TADs are 3D structural units of higher-order chromosome organization in Drosophila

By Szabo, Q. et al. at Science Advances 4, eaar8082 (2018).

黃 宇秀 | 邱 淦均 | 李 柏漢 | 林 穎彥

BioInformatics 113
2025.1.2

Table of Contents

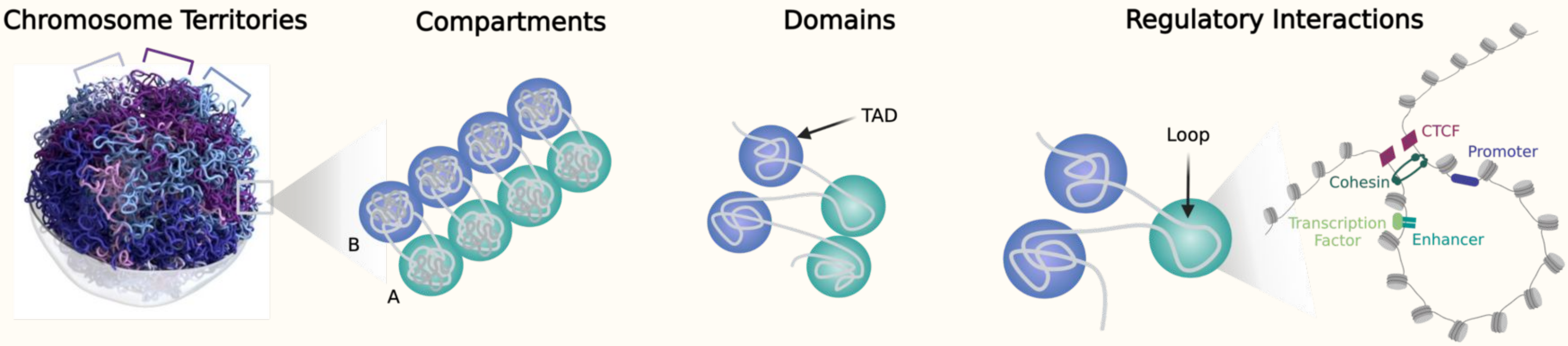
- Paper Introduction
- Experiment
- Experiment Objectives
- Data & Used Tools Description
- Experiment Results
- Cooperation

Paper Introduction

What is Topologically Associating Domains (TADs)?

- Fundamental units of the three-dimensional genome structure

3D Genome Architecture



Chromosome Territories

In the nucleus chromosomes are organized into chromosome territories

Compartments

Chromosomes are divided into cell-specific A/B compartments

Domains

Compartments are organized into topologically associated domains (TADs)

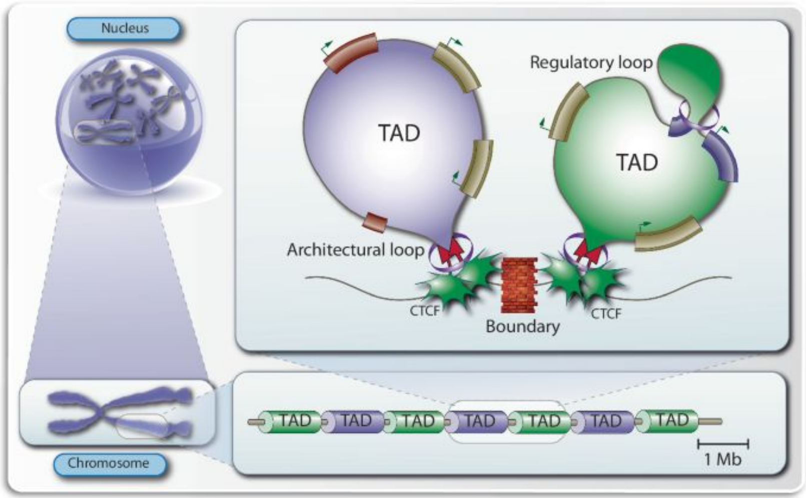
Regulatory Interactions

Within TADs, DNA is looped together with the assistance of architectural proteins and histones

What is Topologically Associating Domains (TADs)?

Key features of TADs:

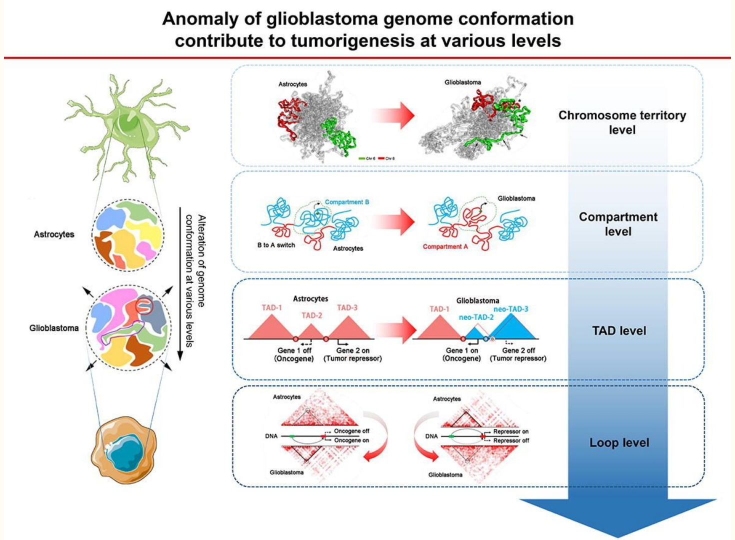
1. **Well-defined boundaries**: TADs are separated by clear boundaries, often marked by specific proteins such as CTCF and structural factors like the cohesin complex.
2. **High internal interactions**: Within a TAD, DNA fragments interact more frequently, facilitating regulatory interactions between genes and elements like enhancers and promoters.
3. **Conservation**: TADs are often conserved across cell types and species, indicating their functional importance in genome organization and gene regulation.

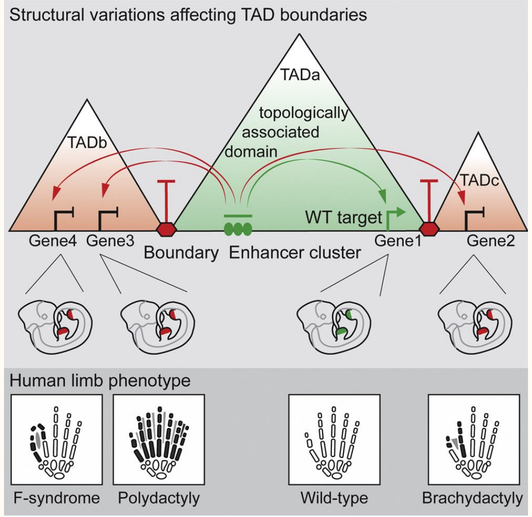


Why Topologically Associating Domains (TADs) so important?

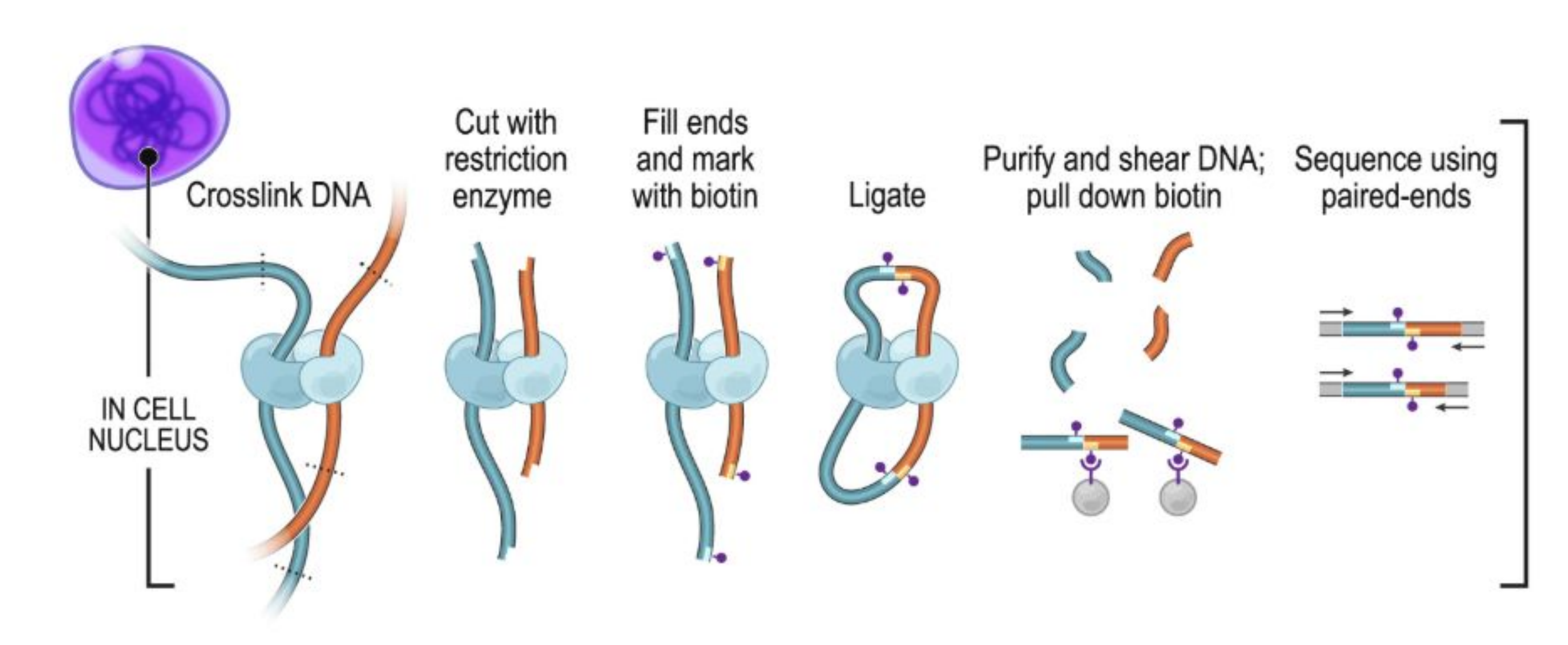
TADs play crucial roles in regulating gene expression, maintaining genome stability, and organizing the chromatin in the nucleus.

Disruptions in TAD boundaries are associated with various diseases, including cancers and developmental disorders.

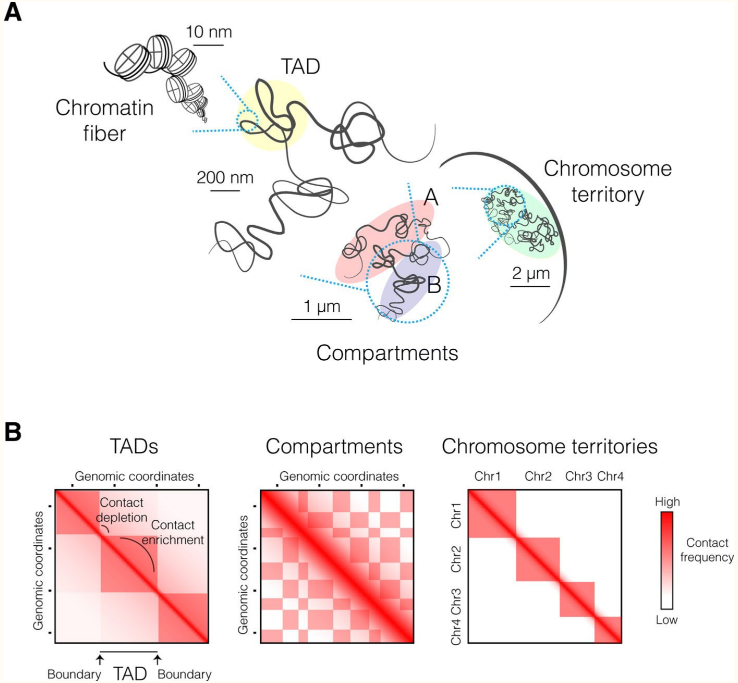


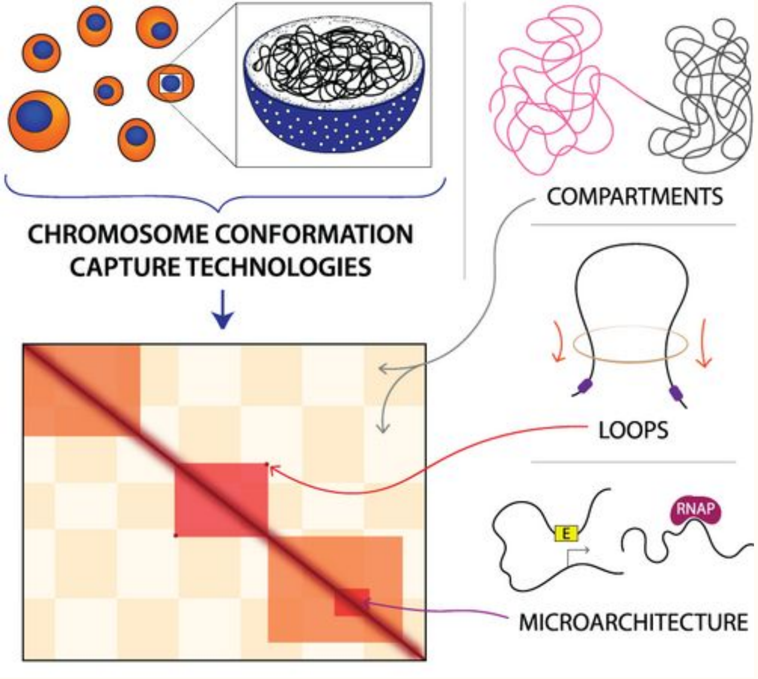


Chromosome Conformation Capture (Hi-C)


Crosslink DNA
IN CELL NUCLEUS
Cut with restriction enzyme
Fill ends and mark with biotin
Ligate
Purify and shear DNA; pull down biotin
Sequence using paired-ends

What can we tell from the HI-C Map

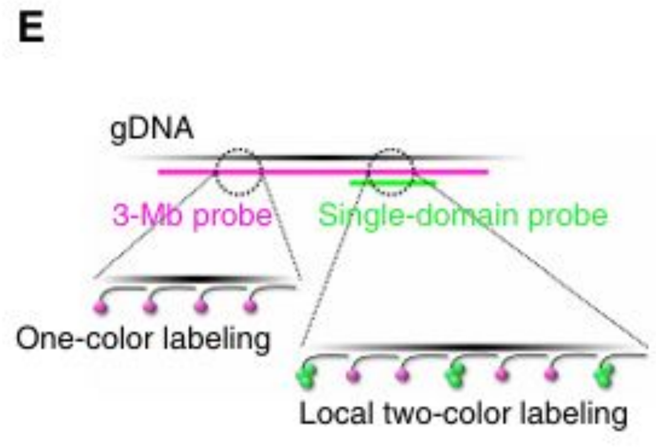




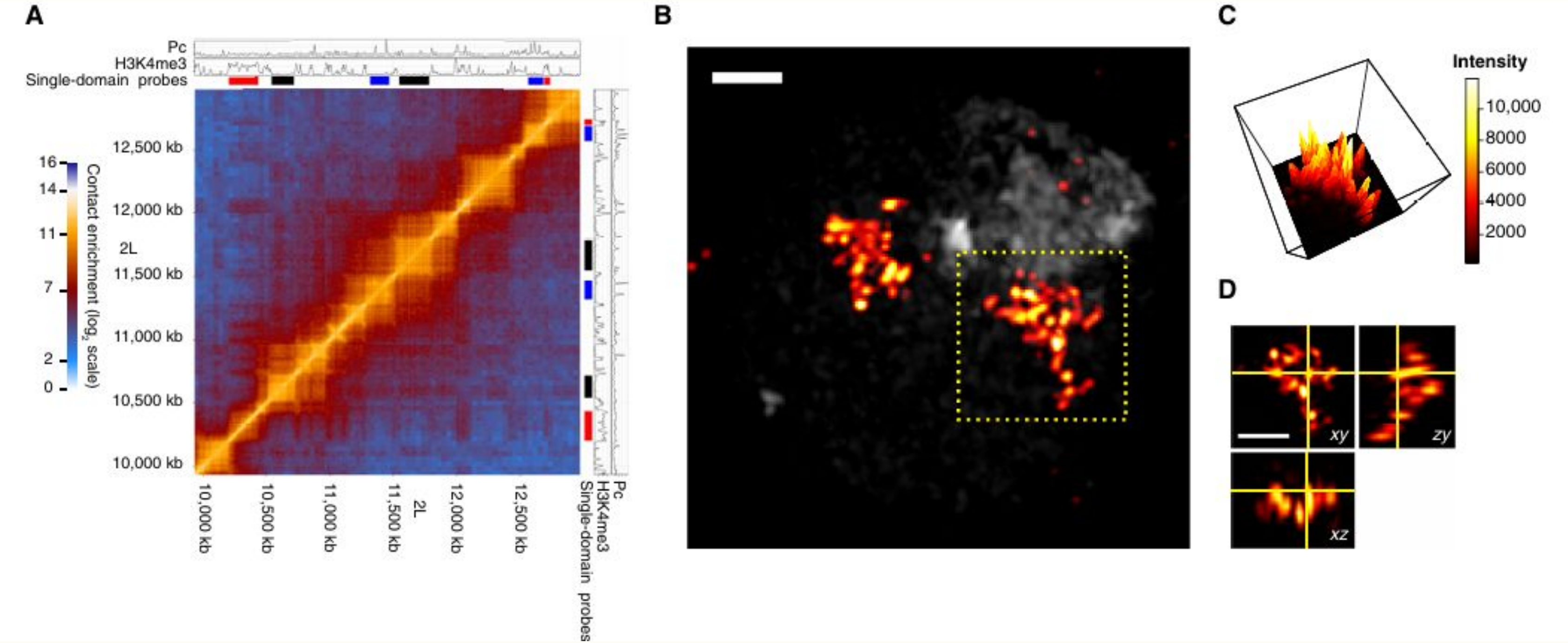
Chromatin is organized in a series of discrete 3D nanocompartments

3-Mb (chr2L: 9935314-12973080) region comprises three main types of Drosophila epigenetic domains:

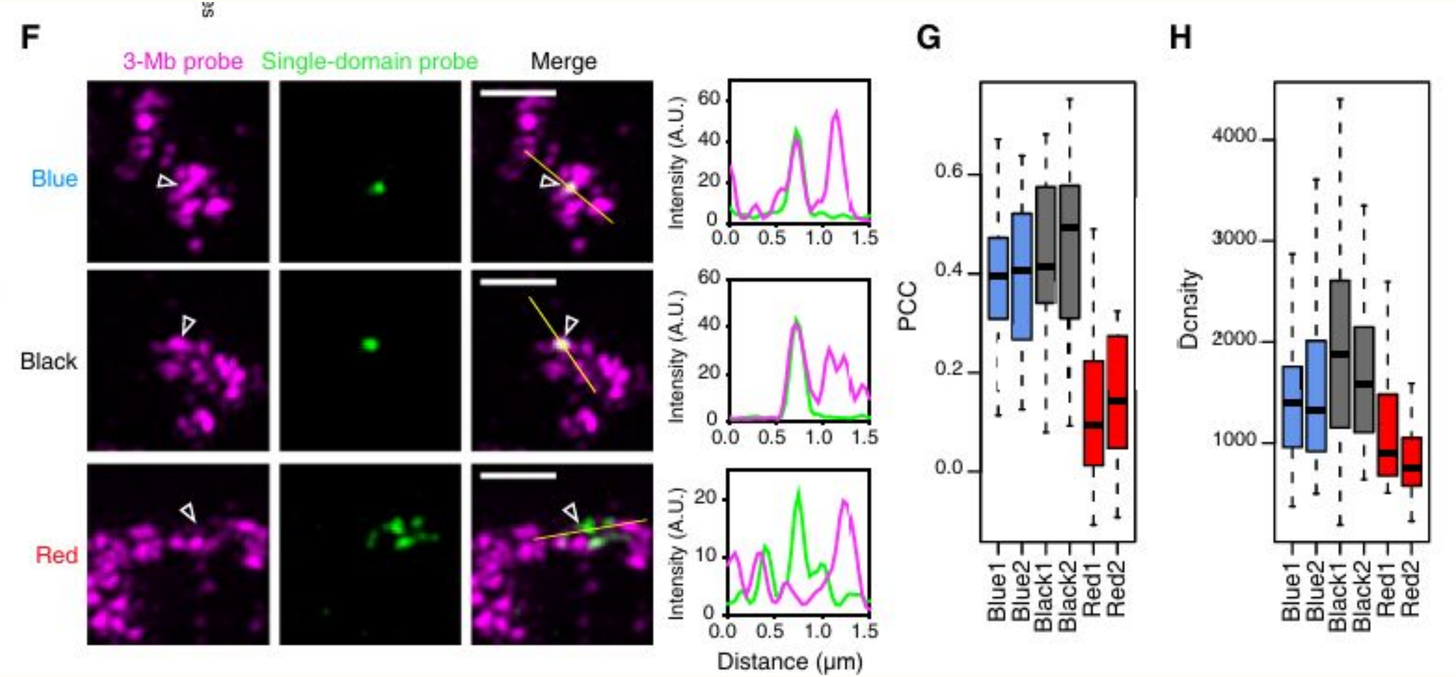
1. active chromatin (Red) enriched in trimethylation of histone 3 lysine 4 (H3K4me3), H3K36me3, and acetylated histones
2. Polycomb group (PcG) protein repressed domains (Blue), defined by the presence of PcG proteins and H3K27me3
3. inactive domains (Black), which are not enriched in specific epigenetic components



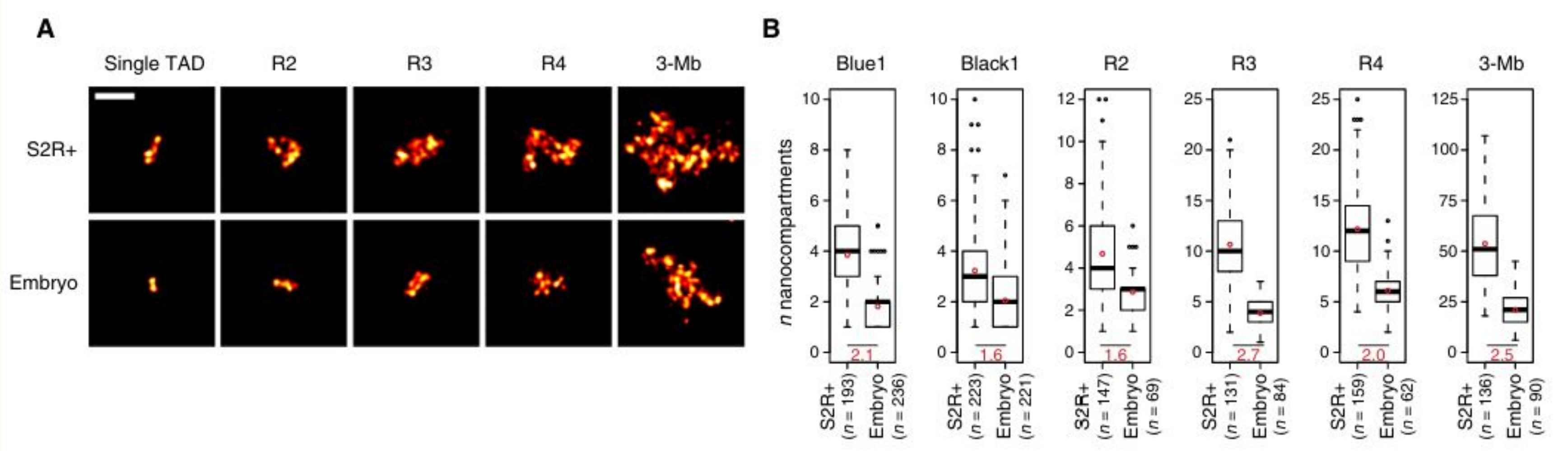
Chromatin is organized in a series of discrete 3D nanocompartments



Chromatin is organized in a series of discrete 3D nanocompartments

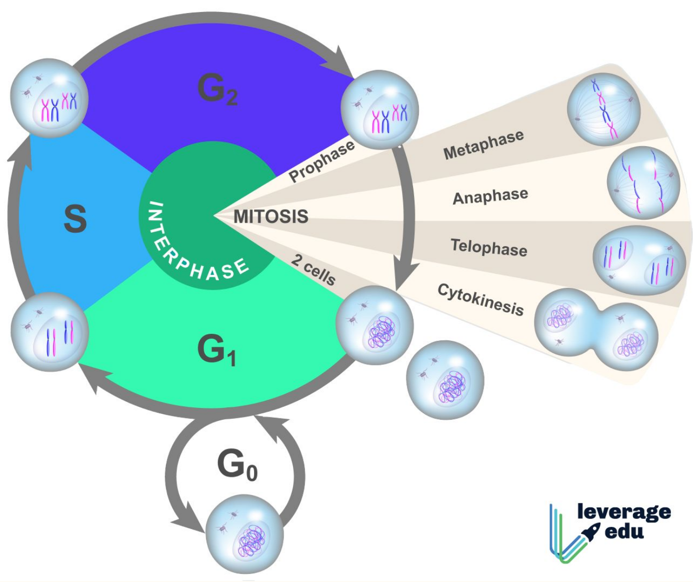


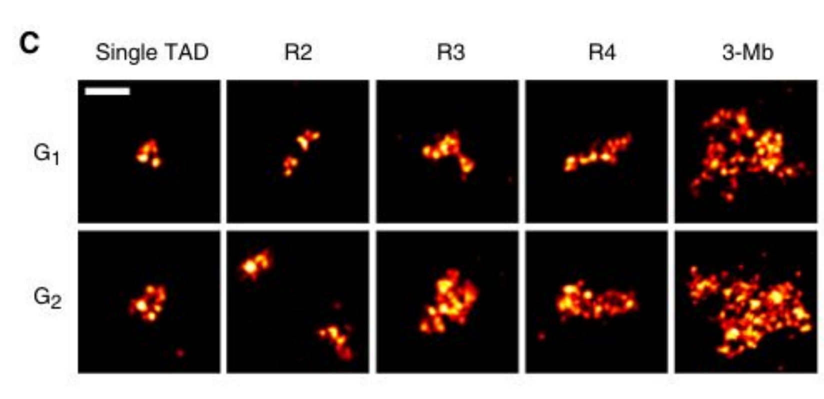
TAD-based 3D nanocompartments undergo dynamic cis and trans contact events

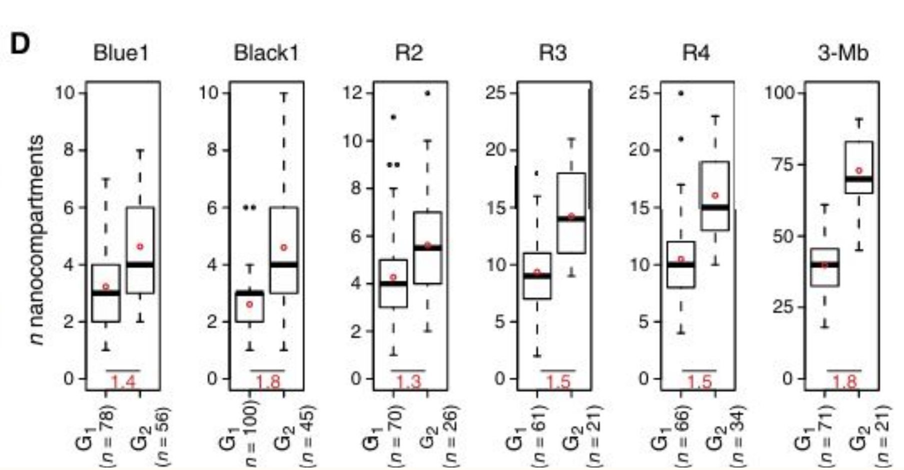


- Tetraploid S2R+ cells versus diploid embryonic (12 to 16 hours) cells
- R2(195kb),R3(805kb),and R4(495kb),covering two,three,and four repressed TADs, respectively

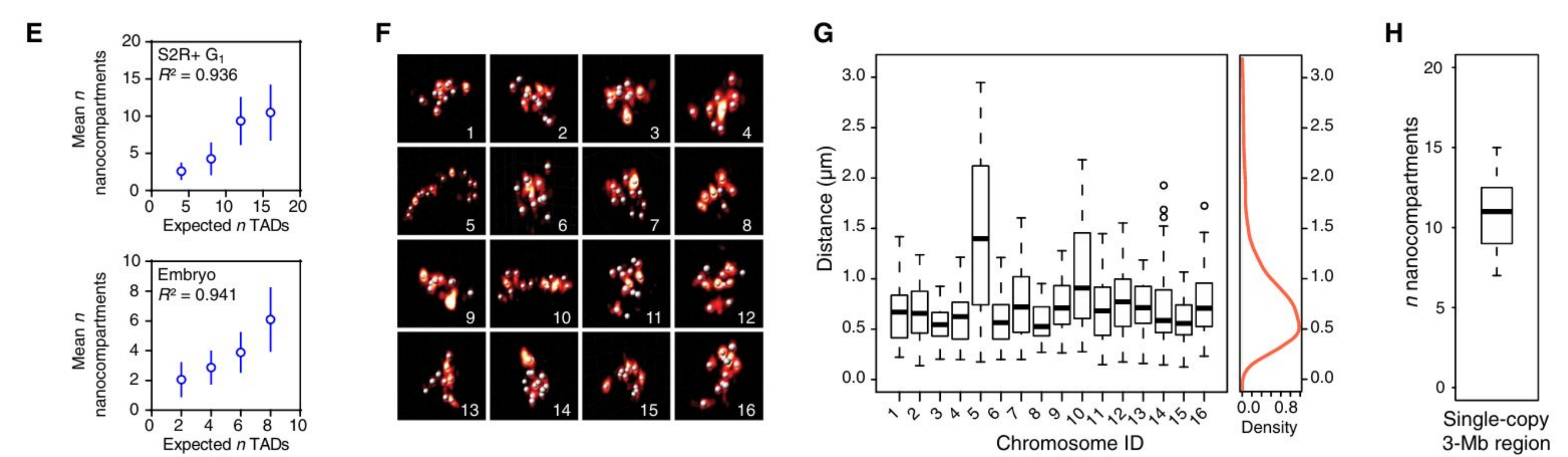
TAD-based 3D nanocompartments undergo dynamic cis and trans contact events







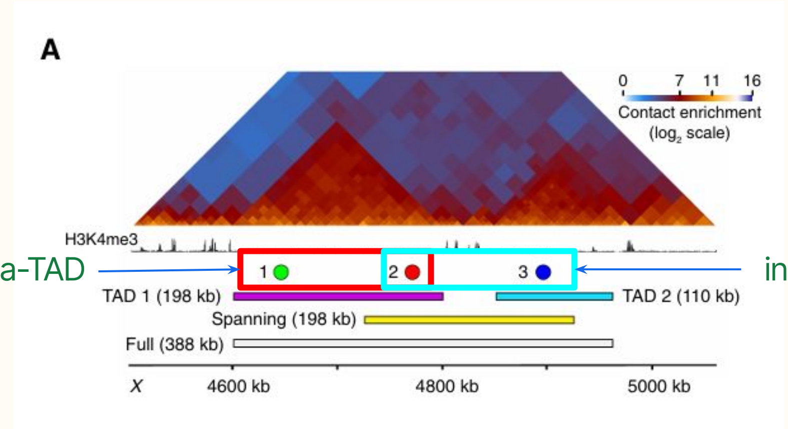
TAD-based 3D nanocompartments undergo dynamic cis and trans contact events



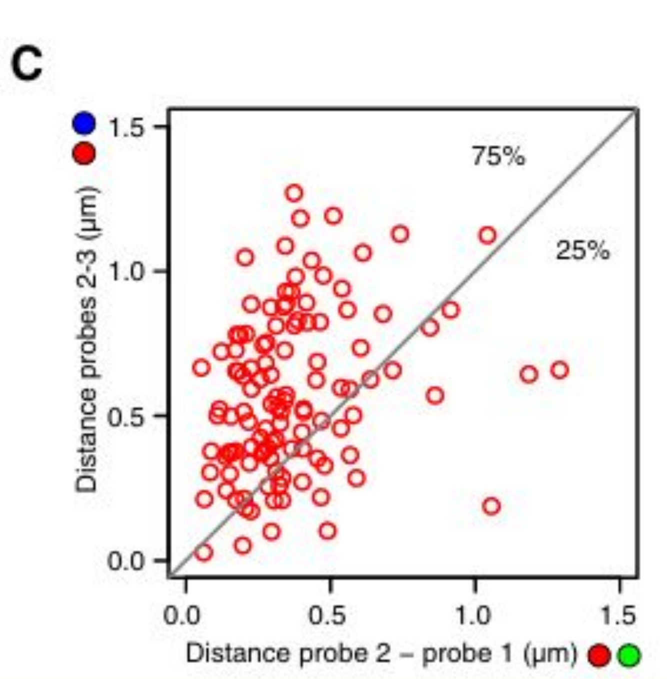
Repressed TADs form physical and structural chromosomal units

1. Single cell analysis revealed that intra-TAD distances are considerably shorter than inter-TAD distances

intra-TAD

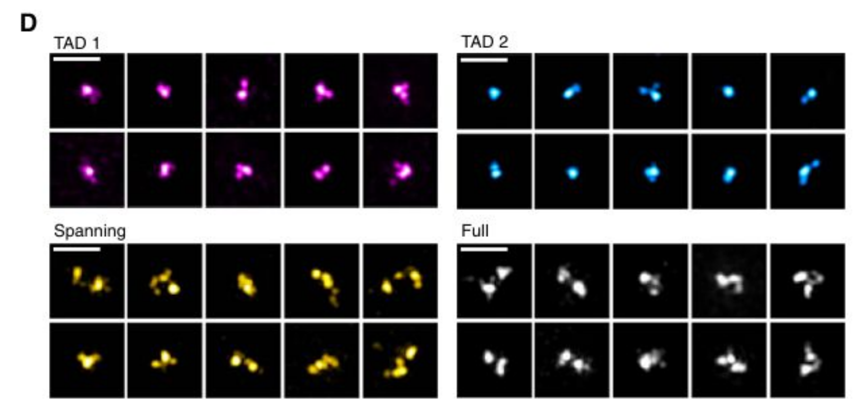


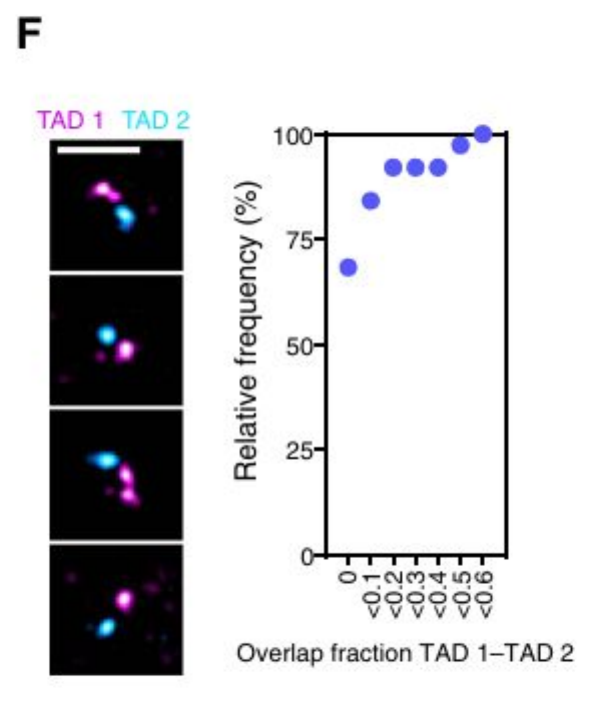
inter-TAD



Repressed TADs form physical and structural chromosomal units

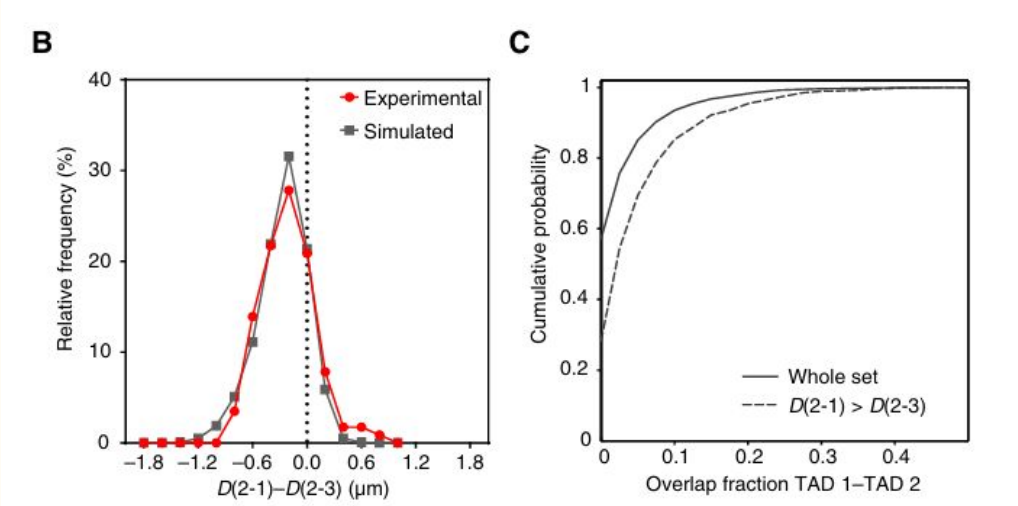
2. Despite variable intra- and inter-TAD contacts in each cell, the physical TAD-based compartmentalization of the chromatin fiber is a general feature of chromosomal domains.

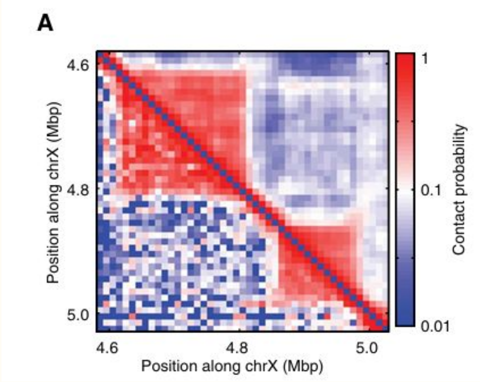




Polymer modeling recapitulates the physical partitioning of chromosomes into TADs

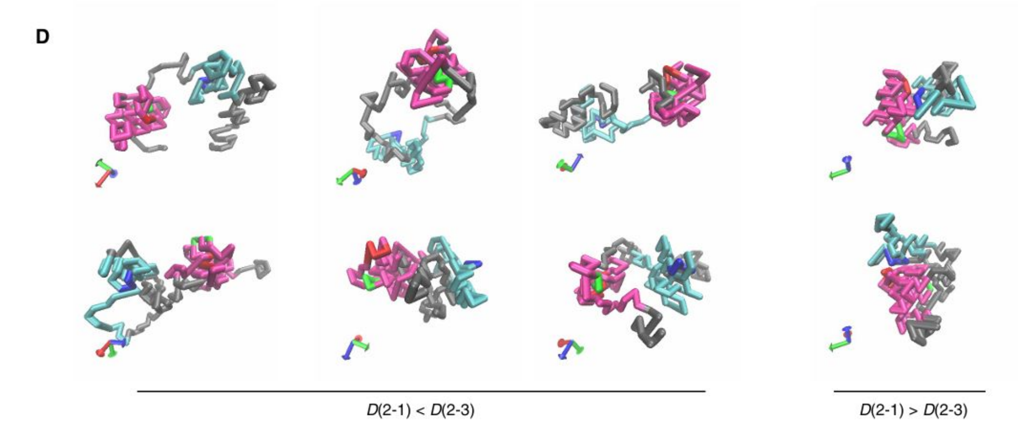
Polymer modeling using parameters that fit Hi-C maps supports the frequent folding of the two TADs into well-separated nanocompartments.





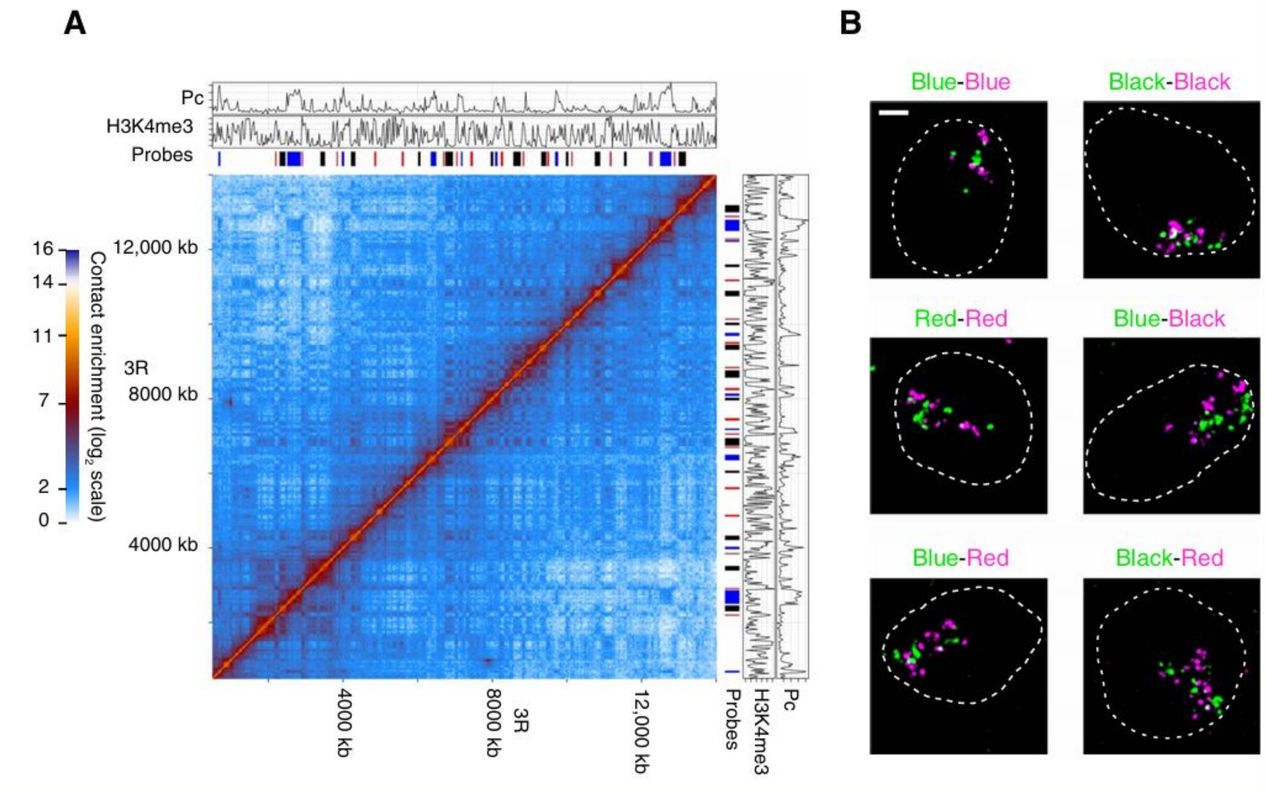
Polymer modeling recapitulates the physical partitioning of chromosomes into TADs

The fraction of intra-TAD distances larger than the inter-TADs counterparts is explained by the dynamic relative positioning of the two TADs.



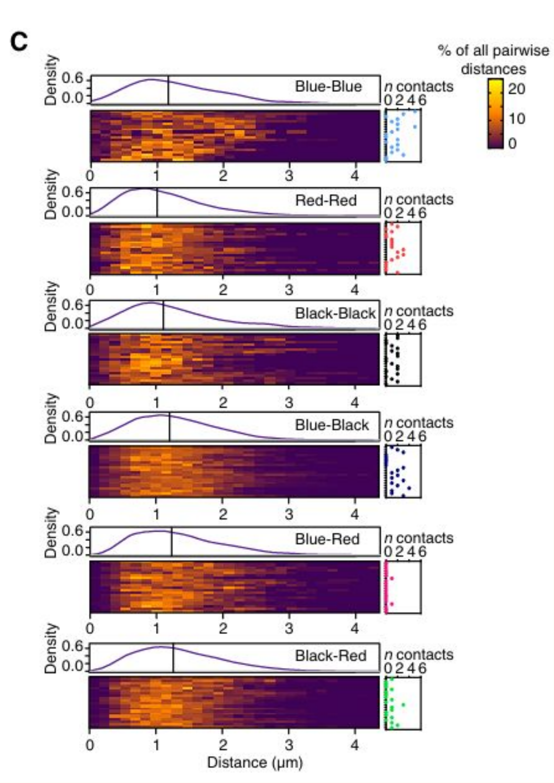
Large-scale chromatin folding reflects highly heterogeneous yet specific, long-range interdomain contacts

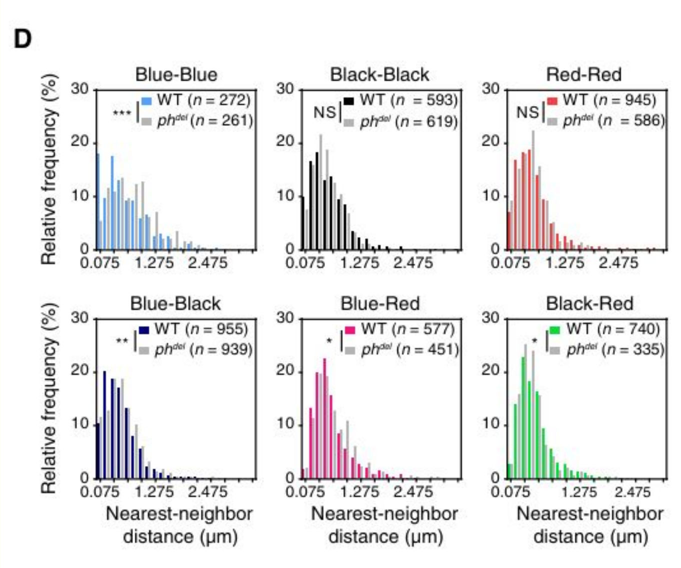
- Sixteen-to 18-hour embryo Hi-C map of a 14-Mb region.
- Labeling chromatin domains of different epigenetic states and studied their relative 3D spatial organization.



Large-scale chromatin folding reflects highly heterogeneous yet specific, long-range interdomain contacts

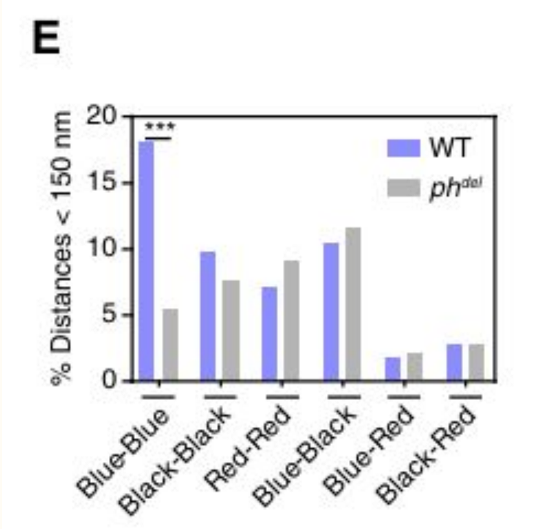
The analysis revealed the presence of discrete interdomain contacts, with preference for contacts among TADs of the same epigenetic type.

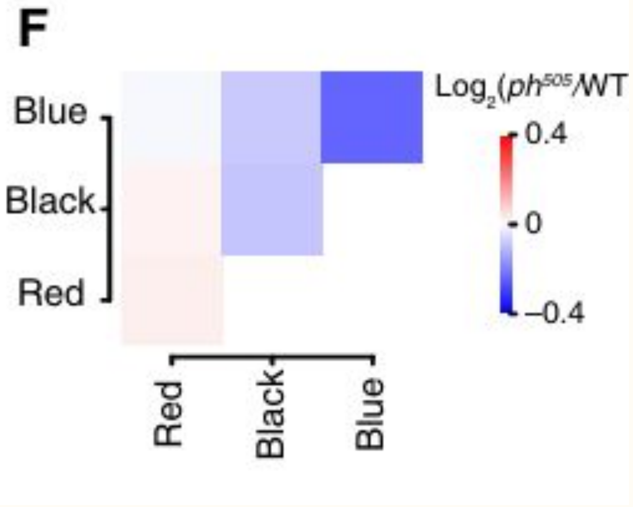


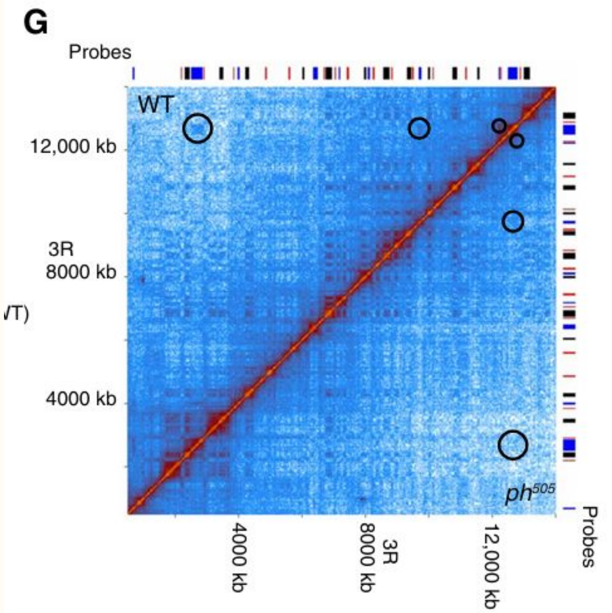


Large-scale chromatin folding reflects highly heterogeneous yet specific, long-range interdomain contacts

The inter-TAD contacts are regulated, as the disruption of the polyhomeotic (ph) PcG gene specifically affects Pc inter-TAD contacts without affecting contacts between other domains.







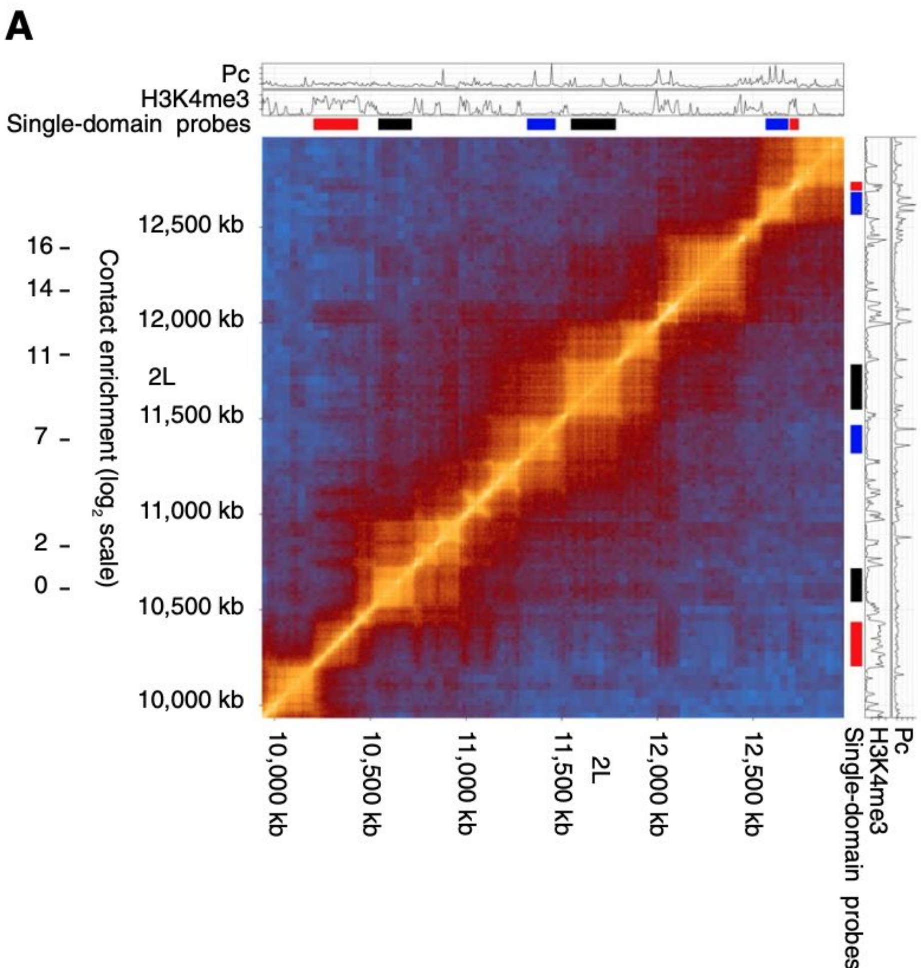
In Summary

This paper **provides an integrative view of chromatin folding in Drosophila:**

1. Repressed TADs form a succession of discrete nanocompartments.
2. Single-cell analysis revealed stable TAD-based chromatin compartmentalization, with some heterogeneity in intra-TAD conformations and cis/trans inter-TAD contact events.

Experiments

TADs are 3D structural units of higher-order chromosome organization in Drosophila

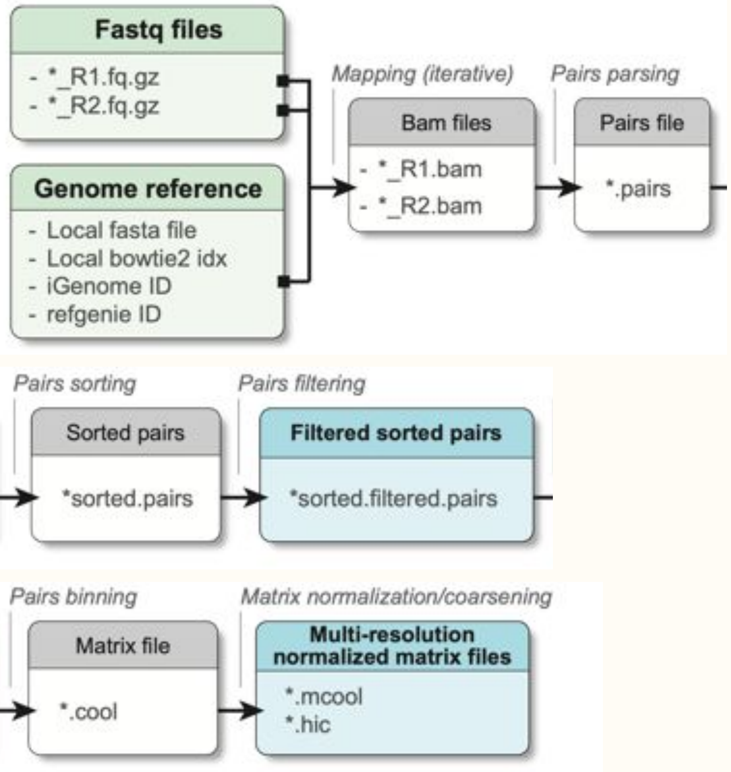


Experiment Objectives:

What we want to recreate?

Figure 1A
Hi-C Contact Map

NGS Workflow



Fastq files
- *_R1.fq.gz
- *_R2.fq.gz

Genome reference
- Local fasta file
- Local bowtie2 idx
- iGenome ID
- refgenie ID

Mapping (iterative)

Bam files
- *_R1.bam
- *_R2.bam

Pairs parsing

Pairs file
*.pairs

Pairs sorting

Sorted pairs
*sorted.pairs

Pairs filtering

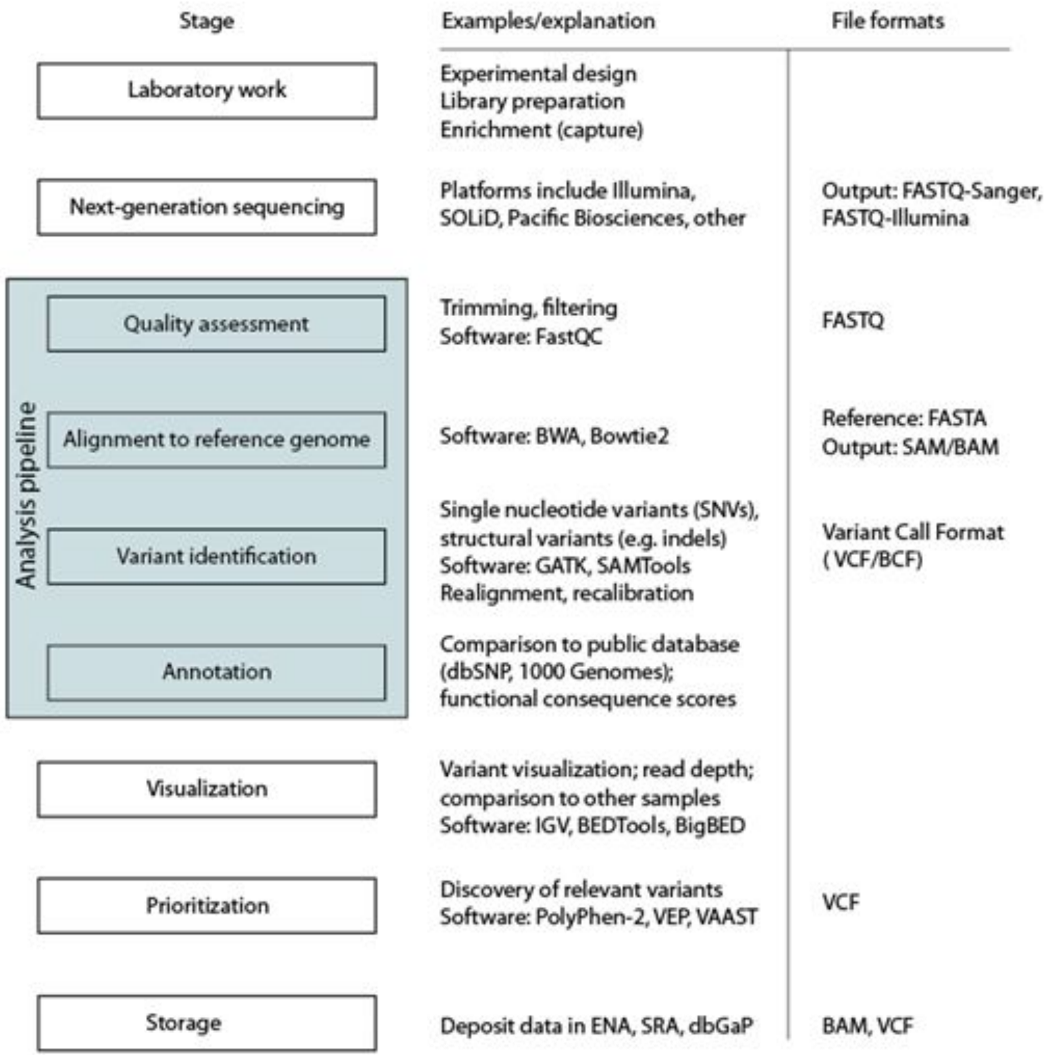
Filtered sorted pairs
*sorted.filtered.pairs

Pairs binning

Matrix file
*.cool

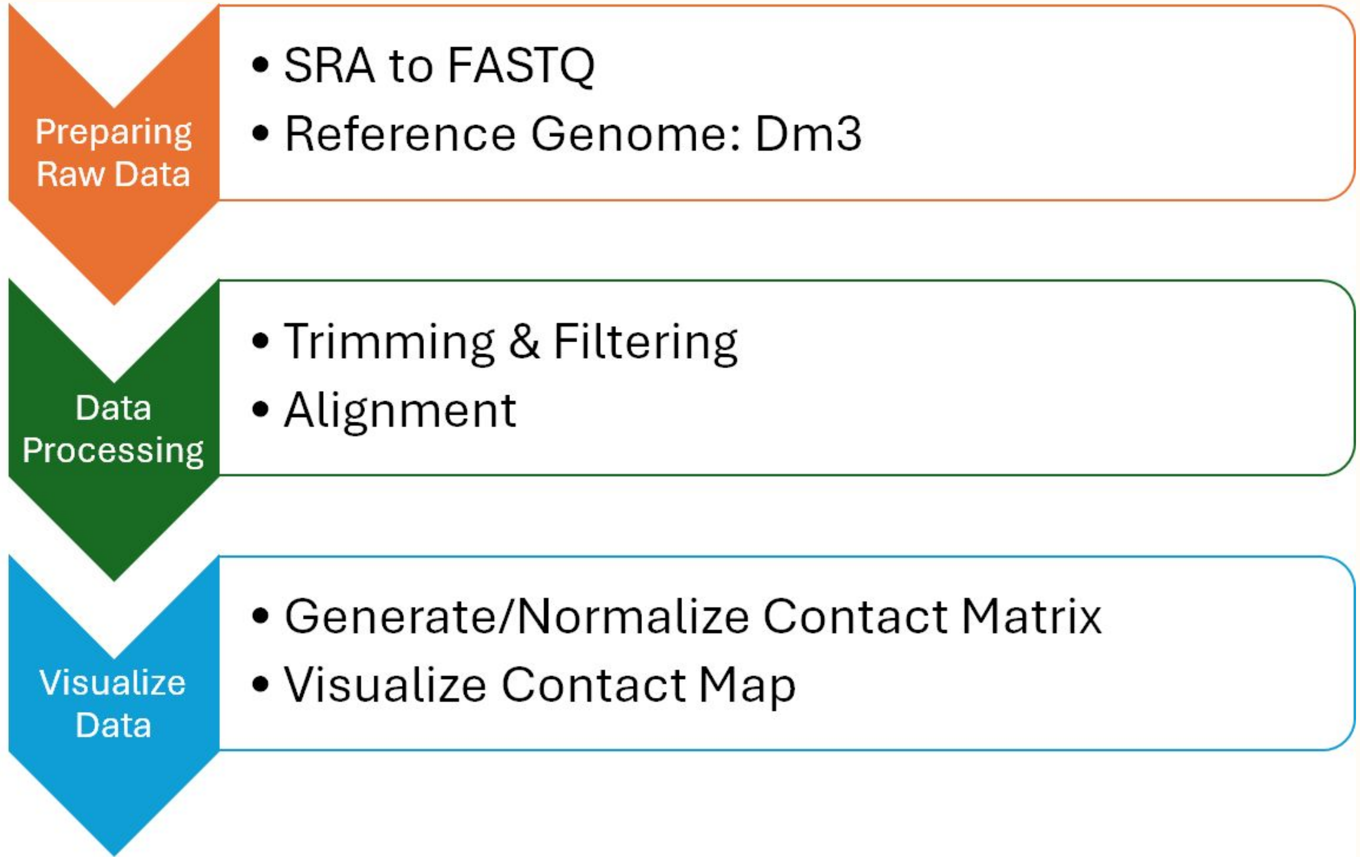
Matrix normalization/coarsening

Multi-resolution normalized matrix files
*.mcool
*.hic



	Stage	Examples/explanation	File formats
	Laboratory work	Experimental design Library preparation Enrichment (capture)	
	Next-generation sequencing	Platforms include Illumina, SOLiD, Pacific Biosciences, other	Output: FASTQ-Sanger, FASTQ-Illumina
Analysis pipeline	Quality assessment	Trimming, filtering Software: FastQC	FASTQ
	Alignment to reference genome	Software: BWA, Bowtie2	Reference: FASTA Output: SAM/BAM
	Variant identification	Single nucleotide variants (SNVs), structural variants (e.g. indels) Software: GATK, SAMTools Realignment, recalibration	Variant Call Format (VCF/BCF)
	Annotation	Comparison to public database (dbSNP, 1000 Genomes); functional consequence scores	
	Visualization	Variant visualization; read depth; comparison to other samples Software: IGV, BEDTools, BigBED	
	Prioritization	Discovery of relevant variants Software: PolyPhen-2, VEP, VAAST	VCF
	Storage	Deposit data in ENA, SRA, dbGaP	BAM, VCF

Overview Data Processing Steps



Preparing Raw Data

- SRA to FASTQ
- Reference Genome: Dm3

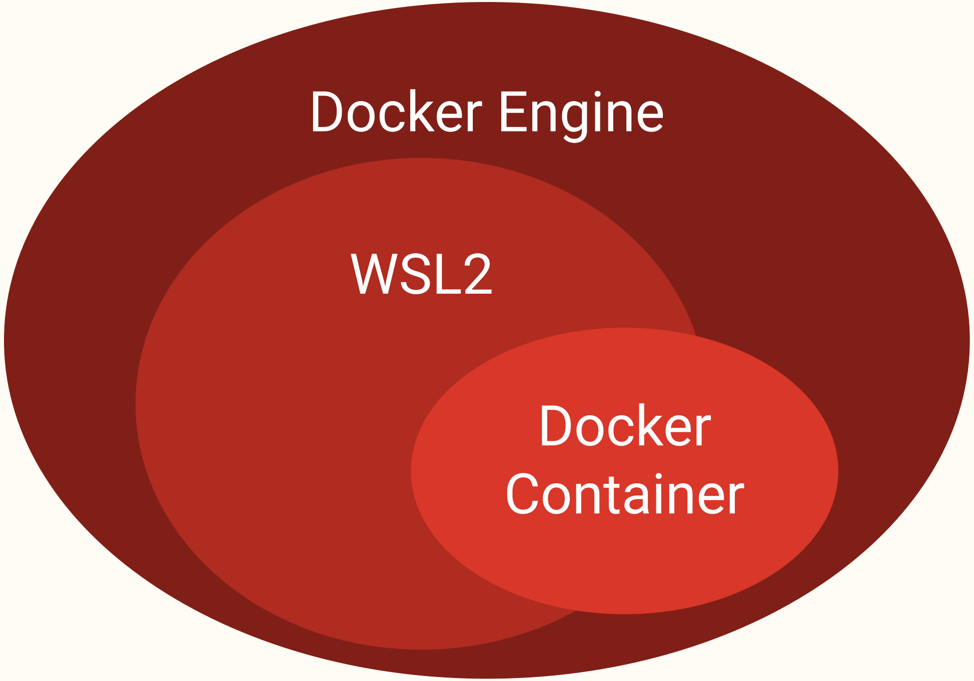
Data Processing

- Trimming & Filtering
- Alignment

Visualize Data

- Generate/Normalize Contact Matrix
- Visualize Contact Map

Environment - Docker with WSL



Docker Engine

WSL2

Docker Container

Preparing Raw Data - 1

Download SRA File

Convert SRA to FASTQ

Quality Control

Docker image:

- `ncbi/sra-tools`

CLI: `prefetch`

- Input: -
- Output: `SRR5579177`

Docker image:

- `ncbi/sra-tools`

CLI: `fasterq-dump`

- Input: `SRR5579177`
- Output: `SRR5579177_1.fastq / SRR5579177_2.fastq`

Docker image:

- `ubuntu:24.04`

CLI: `fastqc`

- Input: `SRR5579177_1.fastq / SRR5579177_2.fastq`
- Output: `SRR5579177_1_fastqc.html / SRR5579177_2_fastqc.html`

Preparing Raw Data - 2

Download Reference Genome

Docker image:

- `ubuntu:24.04`

CLI: `wget / gunzip`

- Input: `dm3.fa.gz`
- Output: `dm3.fa`

(Drosophila melanogaster: fruit fly)

Build Bowtie Index

Docker image:

`ubuntu:24.04`

CLI: `bowtie-build`

- Input: `dm3.fa`
- Output:

```
dm3_index.1.ebwt
dm3_index.2.ebwt
dm3_index.3.ebwt
dm3_index.4.ebwt
dm3_index.rev.1.ebwt
dm3_index.rev.2.ebwt
```

Check Index

Docker image:

`ubuntu:24.04`

CLI: `bowtie-inspect`

- Output:

```
SA-Sample     1 in 32
FTab-Chars    10
Sequence-1    chr2L    23011544
Sequence-2    chr2LHet      368872
Sequence-3    chr2R    21146708
Sequence-4    chr2RHet      3288761
Sequence-5    chr3L    24543557
Sequence-6    chr3LHet      2555491
Sequence-7    chr3R    27905053
......
```

Data Processing - 1

Trimming

Alignment

Build Pairs -
Prepare Size File

Docker image:

- ubuntu:24.04

CLI: cutadapt

- Input: 2 fastq / adapter sequence / score threshold / length threshold
- Output:

trimmed_reads_SRR5579177_1.fastq (forward)
trimmed_reads_SRR5579177_2.fastq (backward)

Docker image:

ubuntu:24.04

CLI: bowtie

- Input: 2 fastq / dm3_index / output SAM format / only unique alignment
- Output: alignment.sam

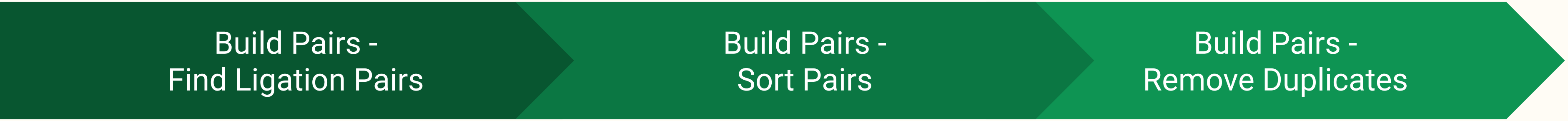
Docker image:

ubuntu:24.04

CLI: wget

- Output: dm3.chrom.sizes

Data Processing - 2



Build Pairs -
Find Ligation Pairs

Build Pairs -
Sort Pairs

Build Pairs -
Remove Duplicates

Docker image:

- `ubuntu:24.04`

CLI: `pairtools parse`

- Input: `dm3.chrom.sizes / alignment.sam`
- Output: `alignment.pairsam`

Docker image:

`ubuntu:24.04`

CLI: `pairtools sort`

- Input: `alignment.pairsam`
- Output: `sort_alignment.pairsam`

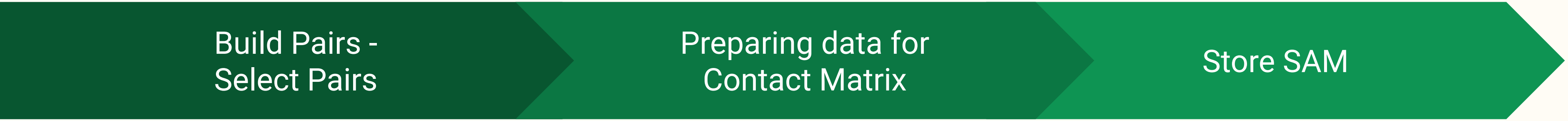
Docker image:

`ubuntu:24.04`

CLI: `pairtools dedup`

- Input: `alignment.pairsam`
- Output: `dedup_alignment.pairsam`

Data Processing - 3



Docker image:

- `ubuntu:24.04`

CLI: `pairtools select`

- Input: `alignment.pairsam / pair type: UU (unique-unique)`
- Output: `alignment.pairs`

Docker image:

`ubuntu:24.04`

Expect Programming: `R`

- Bin: `GSE99104_nm_none_160000.bins.txt`

  Pairs: `alignment.pairs`

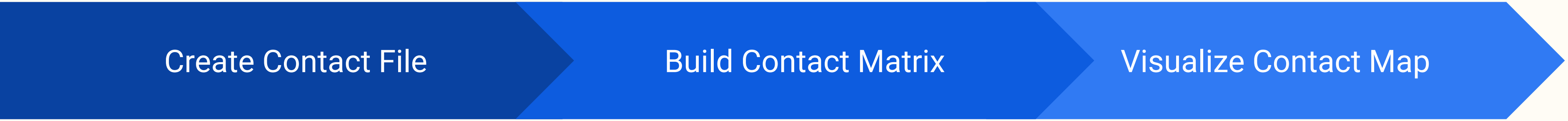
Docker image:

`ubuntu:24.04`

CLI: `samtools view`

- Input: `alignment.sam`
- Output: `alignment.bam`

Visualize Data



Create Contact File

Build Contact Matrix

Visualize Contact Map

Env: windows

Program:

contact_file_generate. R

- Input: GSE99104_nm_none_160000 .bins.txt / alignment.pairs
- Output: n_contact.txt

Env: windows

Program:

contact_file_generate.R

Processing:

- Input: n_contact.txt
- Output: 2L_contact_matrix.txt

Env: windows

Program:

contact_map_generate.R

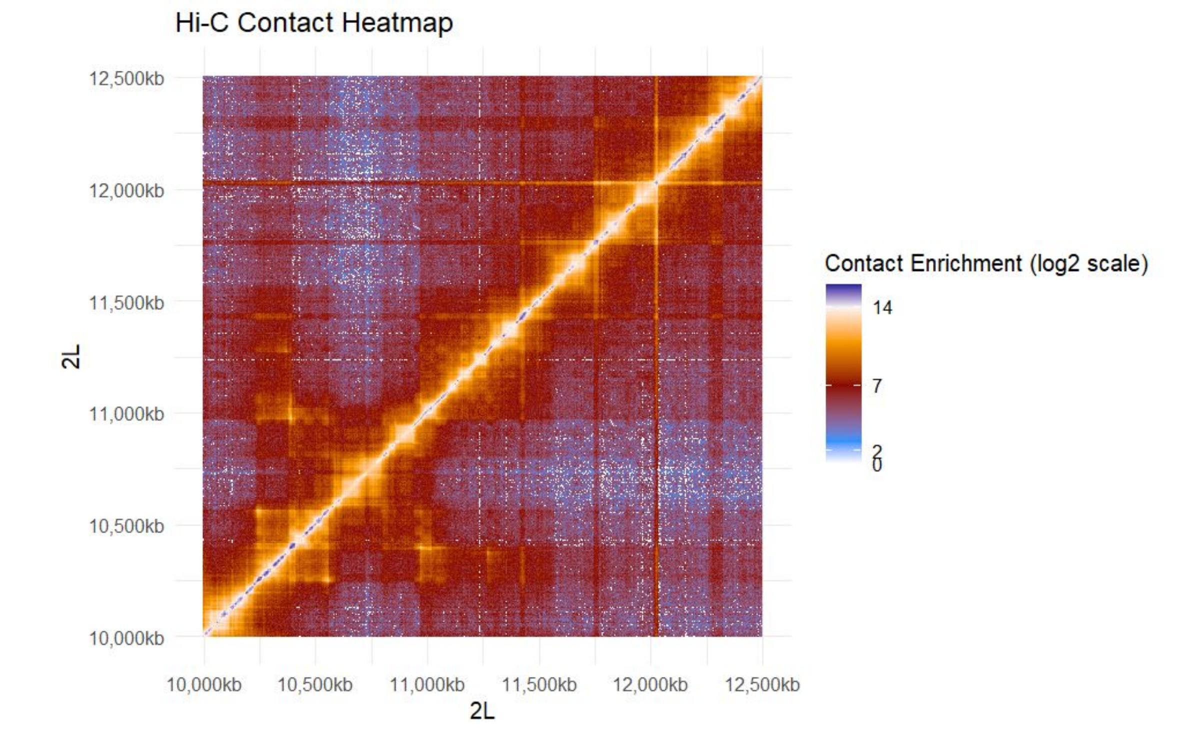
Processing:

Lib: ggplot2 / reshape2

- Input: 2L_contact_matrix.txt
- Output: contact_heatmap.png

Experiment Results

HI-C Contact Map



Data & Used Tools Description

Data Overview - 1

File Types: Source		Actual Files	Sizes
1	SRA: *NCBI/NIH*	`SRR5579177`	● 15.3 GB
2	FASTQ	`SRR5579177_1.fastq` `SRR5579177_2.fastq`	● 68.5 GB Each
3	FASTA: *UCSC Genome Browser*	`dm3.fa`	● 164 MB

4	Bowtie Index	`dm3_index.1.ebwt`
		`dm3_index.2.ebwt`
		`dm3_index.3.ebwt`
		`dm3_index.4.ebwt`
		`dm3_index.rev.1.ebwt`
		`dm3_index.rev.2.ebwt`
			● 1 KB ~ 161 MB

5	SAM	`alignment.sam`	● 115 GB

Data Overview - 2

File Types: Source		Actual Files	Sizes
6	Sizes: *UCSC Genome Browser*	`dm3.chrom.sizes`	● 1 KB

7	PairSAM	`alignment.pairsam`
		`sort_alignment.pairsam`
		`dedup_alignment.pairsam`
			● 133 GB
			● 60.8 GB

8	Pairs	`alignment.pairs`	● 60.8 GB
9	BINS: NCBI/NIH	`GSE99104_nm_none_160000.bins.txt`	● 332 KB

Tools Overview - 1

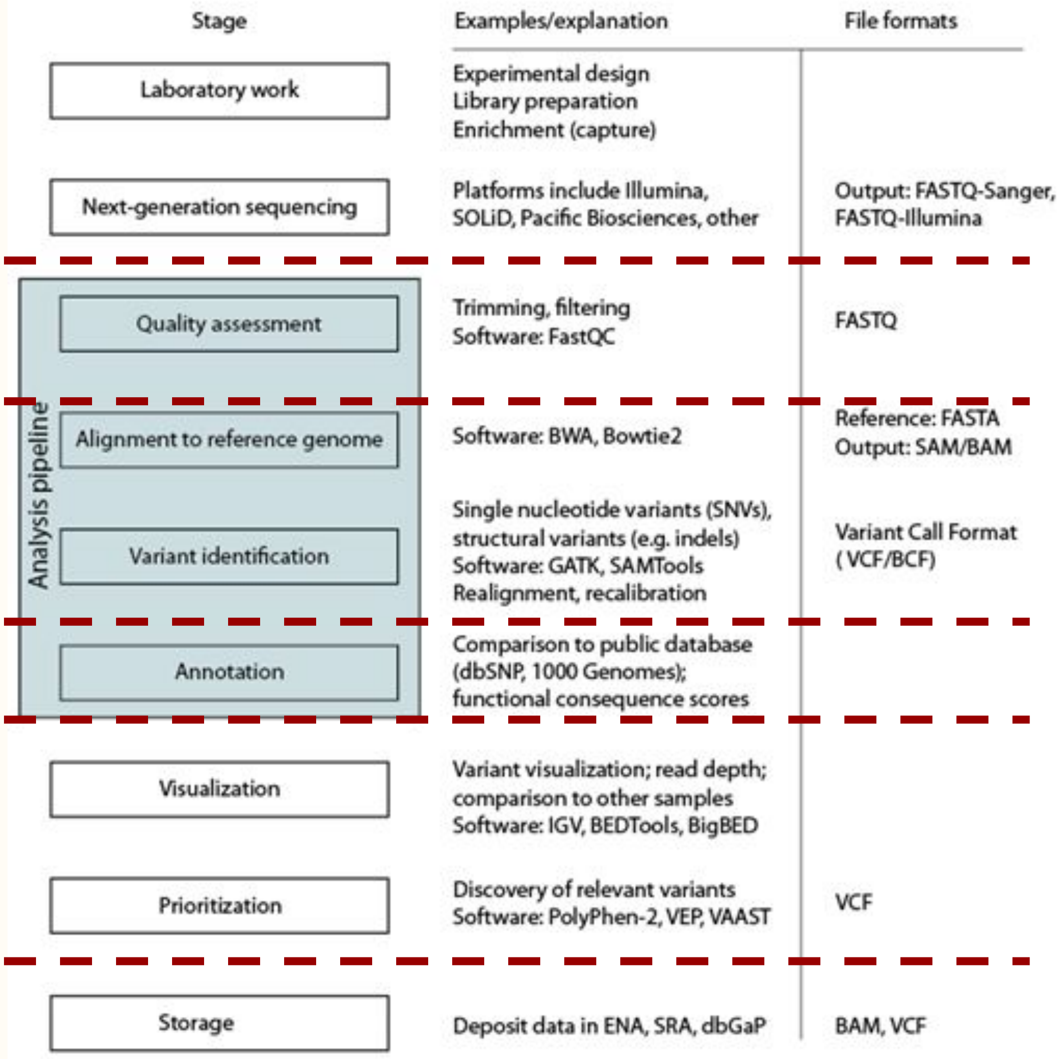
FastQC
cutadapt

Bowtie
samtools

-

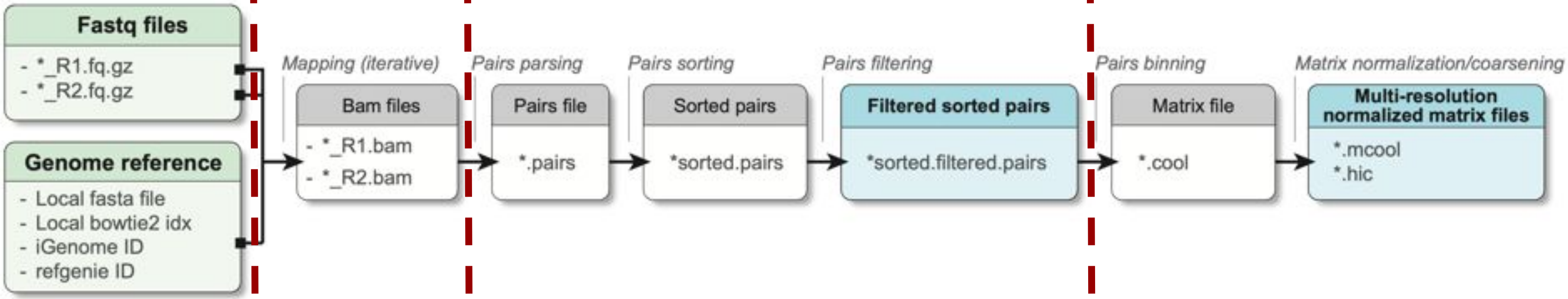
R:
ggplot2
reshape2

samtools



Stage	Examples/explanation	File formats
Laboratory work	Experimental design Library preparation Enrichment (capture)	
Next-generation sequencing	Platforms include Illumina, SOLiD, Pacific Biosciences, other	Output: FASTQ-Sanger, FASTQ-Illumina
Quality assessment	Trimming, filtering Software: FastQC	FASTQ
Alignment to reference genome	Software: BWA, Bowtie2	Reference: FASTA Output: SAM/BAM
Variant identification	Single nucleotide variants (SNVs), structural variants (e.g. indels) Software: GATK, SAMTools Realignment, recalibration	Variant Call Format (VCF/BCF)
Annotation	Comparison to public database (dbSNP, 1000 Genomes); functional consequence scores	
Visualization	Variant visualization; read depth; comparison to other samples Software: IGV, BEDTools, BigBED	
Prioritization	Discovery of relevant variants Software: PolyPhen-2, VEP, VAAST	VCF
Storage	Deposit data in ENA, SRA, dbGaP	BAM, VCF

Tools Overview - 2



sra-tools
wget
bowtie

samtools

pairtools

Cooler CLI
Python cooler Lib

Cooperation

Cooperation

黃 宇秀: Paper, Contact Matrix

邱 淦均: Paper, Contact Map

李 柏漢: Paper, Contact Map

林 穎彥: Data Processing, Docs