Zhang-Lab 生信小课堂 第六期

和趣求真 秉实生信

（张建伟生物信息学课题组 https://zhang.hzau.edu.cn）

Hi-C辅助基因组组装

High-throughput Chromosome Conformation Capture

地点：二综一楼C102 时间：15:00 (2022.10.07)

欢迎大家一起交流学习！

主讲人：韩婉欣

2022/10/07

Zhang-Lab 生信小课堂 第六期

和趣求真 秉实生信

（张建伟生物信息学课题组 https://zhang.hzau.edu.cn）

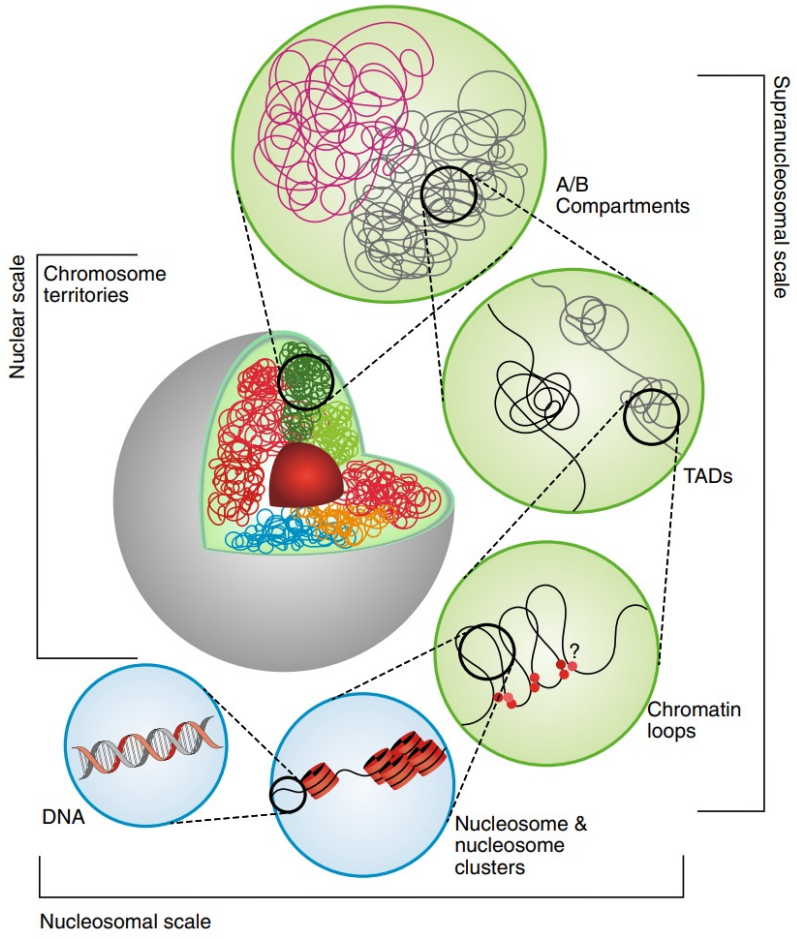
Hi-C辅助基因组组装

High-throughput Chromosome Conformation Capture

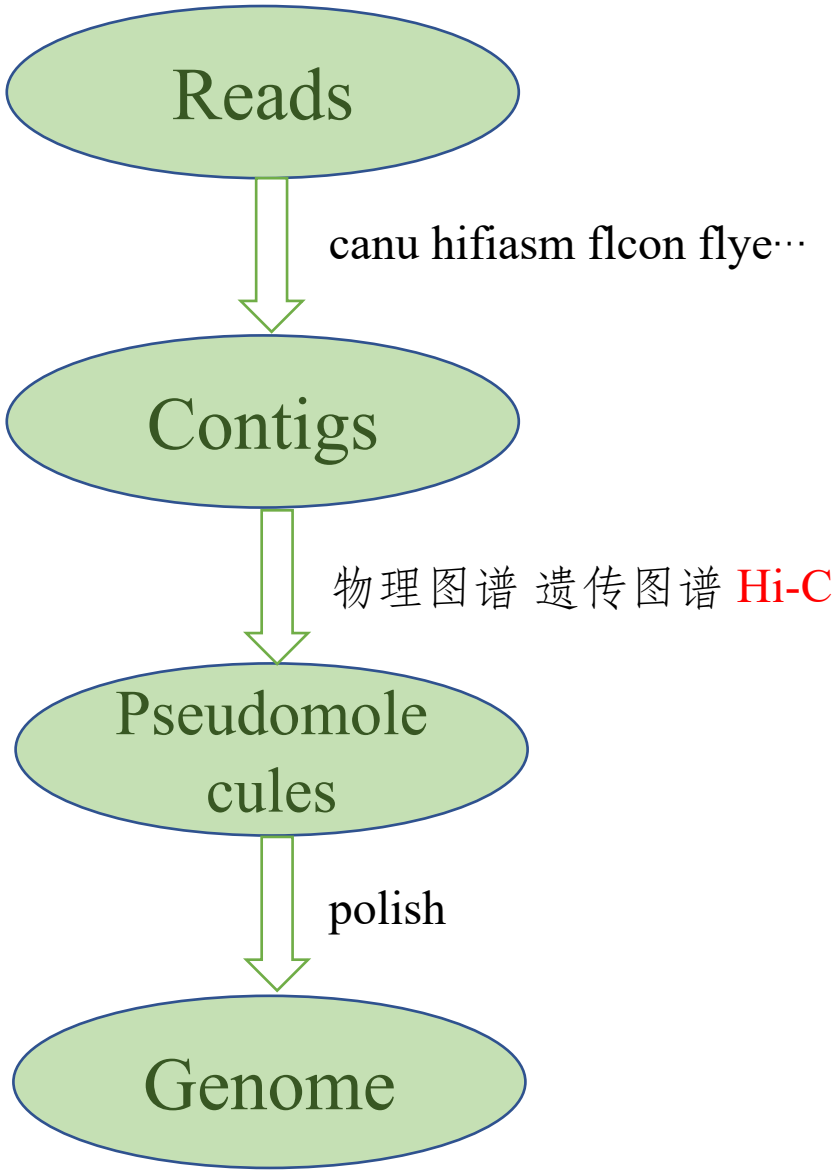
主讲人：韩婉欣

2022/10/07

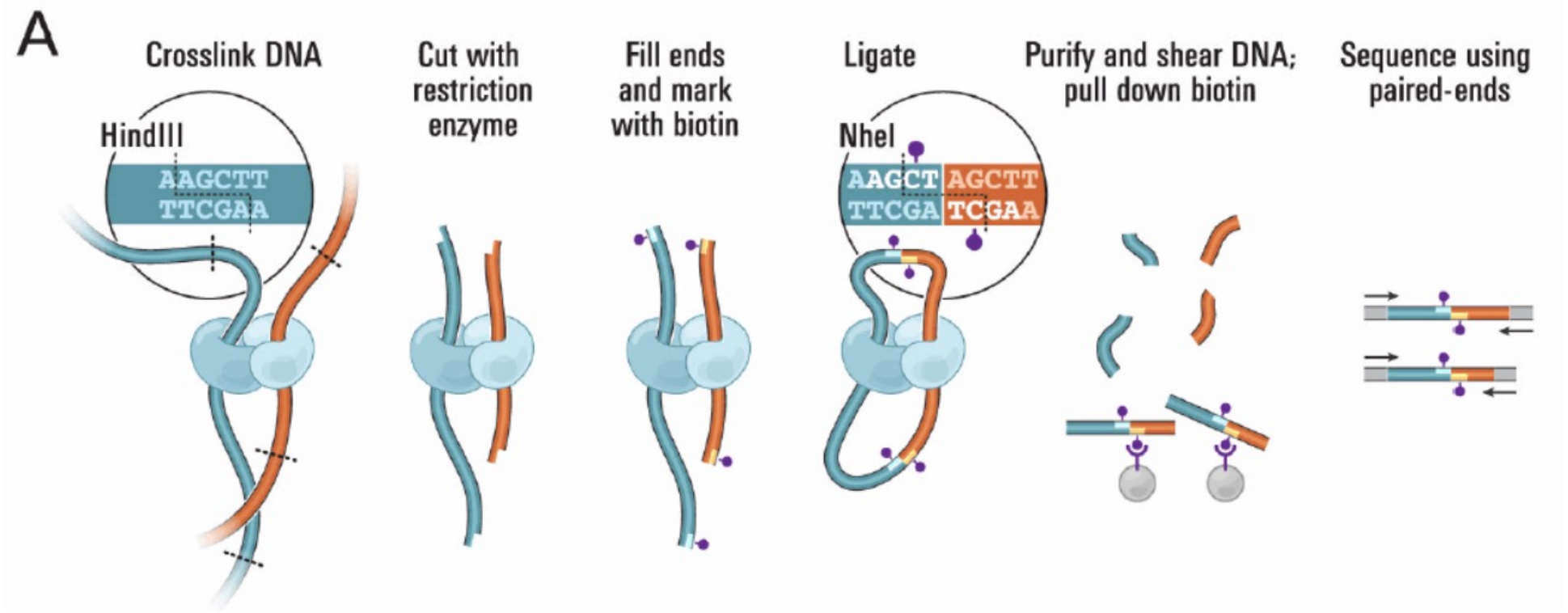
染色体与基因组组装



染色体分级示意图



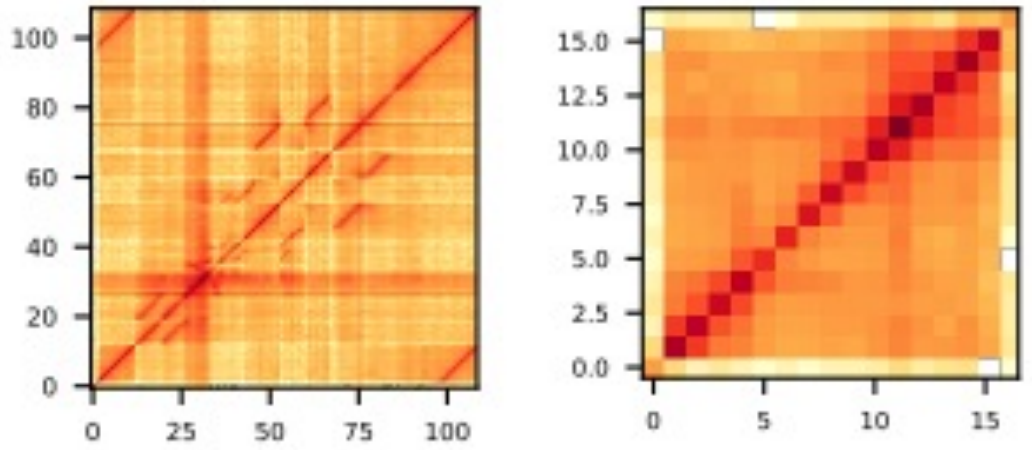
Hi-C技术



Hi-C技术原理

Hi-C互作规律

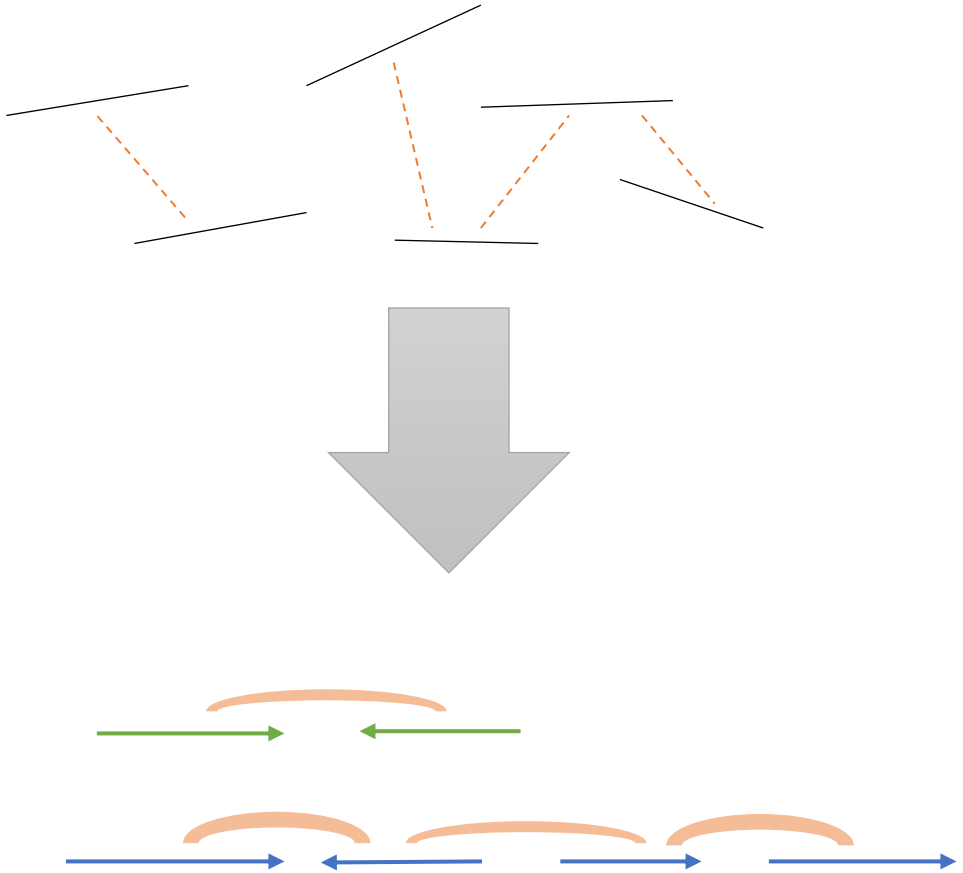
- 染色体内互作富集
- 互作随距离衰减
- 局部互作平滑



可以利用Hi-C互作规律实现将contigs/scaffolds挂载到pseudomolecules水平

Hi-C辅助组装软件

- ✓ 3D-DNA
- ✓ ALLHiC
- ✓ HiC-Pro
- ✓ SALSA
- ✓ LACHESIS



组装软件：3D-DNA+Juicebox

1.软件安装

#Juicer

```
mkdir -p ~/opt/biosoft/
cd ~/opt/biosoft
git clone https://github.com/theaidenlab/juicer.git
cd juicer
ln -s CPU scripts
cd scripts/common
wget https://hicfiles.tc4ga.com/public/juicer/juicer_tools.1.9.9_jcuda.0.8.jar
ln -s juicer_tools.1.9.9_jcuda.0.8.jar  juicer_tools.jar
```

#3D-DNA

```
cd ~/opt/biosoft
git clone https://github.com/theaidenlab/3d-dna.git
```

#Juicebox

安装在windows系统上

2. 软件使用

#建索引
bwa index hap1-2.fa
#根据基因组构建创建酶切位点文件
python /public/home/wxhan/software/juicer-master/misc/generate_site_positions.py DpnII genome hap1-2.fa
#获取每条contig长度
awk 'BEGIN{OFS="\t"}{print $1, $NF}' genome_DpnII.txt > genome.chrom.sizes
#运行Juicer
bash /public/home/wxhan/Santalum_album/chenyang0927003/assembly/hifiasm/Hi-C/3ddna/software/juicer-master/scripts/juicer.sh -d /public/home/wxhan/Santalum_album/chenyang0927003/assembly/hifiasm/Hi-C/3ddna/ -D /public/home/wxhan/Santalum_album/chenyang0927003/assembly/hifiasm/Hi-C/3ddna/software/juicer-master/ -g sandalwood -z reference/hap1-2.fa -y reference/genome_DpnII.txt -p reference/genome.chrom.sizes -s DpnII -t 线程数
bash ~/software/3d-dna-master/run-asm-pipeline.sh ./reference/hap1-2.fa ./aligned/merged_nodups.txt

运行结果文件

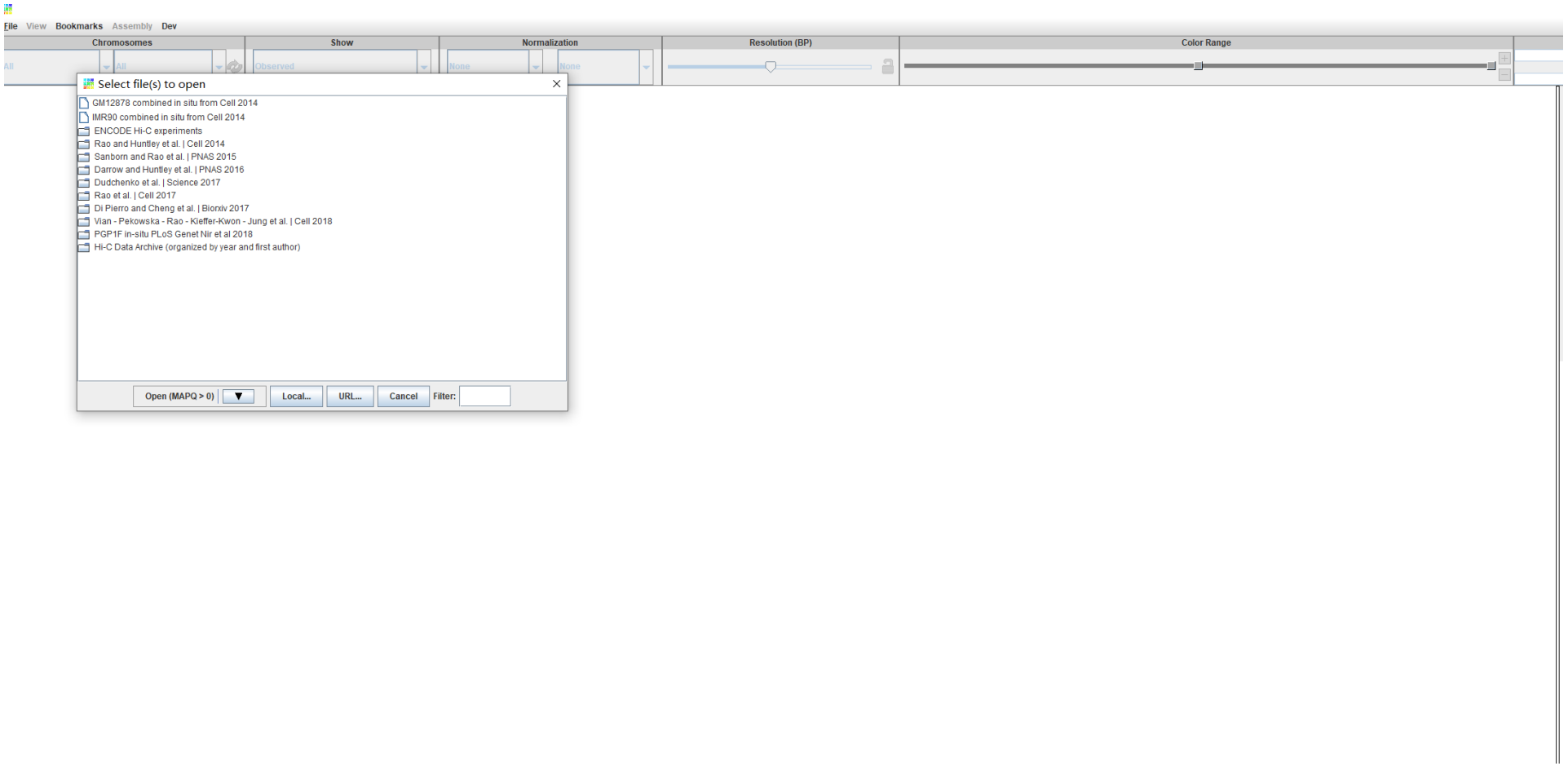
```
S_album.hic.hap1.p_ctg.0.asm
S_album.hic.hap1.p_ctg.0_asm.scaffold_track.txt
S_album.hic.hap1.p_ctg.0_asm.superscaf_track.txt
S_album.hic.hap1.p_ctg.0.assembly
S_album.hic.hap1.p_ctg.0.cprops
S_album.hic.hap1.p_ctg.0.hic
S_album.hic.hap1.p_ctg.0.review2.assembly
S_album.hic.hap1.p_ctg.1.asm
S_album.hic.hap1.p_ctg.1_asm.scaffold_track.txt
S_album.hic.hap1.p_ctg.1_asm.superscaf_track.txt
S_album.hic.hap1.p_ctg.1.assembly
S_album.hic.hap1.p_ctg.1.cprops
S_album.hic.hap1.p_ctg.1.hic
S_album.hic.hap1.p_ctg.2.asm
S_album.hic.hap1.p_ctg.2_asm.scaffold_track.txt
S_album.hic.hap1.p_ctg.2_asm.superscaf_track.txt
S_album.hic.hap1.p_ctg.2.assembly
S_album.hic.hap1.p_ctg.2.cprops
S_album.hic.hap1.p_ctg.2.hic
S_album.hic.hap1.p_ctg.cprops
S_album.hic.hap1.p_ctg.edits.txt
```

将结果文件.hic和.assembly导入Juicebox，进行调整

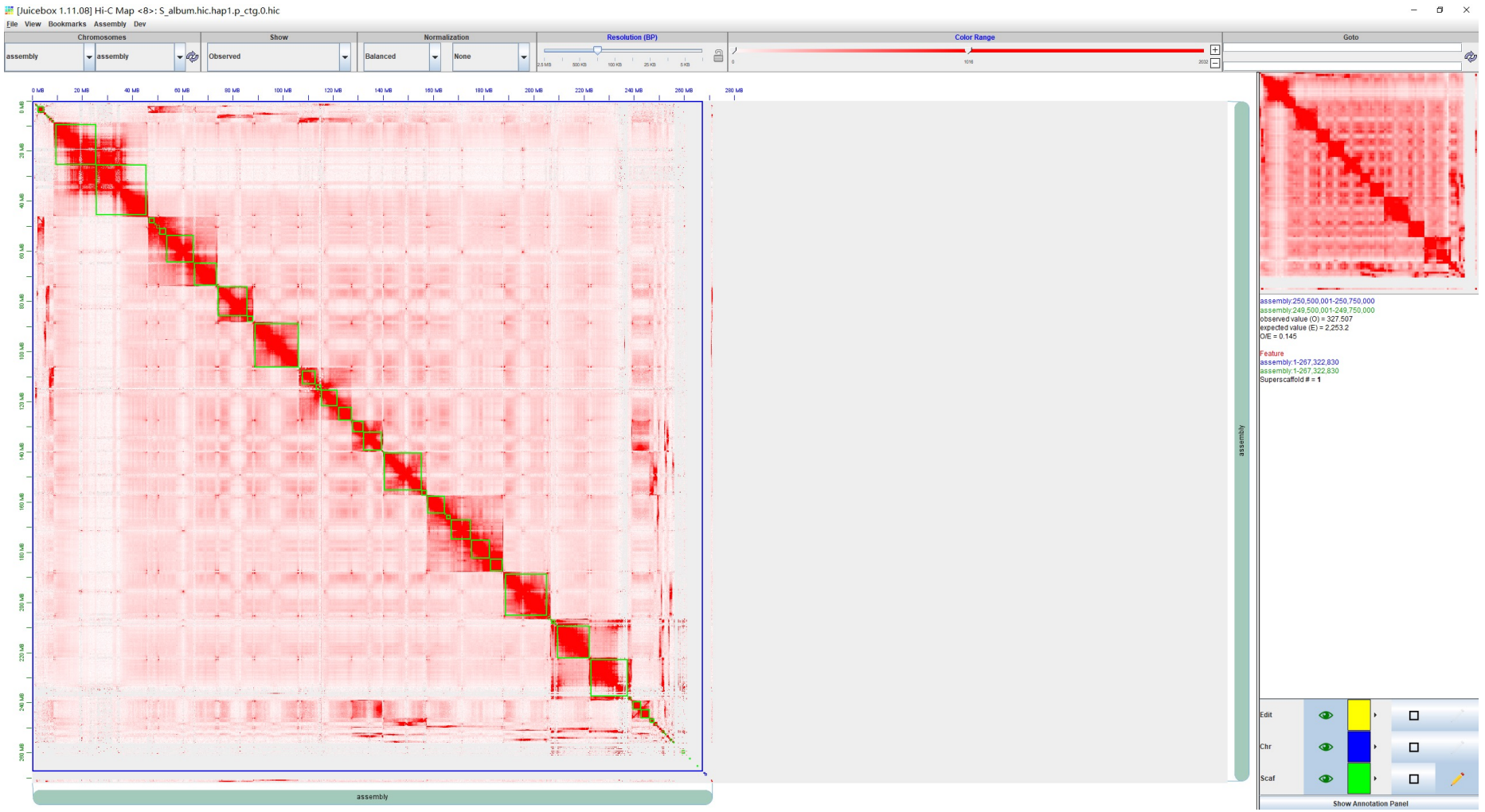
3. Juicebox使用

导入文件

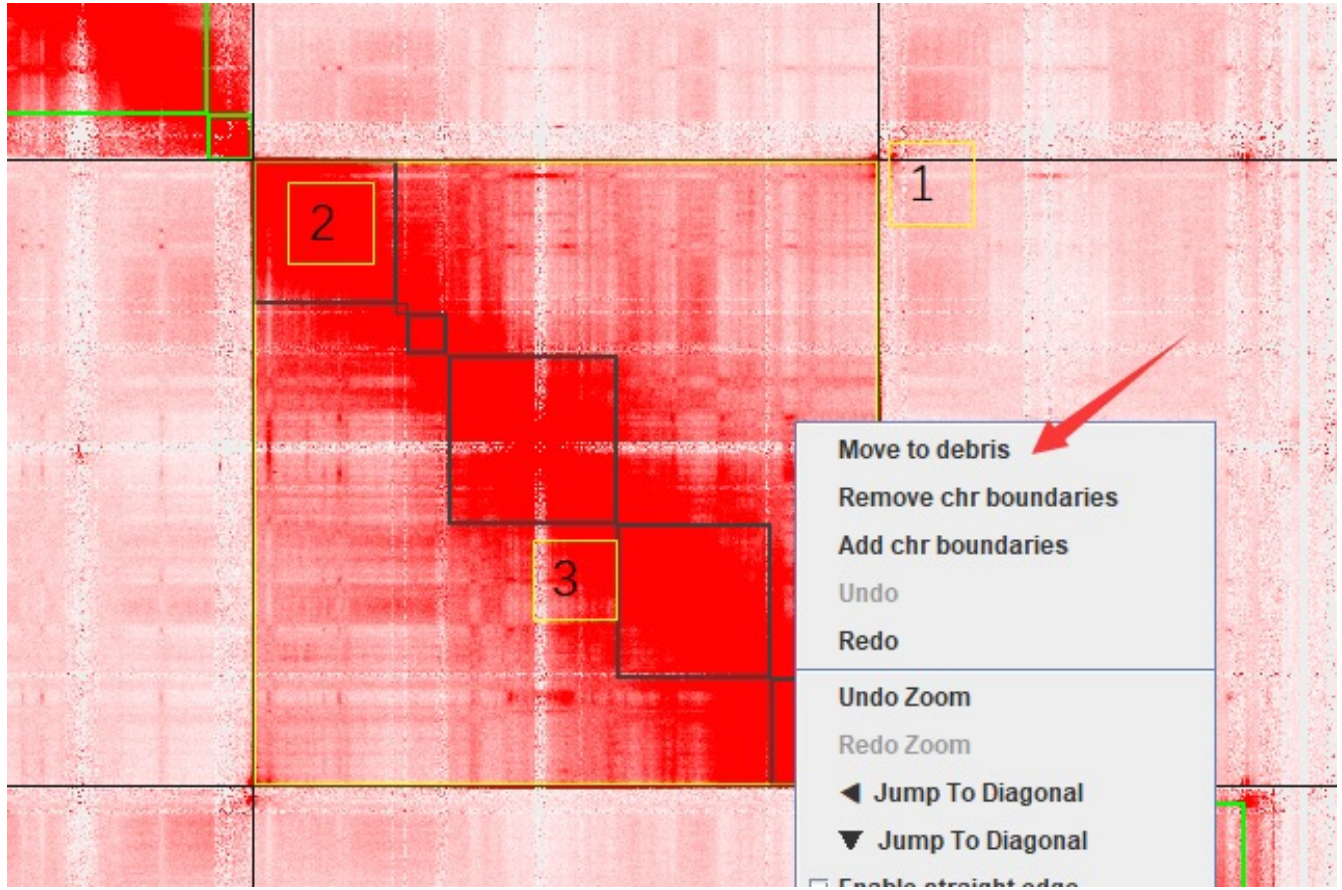
1. .hic: File > Open > Local
2. .assembly: Assembly > Import Map Assembly > Local



导入文件后的界面显示

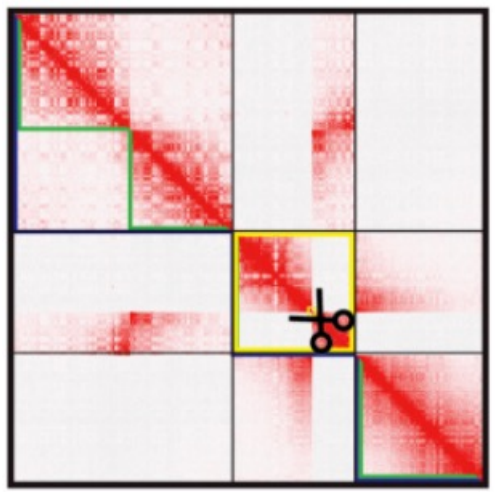

[Juicebox 1.11.08] Hi-C Map <8>: S_album.hic.hap1.p_ctg.0.hic
File View Bookmarks Assembly Dev
Chromosomes
assembly
assembly
Show
Observed
Normalization
Balanced
None
Resolution (BP)
Color Range
Goto
assembly
assembly
assembly:250,500,001-250,750,000
assembly:249,500,001-249,750,000
observed value (O) = 327.507
expected value (E) = 2,253.2
O/E = 0.145
Feature
assembly:1-267,322,830
assembly:1-267,322,830
Superscaffold # = 1
Edit
Chr
Scaf
Show Annotation Panel

操作指南

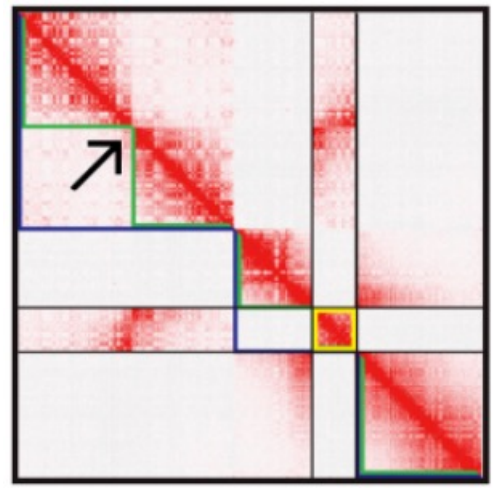


选中序列： Shift + 鼠标拖动
翻转序列： 箭头放在位置1
剪切序列： 箭头放在位置2（对角线上）
合并序列： 箭头放在位置3

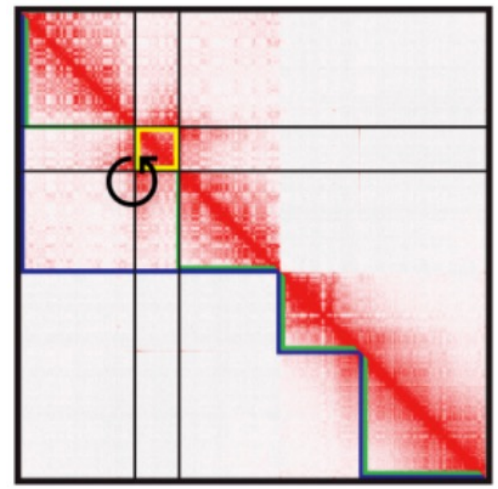
实例



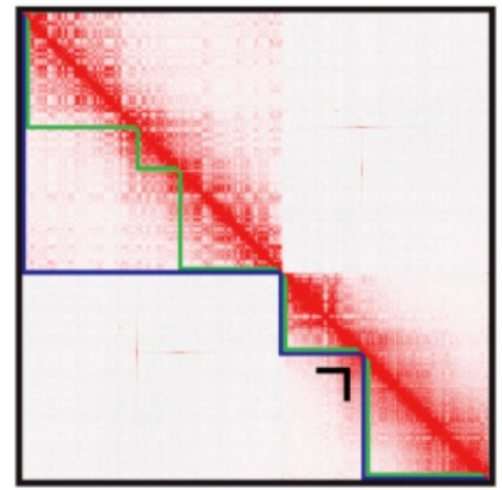
剪切



插入



翻转

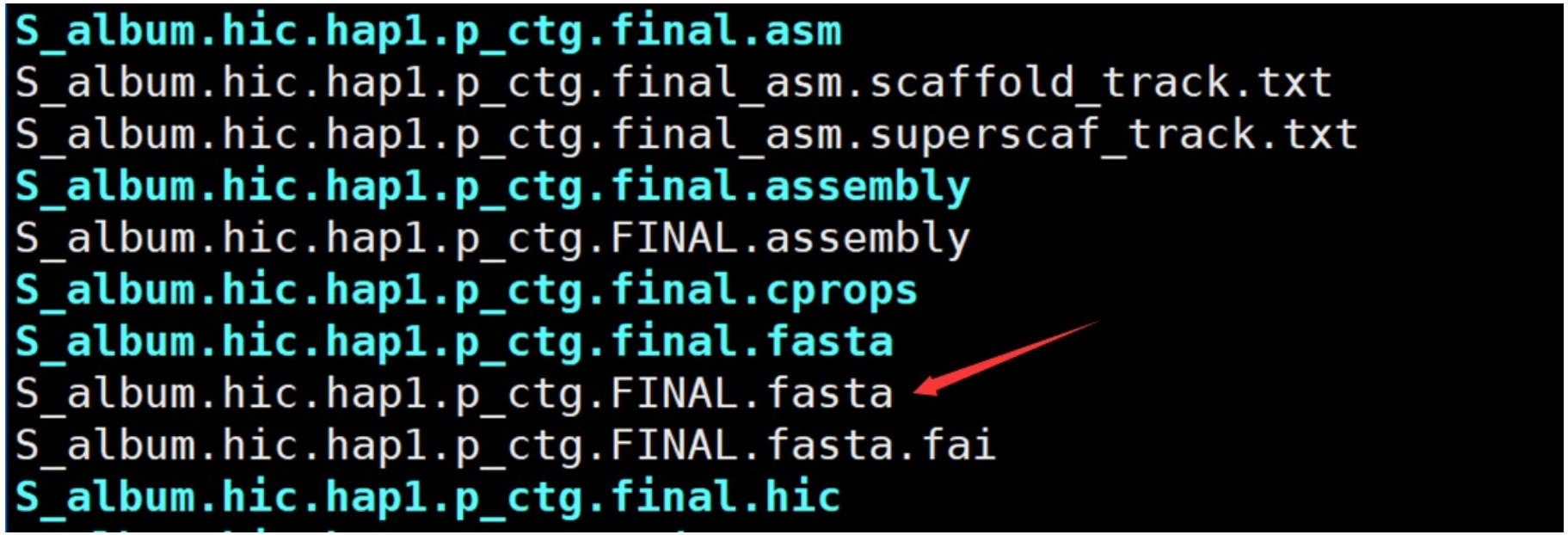


合并

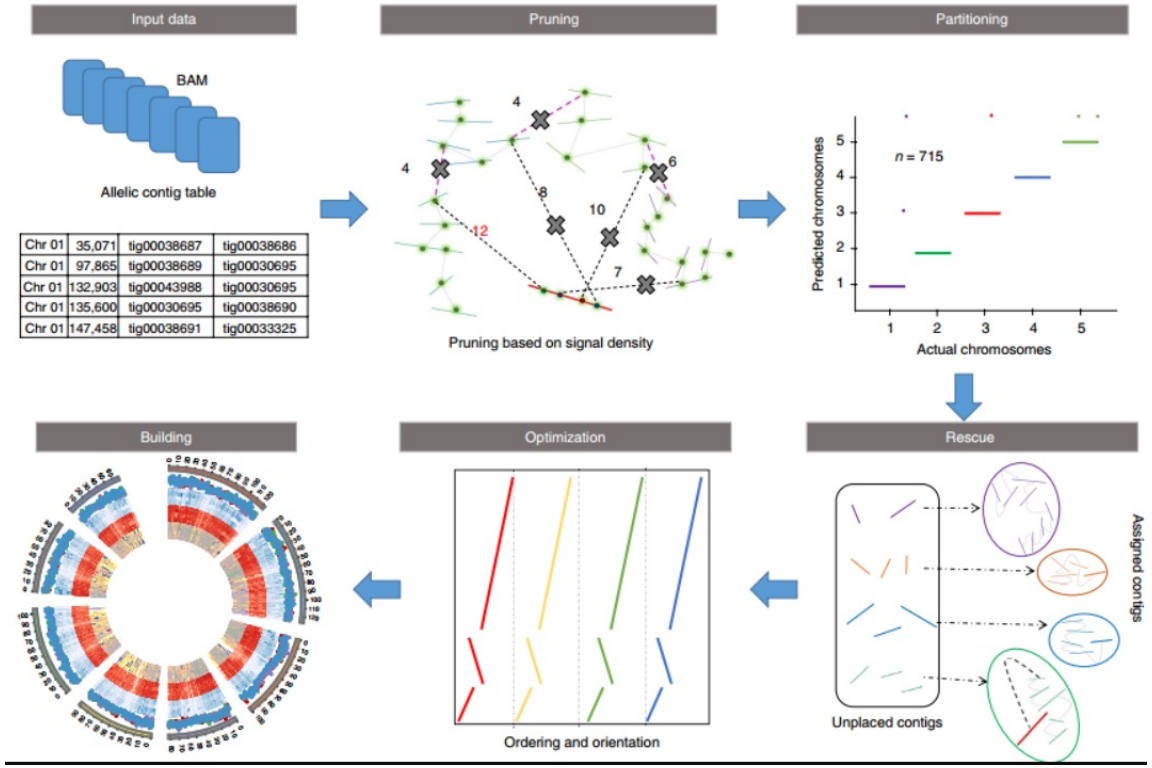
4. 重新运行3D-DNA

#Juicebox纠错之后，再次运行3D-DNA
bash ~/software/3d-dna-master/run-asm-pipeline-post-review.sh -r
S_album.hic.hap2.p_ctg.0.review.assembly ./reference/S_album.hic.hap2.p_ctg.fa aligned/merged_nodups.txt

#运行结束



组装软件：ALLHiC



Prune：修剪，去噪（可选，应用于多倍体基因组）

Partition：划分，凝聚层次聚类算法 对contig进行分组

Rescue：搜索最优互作信息，将未分组的contig继续进行分组

Optimize：确定每组的顺序和方向

Build：得到fasta序列文件

1. 软件安装

git clone https://github.com/tangerzhang/ALLHiC

2. 数据准备

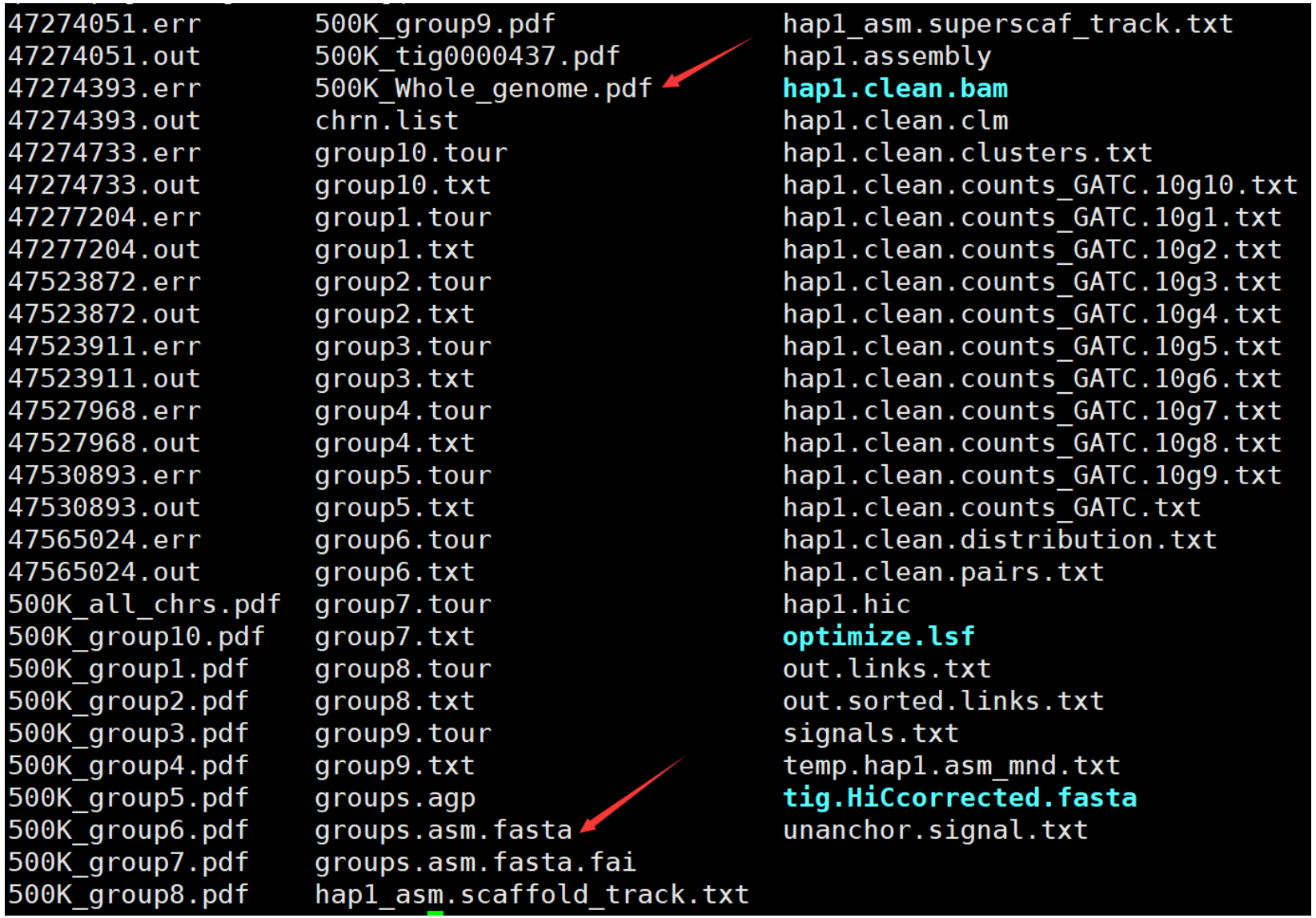
fasta文件（初步组装）
hicreads_R1.fq.gz
hicreads_R2.fq.gz

2. 数据前处理

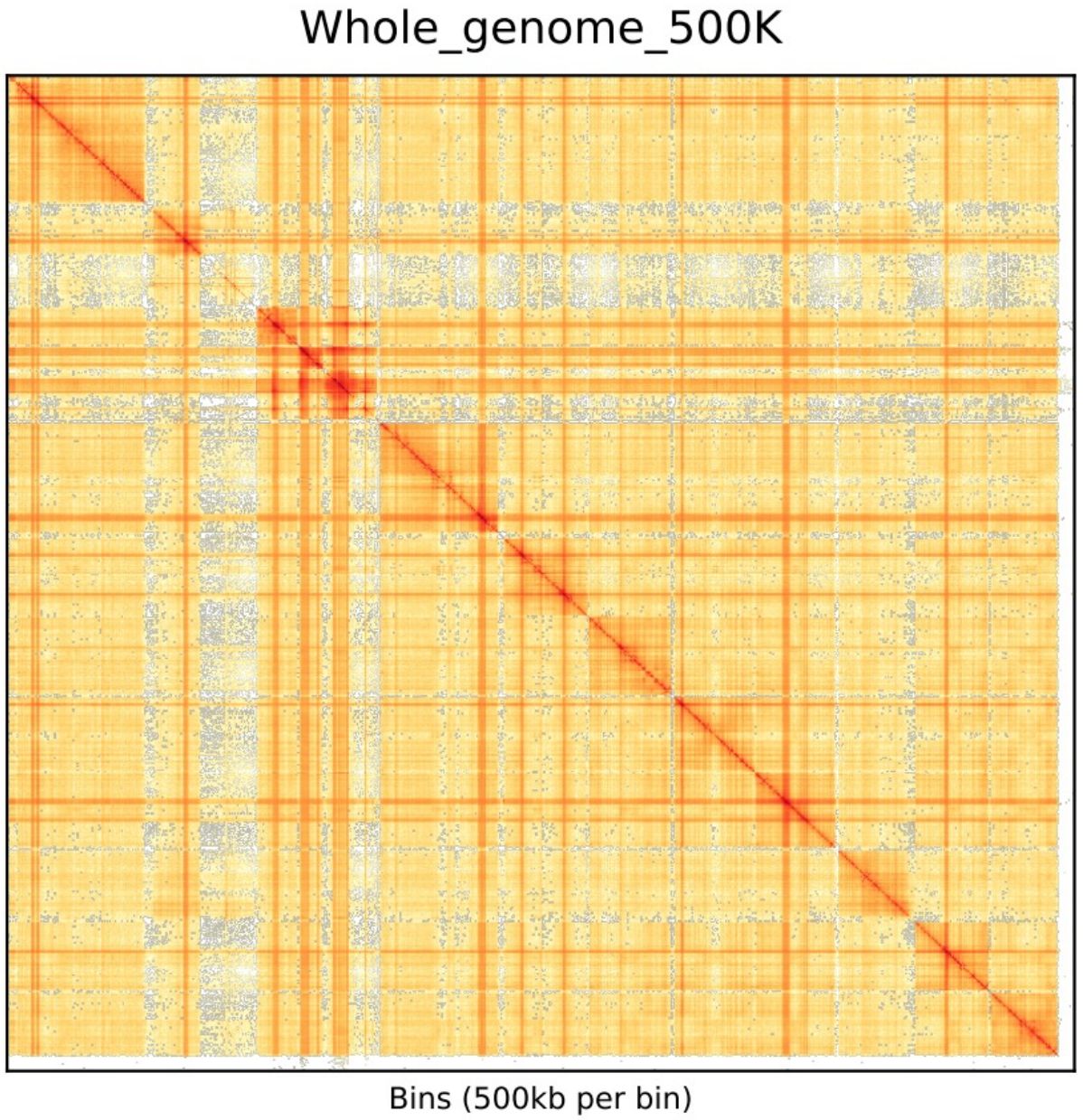
建立索引
samtools faidx hap1-2.fa bwa index -a bwtsw hap1-2.fa
序列回帖
bwa aln -t 4 hap1-2.fa ~/Santalum_album/chenyang0927003/clean/02.hic_clean/BDHC210000210-1A_L1_2_clean.rd.fq.gz > hap1-2_hicReads2.sai
bwa sampe hap1-2.fa hap1-2_hicReads1.sai hap1-2_hicReads2.sai ../../../../clean/02.hic_clean/BDHC210000210-1A_L1_1_clean.rd.fq.gz ../../../../clean/02.hic_clean/BDHC210000210-1A_L1_2_clean.rd.fq.gz | samtools view -bS > hap12-hic.bwa_aln.bam
#筛选bam文件
perl ~/software/ALLHiC-master/scripts/PreprocessSAMs.pl hap12-hic.bwa_aln.bam hap1-2.fa MBOI
~/software/ALLHiC-master/scripts/filterBAM_forHiC.pl hap12-hic.bwa_aln.REduced.paired_only.bam hap12.clean.sam
samtools view -bt hap1-2.fa.fai hap12.clean.sam > hap12.clean.bam

3. 软件使用

```
#聚类划分
ALLHiC_partition -b hap12.clean.bam -r hap1-2.fa -e GATC -k 20
#未划分的contigs/scaffolds继续分组
ALLHiC_rescue -b hap12.clean.bam -r hap1-2.fa -c hap12.clean.clusters.txt -i
hap12.clean.counts_GATC.txt
#优化
allhic extract hap12.clean.bam hap1-2.fa --RE GATC
for i in group*.txt; do     allhic optimize $i hap12.clean.clm; done
#组装
ALLHiC_build hap1-2.fa
#hic热图
samtools faidx groups.asm.fasta
cut -f 1,2 groups.asm.fasta.fai  > chrn.list
ALLHiC_plot hap12.clean.bam groups.agp chrn.list 500k pdf
```


47274051.err
47274051.out
47274393.err
47274393.out
47274733.err
47274733.out
47277204.err
47277204.out
47523872.err
47523872.out
47523911.err
47523911.out
47527968.err
47527968.out
47530893.err
47530893.out
47565024.err
47565024.out
500K_all_chrs.pdf
500K_group10.pdf
500K_group1.pdf
500K_group2.pdf
500K_group3.pdf
500K_group4.pdf
500K_group5.pdf
500K_group6.pdf
500K_group7.pdf
500K_group8.pdf
500K_group9.pdf
500K_tig0000437.pdf
500K_Whole_genome.pdf
chrn.list
group10.tour
group10.txt
group1.tour
group1.txt
group2.tour
group2.txt
group3.tour
group3.txt
group4.tour
group4.txt
group5.tour
group5.txt
group6.tour
group6.txt
group7.tour
group7.txt
group8.tour
group8.txt
group9.tour
group9.txt
groups.agp
groups.asm.fasta
groups.asm.fasta.fai
hap1_asm.scaffold_track.txt
hap1_asm.superscaf_track.txt
hap1.assembly
hap1.clean.bam
hap1.clean.clm
hap1.clean.clusters.txt
hap1.clean.counts_GATC.10g10.txt
hap1.clean.counts_GATC.10g1.txt
hap1.clean.counts_GATC.10g2.txt
hap1.clean.counts_GATC.10g3.txt
hap1.clean.counts_GATC.10g4.txt
hap1.clean.counts_GATC.10g5.txt
hap1.clean.counts_GATC.10g6.txt
hap1.clean.counts_GATC.10g7.txt
hap1.clean.counts_GATC.10g8.txt
hap1.clean.counts_GATC.10g9.txt
hap1.clean.counts_GATC.txt
hap1.clean.distribution.txt
hap1.clean.pairs.txt
hap1.hic
optimize.lsf
out.links.txt
out.sorted.links.txt
signals.txt
temp.hap1.asm_mnd.txt
tig.HiCcorrected.fasta
unanchor.signal.txt

运行结果



全基因组交互热图

后续可将结果文件转化成.hic和.assembly文件，导入juicebox进行调整

转化.hic和.assembly文件：

```
# .assembly
python agp2assembly.py in.agp out.assembly

# .hic
bash ~/software/3d-dna-master/visualize/run-assembly-visualizer.sh -p false hap1.assembly
out.sorted.links.txt
sort -k2,2 -k6,6 out.links.txt > out.sorted.links.txt
matlock bam2 juicer ../hap1.bwa_aln.bam out.links.txt
```

THANKS