Group 3

paper 2 presentation

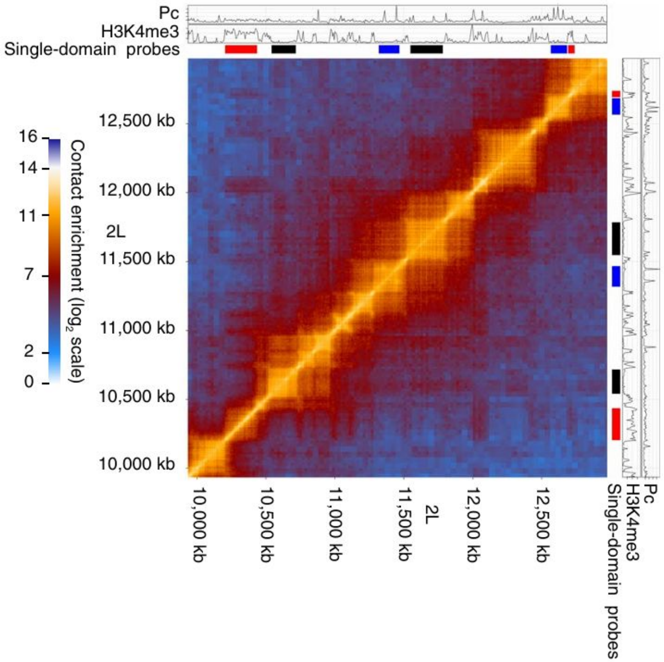
Szabo, Q. *et al.* TADs are 3D structural units of higher-order chromosome organization in Drosophila.
Science Advances 4, eaar8082 (2018).

Motivation: Why Study TADs?

Are TADs Real Physical Units?

- HI-C: Captures contact frequencies between DNA regions in 3D space

- TAD (Topologically Associating Domain):
  - Visible as small triangular domains in Hi-C maps
  - DNA within a TAD interacts frequently
  - Different TADs interact less frequently
  - *Serves as a basic 3D structural unit in this study*



TADs are real 3D structures, not Hi-C artifacts

Polycomb-repressed TADs (Blue)

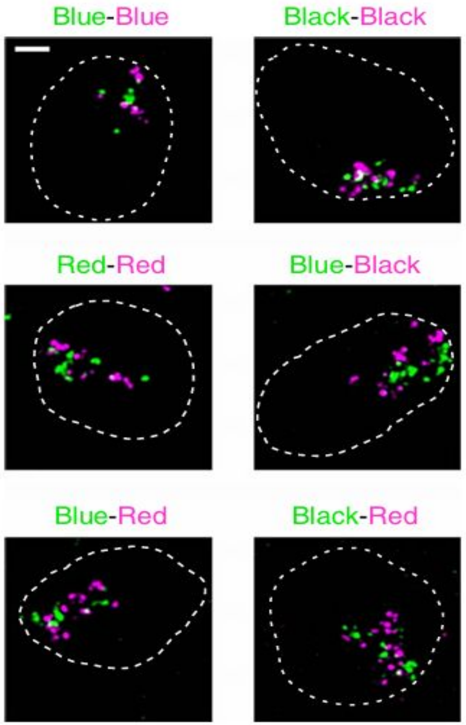
- Very strong Hi-C signal (high contact frequency)
- Highly compact in 3D → forms dense *nanocompartment* spheres

Inactive TADs (Black)

- Intermediate Hi-C intensity
- Moderately compact

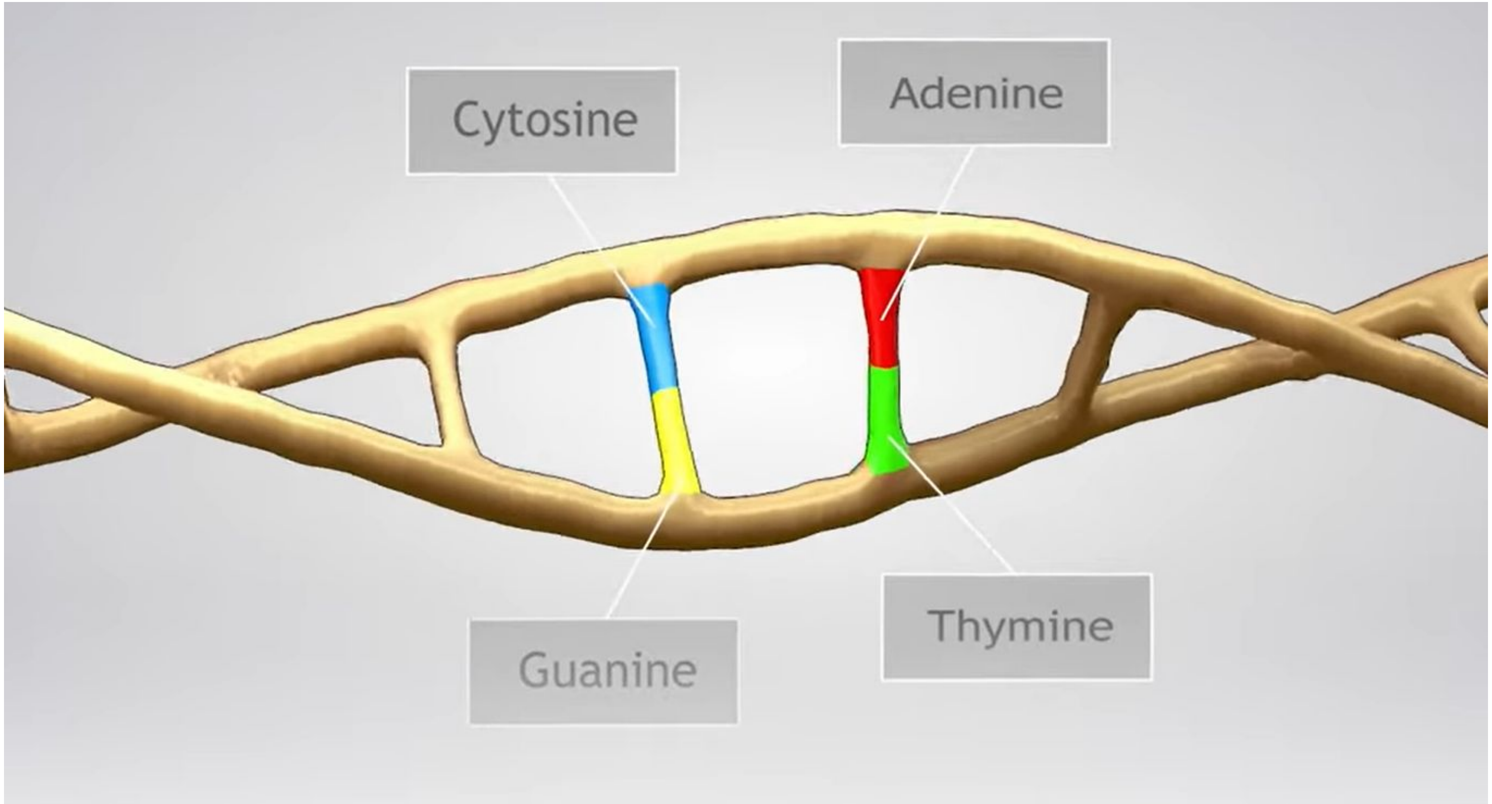
Active TADs (Red)

- Weak Hi-C signal → low density
- In 3D: remain loose and dispersed, never forming spheres

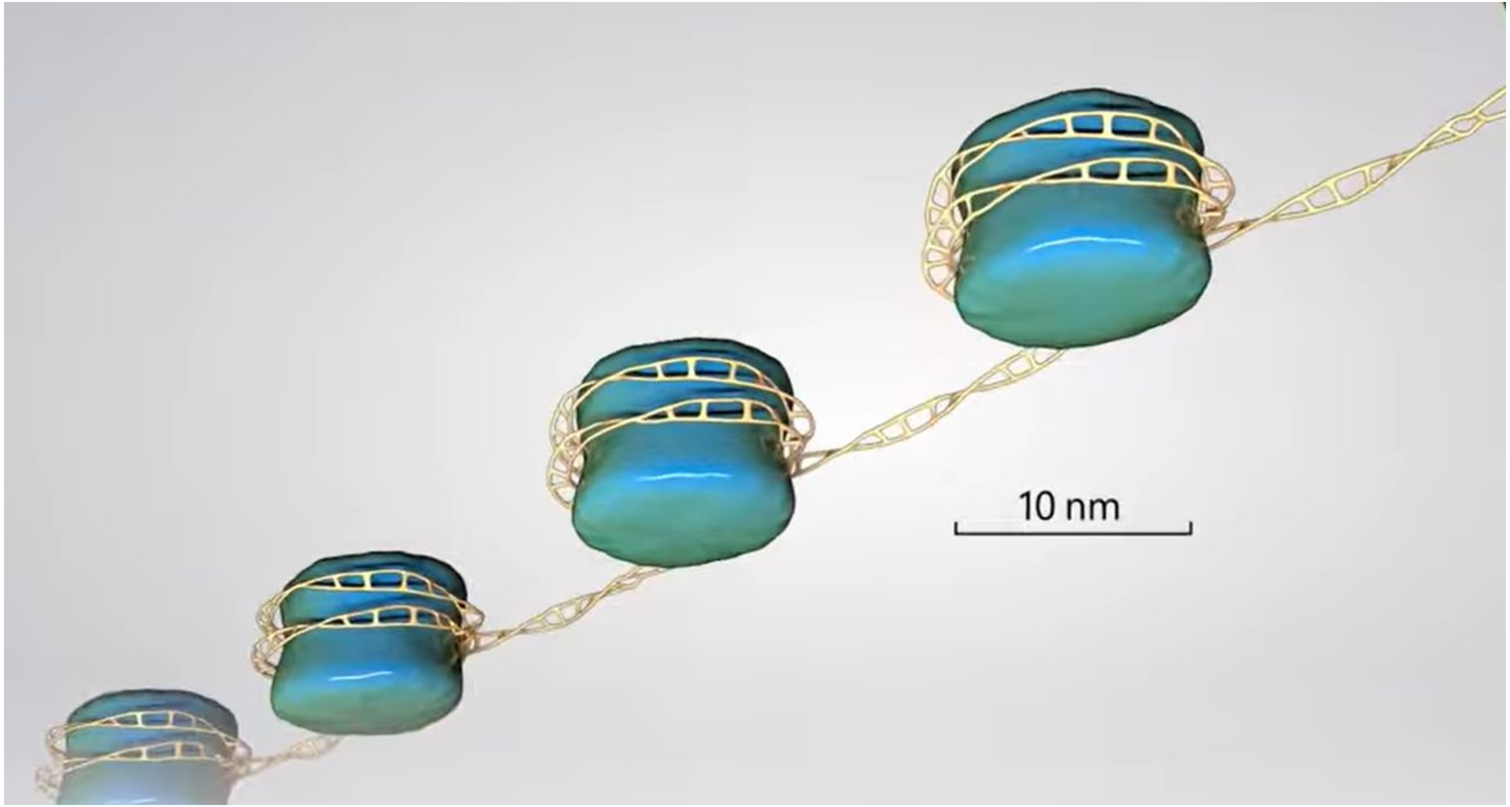


Biological Definition of TADs

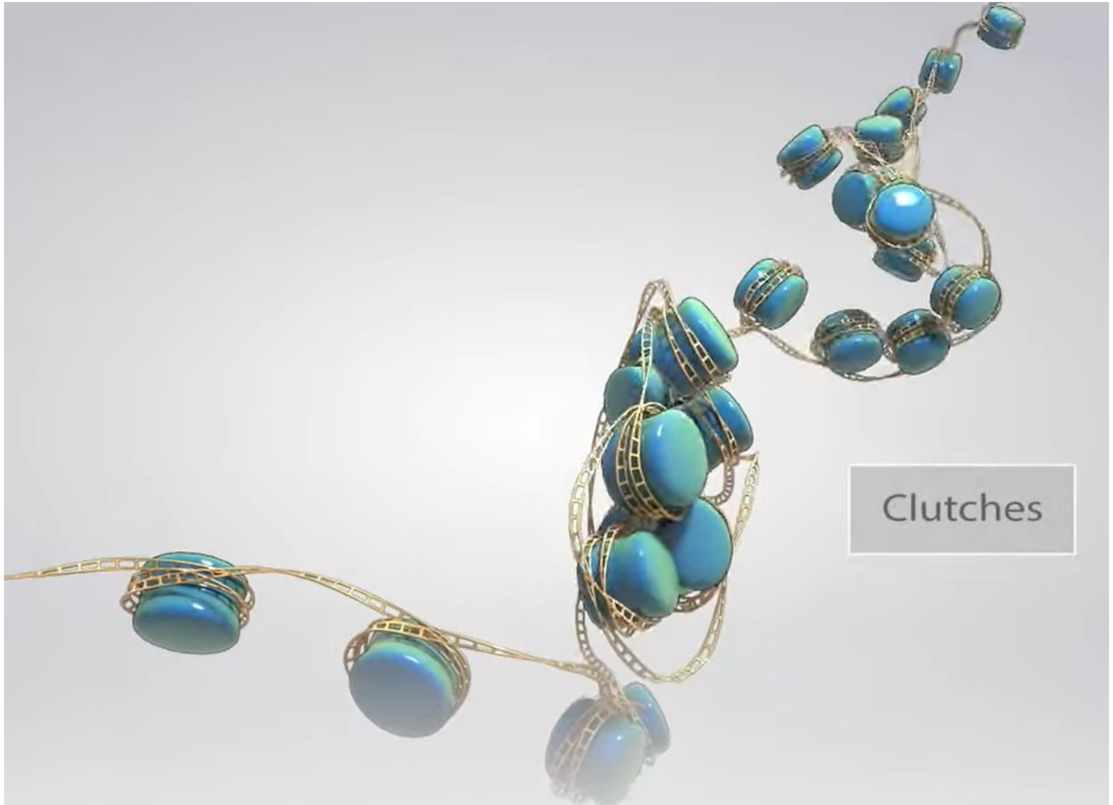
DNA Strand



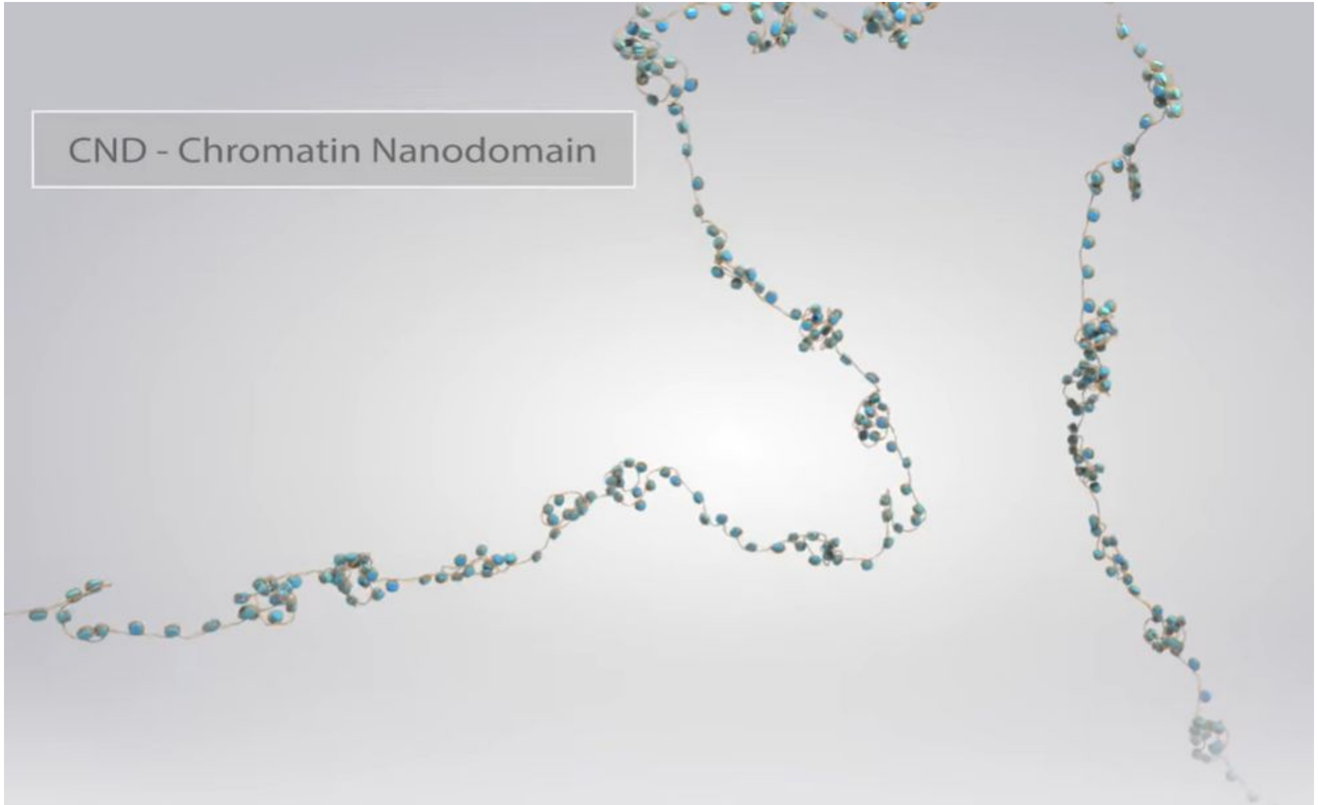
Histone Proteins & Nucleosomes



Clutches

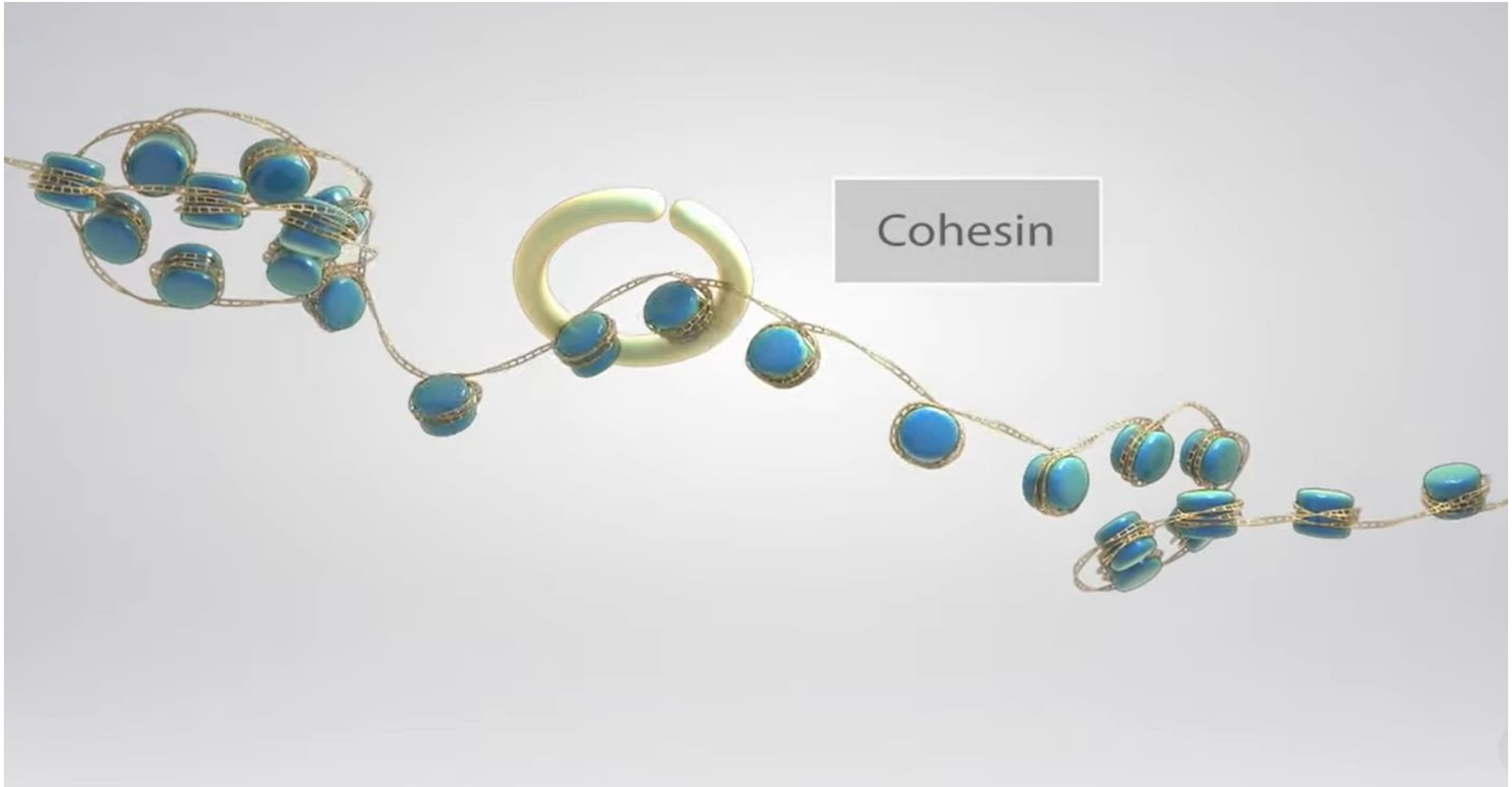


CND

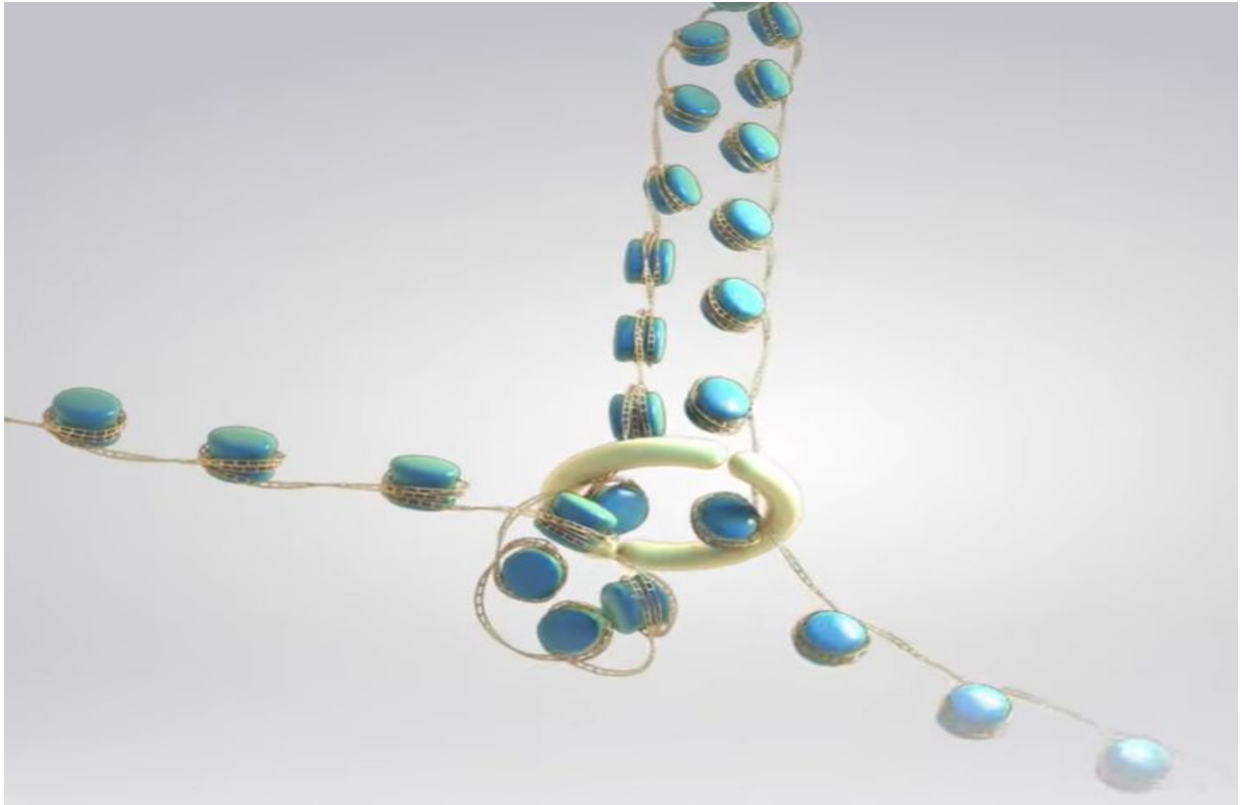


CND - Chromatin Nanodomain

Cohesin

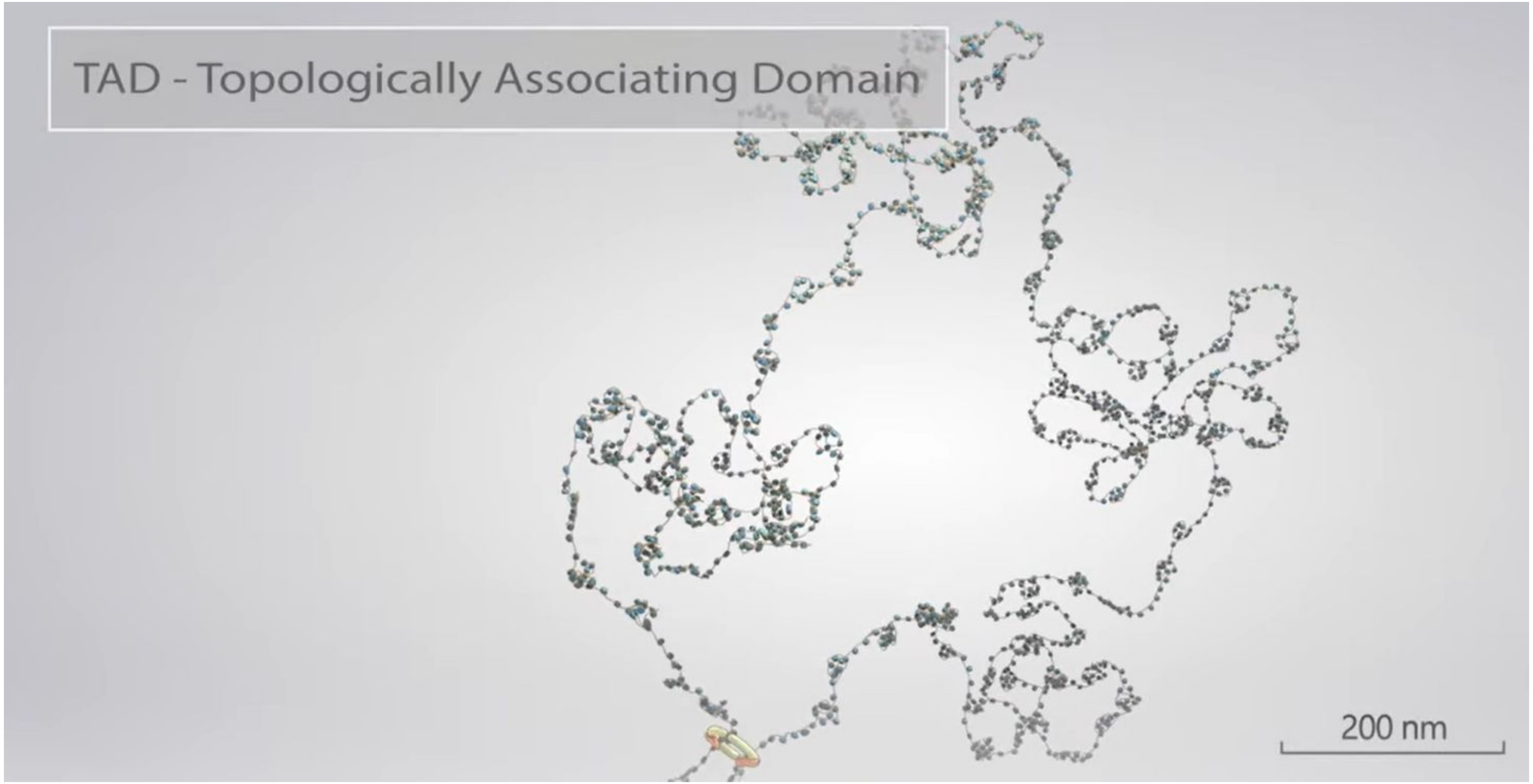


Loop Extrusion



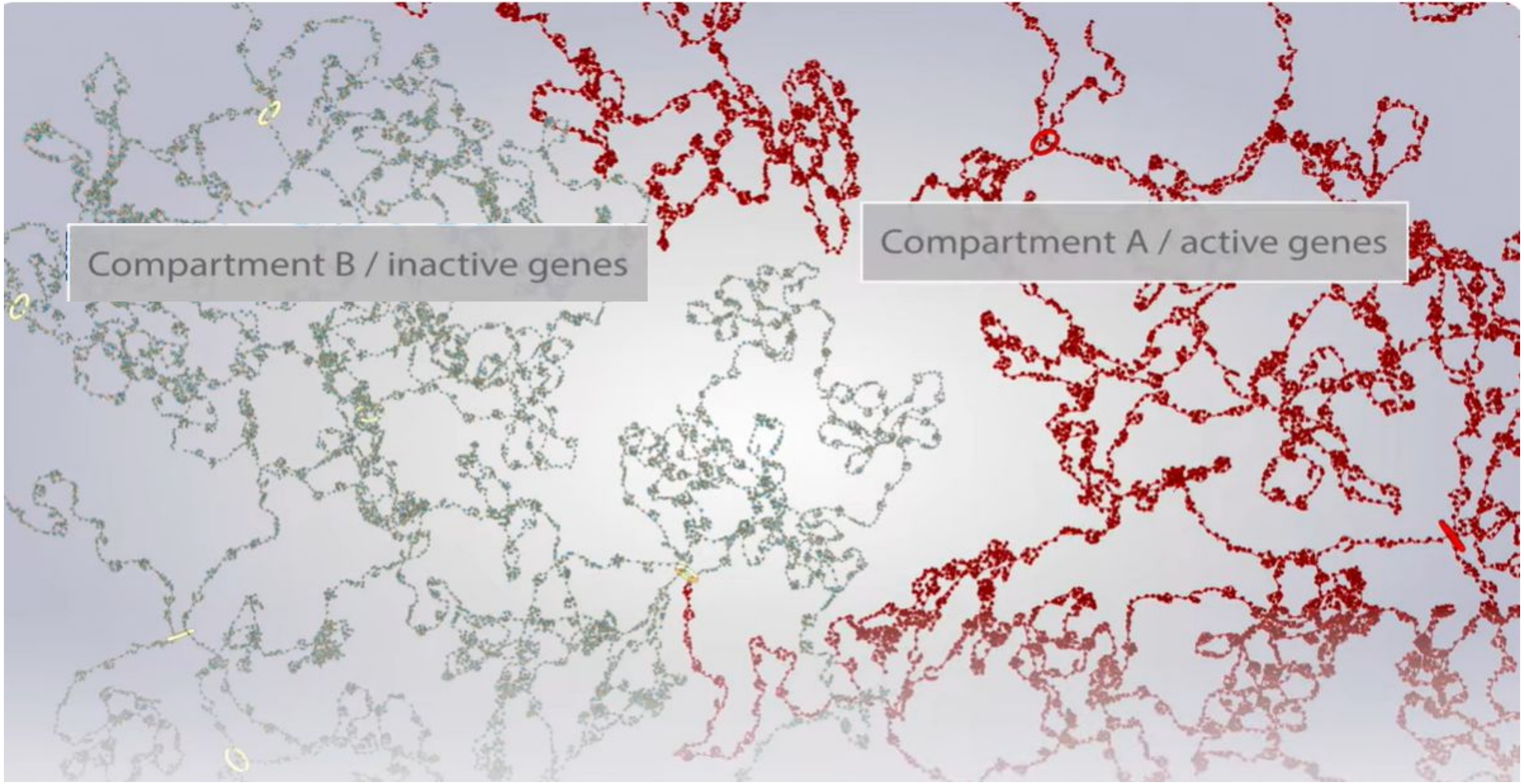
CTCF

TAD



TAD - Topologically Associating Domain

Different Compartment



Reconstructing Hi-C Maps from Contact Matrix

Goal: Reconstruct Figures 1A, 5A, and 5G

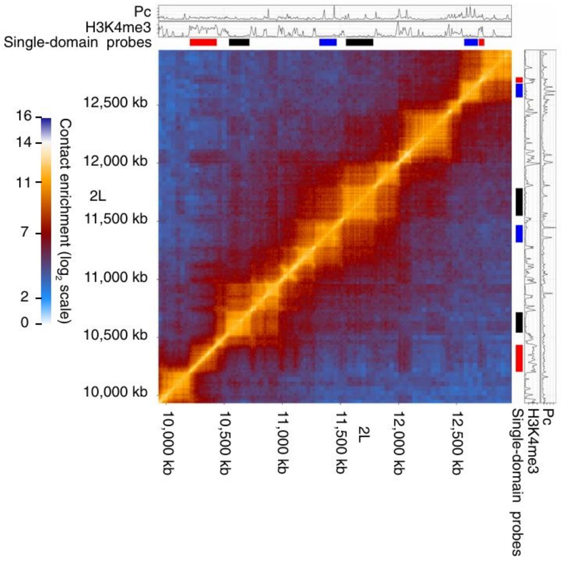


figure 1A

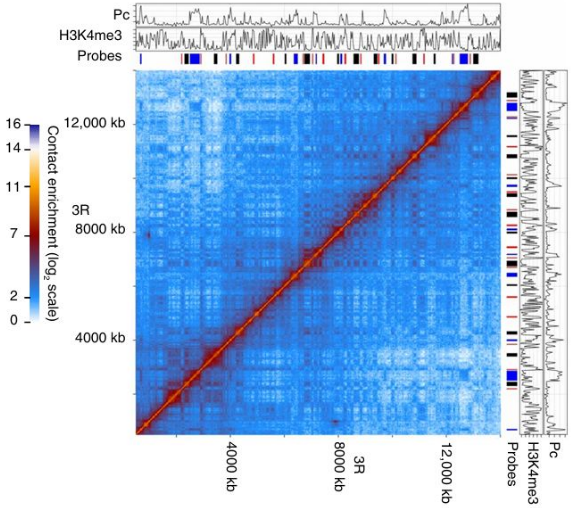


figure 5A

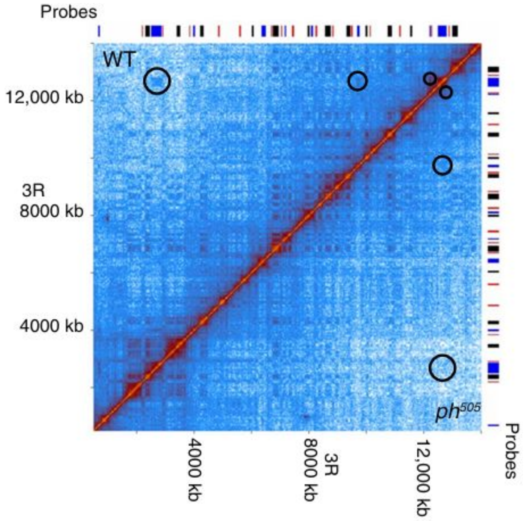
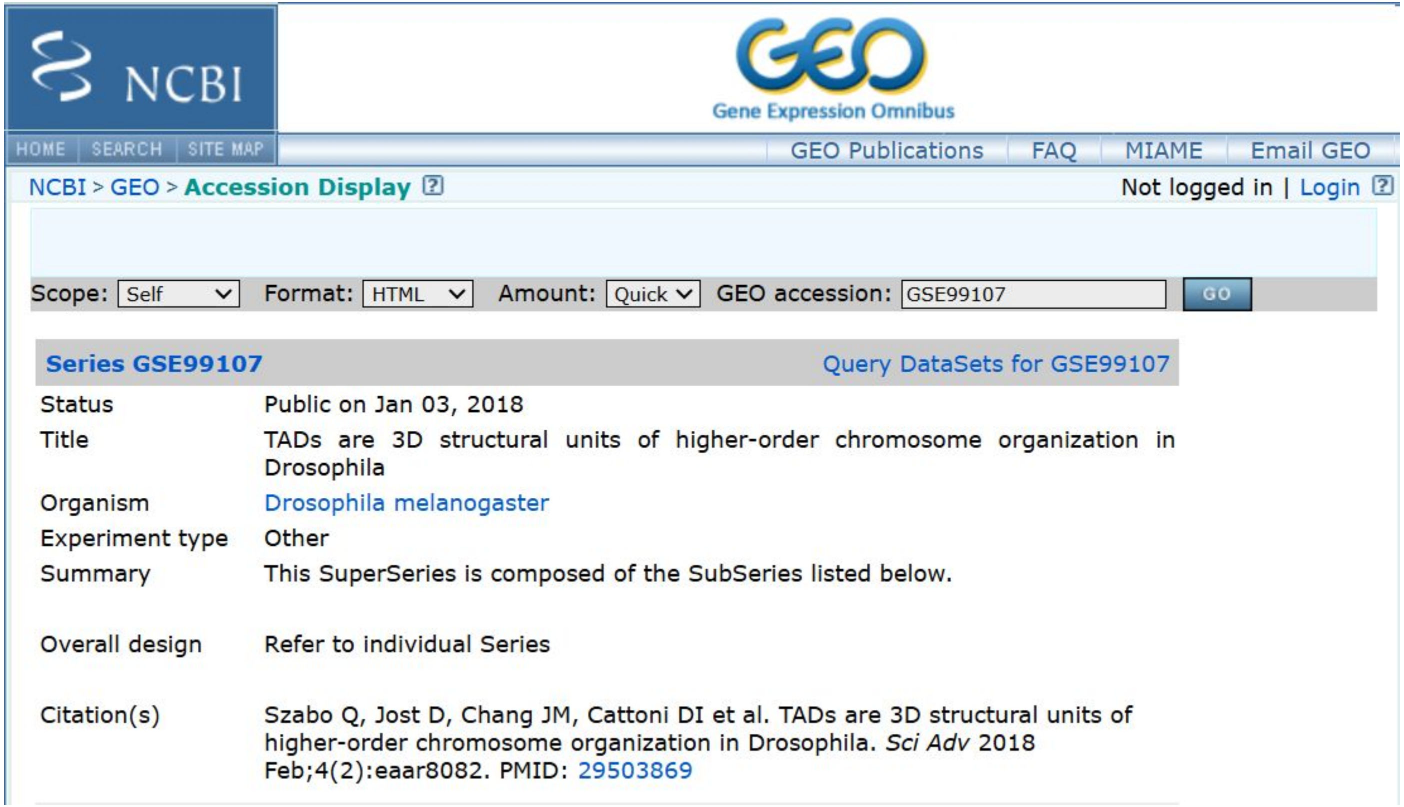


figure 5G

Contact Matrix Dataset Source

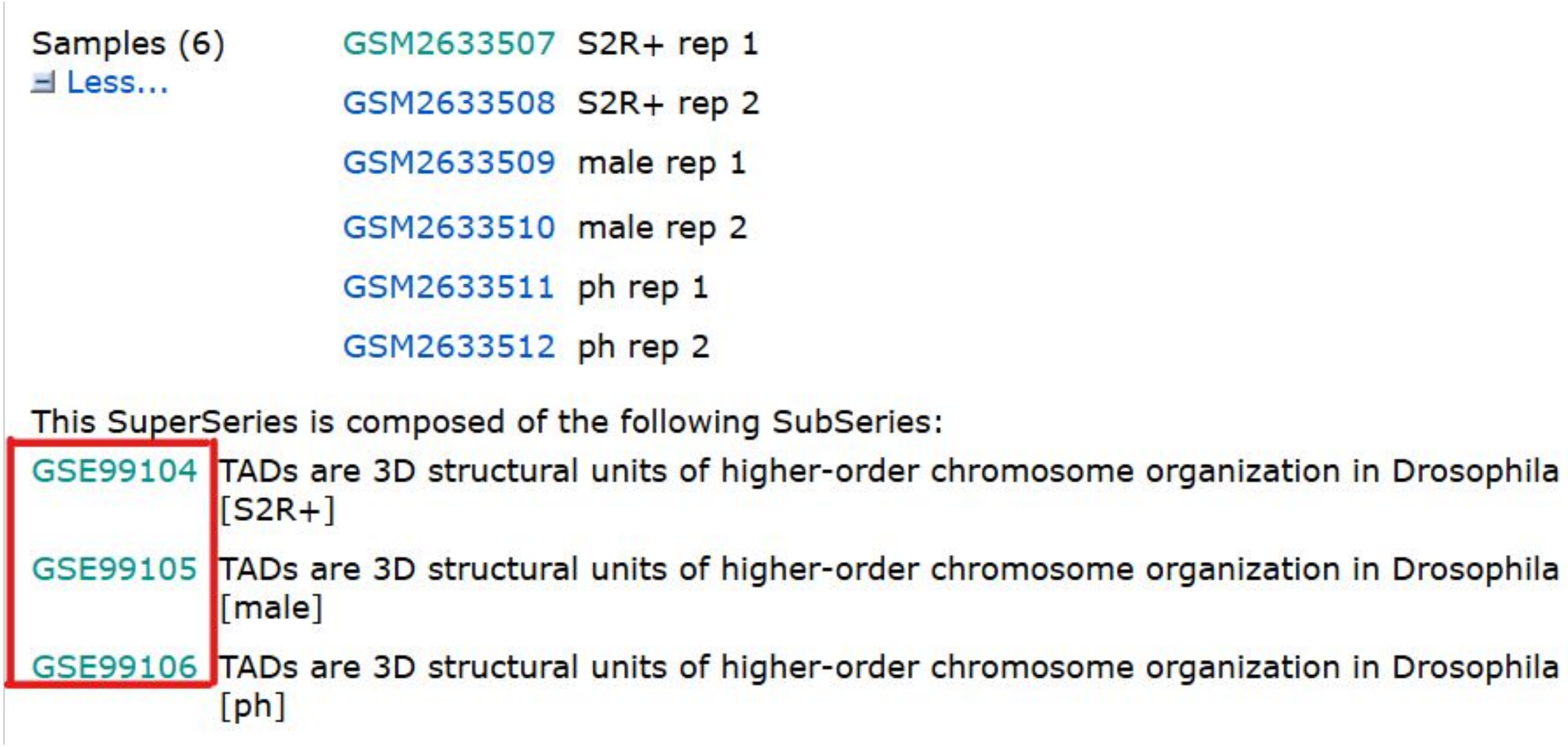


Scope: Self Format: HTML Amount: Quick GEO accession: GSE99107

Series GSE99107 Query DataSets for GSE99107

Status	Public on Jan 03, 2018
Title	TADs are 3D structural units of higher-order chromosome organization in Drosophila
Organism	Drosophila melanogaster
Experiment type	Other
Summary	This SuperSeries is composed of the SubSeries listed below.
Overall design	Refer to individual Series
Citation(s)	Szabo Q, Jost D, Chang JM, Cattoni DI et al. TADs are 3D structural units of higher-order chromosome organization in Drosophila. *Sci Adv* 2018 Feb;4(2):eaar8082. PMID: 29503869

Contact Matrix Dataset Source



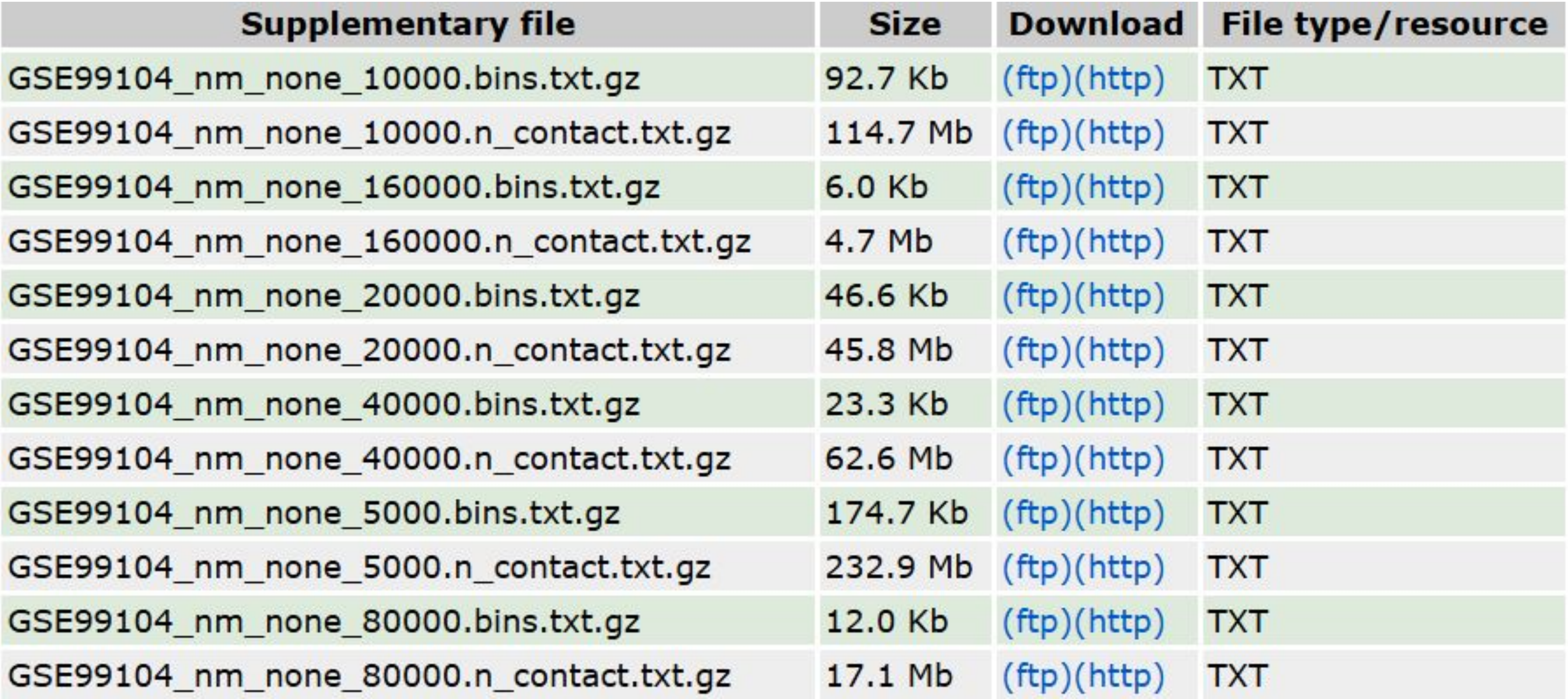
Samples (6)
Less...

GSM2633507	S2R+ rep 1
GSM2633508	S2R+ rep 2
GSM2633509	male rep 1
GSM2633510	male rep 2
GSM2633511	ph rep 1
GSM2633512	ph rep 2

This SuperSeries is composed of the following SubSeries:

GSE99104	TADs are 3D structural units of higher-order chromosome organization in Drosophila [S2R+]
GSE99105	TADs are 3D structural units of higher-order chromosome organization in Drosophila [male]
GSE99106	TADs are 3D structural units of higher-order chromosome organization in Drosophila [ph]

Contact Matrix Dataset Source



Supplementary file	Size	Download	File type/resource
GSE99104_nm_none_10000.bins.txt.gz	92.7 Kb	(ftp)(http)	TXT
GSE99104_nm_none_10000.n_contact.txt.gz	114.7 Mb	(ftp)(http)	TXT
GSE99104_nm_none_160000.bins.txt.gz	6.0 Kb	(ftp)(http)	TXT
GSE99104_nm_none_160000.n_contact.txt.gz	4.7 Mb	(ftp)(http)	TXT
GSE99104_nm_none_20000.bins.txt.gz	46.6 Kb	(ftp)(http)	TXT
GSE99104_nm_none_20000.n_contact.txt.gz	45.8 Mb	(ftp)(http)	TXT
GSE99104_nm_none_40000.bins.txt.gz	23.3 Kb	(ftp)(http)	TXT
GSE99104_nm_none_40000.n_contact.txt.gz	62.6 Mb	(ftp)(http)	TXT
GSE99104_nm_none_5000.bins.txt.gz	174.7 Kb	(ftp)(http)	TXT
GSE99104_nm_none_5000.n_contact.txt.gz	232.9 Mb	(ftp)(http)	TXT
GSE99104_nm_none_80000.bins.txt.gz	12.0 Kb	(ftp)(http)	TXT
GSE99104_nm_none_80000.n_contact.txt.gz	17.1 Mb	(ftp)(http)	TXT

Dataset 1: Genomic Bins for Hi-C Analysis

cbin	chr	from	to
1	2L	0	40,000
2	2L	40,000	80,000
3	2L	80,000	120,000
4	2L	120,000	160,000
5	2L	160,000	200,000

Dataset 2: Hi-C Contact Counts

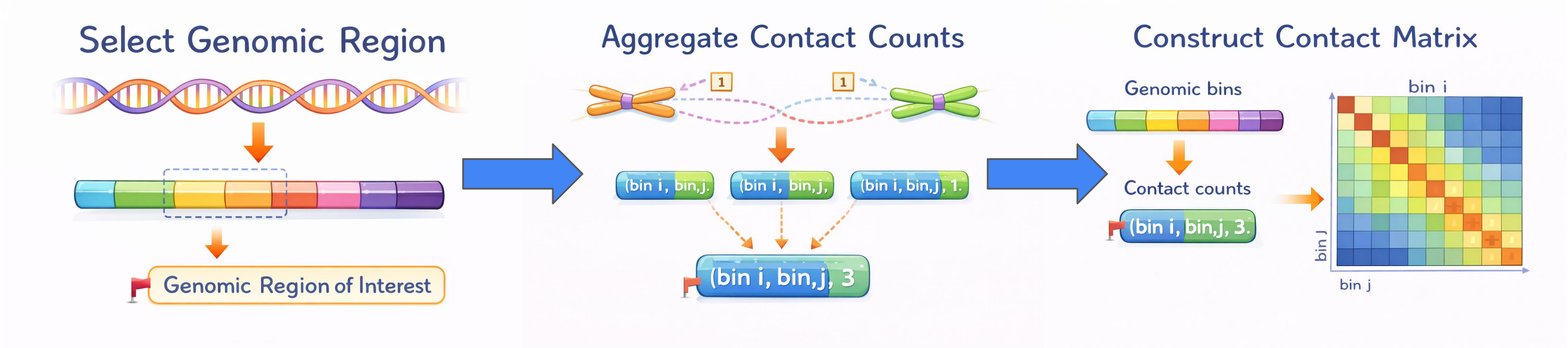
cbin1	cbin2	expected count	observed count
1	1	1.731	1,803
1	2	3.831	1,698
1	3	5.677	457
1	4	5.445	183
1	5	4.283	88

Hi-C Contact Matrix Construction Pipeline

1

2

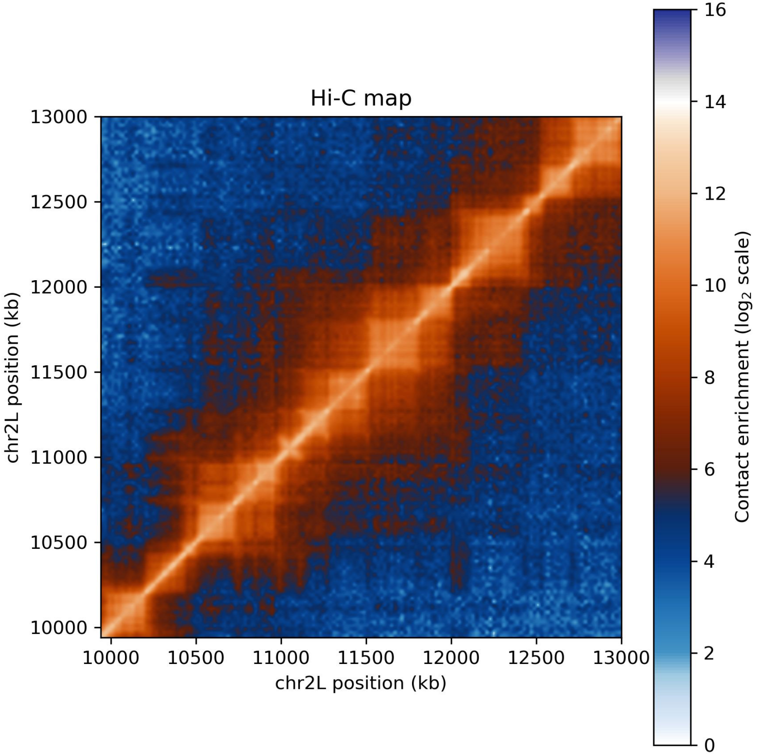
3



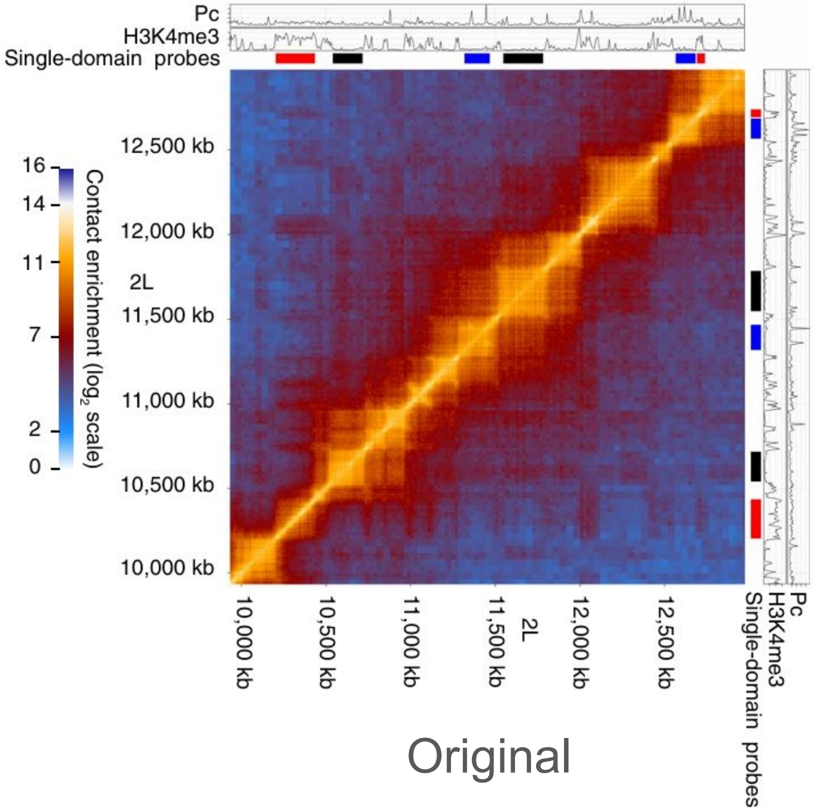
Hi-C Contact Matrix Construction Pipeline

bins	1	2	3	4
1				
2		$\log_2\left(\frac{\text{Observed}}{\text{Expected}}\right)$		
3				
4				

Reconstruction of Figure 1A from S2R+ Hi-C Map

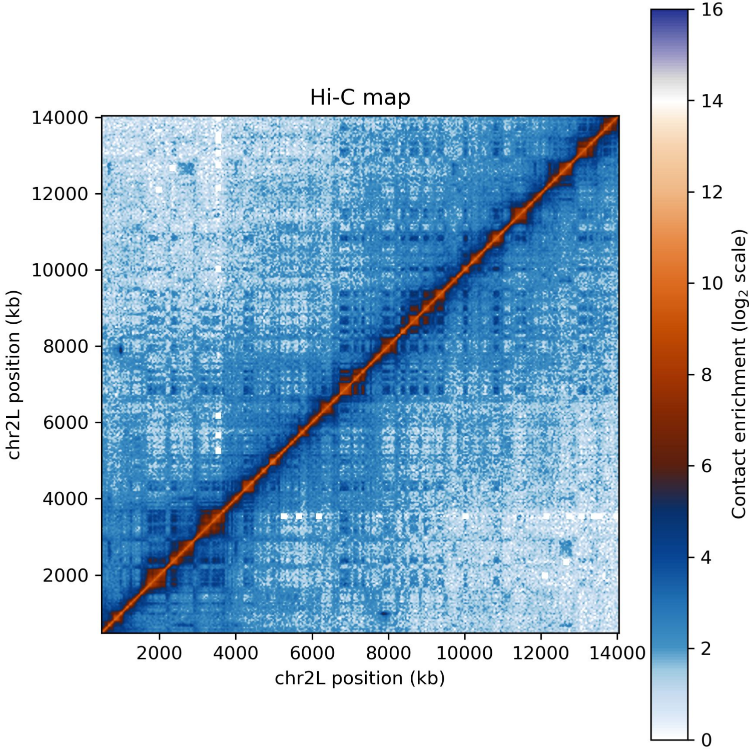


Reconstructed

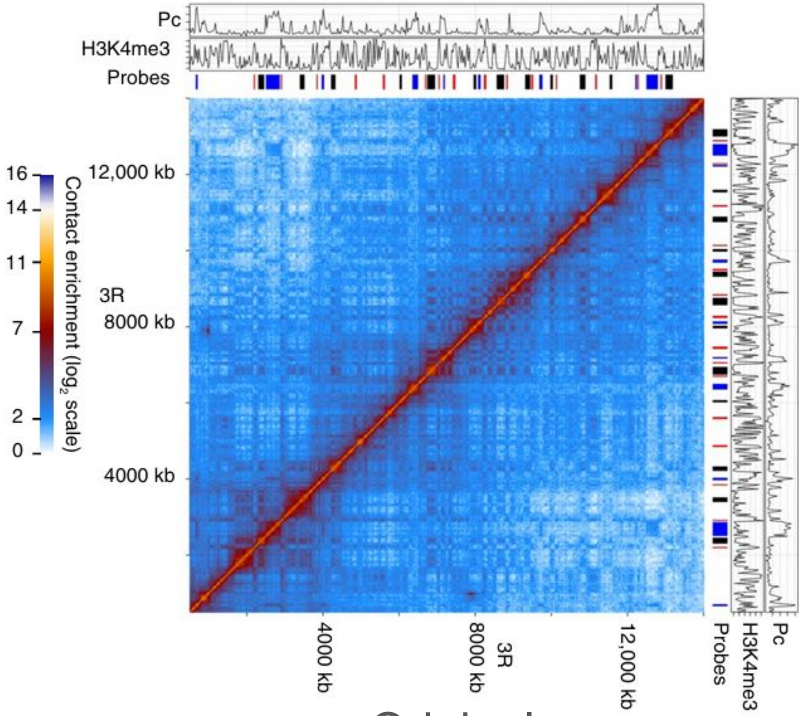


Original

Reconstruction of Figure 5A from Embryonic Hi-C Map

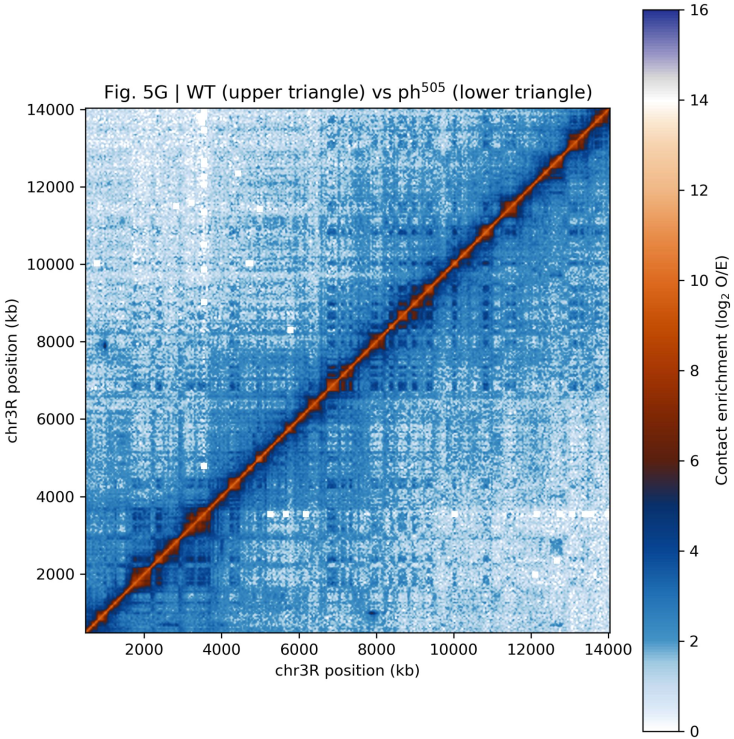


Reconstructed

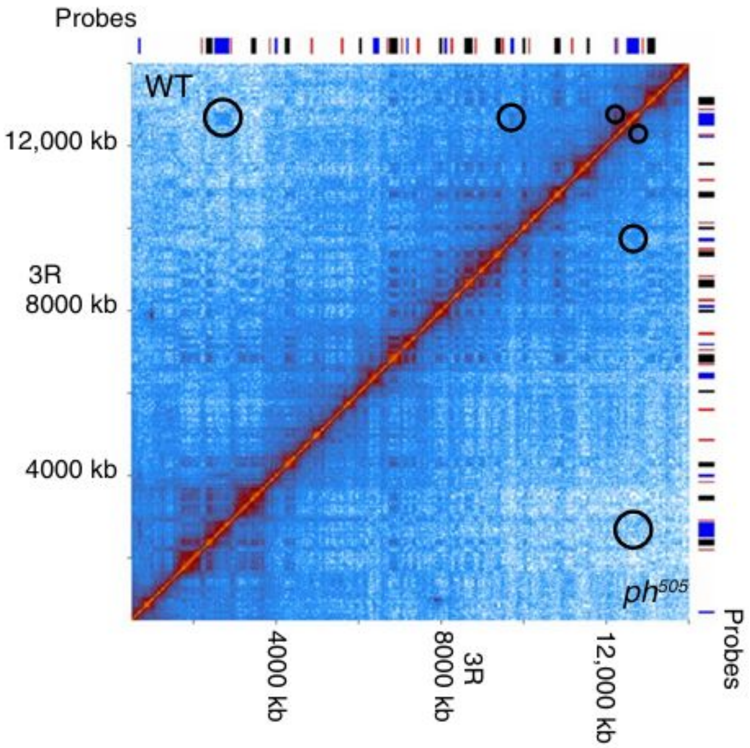


Original

Reconstruction of Figure 5G: Loss of PcG-Mediated (



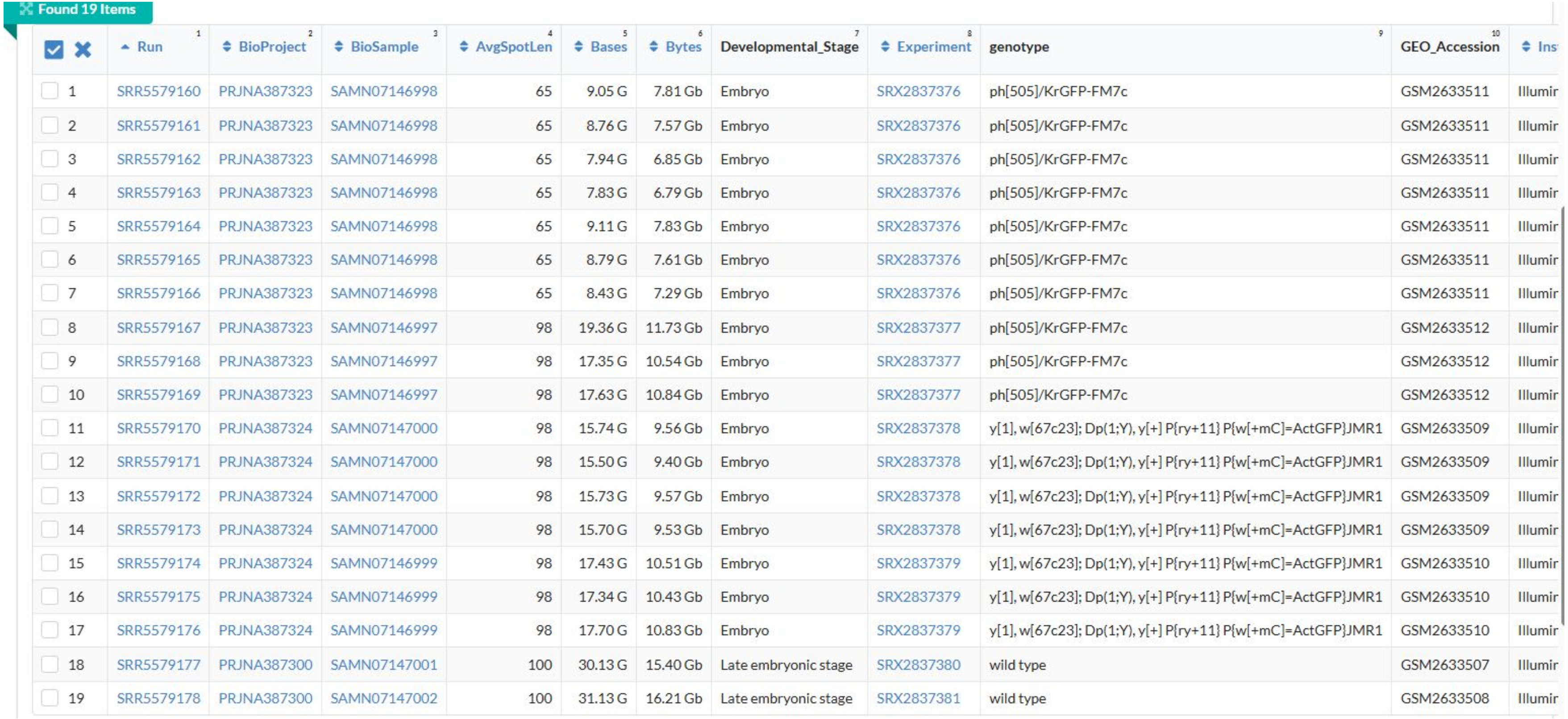
Reconstructed



Original

Step-by-step construction of the HI-C diagram

Contact Matrix Dataset Source



Found 19 Items

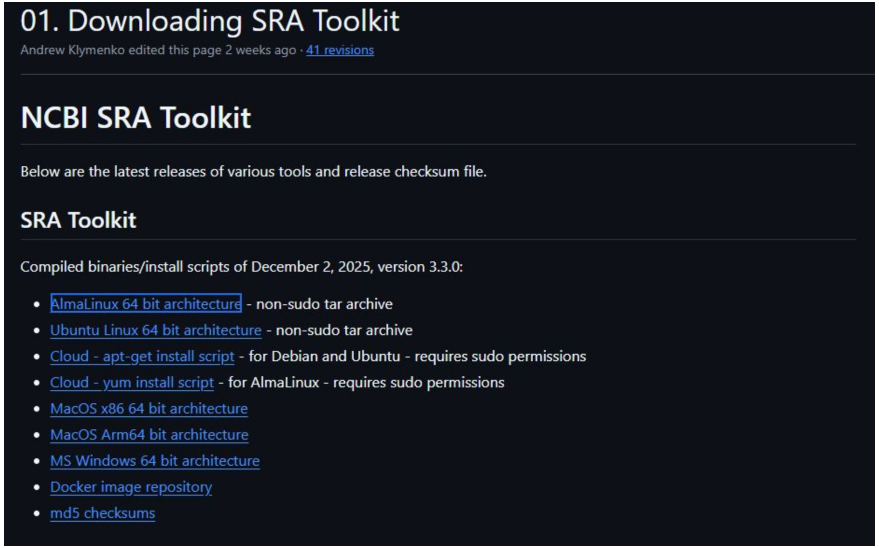
☑ ✖	Run	BioProject	BioSample	AvgSpotLen	Bases	Bytes	Developmental_Stage	Experiment	genotype	GEO_Accession	Ins
□ 1	SRR5579160	PRJNA387323	SAMN07146998	65	9.05 G	7.81 Gb	Embryo	SRX2837376	ph[505]/KrGFP-FM7c	GSM2633511	Illumir
□ 2	SRR5579161	PRJNA387323	SAMN07146998	65	8.76 G	7.57 Gb	Embryo	SRX2837376	ph[505]/KrGFP-FM7c	GSM2633511	Illumir
□ 3	SRR5579162	PRJNA387323	SAMN07146998	65	7.94 G	6.85 Gb	Embryo	SRX2837376	ph[505]/KrGFP-FM7c	GSM2633511	Illumir
□ 4	SRR5579163	PRJNA387323	SAMN07146998	65	7.83 G	6.79 Gb	Embryo	SRX2837376	ph[505]/KrGFP-FM7c	GSM2633511	Illumir
□ 5	SRR5579164	PRJNA387323	SAMN07146998	65	9.11 G	7.83 Gb	Embryo	SRX2837376	ph[505]/KrGFP-FM7c	GSM2633511	Illumir
□ 6	SRR5579165	PRJNA387323	SAMN07146998	65	8.79 G	7.61 Gb	Embryo	SRX2837376	ph[505]/KrGFP-FM7c	GSM2633511	Illumir
□ 7	SRR5579166	PRJNA387323	SAMN07146998	65	8.43 G	7.29 Gb	Embryo	SRX2837376	ph[505]/KrGFP-FM7c	GSM2633511	Illumir
□ 8	SRR5579167	PRJNA387323	SAMN07146997	98	19.36 G	11.73 Gb	Embryo	SRX2837377	ph[505]/KrGFP-FM7c	GSM2633512	Illumir
□ 9	SRR5579168	PRJNA387323	SAMN07146997	98	17.35 G	10.54 Gb	Embryo	SRX2837377	ph[505]/KrGFP-FM7c	GSM2633512	Illumir
□ 10	SRR5579169	PRJNA387323	SAMN07146997	98	17.63 G	10.84 Gb	Embryo	SRX2837377	ph[505]/KrGFP-FM7c	GSM2633512	Illumir
□ 11	SRR5579170	PRJNA387324	SAMN07147000	98	15.74 G	9.56 Gb	Embryo	SRX2837378	y[1], w[67c23]; Dp(1;Y), y[+] P{ry+11} P{w[+mC]=ActGFP}JMR1	GSM2633509	Illumir
□ 12	SRR5579171	PRJNA387324	SAMN07147000	98	15.50 G	9.40 Gb	Embryo	SRX2837378	y[1], w[67c23]; Dp(1;Y), y[+] P{ry+11} P{w[+mC]=ActGFP}JMR1	GSM2633509	Illumir
□ 13	SRR5579172	PRJNA387324	SAMN07147000	98	15.73 G	9.57 Gb	Embryo	SRX2837378	y[1], w[67c23]; Dp(1;Y), y[+] P{ry+11} P{w[+mC]=ActGFP}JMR1	GSM2633509	Illumir
□ 14	SRR5579173	PRJNA387324	SAMN07147000	98	15.70 G	9.53 Gb	Embryo	SRX2837378	y[1], w[67c23]; Dp(1;Y), y[+] P{ry+11} P{w[+mC]=ActGFP}JMR1	GSM2633509	Illumir
□ 15	SRR5579174	PRJNA387324	SAMN07146999	98	17.43 G	10.51 Gb	Embryo	SRX2837379	y[1], w[67c23]; Dp(1;Y), y[+] P{ry+11} P{w[+mC]=ActGFP}JMR1	GSM2633510	Illumir
□ 16	SRR5579175	PRJNA387324	SAMN07146999	98	17.34 G	10.43 Gb	Embryo	SRX2837379	y[1], w[67c23]; Dp(1;Y), y[+] P{ry+11} P{w[+mC]=ActGFP}JMR1	GSM2633510	Illumir
□ 17	SRR5579176	PRJNA387324	SAMN07146999	98	17.70 G	10.83 Gb	Embryo	SRX2837379	y[1], w[67c23]; Dp(1;Y), y[+] P{ry+11} P{w[+mC]=ActGFP}JMR1	GSM2633510	Illumir
□ 18	SRR5579177	PRJNA387300	SAMN07147001	100	30.13 G	15.40 Gb	Late embryonic stage	SRX2837380	wild type	GSM2633507	Illumir
□ 19	SRR5579178	PRJNA387300	SAMN07147002	100	31.13 G	16.21 Gb	Late embryonic stage	SRX2837381	wild type	GSM2633508	Illumir

SRA Toolkit

1.下載壓縮檔(prefetch)

2.解封包(fastQdump)=>forward reverse

.sra → .fastq



01. Downloading SRA Toolkit

Andrew Klymenko edited this page 2 weeks ago · 41 revisions

NCBI SRA Toolkit

Below are the latest releases of various tools and release checksum file.

SRA Toolkit

Compiled binaries/install scripts of December 2, 2025, version 3.3.0:

- AlmaLinux 64 bit architecture - non-sudo tar archive
- Ubuntu Linux 64 bit architecture - non-sudo tar archive
- Cloud - apt-get install script - for Debian and Ubuntu - requires sudo permissions
- Cloud - yum install script - for AlmaLinux - requires sudo permissions
- MacOS x86 64 bit architecture
- MacOS Arm64 bit architecture
- MS Windows 64 bit architecture
- Docker image repository
- md5 checksums

Concatenate the data

一共4個file = { SRR5579177_1.fastq (front), SRR5579177_2.fastq(reverse)

,SRR5579178_1.fastq (front) , SRR5579178_2.fastq(reverse) }

concatenate 成一條 => { front front reverse reverse }

之後壓縮檔案 (gzip + pigz(加速)) 後得到 .gz檔

生成Hi-C: jucier(pipeline)

pipeline有什麼?

pipeline :

Split : 將巨大的 FASTQ 切成小塊 (batch) (~30mins)。

Align : 將 reads 比對到 dm3 基因組(~12hrs)。(得到每一條 read 在基因組上的位置)

Merge & Sort: 合併比對結果並排序。

Chimeric Handling: 處理 Hi-C 特有的嵌合 reads (找Ligation junctions)。

Deduplicate: 移除 PCR 重複 (Duplicates)。

Final: 生成 .hic 檔案 (用於 Juicebox) 和 .hic 統計數據。

Practice

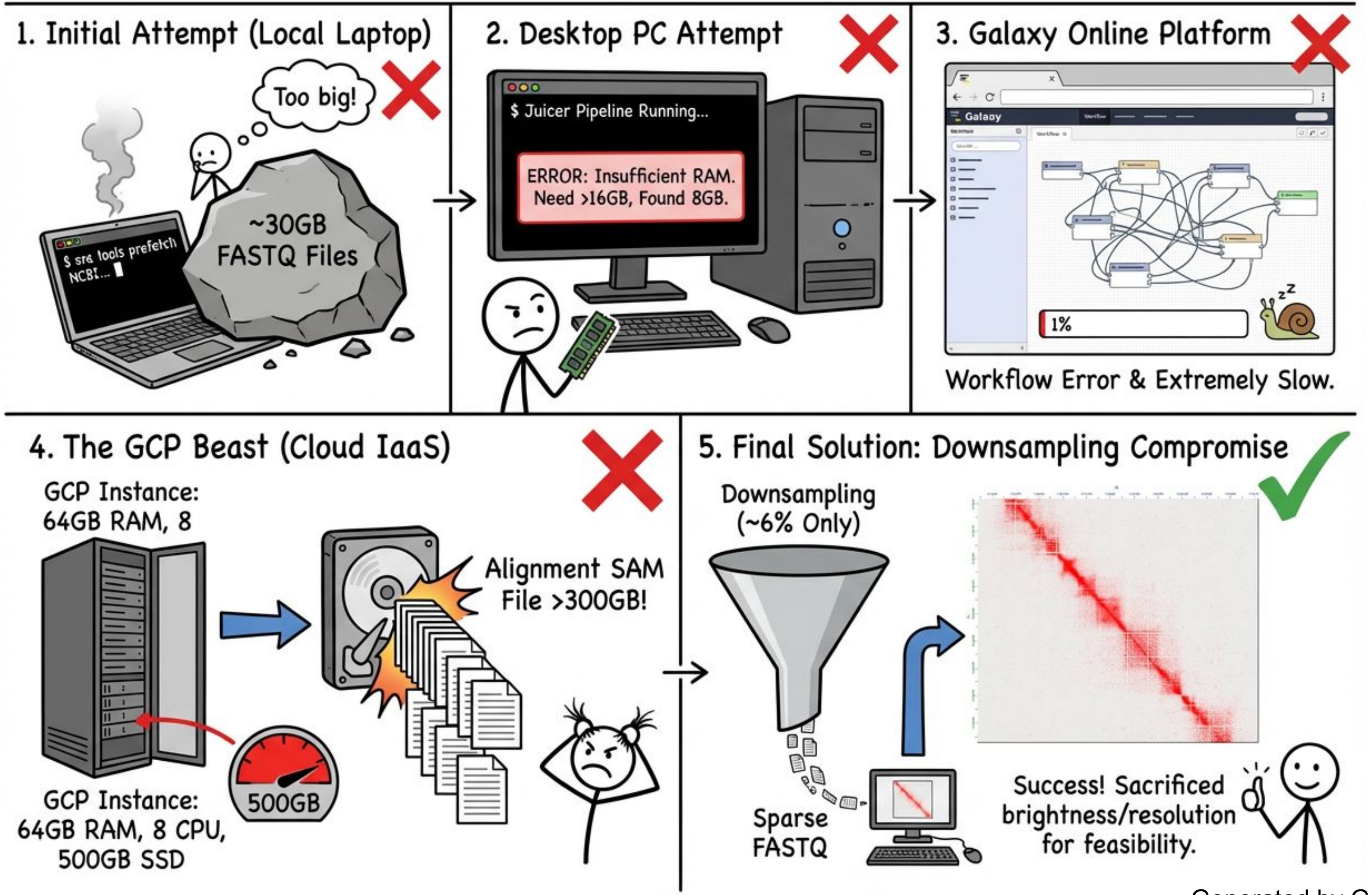
裡用jucier套件 要做的就是引入pipeline所需的額外套件

Align (BWA) => BW transform => .sam => 2L 2R

Chimeric Handling : samtools

Reconstruction Roadmap

My Bioinformatics Final Project Journey: Reproducing Hi-C Contacts from Raw FASTQ.


1. Initial Attempt (Local Laptop)
Too big!
~30GB FASTQ Files
2. Desktop PC Attempt
3. Galaxy Online Platform
Galaxy
1%
4. The GCP Beast (Cloud IaaS)
GCP Instance: 64GB RAM, 8
500GB
5. Final Solution: Downsampling Compromise
Downsampling (~6% Only)
Sparse FASTQ

```
$ sra tools prefetch
NCBI...
```

```
$ Juicer Pipeline Running...
```

```
ERROR: Insufficient RAM.
Need >16GB, Found 8GB.
```

Workflow Error & Extremely Slow.

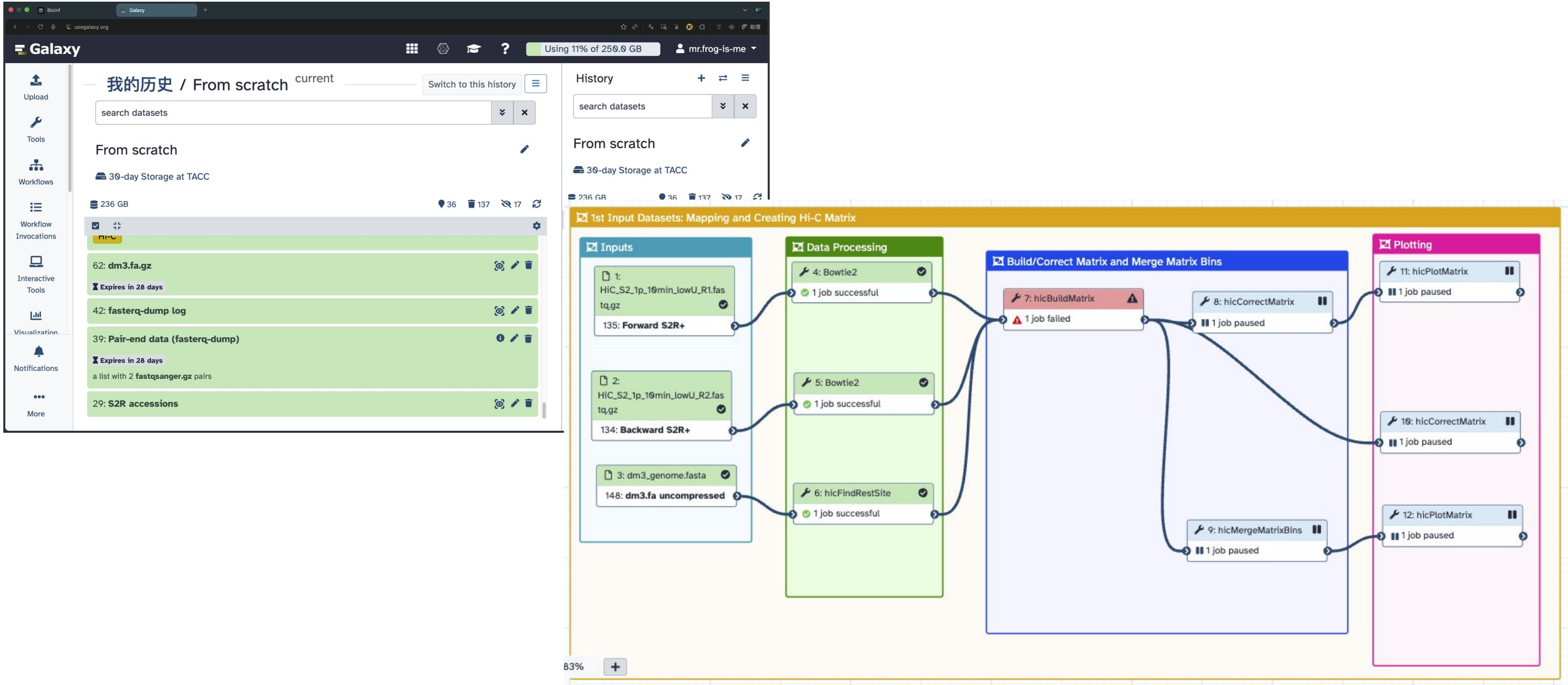
GCP Instance: 64GB RAM, 8 CPU, 500GB SSD

Alignment SAM File >300GB!

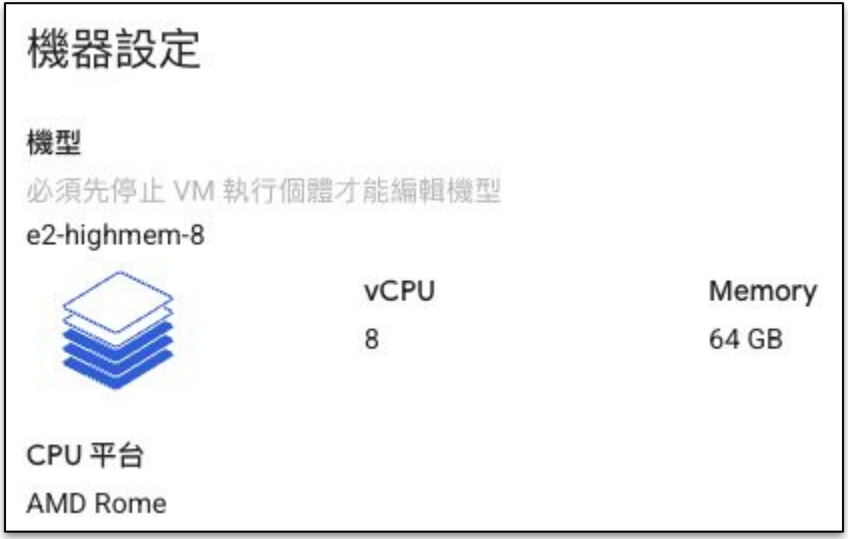
Success! Sacrificed brightness/resolution for feasibility.

Generated by Gemini Enterprise

Galaxy (https://usegalaxy.org/)



GCP 雲端伺服器



機器設定

機型

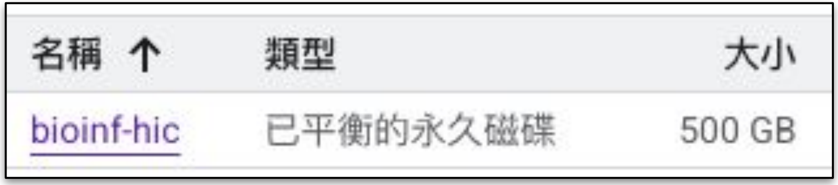
必須先停止 VM 執行個體才能編輯機型

e2-highmem-8

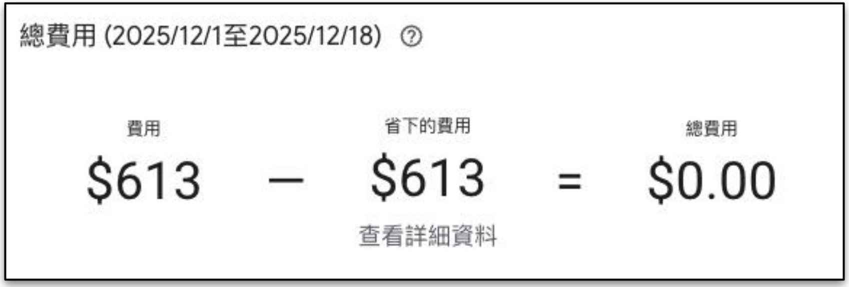
vCPU	Memory
8	64 GB

CPU 平台

AMD Rome



名稱 ↑	類型	大小
bioinf-hic	已平衡的永久磁碟	500 GB



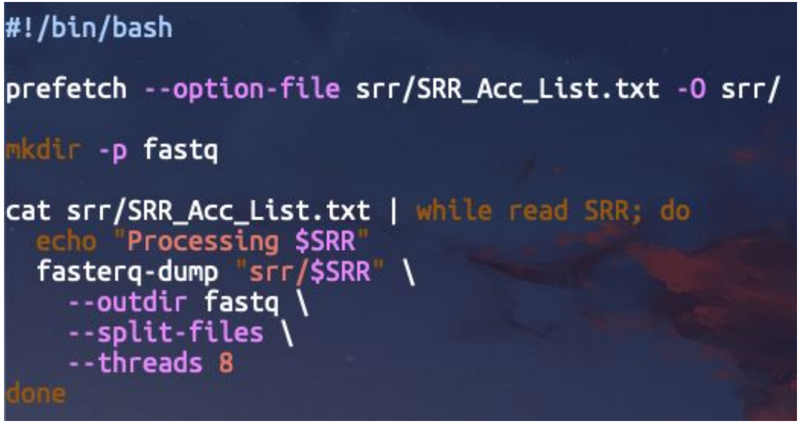
總費用 (2025/12/1至2025/12/18) ⓘ

費用		省下的費用		總費用
$613	−	$613	=	$0.00

查看詳細資料

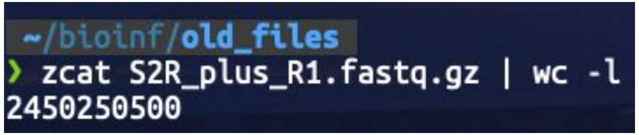
Getting Data From NCBI (RSA tools)

Prefetch (2~3hr) & Dump (~3hr)



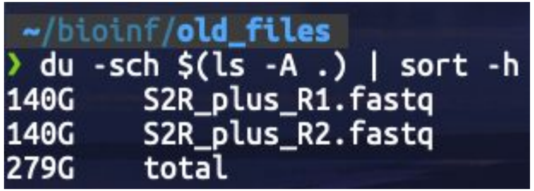
```
#!/bin/bash

prefetch --option-file srr/SRR_Acc_List.txt -O srr/

mkdir -p fastq

cat srr/SRR_Acc_List.txt | while read SRR; do
  echo "Processing $SRR"
  fasterq-dump "srr/$SRR" \
    --outdir fastq \
    --split-files \
    --threads 8
done
```

~2.4B Lines, ~0.6B Reads

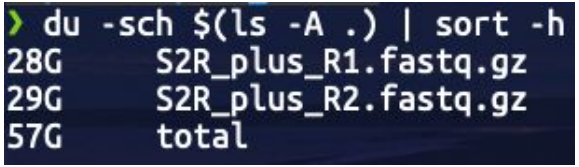


```
 ~/bioinf/old_files
❯ zcat S2R_plus_R1.fastq.gz | wc -l
2450250500
```

Original Fastq size (after concatenate)



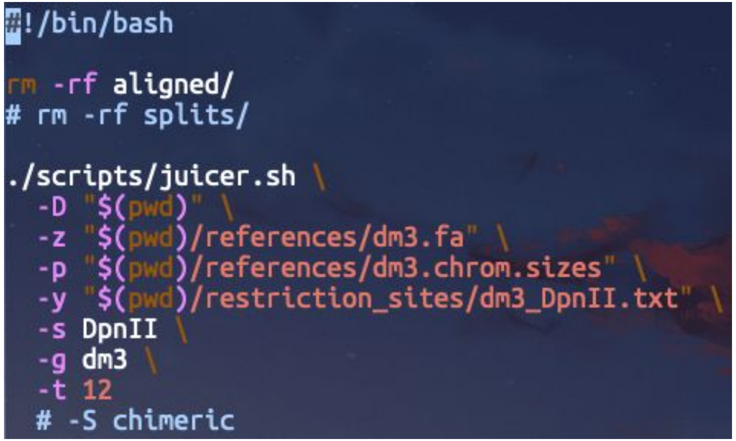
```
 ~/bioinf/old_files
❯ du -sch $(ls -A .) | sort -h
140G     S2R_plus_R1.fastq
140G     S2R_plus_R2.fastq
279G     total
```



```
❯ du -sch $(ls -A .) | sort -h
28G      S2R_plus_R1.fastq.gz
29G      S2R_plus_R2.fastq.gz
57G      total
```

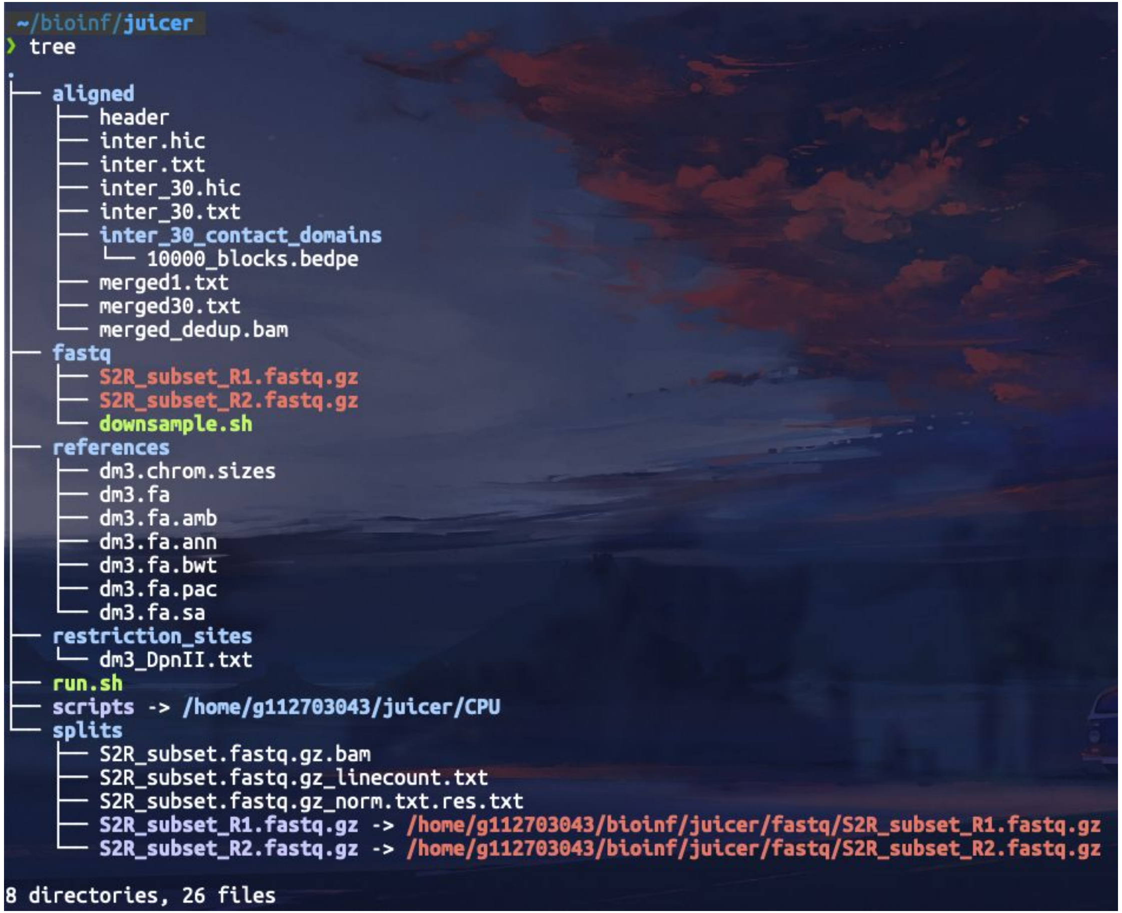
Pipeline - Juicer

Run juicer.sh



```
#!/bin/bash

rm -rf aligned/
# rm -rf splits/

./scripts/juicer.sh \
  -D "$(pwd)" \
  -z "$(pwd)/references/dm3.fa" \
  -p "$(pwd)/references/dm3.chrom.sizes" \
  -y "$(pwd)/restriction_sites/dm3_DpnII.txt" \
  -s DpnII \
  -g dm3 \
  -t 12
  # -S chimeric
```

JucierFile Structure

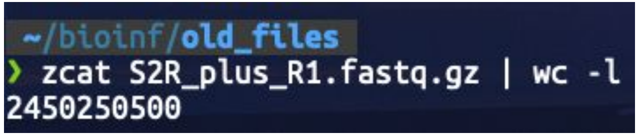


```
 ~/bioinf/juicer
❯ tree
.
├── aligned
│   ├── header
│   ├── inter.hic
│   ├── inter.txt
│   ├── inter_30.hic
│   ├── inter_30.txt
│   ├── inter_30_contact_domains
│   │   └── 10000_blocks.bedpe
│   ├── merged1.txt
│   ├── merged30.txt
│   └── merged_dedup.bam
├── fastq
│   ├── S2R_subset_R1.fastq.gz
│   ├── S2R_subset_R2.fastq.gz
│   └── downsample.sh
├── references
│   ├── dm3.chrom.sizes
│   ├── dm3.fa
│   ├── dm3.fa.amb
│   ├── dm3.fa.ann
│   ├── dm3.fa.bwt
│   ├── dm3.fa.pac
│   └── dm3.fa.sa
├── restriction_sites
│   └── dm3_DpnII.txt
├── run.sh
├── scripts -> /home/g112703043/juicer/CPU
└── splits
    ├── S2R_subset.fastq.gz.bam
    ├── S2R_subset.fastq.gz_linecount.txt
    ├── S2R_subset.fastq.gz_norm.txt.res.txt
    ├── S2R_subset_R1.fastq.gz -> /home/g112703043/bioinf/juicer/fastq/S2R_subset_R1.fastq.gz
    └── S2R_subset_R2.fastq.gz -> /home/g112703043/bioinf/juicer/fastq/S2R_subset_R2.fastq.gz

8 directories, 26 files
```

Bottleneck

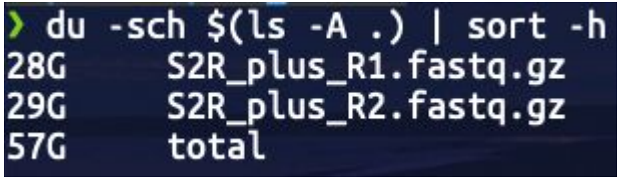
1. 做 Align (bwa mem) 非常吃記憶體(~24GB)，也很花時間(~12hr)
    a. 前前後後嘗試了7~8次
    b. 花了半週的時間等待輸出
2. 做 Align 輸出的 sam 檔案超！極！大！(~300GB)
    a. 內容包括：染色體、位置、CIGAR、插入片段長度、mapping quality、比對分數等。
    b. 把硬碟塞爆，使得後面的 Pipeline 失敗，又要重新來過
    c. 可惜沒有截到圖QQ

~2.4B Lines, ~0.6B Reads



```
~/bioinf/old_files
❯ zcat S2R_plus_R1.fastq.gz | wc -l
2450250500
```

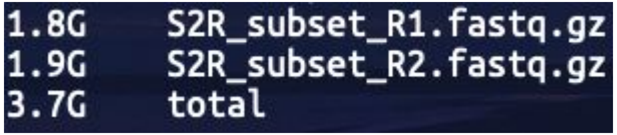
Size before downsampling



```
❯ du -sch $(ls -A .) | sort -h
28G     S2R_plus_R1.fastq.gz
29G     S2R_plus_R2.fastq.gz
57G     total
```

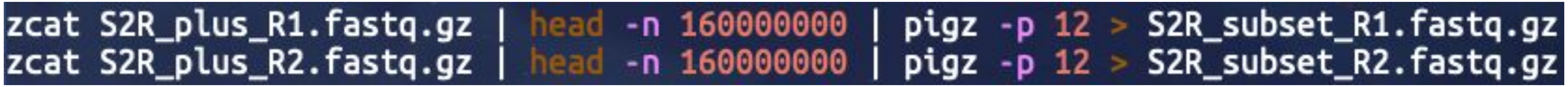
→

Size after downsampling



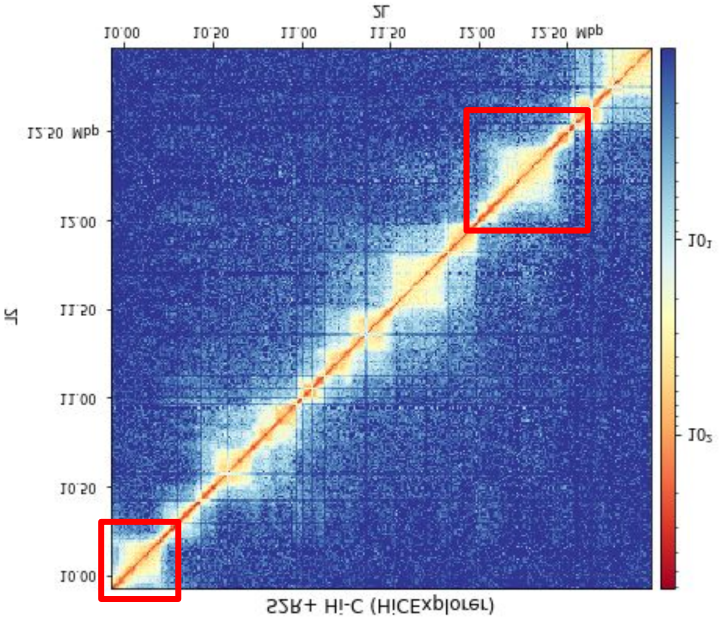
```
1.8G    S2R_subset_R1.fastq.gz
1.9G    S2R_subset_R2.fastq.gz
3.7G    total
```

Downsampling ~6.5% (40M Reads)



```
zcat S2R_plus_R1.fastq.gz | head -n 160000000 | pigz -p 12 > S2R_subset_R1.fastq.gz
zcat S2R_plus_R2.fastq.gz | head -n 160000000 | pigz -p 12 > S2R_subset_R2.fastq.gz
```

Fig 1A. Result (HiC Explorer)



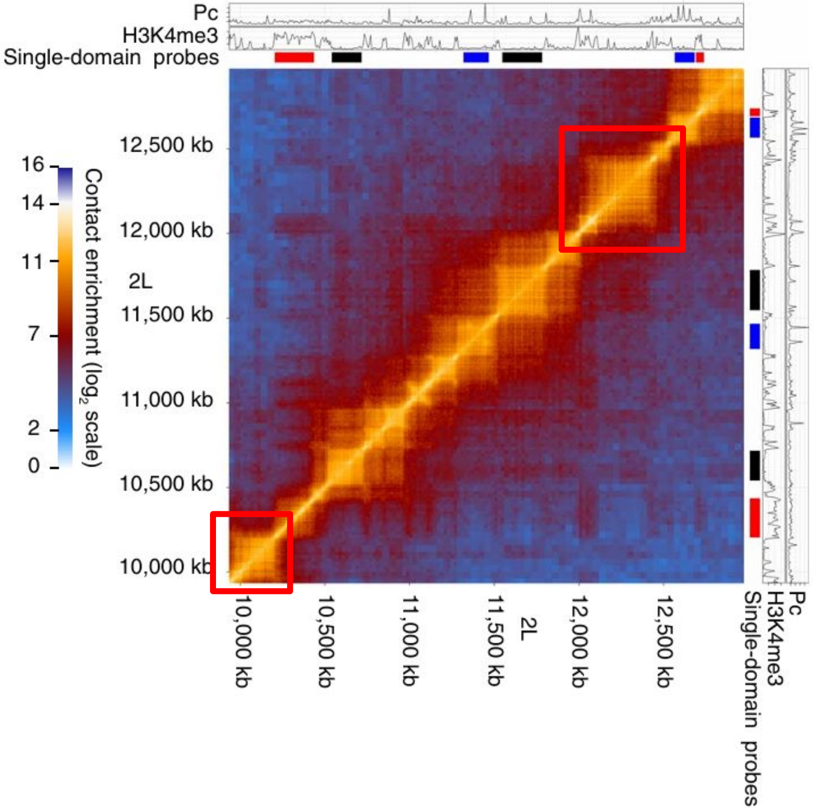
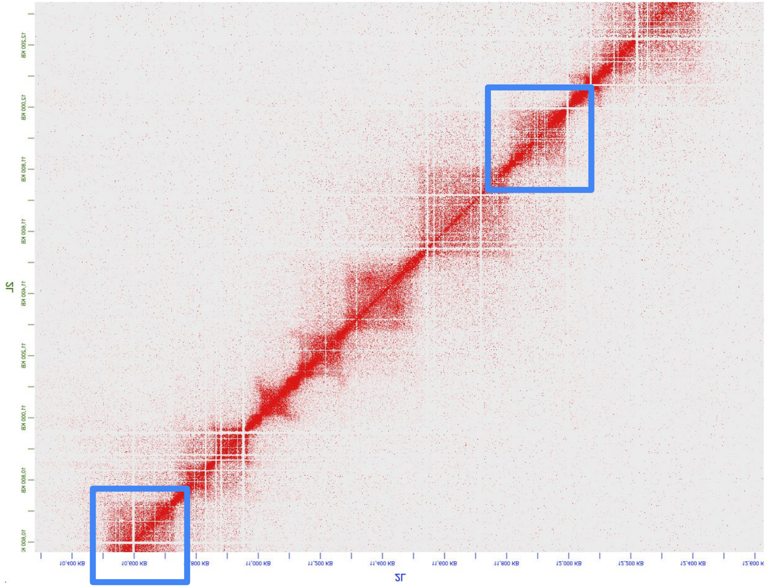
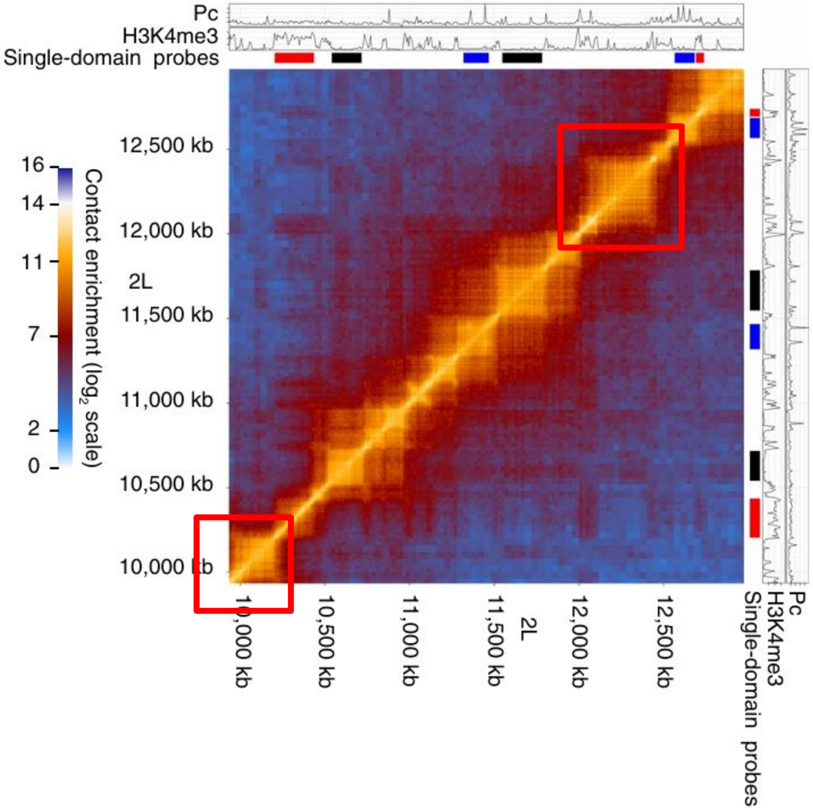


Fig 1A. Result (JuiceBox GUI)





Future Work

1. 如果有足夠的運算資源、硬碟空間和時間，會想要用完整的數據重跑Fig 1A.
2. 花太多時間在Fig 1A上，沒有處理其他圖的序列
   a. Fig 3A: Male Embryos
   b. Fig 5A: WT & Mutant
3. 網站上抓得到 H3K4me3、Pc等等histone mark的資料，會希望也做出論文中圖表旁邊的附加資訊。
   a. https://flybase.org/reports/FBlc0002296

補充

參考資料

我們基因組的三維組織