

Thesis

Chromatin Compartments and Selection on X

Søren Jørgensen

2024-11-28



Chromatin Compartments and Selection on X

How Edges of Active Chromatin Align with Selection Regions in
Primates

Søren Jørgensen



2024-11-28

Submitted in fulfillment of the requirements of the degree of

“ 3-dimensional structure of chromatin brings light onto the mystery of selfish genes.

- Søren Jørgensen //

Table of contents

1	Introduction	1
1.1	Sexual reproduction (spermatogenesis, meiosis)	1
1.2	Selfish genes (and randomness)	1
1.3	High-Throughput Chromosome Conformation Capture (Hi-C)	2
2	Methods	3
2.1	Downloading Data and Project Structure	3
2.2	Handling coolers (Or: preparing coolers)	4
3	Results	7
4	Discussion	9
	Bibliography	11

1 Introduction

1.1 Sexual reproduction (spermatogenesis, meiosis)

The production of gametes in a sexually reproducing organism is a highly complex process that involves numerous 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 et al. 2019), and it is the very basis of male reproduction. The specialized cell division of meiosis neatly handles the pairing, recombination, and segregation of homologous chromosomes, thereby ensuring proper genetic distribution. Deeply understanding the steps of molecular steps of reproduction and how our genetic material is inherited is essential in biology, bringing insight to areas such as speciation, population diversity, and (male) infertility.

1.2 Selfish genes (and randomness)

The conventional story of meiosis in gametogenesis is one of random segregation of the sex chromosomes. They split into haploid gametes, where each chromosome has an equal chance of being passed on to a gamete. That seems like a fair game, but what if some genes are cheating the system by making others less viable. A meiotic driver is a selfish gene element that modulates meiosis and preferentially transmits its own allele through meiosis, regardless of the downstream fitness effects it may have (good or bad) on the organism it is part of. This phenomenon challenges the traditional understanding of selection, extending its scope beyond the fitness effects on an organism to include selective pressures at the molecular level. For example, if some genes on the X chromosome create 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 2001), a result of selfish genetic elements.

Motivated by previous results in the Munch Research group (Munch 2024) on hybrid incompatibility and extended common haplotypes (Skov et al. 2023; Sørensen et al. 2023) that could be explained by meiotic drive, we wanted to investigate how these patterns correlate with chromatin compartments.

1 Introduction

1.3 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 (high-throughput 3C) method (Lieberman-Aiden et al. 2009) opened new possibilities for exploring the three-dimensional organization of the genome.

2 Methods

In this project, we formulate two objectives:

A: Reproduce the Hi-C interaction maps and eigendecomposition from (Wang et al. 2019), with some modifications. We briefly use *HiCExplorer*, but change the analyses to use the *Open2C Ecosystem* (Open Chromosome Collective 2024) which have a Python 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.

All computations were performed on GenomeDK (GDK) [ref], an HPC cluster located on Aarhus University, and most of the processing of the data was nested into a *gwf* workflow [ref], a workflow manager developed at GDK. I would like to thank GDK and Aarhus University for providing computational resources and support that contributed to these research results.

The whole of this project is carried out with reproducibility in mind, so an effort (and quite a significant amount of time) has been put into documenting code and organizing the project for readability and transparency through a Quarto project [ref]. Therefore, all code, virtual environments and text is made available as a Quarto book, rendered directly from the GitHub repository with GitHub Pages []. To make this possible, the Quarto documentation has been extensively studied and discussed with *KMT* [ref, acknowledge].

2.1 Downloading Data and Project Structure

To reproduce the results from (Wang et al. 2019), I chose to use their raw data directly from the SRA portal [ref]. I filtered the data to contain all their paired-end Hi-C reads, and included only macaque samples. The data set also contains RNAseq data, and the same tissues for both macaque and mouse. The meta data for the data set was extracted into a runtable `SRA-runtable.tsv`. To get an overview of the data accessions used in this analysis, we will first summarize the runtable that contains the accession numbers and some metadata for each sample (Table 2.1). It adds up to ~1Tb of compressed `fastq` files, holding ~9.5 billion reads, roughly evenly spread on the 5 tissue types.

2 Methods

Table 2.1: Summary of the data accessions used in this analysis

Table 2.1

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

2.2 Handling coolers (Or: preparing coolers)



Figure 2.1: A flowchart showing the pipeline from `.fastq` to `.mcool`. The first 6 steps were done with a Probably BioRender or Inkscape.

2.2.1 The `gwf` workflow targets

A `gwf` workflow was created to handle the first part of the data processing, and each accession number (read pair, mate pair) from the Hi-C sequencing was processed in parallel, so their execution was independent from each other.

2.2.1.1 Downloading the reads

The reads were downloaded from NCBI SRA portal [ref] directly to GDK using `sra-downloader` [ref] through docker [ref] as `.fastq.gz` files.

2.2.1.2 Handling the reference

The latest reference genome for rhesus macaque (*macaca mulata*), *rheMac10* (or *Mmul_10*, UCSC or NCBI naming conventions, respectively) was downloaded to GDK from UCSC web servers with `wget` [ref]. To use `bwa` (Burrow Wheeler's Aligner) [ref] for mapping, *rheMac10* needs

2.2 Handling coolers (Or: preparing coolers)

to be indexed with both `bwa index` with the `--bwtsw` option and `samtools faidx`, which results in six indexing files for `bwa mem` to use.

2.2.1.3 Mapping paired-end reads

2.2.1.4 Pair and sort the reads

2.2.1.5 Filter (deduplicate) pairs

2.2.1.6 Create interaction matrices (coolers)

2.2.2 Notebook edits

As `cooler` and `cooltools` have a Python API, the more experimental parts of the analysis were moved to Jupyter Notebooks (still running on GenomeDK)

2.2.2.1 Pooling samples (Merging coolers)

The samples are grouped into *replicates* with a unique **BioSample** ID, but we chose to pool all the interaction matrices for each cell type. We argue that when Wang et al. (2019) determine compartments to be highly reproducible between replicates, by merging the replicates we can get a more robust signal.

`cooler merge` was used to merge all samples in each sub-folder (cell type) to just one interaction matrix for each cell type. The function merges matrices of the same dimensions by simply adding the interaction frequencies of each genomic position together, resulting in less empty positions by chance.

2.2.2.2 Create multi-resolution coolers (zoomify)

A feature of working inside the ecosystem of *Open2C* [ref] is that it natively provides support for storing sparse interaction matrices in multiple resolutions in the same file by adding groups to the hdf5-file [ref]. We can then efficiently store resolutions (i.e., different bin sizes) that is multiples of the smallest bin size. We chose to use 10kb, 50kb, 100kb, and 500kb bins.

2 Methods

2.2.3 Matrix balancing (Iterative correction)

Finally, we balance the matrices using the cooler CLI. We use the `cooler balance` command with the default options using 32 cores, which iteratively balances the matrix (Iterative Corection). It is first described as a method for bias correction of Hi-C matrices in (Imakaev et al. 2012), where it is paired with eigenvector decomposition (ICE). Here the eigenvector decomposition of the obtained maps is shown to provide insights into local chromatin states.

According to `cooler` documentation, we have to balance the matrices on each resolution, and thus it cannot be done prior to zoomifying. The argument is that the balancing weights are resolution-specific and will no longer retain its meaning when binned with other weights. Therefore, we use a nested for-loop that iterates through all the `.mcools` and all the resolutions in each `.mcool`. `cooler balance` will create a new column in the `bins` group of each cooler, `, weight`, which can then be included or not in the downstream analysis. This means we will have access to both the balanced and the unbalanced matrix.

The default mode uses genome-wide data to calculate the weights for each bin. It would maybe be more suitable to calculate the weights for *cis* contacts only, and that is possible through the `--cis-only` flag, and that can be added to another column, so that we can compare the difference between the two methods easily. However, we will only use the default mode for now.

2.2.4 Eigendecomposition

3 Results

Here are the glorious results

4 Discussion

Here is the discussion

Bibliography

- Imakaev, Maxim, Geoffrey Fudenberg, Rachel Patton McCord, Natalia Naumova, Anton Goloborodko, Bryan R Lajoie, Job Dekker, and Leonid A Mirny. 2012. "Iterative Correction of Hi-C Data Reveals Hallmarks of Chromosome Organization." *Nature Methods* 9 (10): 999–1003. <https://doi.org/10.1038/nmeth.2148>.
- Jaenike, John. 2001. "Sex Chromosome Meiotic Drive." *Annual Review of Ecology and Systematics* 32 (1): 25–49. <https://doi.org/10.1146/annurev.ecolsys.32.081501.113958>.
- Lieberman-Aiden, Erez, Nynke L. Van Berkum, Louise Williams, Maxim Imakaev, Tobias Ragoczy, Agnes Telling, Ido Amit, et al. 2009. "Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome." *Science* 326 (5950): 289–93. <https://doi.org/10.1126/science.1181369>.
- Munch, Kasper. 2024. "Munch-Group." <https://munch-group.org/research.html>.
- Open Chromosome Collective. 2024. "Open Chromosome Collective (Open2C)." Resource. *Open2C*. <https://open2c.github.io/>.
- Skov, Laurits, Moisès Coll Macià, Elise Anne Lucotte, Maria Izabel Alves Cavassim, David Castellano, Mikkel Heide Schierup, and Kasper Munch. 2023. "Extraordinary Selection on the Human X Chromosome Associated with Archaic Admixture." *Cell Genomics* 3 (3): 100274. <https://doi.org/10.1016/j.xgen.2023.100274>.
- Sørensen, Erik F., R. Alan Harris, Liye Zhang, Muthuswamy Raveendran, Lukas F. K. Kuderna, Jerilyn A. Walker, Jessica M. Storer, et al. 2023. "Genome-Wide Coancestry Reveals Details of Ancient and Recent Male-Driven Reticulation in Baboons." *Science* 380 (6648): eabn8153. <https://doi.org/10.1126/science.abn8153>.
- Wang, Yao, Hanben Wang, Yu Zhang, Zhenhai Du, Wei Si, Suixing Fan, Dongdong Qin, et al. 2019. "Reprogramming of Meiotic Chromatin Architecture During Spermatogenesis." *Molecular Cell* 73 (3): 547–561.e6. <https://doi.org/10.1016/j.molcel.2018.11.019>.

Bioinformatics Research Centre
Department of Molecular Biology and Genetics
Aarhus University
Universitetsbyen 81
8000 Aarhus C
Denmark

