Manuscript

Søren Jørgensen

November 19, 2024 13:23:43 (UTC +01:00)

This is a dummy abstract, dreamt up by chatGPT. This thesis investigates the 3D chromatin architecture of the X chromosome in baboons, macaques, and humans, focusing on chromatin compartments during spermatogenesis. Using publicly available Hi-C data, interaction maps were created to identify Principal Component 1 (PC1) compartments, revealing distinct compartmentalization patterns among species. The analysis included transition zones, where chromatin shifts between compartment types, and their correlation with positively selected regions. By comparing these zones with evolutionarily significant regions, the study explores how chromatin structure influences evolutionary pressures. Key findings include conserved chromatin features that may help retain non-advantageous alleles, suggesting a role for selfish genetic elements in genome evolution. This research offers new insights into the relationship between chromatin architecture and evolutionary dynamics across primate species.

Table of contents

Chromatin Compartments and Selection on *X*

Introduction

= draft block

Sexual reproduction (spermatogenesis, meiosis)

The production of gametes in a sexually reproducing organism is a highly complex process that involves numeruous elements. Spermatogenesis, the process of forming male gametes, involves four stages of differentiation from a germ cell through *spermatogonia*, *pachytene spermatocyte*, and *round spermatids* to *spermatozoa*, or *sperm* (**wang_reprogramming_2019?**).

Selfish genes

The conventional story of meiosis in gametogenesis is one of random segregation of the sex chromosomes. They split into haploid gametes that are each responsible for their own survival, and nothing more. That seems like a fair game, but what if some genes are cheating the system by making other participants less viable. Say, some genes on the X chromosome creates a disadvantage for gametes that *do not* contain those genes, making sure the Y chromosome is not as viable as the X, resulting in a sex imbalance and possibly numerous other downstream effects. That is exactly what is coined *sex chromosome meiotic drive* (**jaenike_sex_2001?**), a result of selfish genetic elements.

Motivated by previous results in the Munch Research group (**munch_group_2024?**) on hybrid incompatibility and extended common haplotypes (**skov_extraordinary_2023?**; **sorensen_genome_wide_2023?**) that could be explained by meiotic drive, we wanted to investigate how these patterns correlate with chromatin compartments.

High-Throughput Chromosome conformation capture (Hi-C)

Our DNA can be divided into different orders of structure. *3C* focus on identifying the highest orders of organization inside the nucleus, that is, when the 30 nm thick coil of chromatin fibers folds into loops, Topologically Associating Domains (TADs), and chromatin compartments. Here, we narrow our focus on the largest of the structures, *compartments*, that is known to determine availability to transcription factors, thus making an *A* compartment *active*—and the *B* compartment *inactive*. The introduction of the Hi-C method (**lieberman-aiden_comprehensive_2009?**) (high-throughput 3C) opened new possibilities for exploring the three-dimensional organization of the genome.

Methods

In this project, we formulate two objectives:

A: Reproduce the Hi-C interaction maps and eigendecomposition from (**wang_reprogramming_2019?**), with some modifications. We briefly use *HiCExplorer*, but change the analyses to use the *Open2C Ecosystem* (**open2c?**) which have a Pyton API as well as command-line functions, which can be paired very well with Jupyter Notebooks. The majority of the data analysis was run with a *gwf* workflow, and the commands that were visually inspected were run in Jupyter Notebooks.

B Compare with regions of selection that are found in *papio anubis*, and maybe in *human* too. Investigate the biological meaning of the results.

Data and structure

To get an overview of the data accessions used in this analysis, we will first summarize the `SRA-runtable.tsv` that contains the accession numbers and some metadata for each sample (Table 1).

Table 1: Summary of the data accessions used in this analysis

Table 1

	source_name	GB	Bases	Reads
0	fibroblast	211.403275	553,968,406,500	1,846,561,355
1	pachytene spermatocyte	274.835160	715,656,614,700	2,385,522,049
2	round spermatid	243.128044	655,938,457,200	2,186,461,524
3	sperm	164.131640	428,913,635,400	1,429,712,118
4	spermatogonia	192.794420	518,665,980,300	1,728,886,601

For ease of mind, here is the folder structure of the project. `../steps/bwa/PE/` is the base directory artificially defined in the `master_workflow.py`. It ccould be any other directory inside `steps`. It is defined relative to the `master_workflow.py` file (inside the worklow), and converted to an absolute path by python.

The following section contains the visualization of the matrices and includes the calculation of the E1 compartments and their visualization. I discuss differents methods for construction and try to get both methodologically and result-wise close to (**wang_reprogramming_2019?**).

First, we will work on matrices in 500kb resolution, as documentation [maybe it was HiCExplorer??] states it is sufficient for chromosome-wide analysis and plotting. We will follow the `cooltools`-recommended pipeline (except that they use a 100kb cooler) for visualization and compartment calling with one example cooler (the merged fibroblast cooler at 500kb resolution).

Then, the most relevant parts will be generalized to run in a loop over all the coolers. First at 500kb resolution, then at 100kb resolution.

To accomodate the approach (barely) described in the paper, we will discuss a smoothing step to the observed/expected matrices before compartment calling, and we will apply a smoothing step to the E1 compartments and compare to the raw compartments.

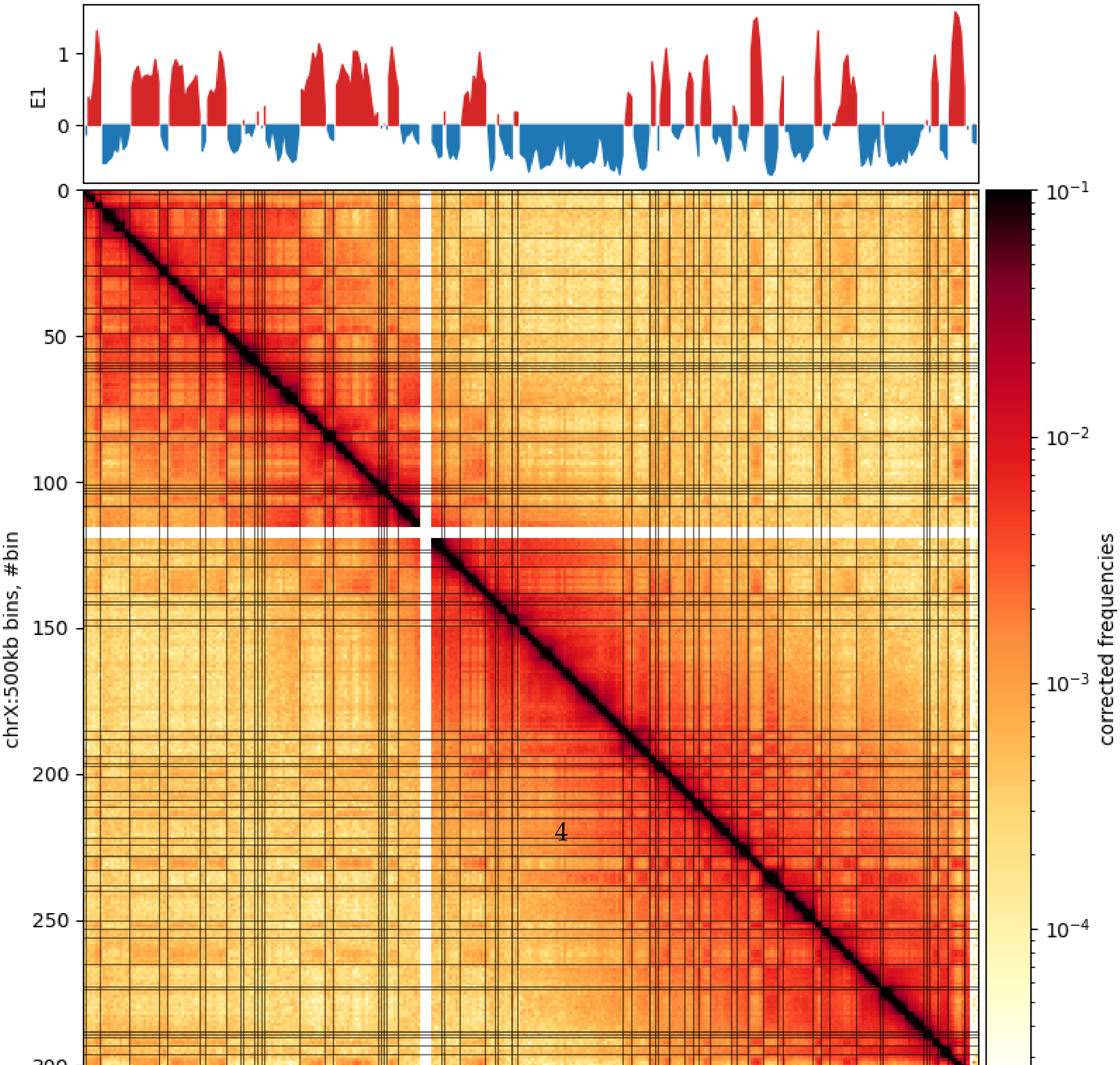
First, I will explore the visualization pipeline for a single cooler at 500kb resolution. I will modify the plot to be 'stairs' in stead of just a regular line plot, as it is both a more accurate representation of the data and it is more aesthetically pleasing with less spiky lines and holes.

In practice, the length of the dataframe is doubled, as it now contains an E1 value for both the start and end position for each bin in stead of only for the start. However, I first make the regular line plot to show the difference.

```
../steps/bwa/PE
 bamfiles
ăă  fibroblast
ăă  pachytene_spermatocyte
ăă  round_spermatid
ăă  sperm
ăă  spermatogonia
 cool
ăă  fibroblast
ăă  pachytene_spermatocyte
ăă  round_spermatid
ăă  sperm
ăă  spermatogonia
 pairs
     fibroblast
     pachytene_spermatocyte
     round_spermatid
     sperm
     spermatogonia

18 directories
```

Figure 1



Here we see that Figure 2. . .

Results

Here are the glorious results

Discussion

Here is the discussion

Bibliography