are weakened or restructured during prophase.

Specific B compartments, enriched in LINE elements and heterochromatin, show elevated pairing dynamics.

Suggests B compartment flexibility supports pairing more than A.

3. Role of Repeat Elements Repeat-rich regions, especially LINEs and LTRs, are preferentially localized to B compartments and correlate with stronger interhomolog interactions.

Repeats may play a structural role in mediating nuclear architecture during meiosis.

4. Sex Differences in Chromosome Pairing The authors find distinct dynamics between male and female germ cells, including:

Differences in compartment switching speed.

Asynchronous timing in homolog pairing progression.

Oocytes retain more stable open chromatin than spermatocytes during mid-prophase.

(**article?**){He2023, year = {2023}, title = {{3D genome remodeling and homologous pairing during meiotic prophase of mouse oogenesis and spermatogenesis}}, author = {He, Jianrong and Yan, Ao and Chen, Bohan and Huang, Jin and Kee, Kehkooi}, journal = {Developmental Cell}, doi = {10.1016/j.devcel.2023.07.004}, issn = {1534-5807}, abstract = {{By applying Hi-C and super-resolution microscopy, this study tracks chromatin compartment transitions and homolog pairing in meiosis. Repeat-enriched B compartments show increased pairing dynamics and regulate meiotic architecture.}}, pages = {}, number = {}, volume = {}, pmid = {}, pmcid = {}, keywords = {}, local-url = {https://www.cell.com/developmental-cell/abstract/S1534-5807(23)00553-1} }

8.0.4 DICER regulates repeat RNA and chromosome segregation (maybe???)

- Satellite repeats involved in chromatin architecture, A/B compartment boundary insulation via RNA-based silencing pathways.

(**article?**){He2023, year = {2023}, title = {{3D genome remodeling and homologous pairing during meiotic prophase of mouse oogenesis and spermatogenesis}}, author = {He, Jianrong and Yan, Ao and Chen, Bohan and Huang, Jin and Kee, Kehkooi}, journal = {Developmental Cell}, doi = {10.1016/j.devcel.2023.07.004}, issn = {1534-5807}, abstract = {{By applying Hi-C and super-resolution microscopy, this study tracks chromatin compartment transitions and homolog pairing in meiosis. Repeat-enriched B compartments show increased pairing dynamics and regulate meiotic architecture.}}, pages = {}, number = {}, volume = {}, pmid = {}, pmcid = {}, keywords = {}, local-url = {https://www.cell.com/developmental-cell/abstract/S1534-5807(23)00553-1} }

9 whats next

can we show this 3. Mutation Rate Variation and Evolution Repetitive DNA in closed chromatin accumulates mutations at different rates compared to open regions, contributing to genomic variation and evolution (Makova & Hardison, 2015).

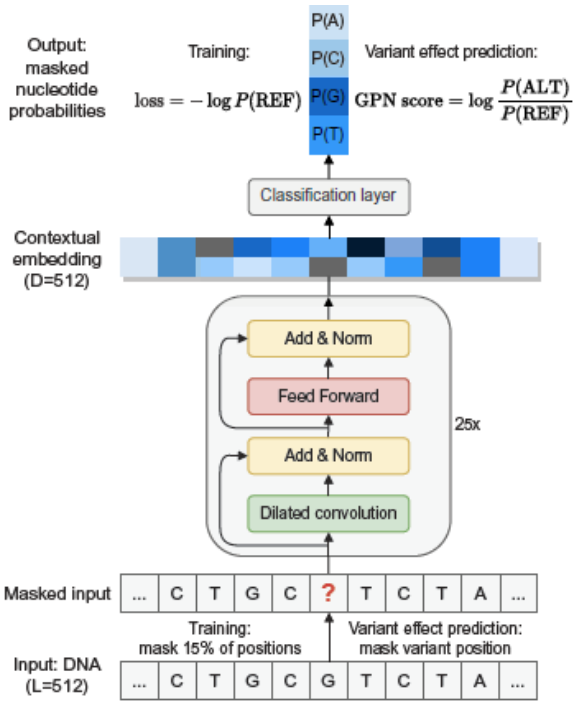

Figure 9.1

10 Genome-wide coancestry reveals details of ancient and recent male-driven reticulation in baboons

(Knuth 1984)

```
# Here is example python code
print("Hello world")
print("i need to make a update")
```

Here is a reference (Nielsen and Slatkin 2016)

11 References

Benegas, Gonzalo, Sanjit Singh Batra, and Yun S. Song. 2024. "DNA Language Models Are Powerful Predictors of Genome-Wide Variant Effects." *PNAS*. https://doi.org/10.1073/pnas.2311219121.

Knuth, Donald E. 1984. "Literate Programming." *Comput. J.* 27 (2): 97–111. https://doi.org/10.1093/comjnl/27.2.97.

Nielsen, Rasmgb, and Montgomery Slatkin. 2016. *An Introduction to Population Genetics: Theory and Applications*.