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

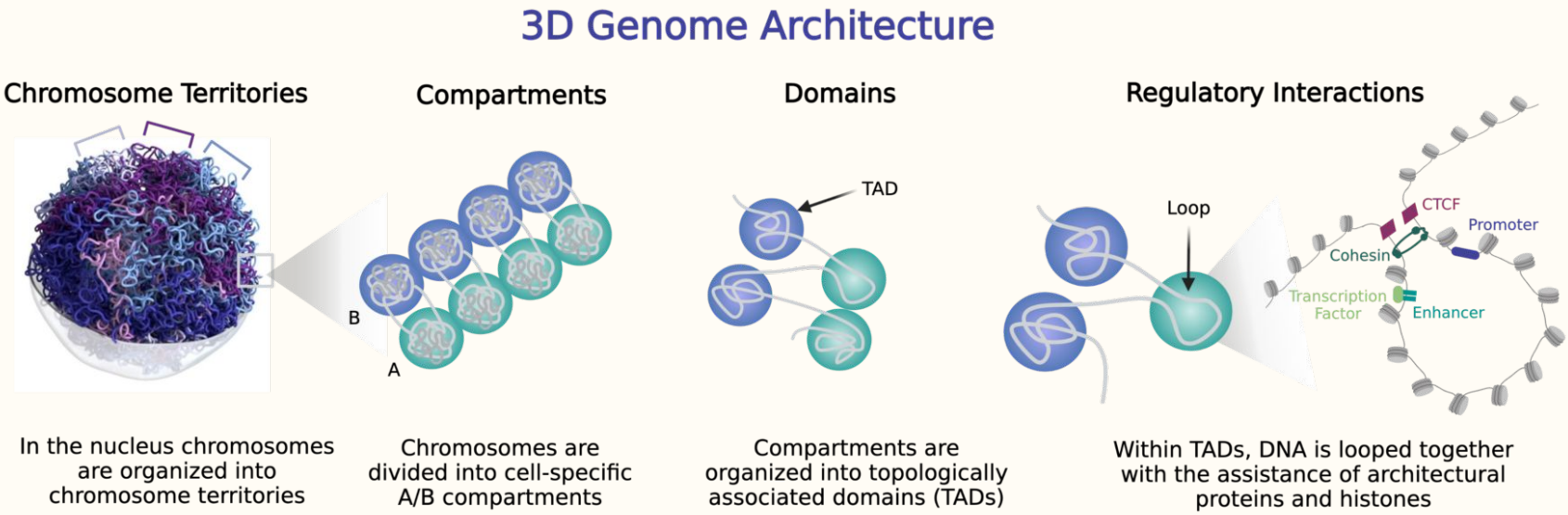
BioInformatics 113
2025.1.2

Table of Contents

Paper Introduction

What is Topologically Associating Domains (TADs)?

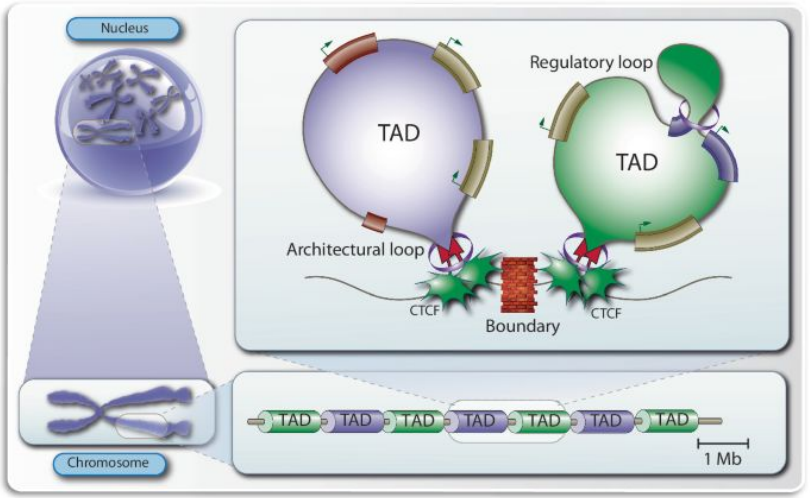
- Fundamental units of the three-dimensional genome structure



What is Topologically Associating Domains (TADs)?

Key features of TADs:

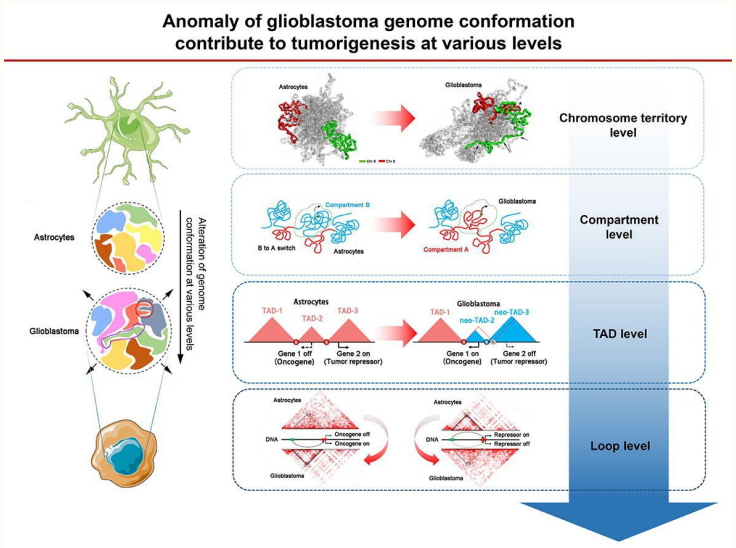
1. **Well-defined boundaries**: TADs are separated by clear boundaries, often marked by specific proteins such as CTCF and structural factors like the cohesin complex.
2. **High internal interactions**: Within a TAD, DNA fragments interact more frequently, facilitating regulatory interactions between 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.

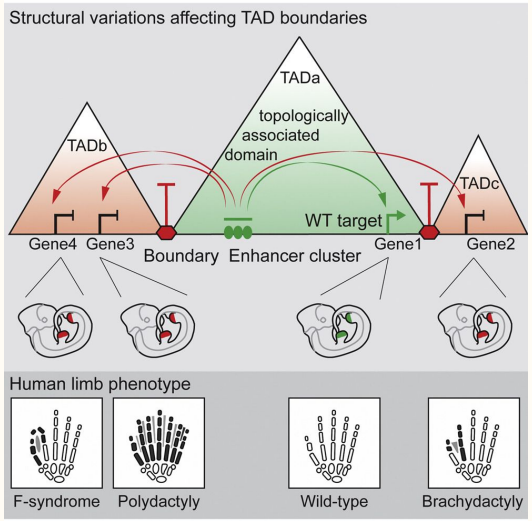


Why Topologically Associating Domains (TADs) so important?

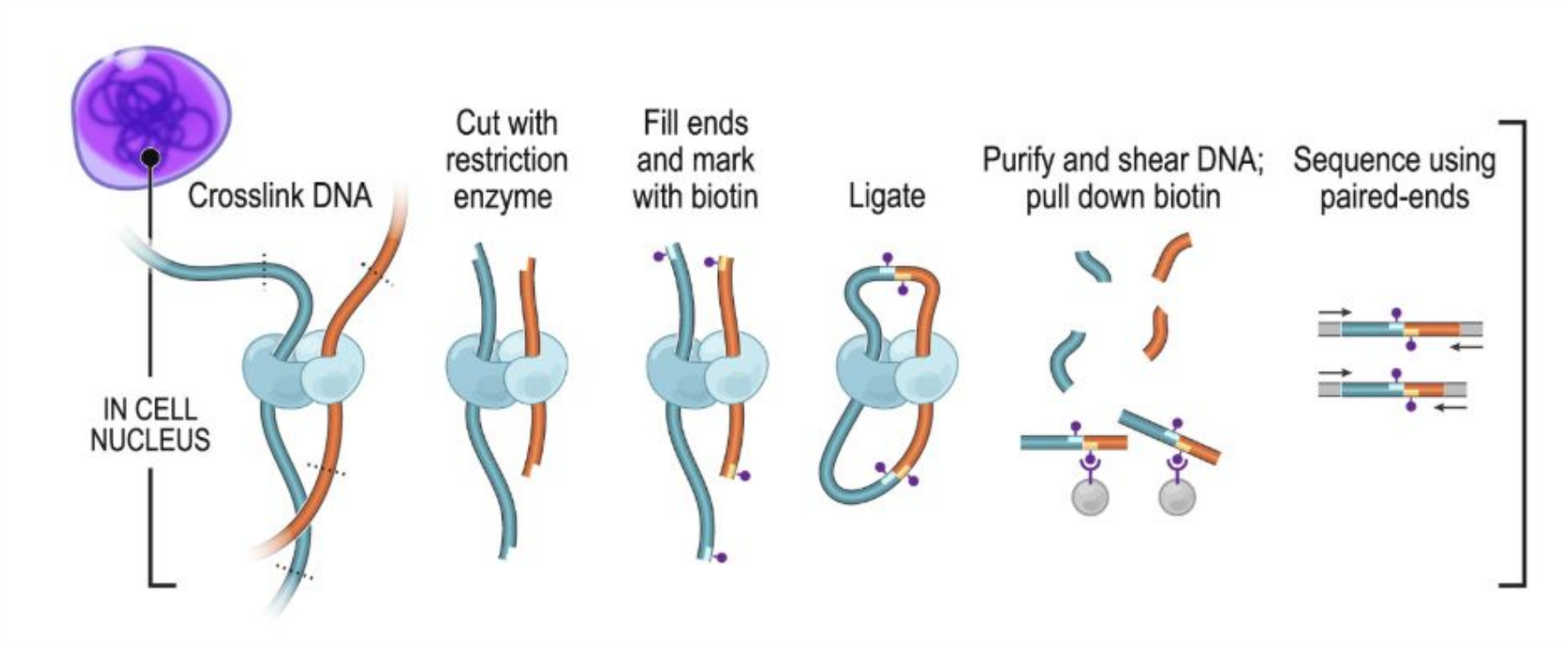
TADs play crucial roles in regulating gene expression, maintaining genome stability, and organizing the chromatin in the nucleus.

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

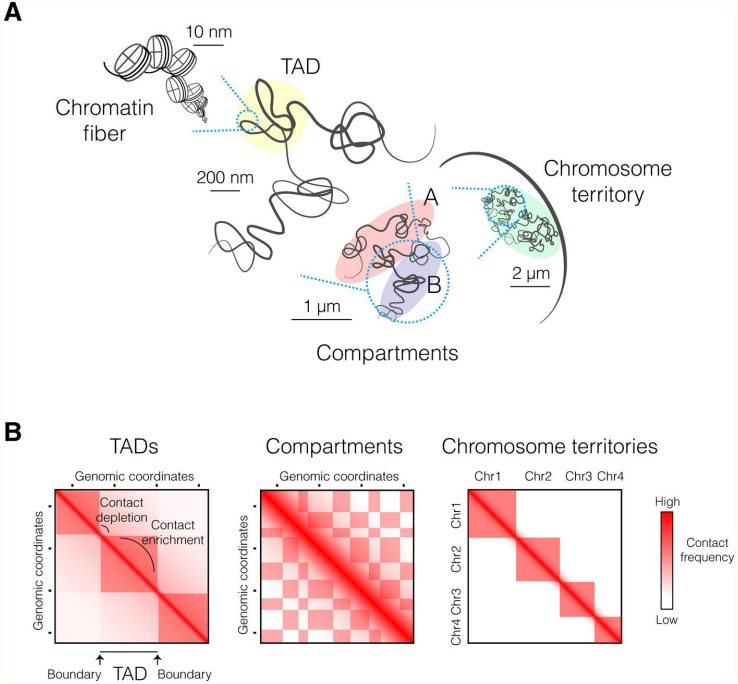


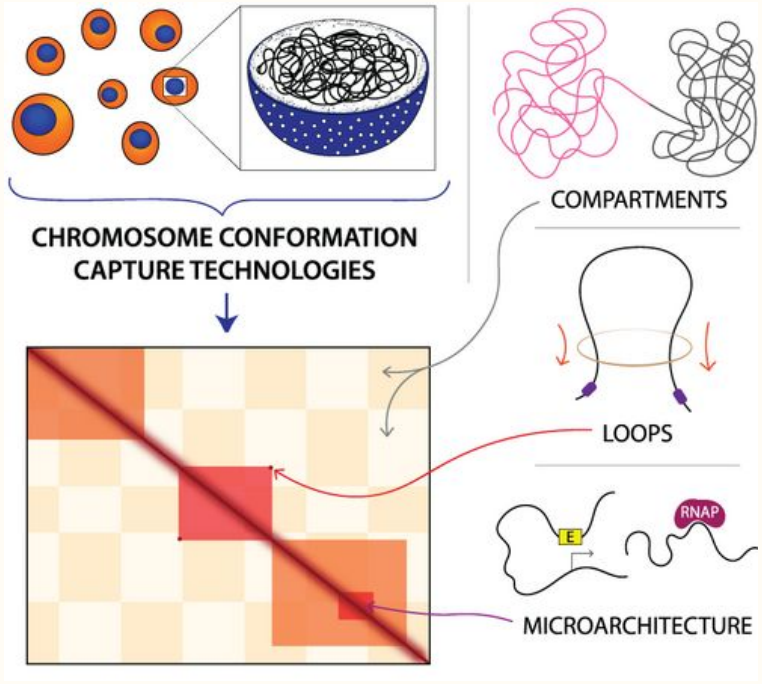


Chromosome Conformation Capture (Hi-C)



What can we tell from the HI-C Map

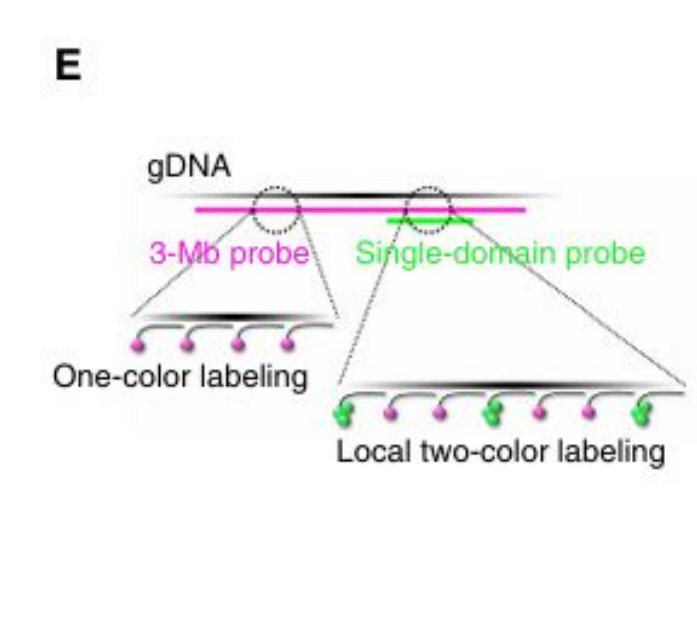




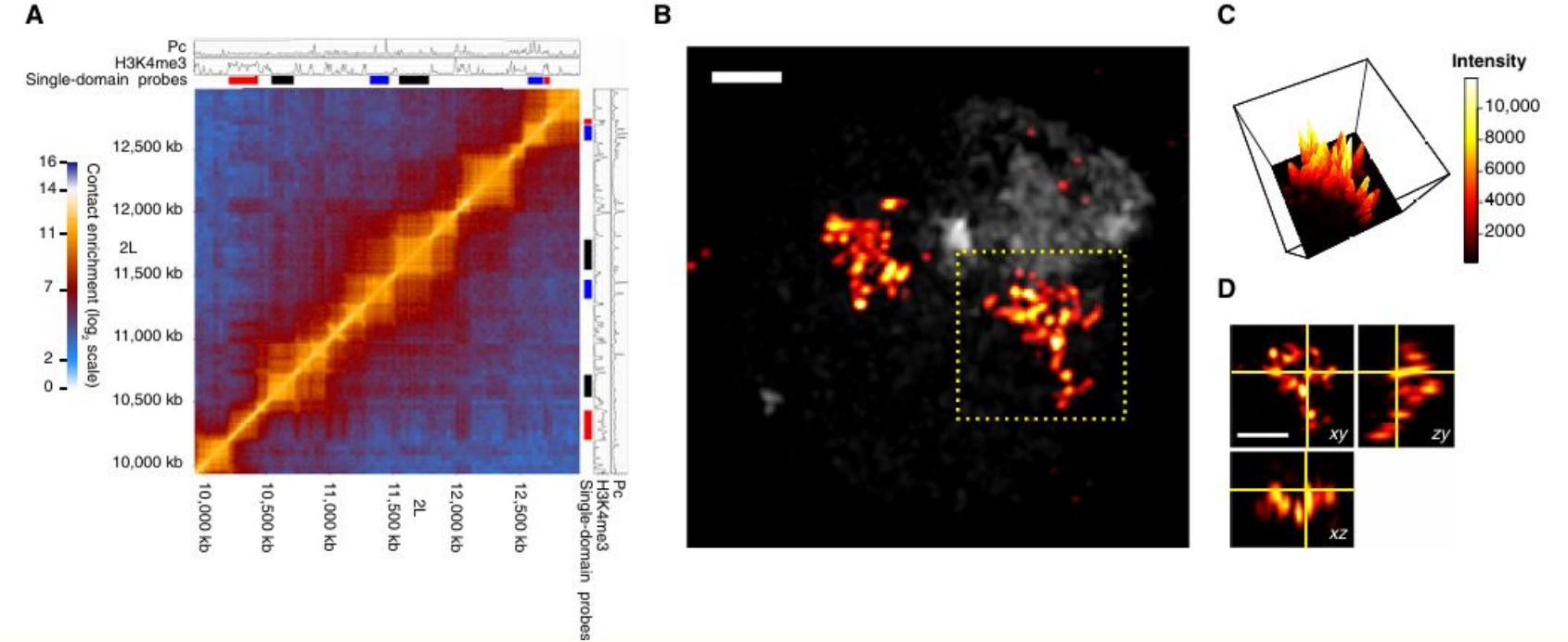
Chromatin is organized in a series of discrete 3D nanocompartments

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

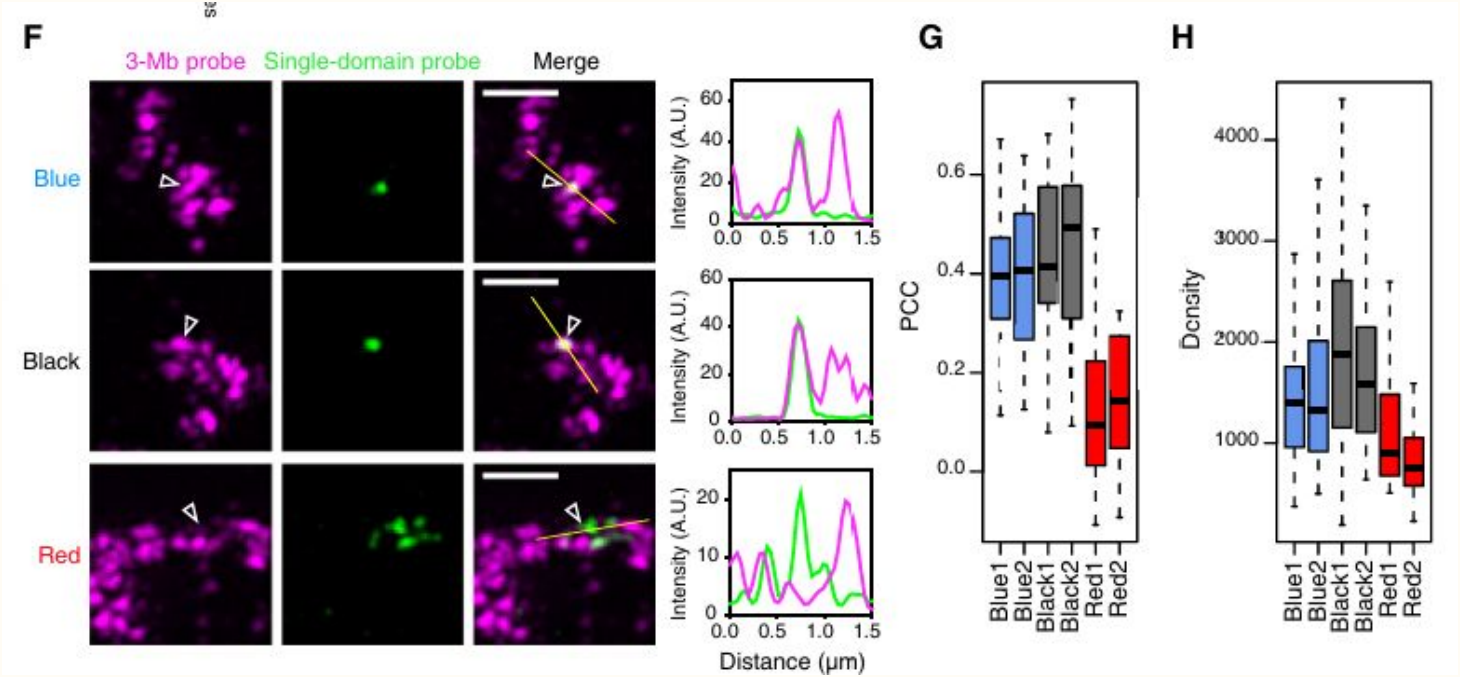
1. active chromatin (Red) enriched in trimethylation of histone 3 lysine 4 (H3K4me3), H3K36me3, and acetylated histones
2. Polycomb group (PcG) protein repressed domains (Blue), defined by the presence of PcG proteins and H3K27me3
3. inactive domains (Black), which are not enriched in specific epigenetic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



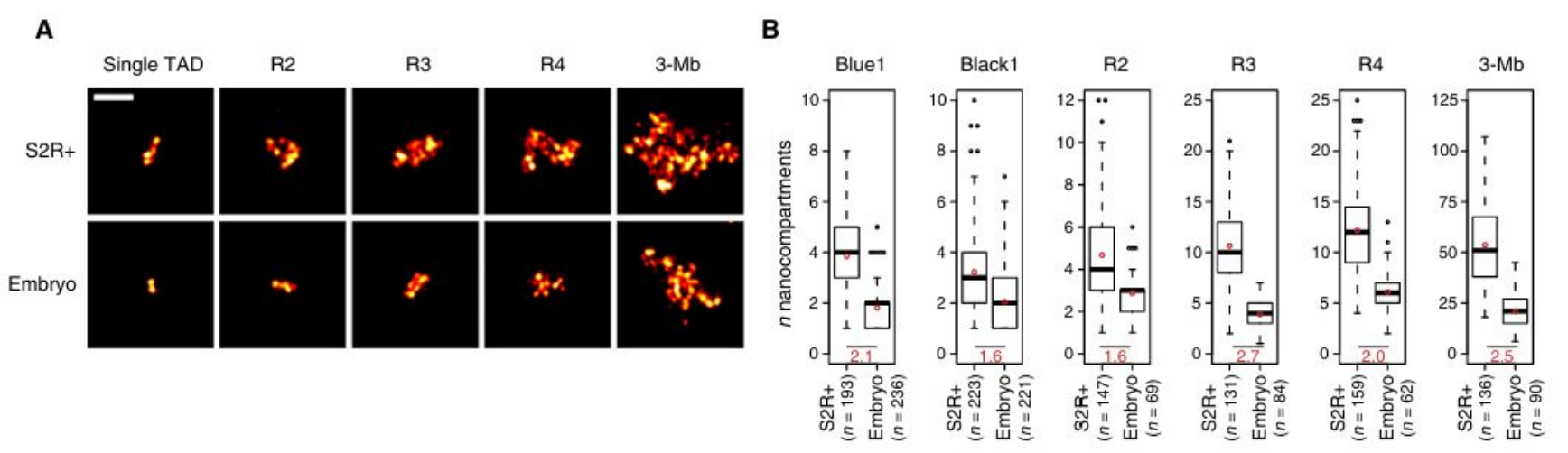
Chromatin is organized in a series of discrete 3D nanocompartments



Chromatin is organized in a series of discrete 3D nanocompartments

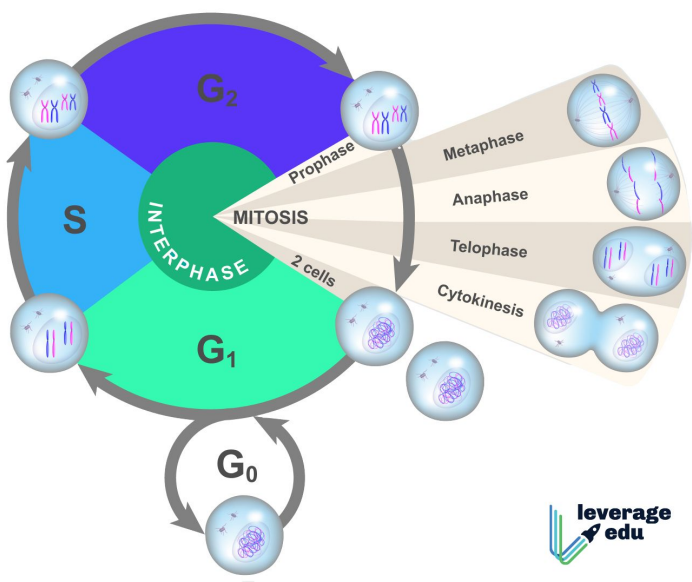


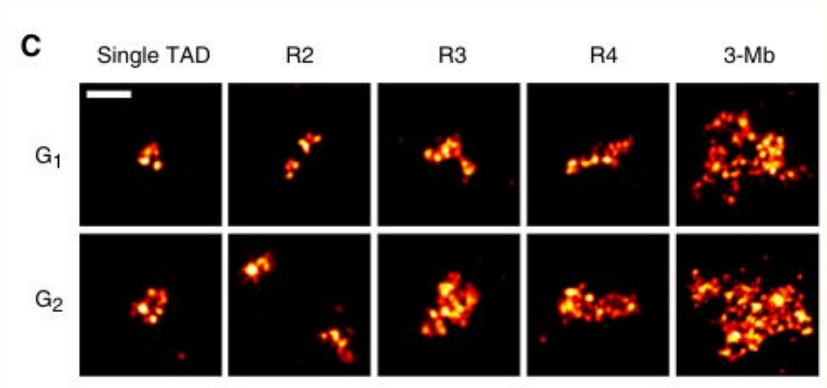
TAD-based 3D nanocompartments undergo dynamic cis and trans contact events

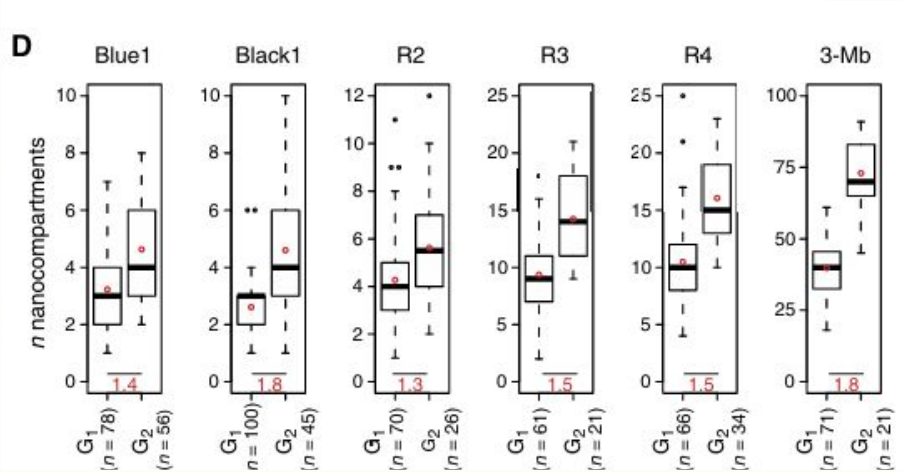


- Tetraploid S2R+ cells versus diploid embryonic (12 to 16 hours) cells
- R2(195kb),R3(805kb),and R4(495kb),covering two,three,and four repressed TADs, respectively

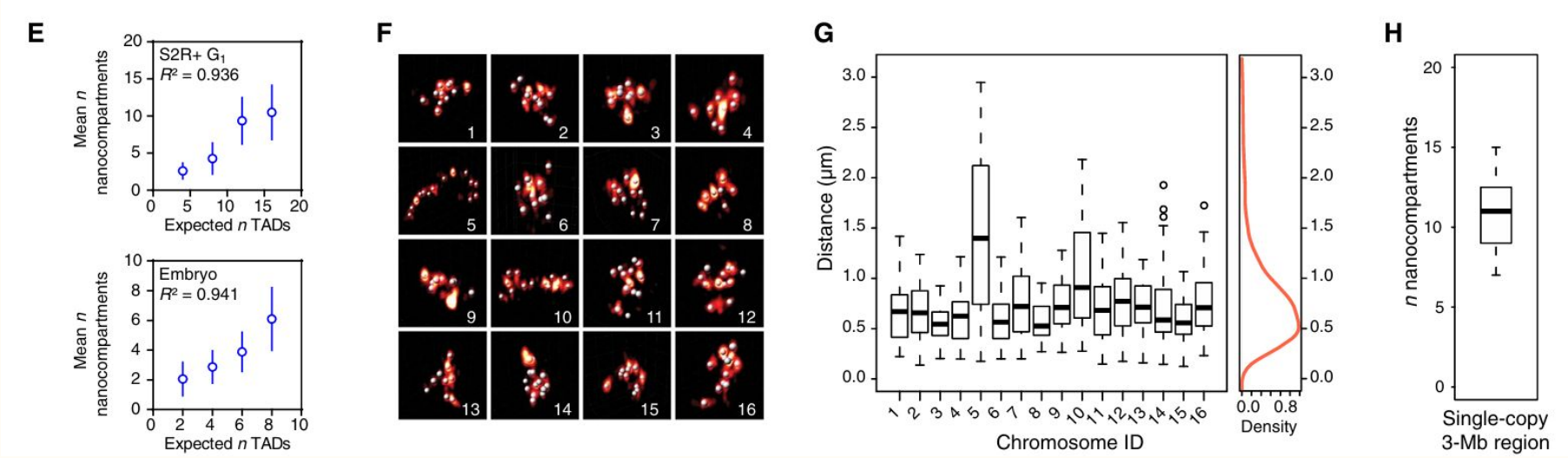
TAD-based 3D nanocompartments undergo dynamic cis and trans contact events





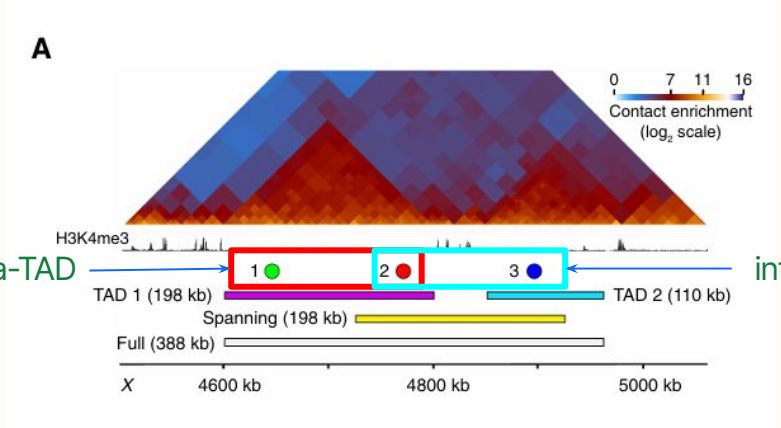


TAD-based 3D nanocompartments undergo dynamic cis and trans contact events



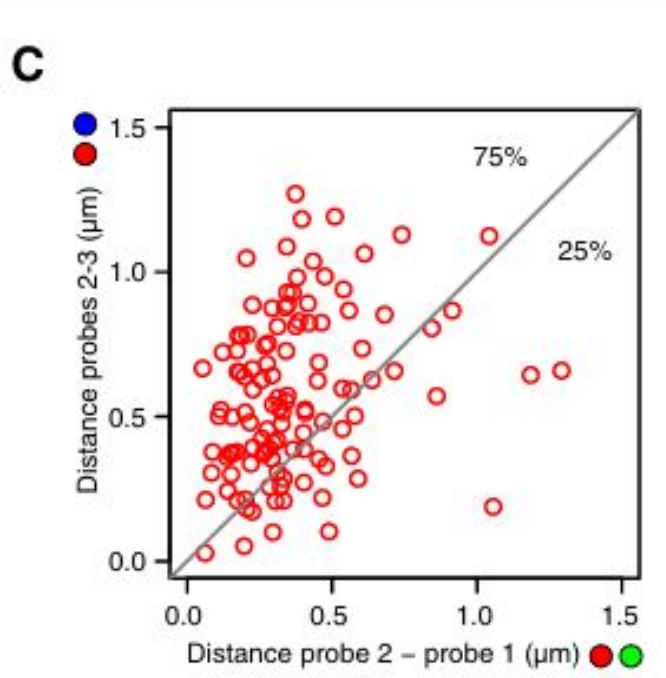
Repressed TADs form physical and structural chromosomal units

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



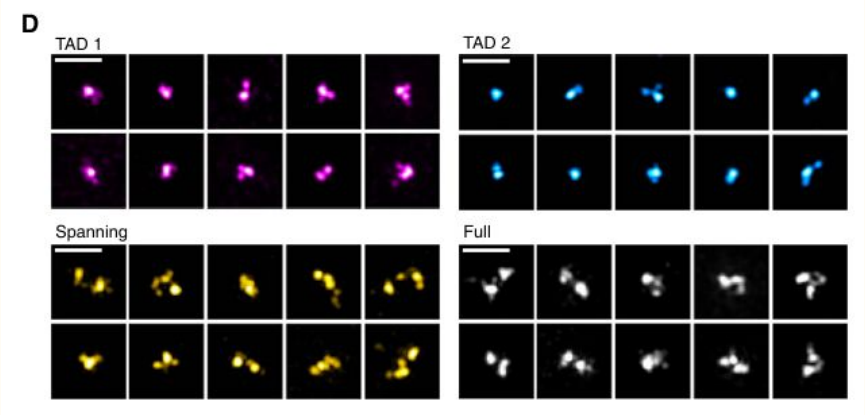
intra-TAD

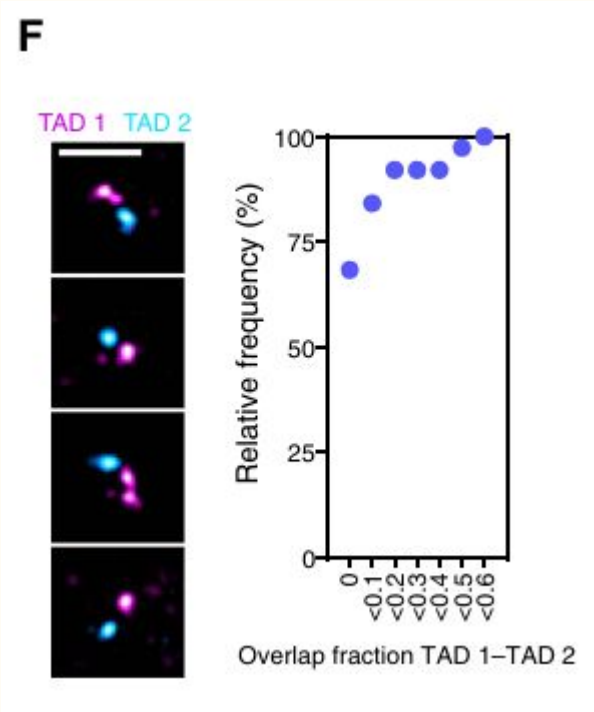
inter-TAD



Repressed TADs form physical and structural chromosomal units

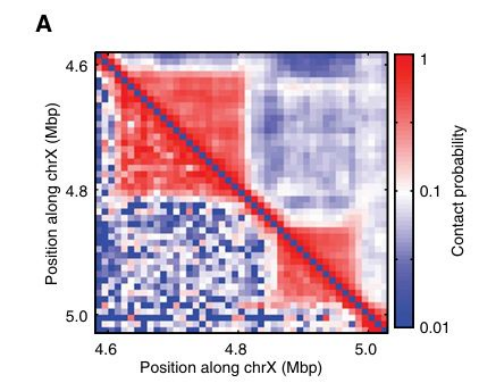
2. Despite variable intra- and inter-TAD contacts in each cell, the physical TAD-based compartmentalization of the chromatin fiber is a general feature of chromosomal domains.

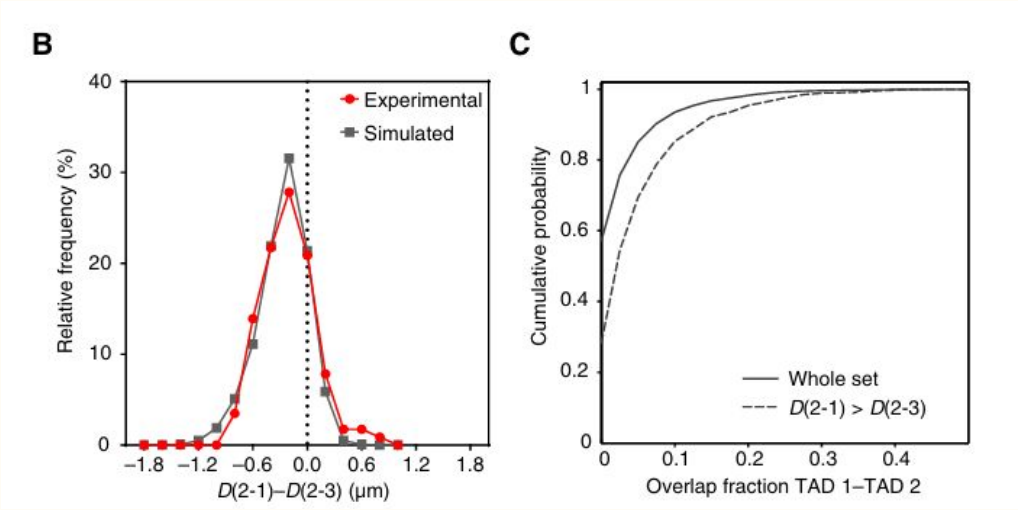




Polymer modeling recapitulates the physical partitioning of chromosomes into TADs

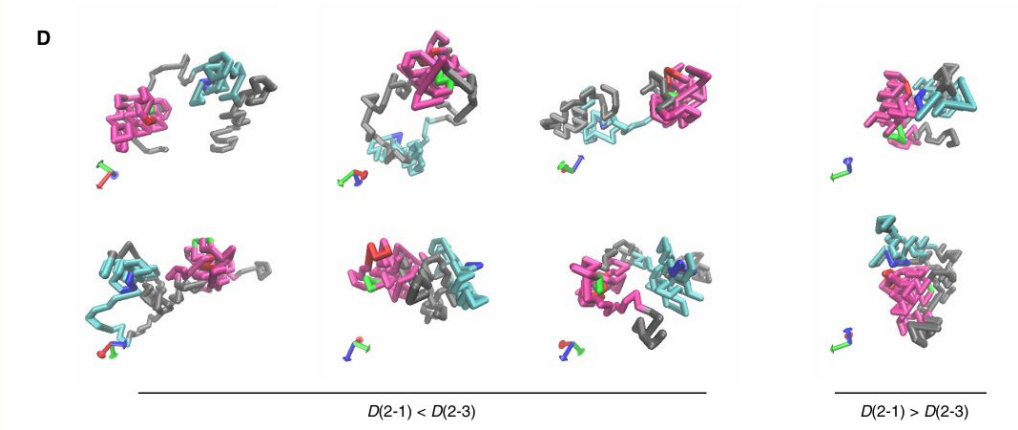
Polymer modeling using parameters that fit Hi-C maps supports the frequent folding of the two TADs into well-separated nanocompartments.





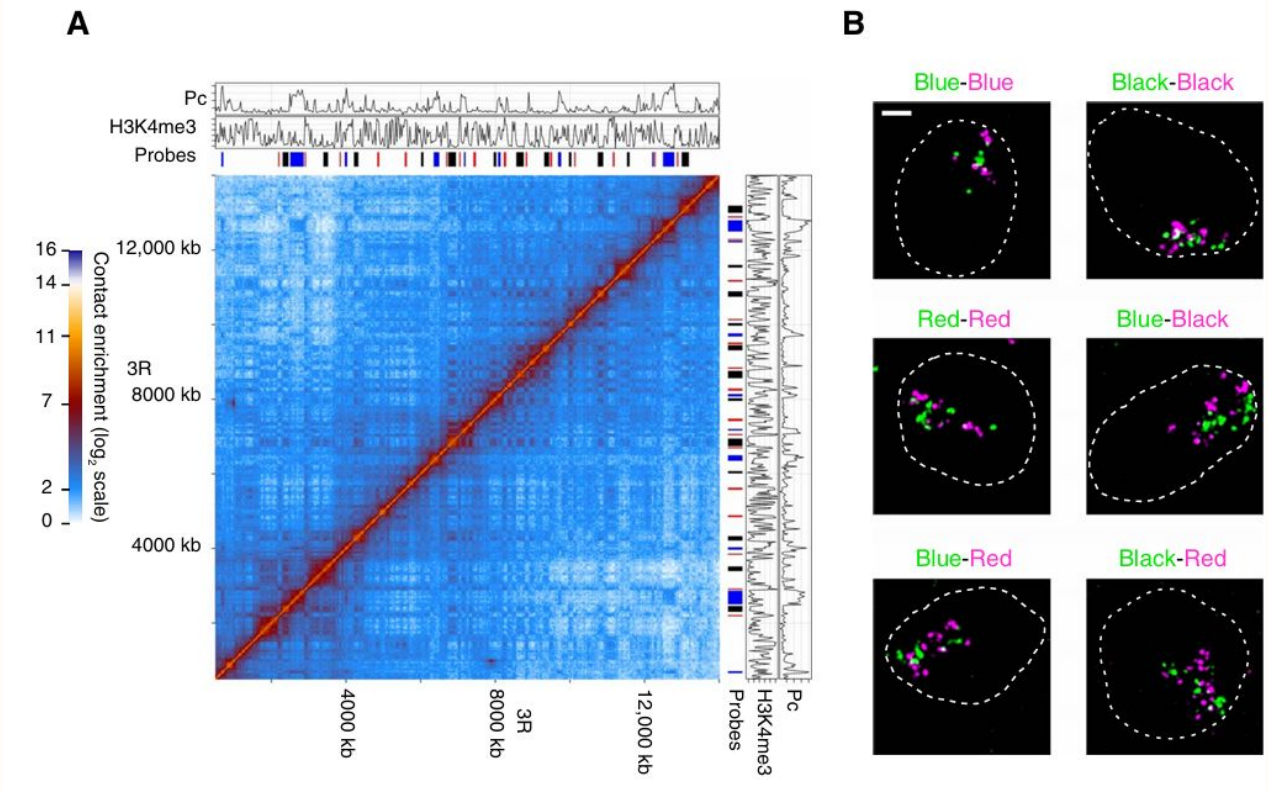
Polymer modeling recapitulates the physical partitioning of chromosomes into TADs

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



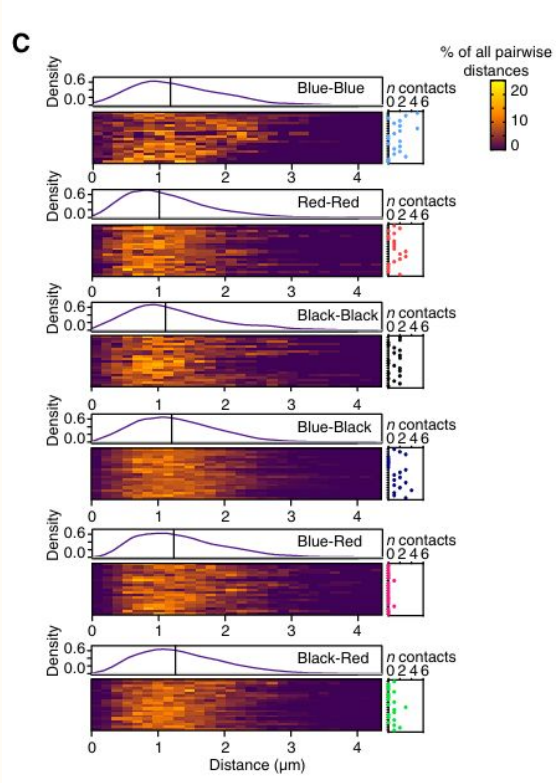
Large-scale chromatin folding reflects highly heterogeneous yet specific, long-range interdomain contacts

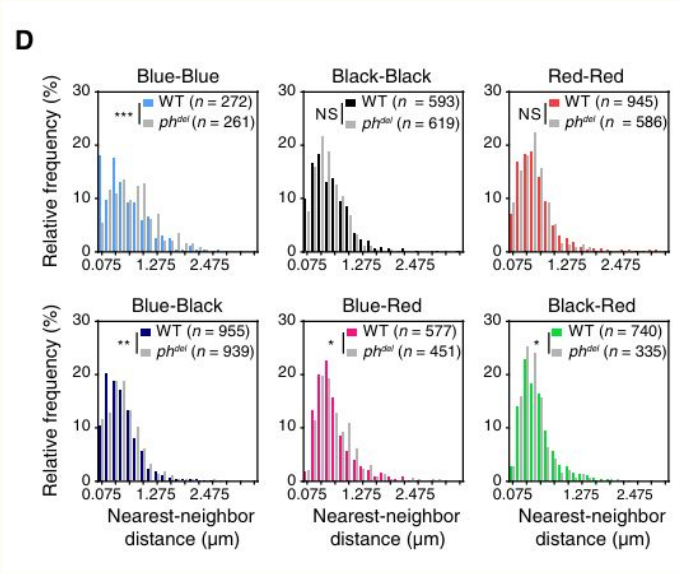
- Sixteen-to 18-hour embryo Hi-C map of a 14-Mb region.
- Labeling chromatin domains of different epigenetic states and studied their relative 3D spatial organization.



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

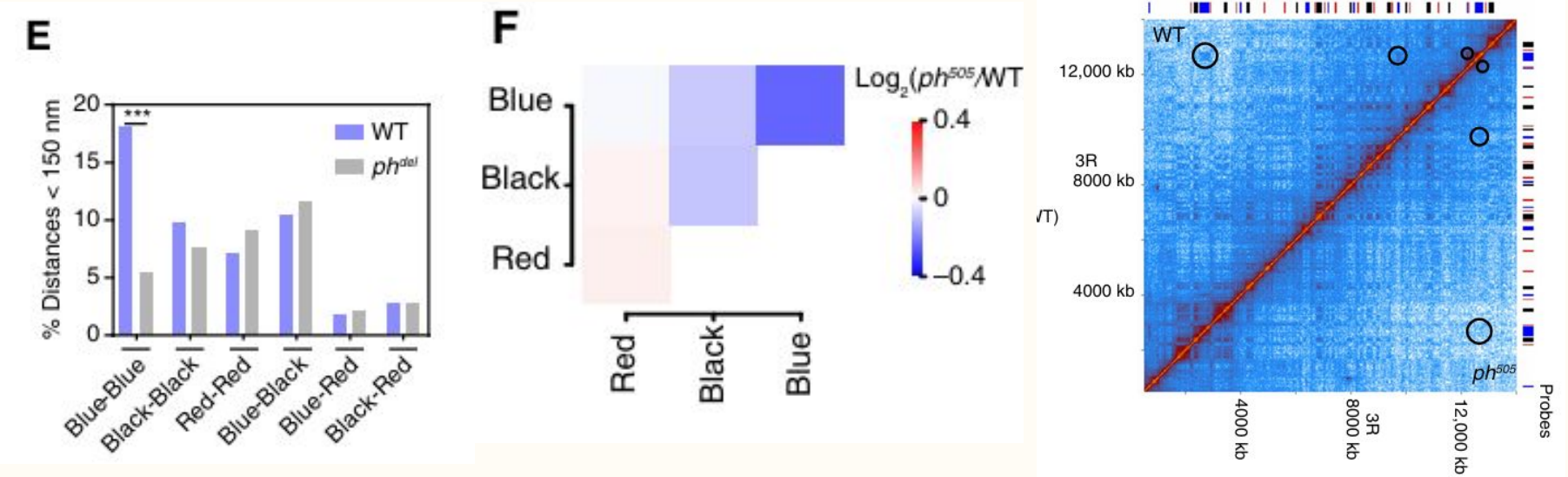
The analysis revealed the presence of discrete interdomain contacts, with preference for contacts among TADs of the same epigenetic type.





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:**

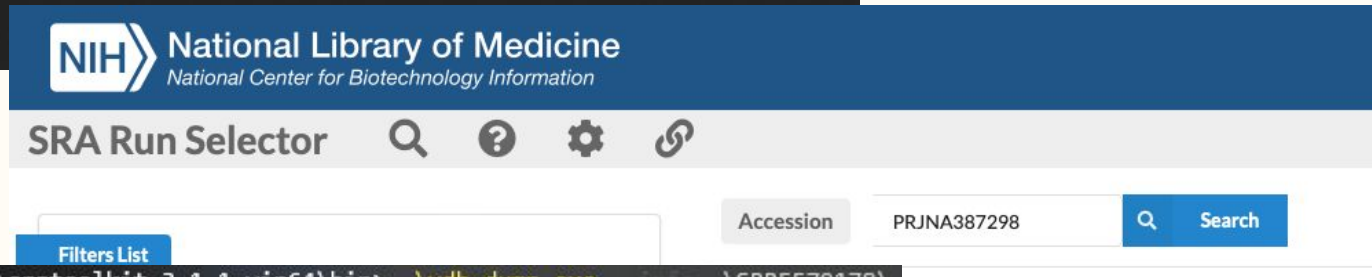
1. Repressed TADs form a succession of discrete nanocompartments.
2. Single-cell analysis revealed stable TAD-based chromatin compartmentalization, with some heterogeneity in intra-TAD conformations and cis/trans inter-TAD contact events.

Sequencing data

Data we are using

Where to download it?

```
PS D:\yy\sratoolkit.3.1.1-win64\sratoolkit.3.1.1-win64\bin> .\fasterq-dump.exe .\SRR5579178\
spots read      : 311,302,433
reads read      : 622,604,866
reads written   : 622,604,866
```




```
PS D:\yy\sratoolkit.3.1.1-win64\sratoolkit.3.1.1-win64\bin> .\vdb-dump.exe --info .\SRR5579178\
acc     : ./SRR5579178/
path    : ./SRR5579178/
type    : Table
platf   : SRA_PLATFORM_ILLUMINA
SEQ     : 311,302,433
SCHEMA  : NCBI:SRA:Illumina:tbl:phred:v2#1.0.4
TIME    : 0x00000000591f4e66 (05/19/2017 19:58)
FMT     : Fastq
FMTVER  : 2.8.2
LDR     : fastq-load.2.8.2
LDRVER  : 2.8.2
LDRDATE: Mar  2 2017 (3/2/2017 0:0)
```

How big in terms of GB? in terms of reads?

	Run	BioProject	BioSample	AvgSpotLen	Bases	Bytes
1	SRR5579160	PRJNA387323	SAMN07146998	65	9.05 G	7.81 Gb
2	SRR5579161	PRJNA387323	SAMN07146998	65	8.76 G	7.57 Gb
3	SRR5579162	PRJNA387323	SAMN07146998	65	7.94 G	6.85 Gb
4	SRR5579163	PRJNA387323	SAMN07146998	65	7.83 G	6.79 Gb
5	SRR5579164	PRJNA387323	SAMN07146998	65	9.11 G	7.83 Gb
6	SRR5579165	PRJNA387323	SAMN07146998	65	8.79 G	7.61 Gb
7	SRR5579166	PRJNA387323	SAMN07146998	65	8.43 G	7.29 Gb
8	SRR5579167	PRJNA387323	SAMN07146997	98	19.36 G	11.73 Gb
9	SRR5579168	PRJNA387323	SAMN07146997	98	17.35 G	10.54 Gb
10	SRR5579169	PRJNA387323	SAMN07146997	98	17.63 G	10.84 Gb
11	SRR5579170	PRJNA387324	SAMN07147000	98	15.74 G	9.56 Gb
12	SRR5579171	PRJNA387324	SAMN07147000	98	15.50 G	9.40 Gb
13	SRR5579172	PRJNA387324	SAMN07147000	98	15.73 G	9.57 Gb
14	SRR5579173	PRJNA387324	SAMN07147000	98	15.70 G	9.53 Gb
15	SRR5579174	PRJNA387324	SAMN07146999	98	17.43 G	10.51 Gb
16	SRR5579175	PRJNA387324	SAMN07146999	98	17.34 G	10.43 Gb
17	SRR5579176	PRJNA387324	SAMN07146999	98	17.70 G	10.83 Gb
18	SRR5579177	PRJNA387300	SAMN07147001	100	30.13 G	15.40 Gb
19	SRR5579178	PRJNA387300	SAMN07147002	100	31.13 G	16.21 Gb

Any other data besides sequencing data?

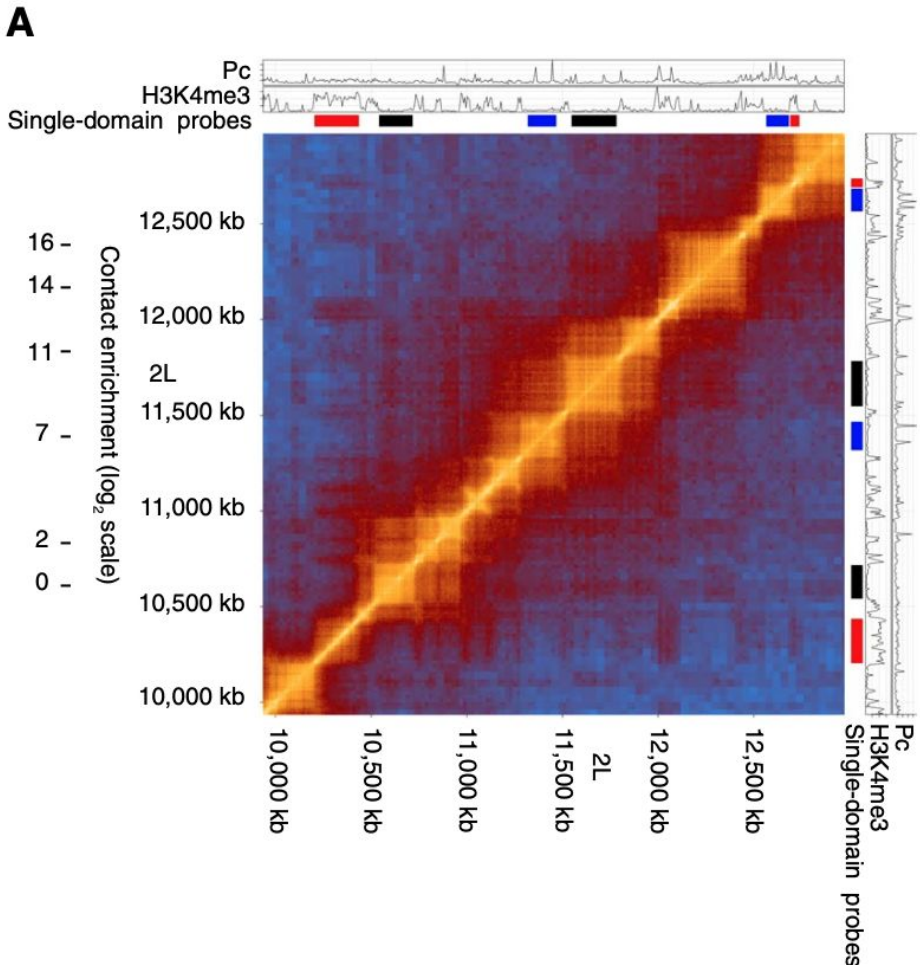
Supplementary file	Size
GSE99106_nm_none_10000.bins.txt.gz	92.7 Kb
GSE99106_nm_none_10000.n_contact.txt.gz	115.6 Mb
GSE99106_nm_none_160000.bins.txt.gz	6.0 Kb
GSE99106_nm_none_160000.n_contact.txt.gz	4.9 Mb
GSE99106_nm_none_20000.bins.txt.gz	46.6 Kb
GSE99106_nm_none_20000.n_contact.txt.gz	46.4 Mb
GSE99106_nm_none_40000.bins.txt.gz	23.3 Kb
GSE99106_nm_none_40000.n_contact.txt.gz	65.4 Mb
GSE99106_nm_none_5000.bins.txt.gz	174.7 Kb
GSE99106_nm_none_5000.n_contact.txt.gz	233.9 Mb
GSE99106_nm_none_80000.bins.txt.gz	12.0 Kb
GSE99106_nm_none_80000.n_contact.txt.gz	17.9 Mb

SRA Run Selector

cbin	chr	from.coord	to.coord	count
1	2L	5000	10000	16
2	2L	10000	15000	34
3	2L	15000	20000	11
4	2L	20000	25000	15
5	2L	25000	30000	26
6	2L	30000	35000	30
7	2L	35000	40000	16
8	2L	40000	45000	20
9	2L	45000	50000	3
10	2L	50000	55000	11
11	2L	55000	60000	18
12	2L	60000	65000	9
13	2L	65000	70000	37
14	2L	70000	75000	12
15	2L	75000	80000	26
16	2L	80000	85000	22
17	2L	85000	90000	26
18	2L	90000	95000	21
19	2L	95000	100000	45
20	2L	100000	105000	36
21	2L	105000	110000	26
22	2L	110000	115000	26
23	2L	115000	120000	26
24	2L	120000	125000	27
25	2L	125000	130000	14
26	2L	130000	135000	23
27	2L	135000	140000	35
28	2L	140000	145000	44
29	2L	145000	150000	26
30	2L	150000	155000	34
31	2L	155000	160000	32
32	2L	160000	165000	8
33	2L	165000	170000	12
34	2L	170000	175000	24
35	2L	175000	180000	8
36	2L	180000	185000	23
37	2L	185000	190000	38

cbin1	cbin2	expected_count	observed_count
1	1	0.018053	41
1	2	0.077088	106
1	3	0.029404	42
1	4	0.036005	12
1	5	0.07268 15	
1	6	0.077878	19
1	7	0.03885 16	
1	8	0.059693	12
1	9	0.009261	5
1	10	0.029263	6
1	11	0.049696	10
1	12	0.027688	6
1	13	0.093496	10
1	14	0.035852	9
1	15	0.066849	8
1	16	0.061525	5
1	17	0.066887	8
1	18	0.056337	5
1	19	0.114289	4
1	20	0.090037	4
1	21	0.063825	6
1	22	0.06549 3	
1	23	0.068814	4
1	24	0.078689	6
1	25	0.040473	4
1	26	0.051058	2
1	27	0.087299	1
1	28	0.1121 3	
1	29	0.063572	3
1	30	0.082776	3
1	31	0.087351	2

Experiments

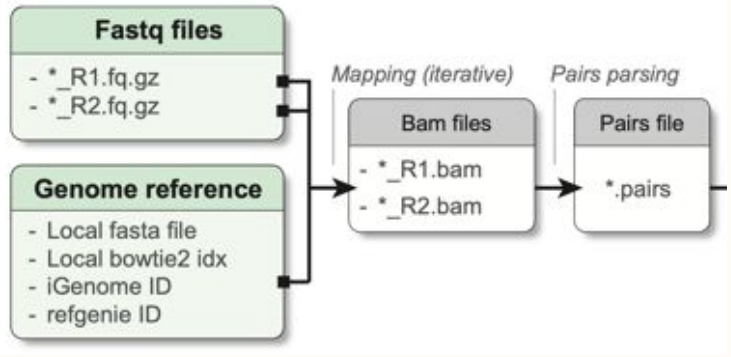


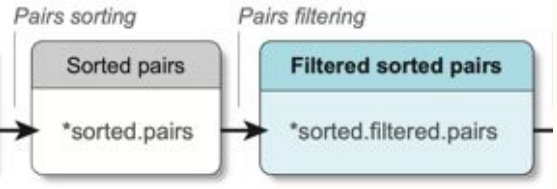
Experiment Objectives:

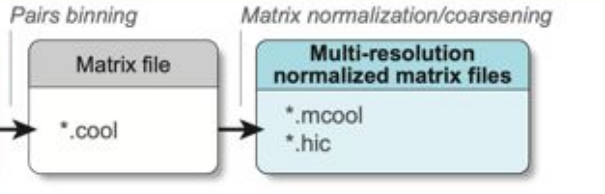
What we want to recreate?

Figure 1A
Hi-C Contact Map

NGS Workflow

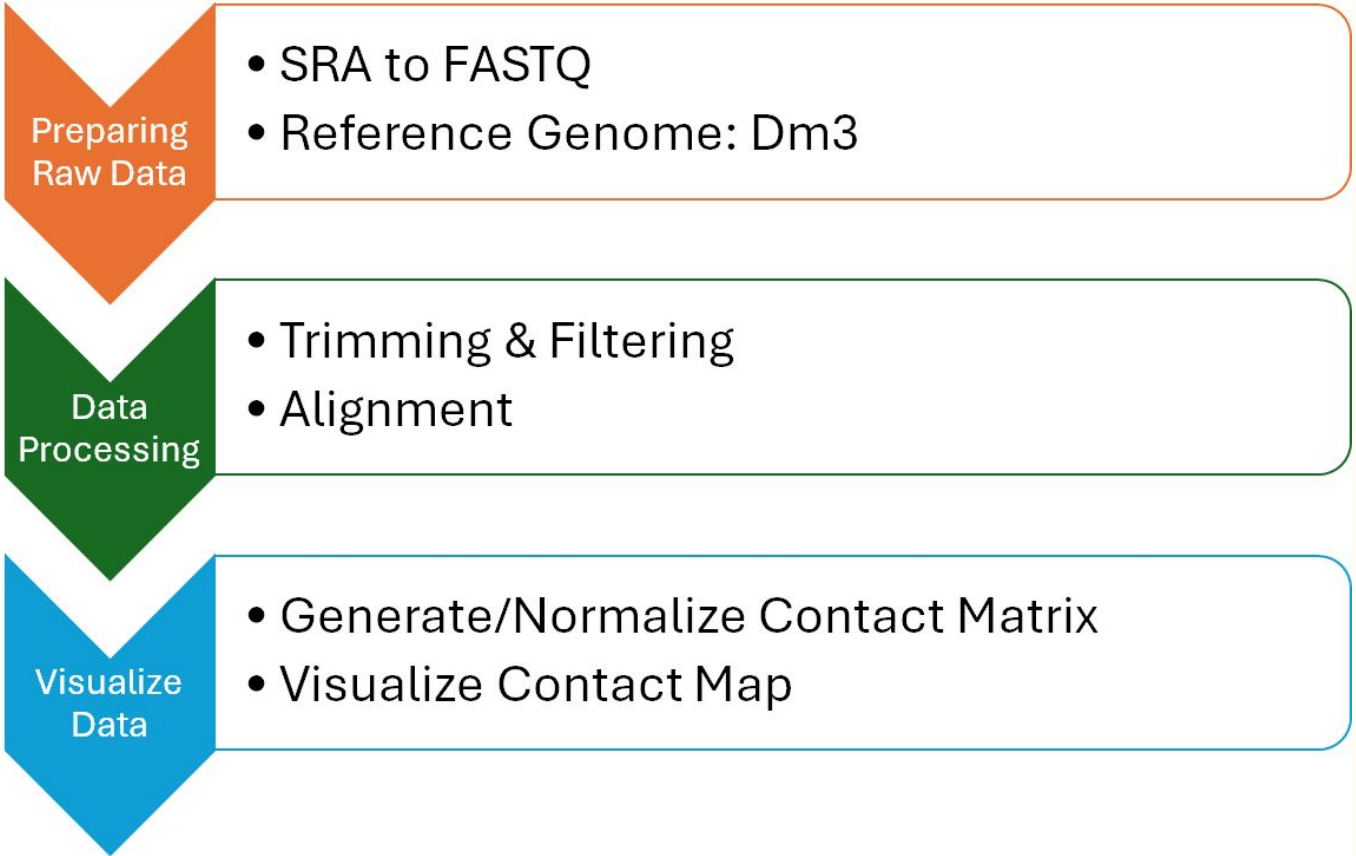




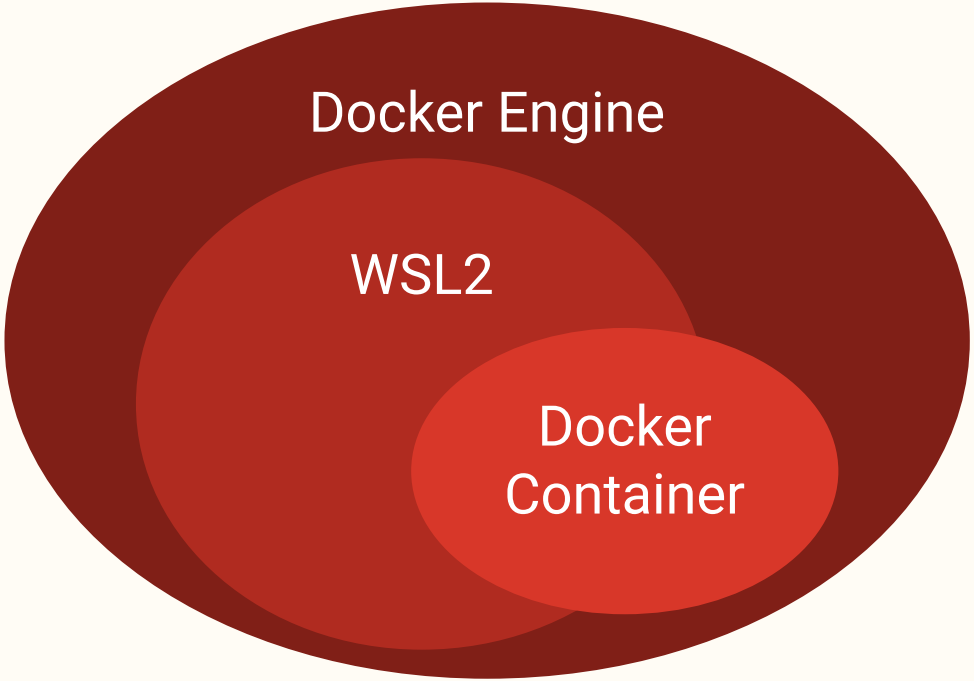


Stage	Examples/explanation	File formats
Laboratory work	Experimental design Library preparation Enrichment (capture)	
Next-generation sequencing	Platforms include Illumina, SOLiD, Pacific Biosciences, other	Output: FASTQ-Sanger, FASTQ-Illumina
Analysis pipeline: Quality assessment	Trimming, filtering Software: FastQC	FASTQ
Analysis pipeline: Alignment to reference genome	Software: BWA, Bowtie2	Reference: FASTA Output: SAM/BAM
Analysis pipeline: Variant identification	Single nucleotide variants (SNVs), structural variants (e.g. indels) Software: GATK, SAMTools Realignment, recalibration	Variant Call Format (VCF/BCF)
Analysis pipeline: 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



Environment - Docker with WSL



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 → Build Bowtie Index → Check Index

Docker image:
- `ubuntu:24.04`

CLI: `wget / gunzip`
- Input: `dm3.fa.gz`
- Output: `dm3.fa`

(Drosophila melanogaster: fruit fly)

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`

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

Docker image:
- `ubuntu:24.04`

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

Build Pairs - Sort Pairs

Docker image:
`ubuntu:24.04`

CLI: `pairtools sort`
- Input: `alignment.pairsam`
- Output: `sort_alignment.pairsam`

Build Pairs - Remove Duplicates

Docker image:
`ubuntu:24.04`

CLI: `pairtools dedup`
- Input: `alignment.pairsam`
- Output: `dedup_alignment.pairsam`

Data Processing - 3

Build Pairs - Select Pairs

Docker image:
- `ubuntu:24.04`

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

Preparing data for Contact Matrix

Docker image:
`ubuntu:24.04`

Expect Programming: `R`
- Bin: `GSE99104_nm_none_160000.bins.txt`

Pairs: `alignment.pairs`

Store SAM

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

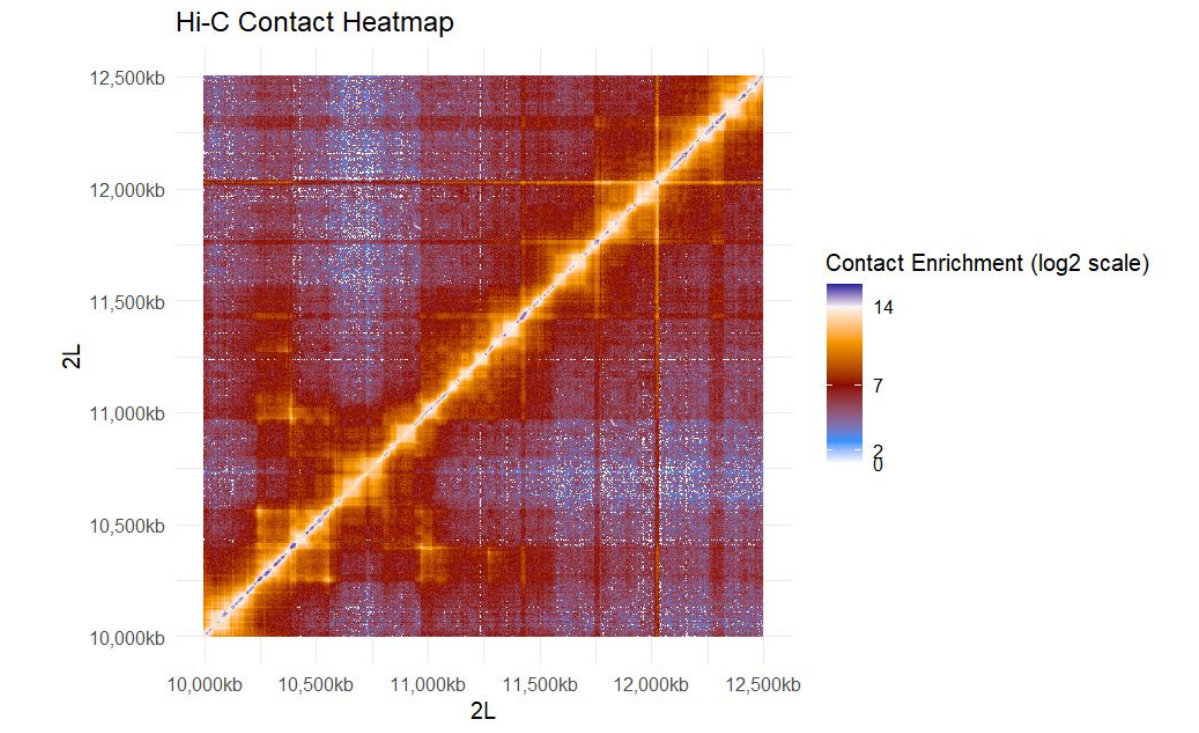
Env: windows
Program:
contact_file_generate. R
- Input:
GSE99104_nm_none_160000
.bins.txt /
alignment.pairs
- Output:
n_contact.txt

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

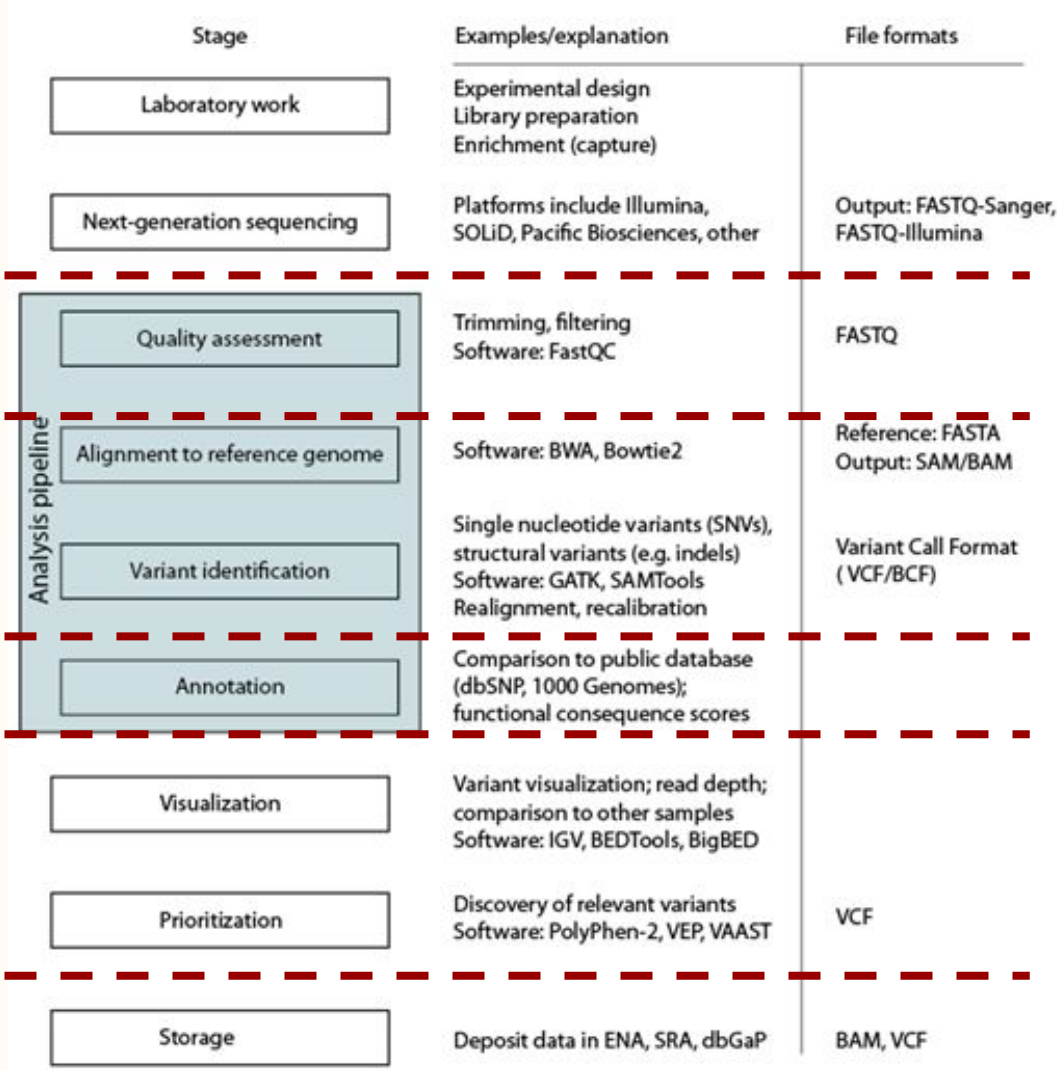
FastQC
cutadapt

Bowtie
samtools

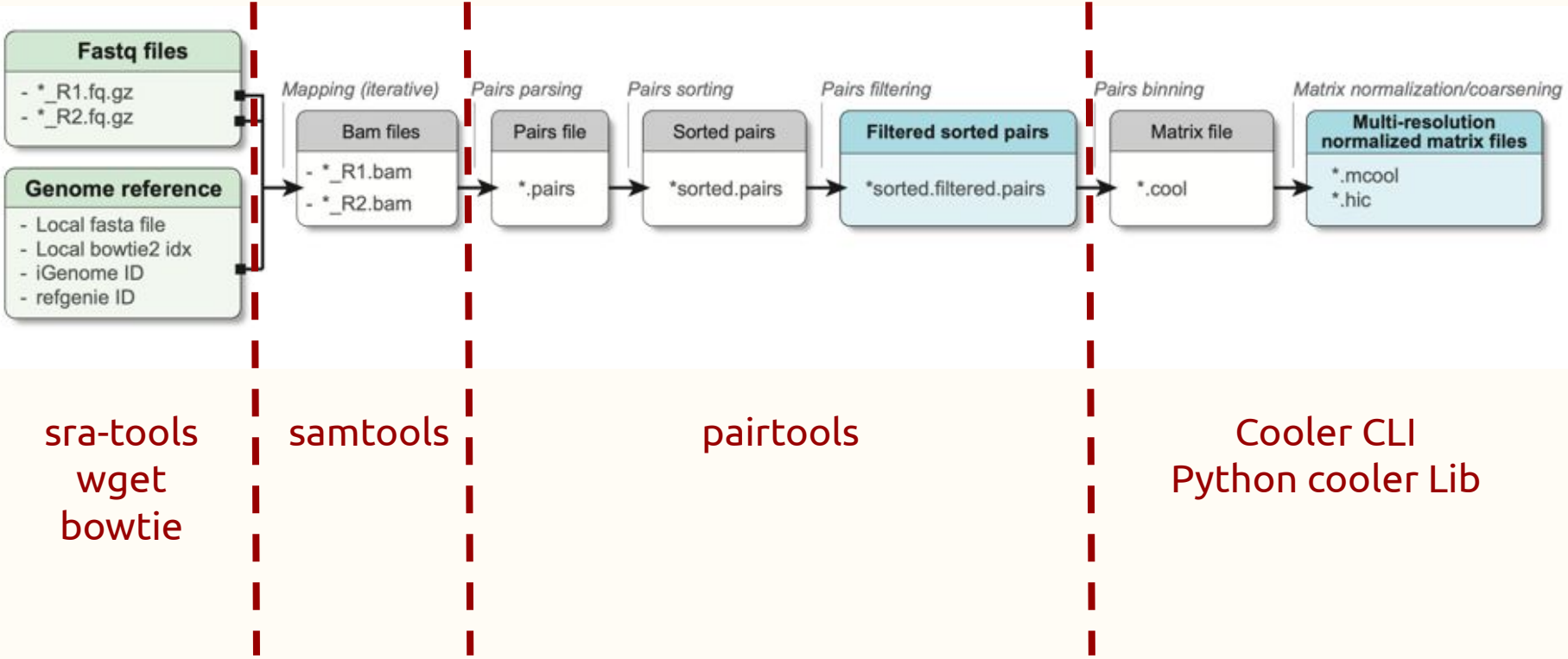
-

R:
ggplot2
reshape2

samtools



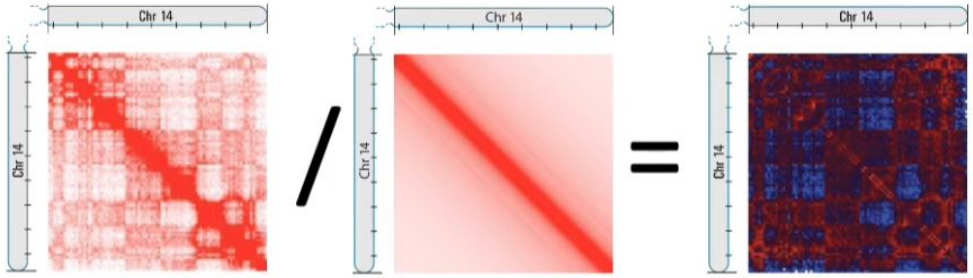
Tools Overview - 2



Paper Data Processing

bowtie → discard Unmapped or non-unique mapped reads → divide to 4 groups
SS: reads mapping to the same restriction fragment, or to two adjacent restriction fragments)
S2: reads in which both ends mapped precisely on the DpnII site
S1: reads in which one of the ends mapped precisely on the DpnII site
S0: reads in which both ends did not map precisely to a DpnII site
→ discard SS, S2, S1 →

Interpreting Hi-C data: normalization



- The observed read counts (*O*) are typically normalized using a second matrix with expected read counts (*E*).
- Matrix *E* is derived by calculating average read counts as a function of genomic distance.
- This results in observed/expected ratios (*R*), indicating which interactions are enriched/depleted in the data.

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