基因组学培训

生信帮
云南农业大学
云南省生物大数据重点实验室
2023年8月19日

目录

第一章 基因组调研图

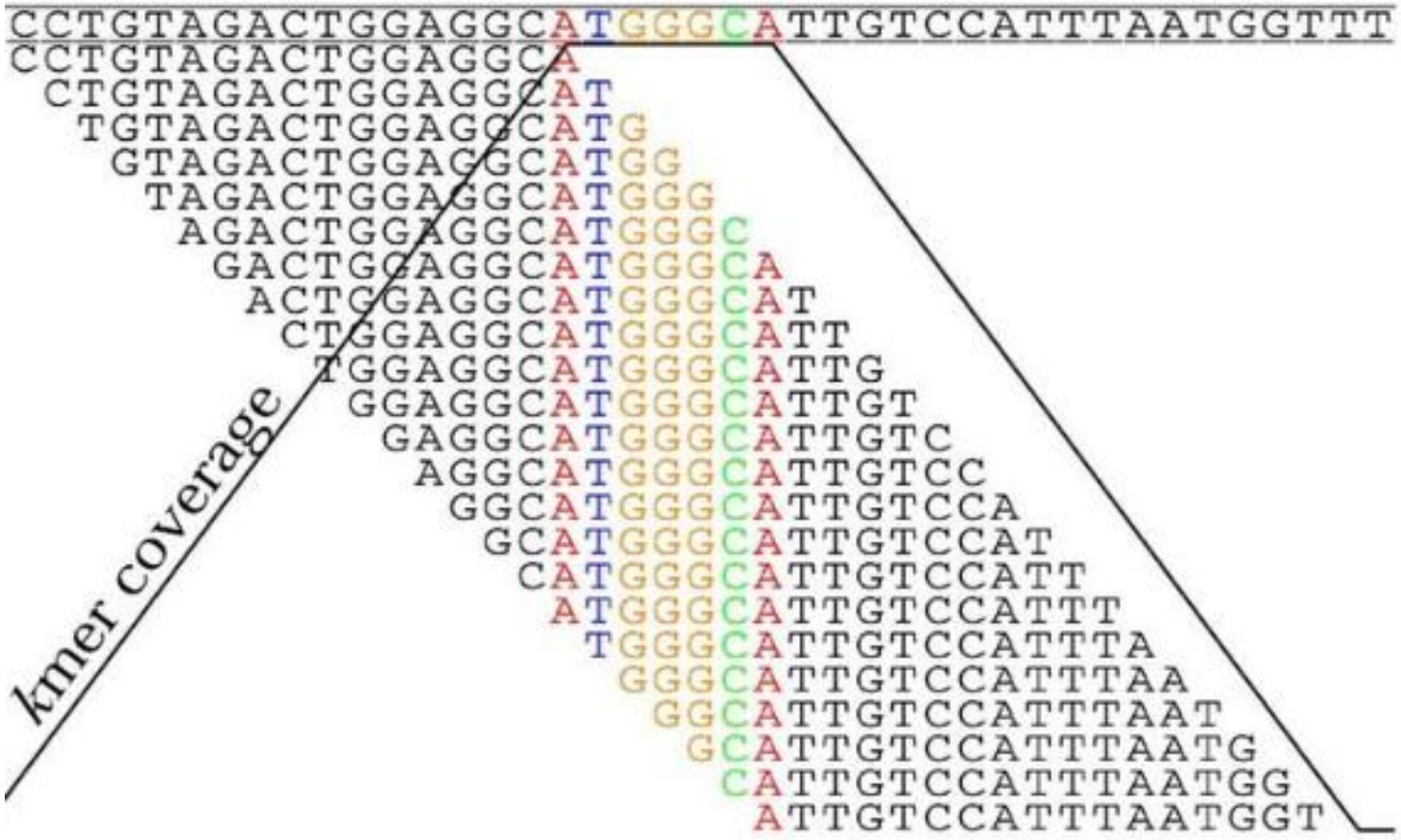
基因组调研图简介

- 基因组的杂合度和重复序列是影响后续基因组组装结果的重要因素。在进行基因组测序之前，了解目标生物体基因组的特征（重复和杂合）非常重要，可以帮助确定适合的测序方案和周期，同时也可以间接评估最终基因组大小是否符合预期等。
- 目前常用的调研手段有三种：
  - 用流式细胞仪测定细胞核内的DNA总量，估计基因组大小。
  - 用核型分析方法，识别染色体数量、倍性
  - 用Kmer，通过二代测序(30X)，估算基因组大小、杂合度、重复序列比例、GC含量等。
- 分析软件：GenomeScope 2.0和smudgeplot

K-mer

K-mer指的是是由K个相邻的碱基组成的片段。如将一条read,连续按照指定Kmer切割，得到的LEN-K+1条核苷酸序列。

K-mer深度：相同k-mer序列重复的次数。

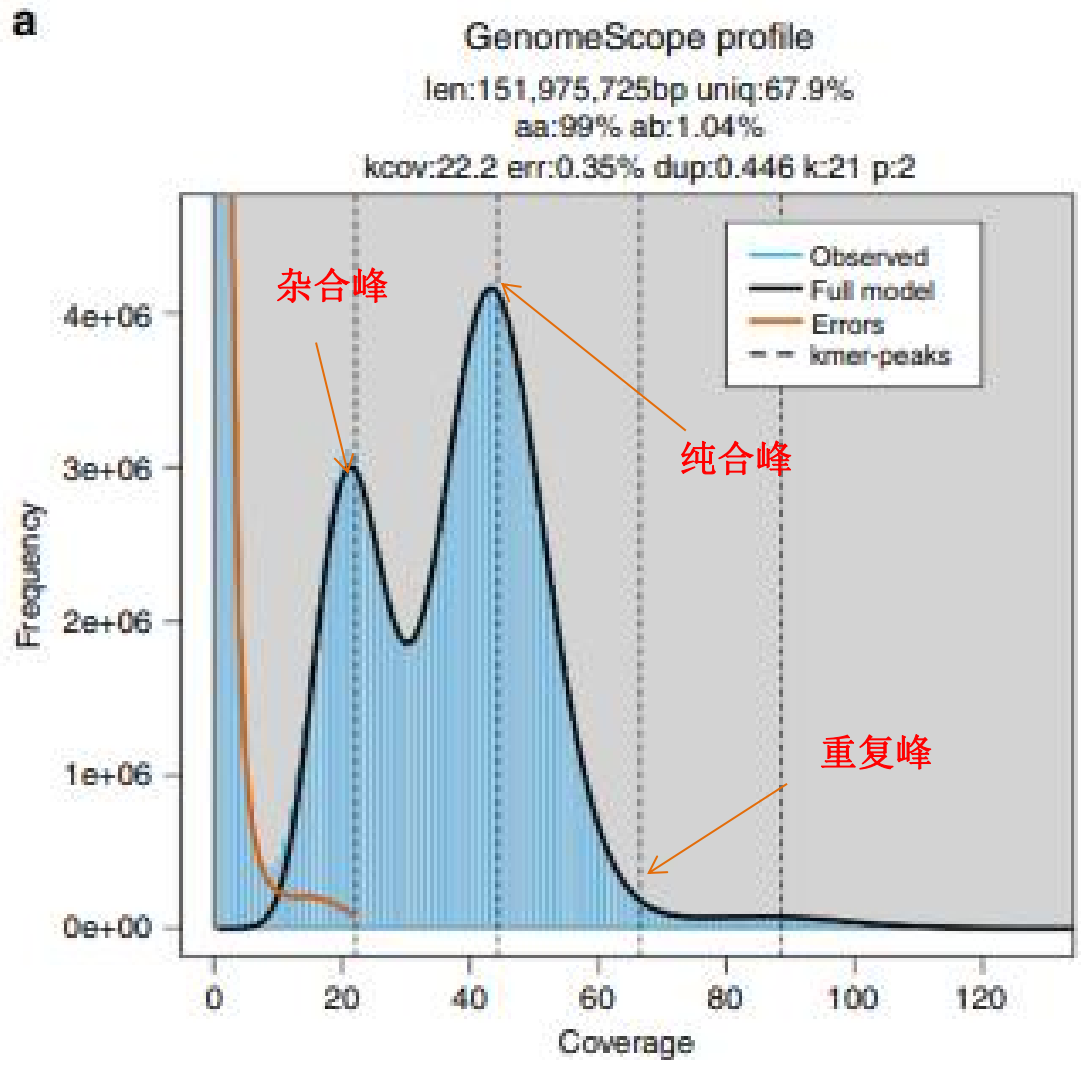


K-mer

- Kmer值得选取公式 k~=log(200G)/log(4)，当基因组大小<100M选17，<1G选19，>1G选21。
- 如果K-mer太短，会导致多数K-mer在基因组中出现多次，从而使主峰的深度估计偏大。这可能会给后续的分析带来误导。另一方面，较长的K-mer可以跨越更长的重复片段，提高了对复杂基因组的解析能力。
- 通常，K-mer的选择范围为15到21的奇数，这既可以确保K-mer的种类能够覆盖整个基因组，又足够小以避免错误的影响。在选择具体的K-mer长度时，需要根据具体的测序数据和基因组特点进行评估和调整。
- 基因组中确实存在一些重复序列，它们可能会导致K-mer的重复出现。尽管这些重复序列会降低主峰的高度，但不会改变主峰的位置。因此，在K-mer曲线上主峰后出现的小峰往往表示重复序列的存在。
- 在评估杂合率和重复序列时，除了观察K-mer曲线的特征，还可以使用其他分析方法和工具，如比对和组装算法、重复序列识别等，以综合考虑不同因素，并得出更准确的结果。这些方法的选择和应用取决于具体的研究目的和数据特点。

参考：http://www.cbcb.umd.edu/software/quake/faq.html

k-mer分布图



- 在高杂合基因组中，由于存在杂合位点，部分K-mer会被分为两个亚型，进而导致覆盖深度减半，使得K-mer计数也减半。因此，在K-mer曲线上，主峰前约在横坐标的1/2处会出现一个小峰，该小峰对应的频率值大约是主峰处频率值的一半。因此，K-mer曲线上在主峰前出现的小峰可以被视为基因组杂合度的指示器。当基因组的杂合度越高时，这个小峰就越明显。
- 主峰：通常是频率最高的峰，其对应的深度通常代表覆盖基因组的平均深度。Kmer分析中此值用于确定基因组大小，是基因组调研图中最重要的一个值。

横轴Coverage：K-mer测序深度；纵轴Frequency：K-mer占比；蓝色柱子是kmer的观测值；橙红色拟合线对应深度过低的kmer（测序错误）；黑色拟合线是除去错误后剩下的可靠的kmer数据；垂直的黑色虚线为预测最低深度峰的整数倍覆盖度；

K-mer评估基因组大小

前提：测序的数据量要足够大，使得read尽量均匀的覆盖整个基因组.然后将reads全部转化为K-mer,设得到K-mer总数为M，再通过统计获得K-mer的平均深度，设为H，那么评估基因组大小（G）：

$$M = H * (G - K + 1)$$

其中，M表示最后得到的K-mer数量，H表示测得的reads平均乘数，G表示基因组的大小（长度），K表示K-mer的长度。

制定合理的基因组测序方案

- 了解目标生物体基因组的特征是进行基因组测序前的重要准备工作。通过综合考虑杂合度、重复序列和其他相关因素，可以选择适当的测序方案和周期，并在后续数据分析中获得更准确的基因组组装结果。根据目标研究的目的和预算，确定所需的测序覆盖度。较高的测序覆盖度可以提供更准确的基因组组装结果，但相应地会增加成本和数据量。具体应根据杂合度和重复含量确定。
- 杂合度：高杂合度的基因组往往具有更多的等位基因和变异，这可能导致在组装过程中无法合并姊妹染色体。为了解决这个问题，可以使用长读长（Long Reads）测序技术，如PacBio HiFI或Oxford Nanopore超长等，这些技术能够产生较长的读长，有助于跨越高度一致区域并解决杂合度问题。

- 重复序列：重复序列是基因组中存在的相似或完全相同的DNA序列片段。在基因组组装过程中，重复序列容易造成混淆和错误。针对重复序列含量高的物种，可以增加测序覆盖度和使用长读长测序来减少组装错误。

- 数据分析提供参考。基因组测序后，还需要进行数据处理和分析。这包括质量控制、去除重复序列、错误校正和组装等步骤，以获得最终的基因组组装结果。

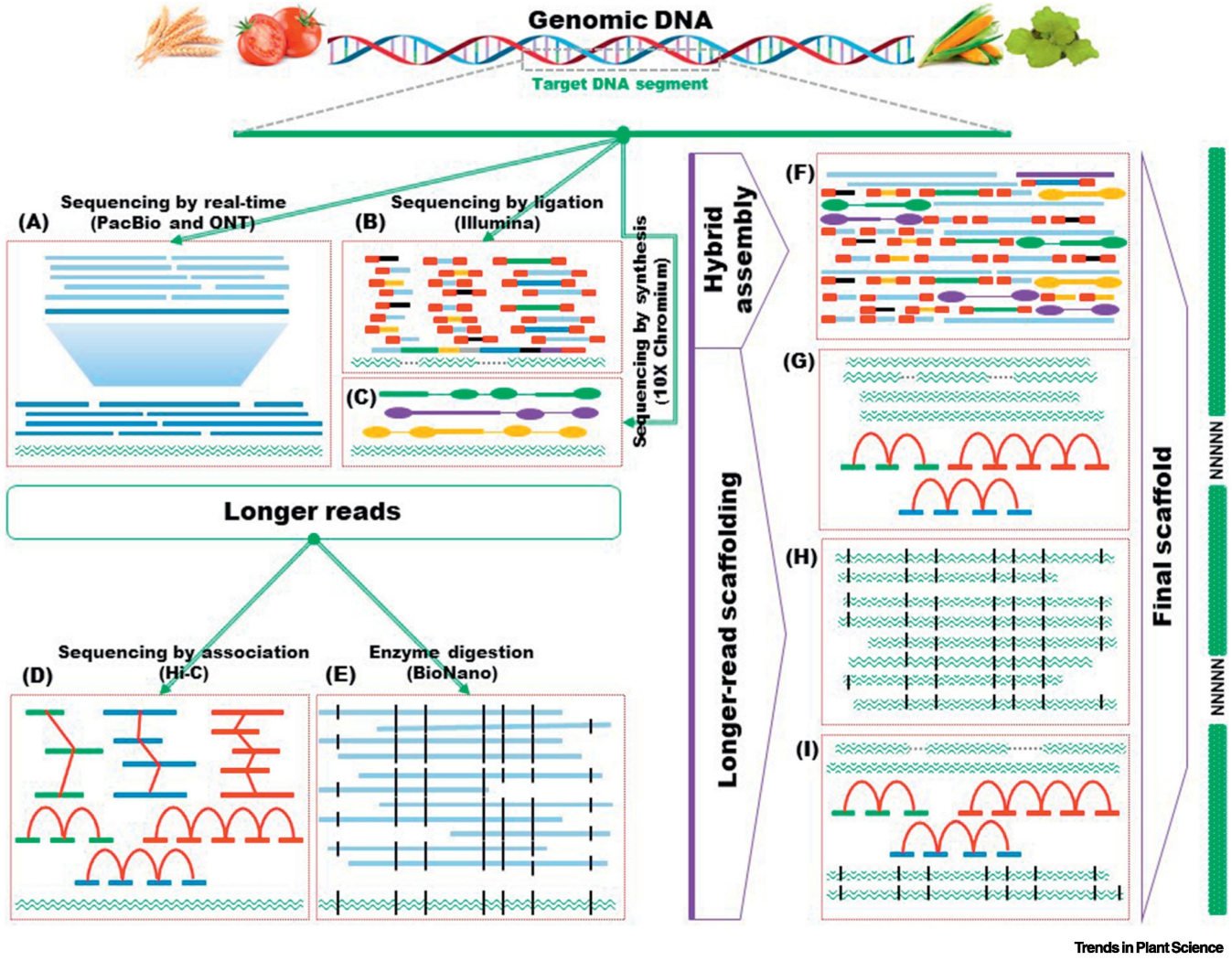
第二章 基因组组装

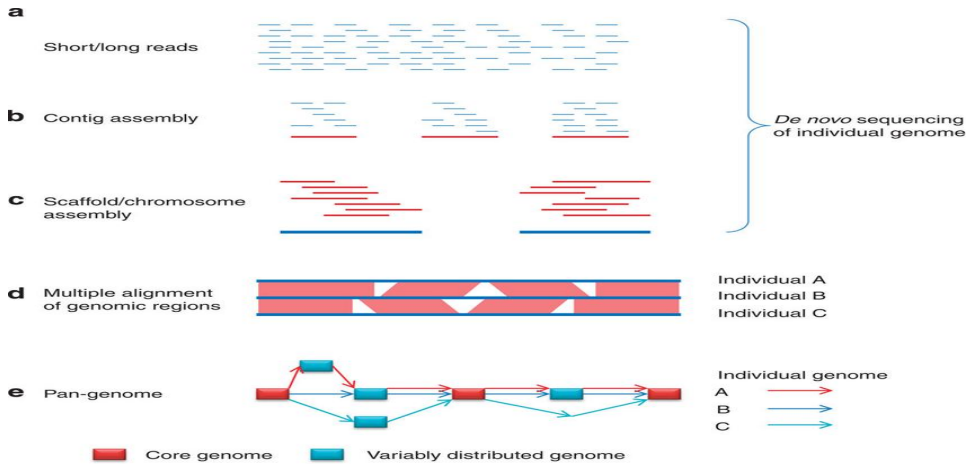
基因组denovo组装

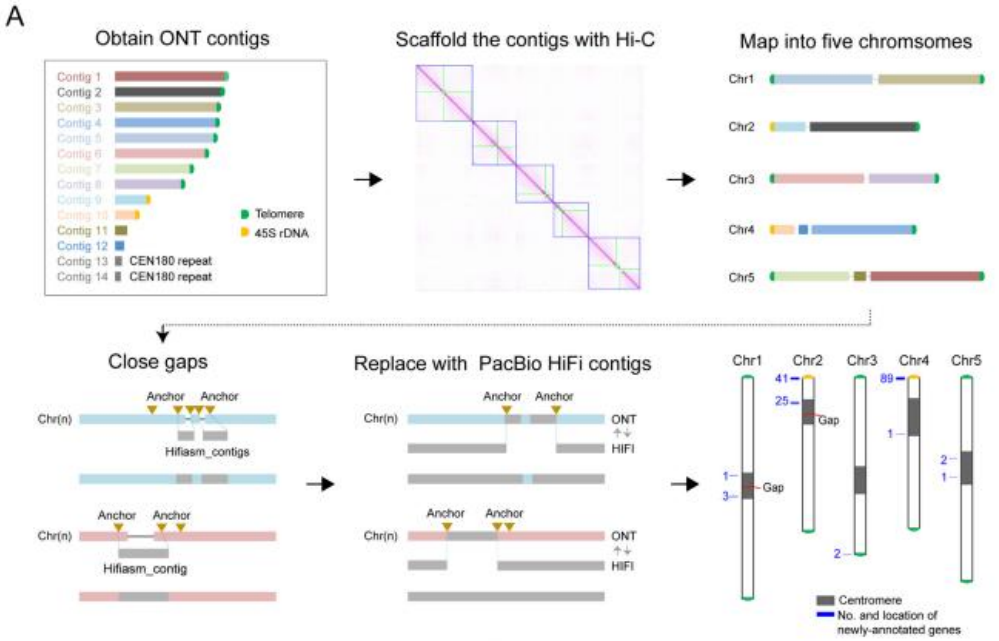
基因组 de novo 组装是指在没有参考基因组序列的情况下，从测序数据中重建出一个完整的基因组序列的过程。这个过程主要包括两个步骤：序列拼接和序列重组。

- 数据预处理：首先，需要对原始测序数据进行预处理，包括去除低质量的碱基、修剪接头序列等。这可以使用一些专门的工具或流程来完成，如Trimmomatic、FastQC等。
- 序列拼接：拼接是将原始测序数据中的短序列片段（reads）合并起来，以重建出原始基因组的完整序列。这一步骤通常使用拼接软件，如Hifiasm等，它们会根据短序列的重叠信息将其组装成较长的连续序列（contigs）。
- 长序列构建：由于基因组中可能存在重复序列或染色体间的相似区域，拼接得到的连续序列可能并不完整。为了解决这个问题，需要利用更长的序列来构建完整的基因组。这可以通过生成跨连接信息（paired-end、mate-pair reads，光学图谱）来实现，这些信息能帮助确定不同序列的相对位置和方向。
- 组装评估与修正：得到的组装结果可能包含错误或不完整的部分，在这一步骤中需要对组装结果进行检查和评估。常见的评估指标包括N50值、统计分析等。
- 需要注意的是，基因组 de novo 组装是一个复杂且计算密集的过程，其结果受到测序数据的质量和数量、物种的复杂性、参数的选择等多种因素的影响。在实际应用中，可能需要多次尝试不同的软件和参数组合，以获得最佳的组装结果。

组装原理图

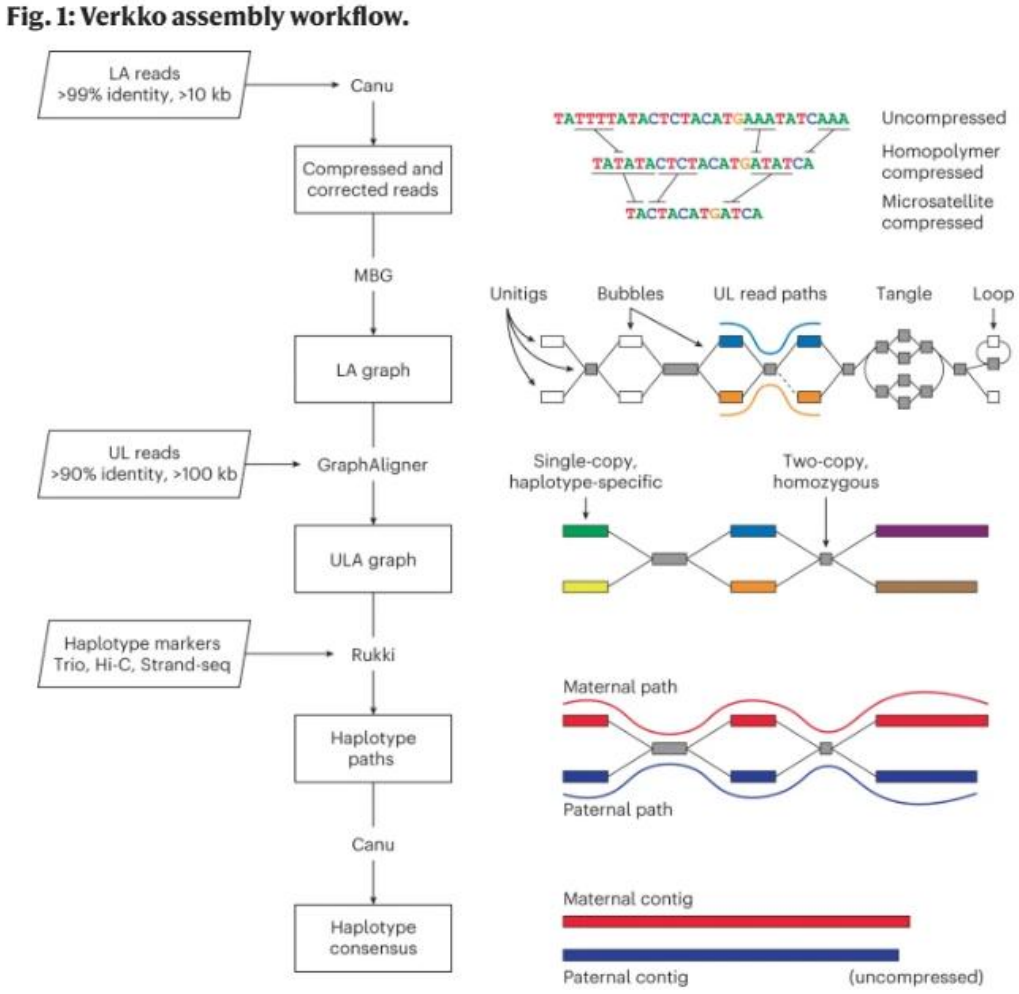

Genomic DNA
Target DNA segment
Sequencing by real-time (PacBio and ONT)
(A)
Sequencing by ligation (Illumina)
(B)
(C)
Sequencing by synthesis (10X Chromium)
Longer reads
Sequencing by association (Hi-C)
(D)
Enzyme digestion (BioNano)
(E)
Hybrid assembly
Longer-read scaffolding
(F)
(G)
(H)
(I)
Final scaffold
NNNNN
NNNNN
Trends in Plant Science


a
Short/long reads
b
Contig assembly
c
Scaffold/chromosome assembly
De novo sequencing of individual genome
d
Multiple alignment of genomic regions
Individual A
Individual B
Individual C
e
Pan-genome
Individual genome
A
B
C
Core genome
Variably distributed genome

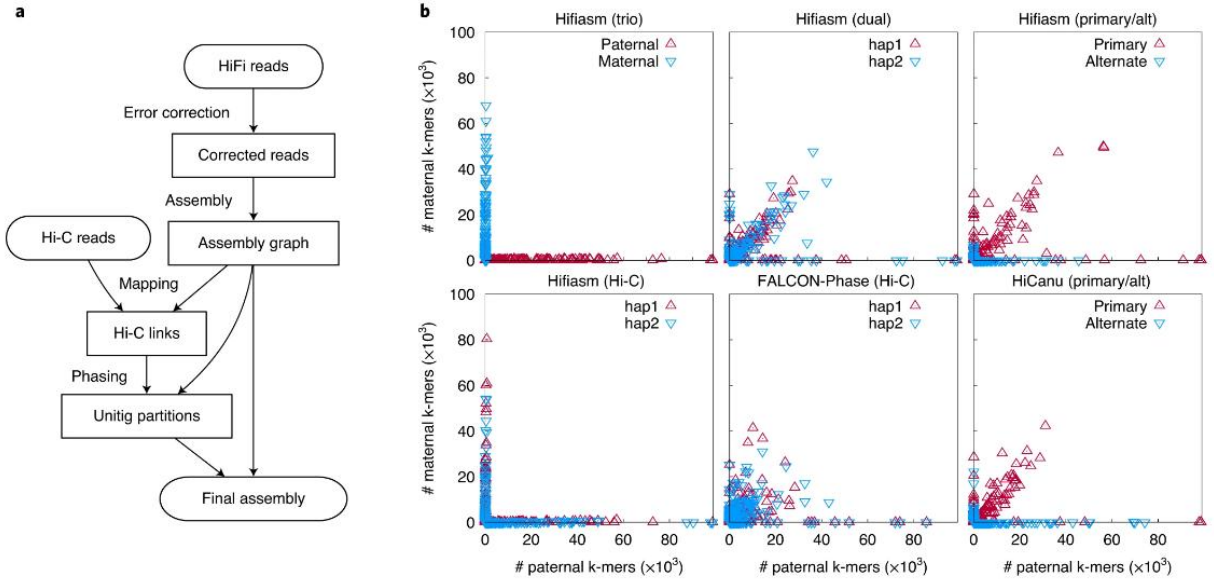

A
Obtain ONT contigs
Scaffold the contigs with Hi-C
Map into five chromsomes
Telomere
45S rDNA
CEN180 repeat
Close gaps
Replace with PacBio HiFi contigs
Anchor
Chr(n)
Hifiasm_contigs
Hifiasm_contig
ONT
HIFI
Chr1
Chr2
Chr3
Chr4
Chr5
Gap
Centromere
No. and location of newly-annotated genes

常用的组装软件

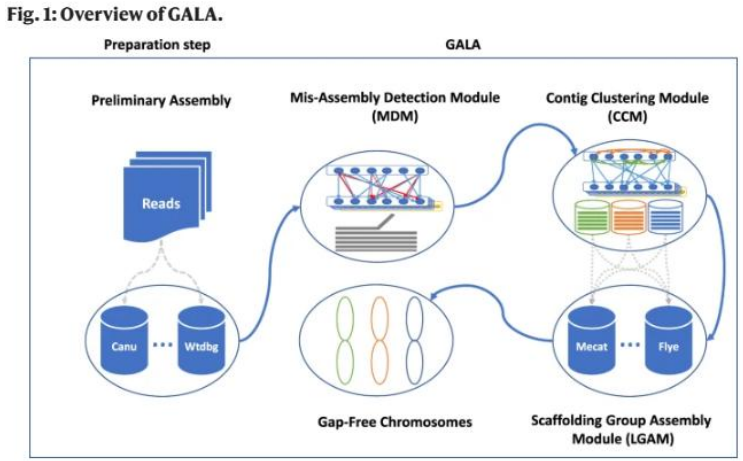
云南农业大学 Yunnan Agricultural University
云南省生物大数据重点实验室
生信帮



2023.Nature Biotechnology



2022.Nature Biotechnology



2023.Nature Communication

测序原理

测序技术	测序原理	Read读长（bp）	优点	缺点
第一代	Sanger	>1000	组装准确性高	成本高，通量低
第二代	**Illumina**	<500	成本低，通量高	组装准确性低
第三代	单分子	>10K	测序数据比较长	存在Indel

基本名词

Read：测序读到的碱基序列，组装的最小单位

Pair ends：双端测序产生的成对reads

Contig：由reads组装成的没有gap的序列

Scaffold：通过pair ends、光学图谱信息确定出的contig排列，中间有gap

Coverage：覆盖度

N50：将一组序列按长度从长到短排序，累加到长度和超过总长的50%时的那个contig的长度。N50是衡量组装完整性的指标

gap：连接scaffold过程中添加的N

组装注意事项

- 重复序列和杂合----组装的大敌
  - 可能导致大面积错拼
  - 导致组装中产生过多的repeat-gap
  - 导致后期分析和实验的困难，比如难于设计引物和PCR失败

- 样品要求：重复序列及杂合率与基因组组装的难易有直接的联系，样品越纯合越好。

基因组评估

进行基因组组装的评估是了解组装质量和完整性的重要步骤。以下是一些常用的基因组组装评估方法，统计指标评估：

- N50: N50值表示将所有contigs按照长度排序，累加到长度和超过总长度的50%时的那个contig的长度。N50值较大表示组装结果中包含更长的连续序列片段。
- 最大/最小contig长度：评估组装结果中最大和最小的contig长度，这可以帮助判断组装的覆盖度和连续性。
- 基因组大小估计：通过比对已知参考基因组或使用其他方法（调研图，流式）与预期的基因组大小相比较可以评估组装的完整性。
- Mapping评估：
  - 比对基因组测序reads回基因组：将原始的测序reads与组装结果进行比对，评估比对的覆盖度和一致性。工具如Bowtie2、BWA等可以进行比对。
  - 比对转录组测序reads回基因组：评估基因区域组装的准确性和连续性，使用minimap2,gmap,blat等。
- Reference-based评估：使用已知的参考基因组进行比对，评估组装结果与参考基因组的一致性和相对位置关系。
- Repeat分析：使用LAI分析组装结果中的重复序列，评估重复序列组装质量。
- 组装图谱评估：使用软件如QUAST、MUMmer等进行整体比对和评估。
- 可靠性评估：内部一致性：检查同一样本的不同测序数据集之间的一致性，例如不同测序平台、不同文库制备方法（如BAC验证）等；外部验证：使用其他独立的实验方法进行验证，例如PCR扩增、Sanger测序等。
- 需要注意，组装评估是一个复杂的过程，没有单一的评估指标可以全面衡量组装质量。建议结合多个评估方法和指标来综合评估组装结果的准确性、连续性和完整性。

基因组评估

Earth Biogenome Project（EBP）在2021年3月发布了4.0版本的核基因组组装质量标准报告。

Quality	Metric	Methods
Continuity	Contig (NG50)	Genometools `seqstat` or equivalent NG50 requires a genome size estimate, obtainable from `GenomeScope`, `PreQC`, or other tools
	Scaffolds (NG50)	
	Gaps / Gbp	
Structural accuracy	False duplications	`Merqury`
	Reliable blocks	Raw data mapping + `Asset` + script to accumulate desired blocks
	Curation improvements	False duplication removal: `purge_dups`
Base accuracy	Base pair QV	`Merqury`
	k-mer completeness	`Merqury`
Haplotype phasing	*Phased block (NG50)*	*`Merqury`, with availability of parental / siblings / population illumina data*
Functional completeness	Genes	`BUSCO`
	Transcript mappability	RNA-seq preprocessing + `STAR` or other dedicated transcript mappers
Chromosome status	Assigned %	Assigned during curation
	Sex chromosomes	Assigned during curation
	Organelles (e.g. MT)	Assigned during curation

- 连续性（contiguity）：contig N50和scaffold N50达到Mbp级别
- 错误率（error rate）：小于1/10000的错误率
- <5% false duplications
- >90% kmer completeness
- >90%的序列可以align到候选染色体
- >90%的单拷贝保守基因是完整且单拷贝的，例如用BUSCO评估来自同一
- >90%转录本可以mapping到基因组

https://www.earthbiogenome.org/assembly-standards

基因组评估

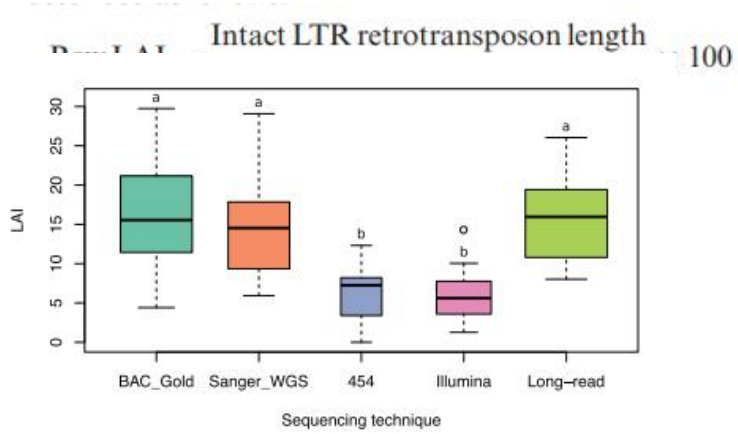
Dimension	Metric	Score for a finished assembly
I. Contiguity	N50	Chromosome N50
	CC ratio[a]	1
II. Completeness	Overall completeness: *k*-mer-based	100%
	Gene space completeness: BUSCO[b]	Near 100%[c]
	Tandem repeat completeness: telomeric and subtelomeric satellite arrays, centromeric satellite arrays, ribosomal DNA loci	100%
	Complete organelle genomes	100%
III. Correctness	Base-level error rate	0%
	Structural error: collapse, inversion, false duplication, chimeric joins	0%
IV. Organellar genomes	Contiguity: (organelle contig)/(organelle genomes)	1
	Completeness	100%
	Correctness: error rate	0%
V. Heterozygosity	Contiguity: Phase block N50	Chromosome N50[d]
	Completeness	100%
	Correctness: error rate	0%

Wang P, Wang F. A proposed metric set for evaluation of genome assembly quality. Trends Genet. 2023 Mar;39(3):175-186. doi: 10.1016/j.tig.2022.10.005. Epub 2022 Nov 17. PMID: 36402623.

基因组评估-完整性（Completeness）

基因组评估-完整性（Completeness）

- BUSCO在考察裸子植物及一些具有特殊生境或生活习性的生物时，该指标往往低估基因组编码区域的完整性。更大的问题在于，该指标仅考察了蛋白编码区域的完整性，然而事实上整合生物基因组中大部分为非编码区域。
- 基因组测序回比；转录组测序回比
- 在当前的基因组组装技术水平上，组装最有难度的区域往往是串联重复区域，包括端粒、着丝粒、rDNA区域。因此，相较于蛋白编码区域，考察重复序列区域，尤其是串联重复序列区域，更能反映一个基因组组装的完整性水平。
- LAI：计算基因组装的LAI分为四个步骤：(i)获取LTR反转录序列候选集；(ii)通过过滤掉假阳性候选集保留所有完整的LTR-RTs；(iii)对整个基因组进行LTR-RT注释；(iv)计算LAI：LAI = raw LAI + 2.8138 × (94 – whole genome LTR identity)

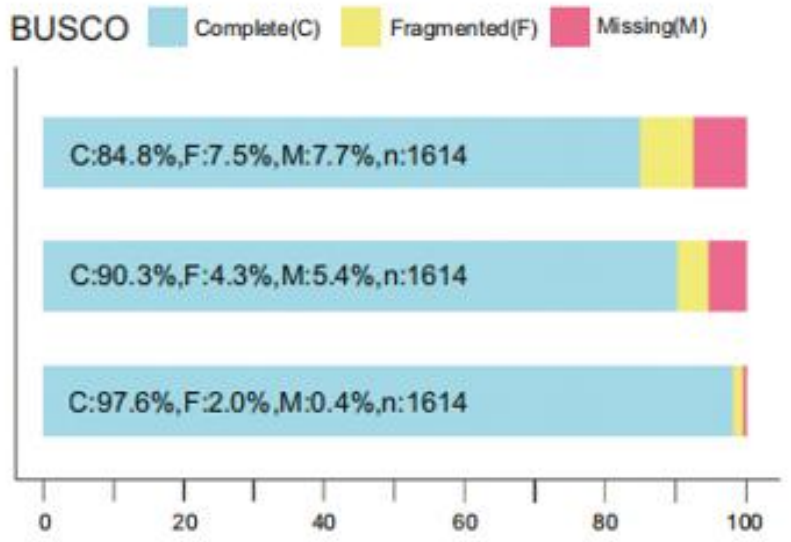


Category	LAI	Examples
Draft	0 ≤ LAI < 10	Apple (v1.0), Cacao (v1.0)
Reference	10 ≤ LAI < 20	Arabidopsis (TAIR10), Grape (12X)
Gold	20 ≤ LAI	Rice (MSUv7), Maize (B73 v4)

基因组评估-BUSCO

- BUSCO利用特定类群的信息在分析的序列中识别BUSCO基因。数据库序列的集合（或称为谱系数据集）是用于BUSCO基因识别的重要资源。每个谱系数据集都是针对特定类群（例如细菌、真菌、植物等）进行整理和优化的。这些数据集包含了已知的核心基因组成，用于评估给定序列的完整性和准确性。通过使用适当的谱系数据集，BUSCO可以根据输入序列的相似性和匹配度来识别和分类基因，并提供关于序列完整性和质量的评估结果。总之，谱系数据集是BUSCO用于基因识别和序列评估的重要组成部分，它提供了与特定类群相关的关键基因序列，有助于评估给定序列的完整性和准确性。
- 两种模式
- 基因组模式：评估基因组组装
- 蛋白质模式：评估基因集

Mode	Type	Col-PEK Number	TAIR10 Number	Col-PEK Percent (%)	TAIR10 Percent (%)
Genome	Complete BUSCOs (C)	1,603	1,603	99.4	99.3
	Complete and single-copy BUSCOs (S)	1,589	1,590	98.5	98.5
	Complete and duplicated BUSCOs (D)	14	13	0.9	0.8
	Fragmented BUSCOs (F)	3	3	0.2	0.2
	Missing BUSCOs (M)	8	8	0.4	0.5
	Total BUSCO groups searched	1,614	1,614	100.00	100.00
Protein	Complete BUSCOs (C)	1,608	1,607	99.6	99.6
	Complete and single-copy BUSCOs (S)	904	900	56.0	55.8
	Complete and duplicated BUSCOs (D)	704	707	43.6	43.8
	Fragmented BUSCOs (F)	1	1	0.1	0.1
	Missing BUSCOs (M)	5	6	0.3	0.3
	Total BUSCO groups searched	1,614	1,614	100.00	100.00



基因组评估-准确性（Correctness）

- Merqury是一种基于k-mer的基因组装评估工具，用于评估基因组装的质量。通过比较高质量测序reads所产生的k-mer与基因组装结果之间的差异，Merqury可以生成一系列的评估指标。相对于现有的方法，如BUSCO基因完整性和基于比对的测量方法，Merqury的k-mer-based评估方法在生成的结果上表现出可比较或更好的表现。
- Merqury的评估指标包括**共识质量(QV)、k-mer完整性**以及对于已知的亲代基因组序列，Merqury还可以输出单倍体完整性、相位区块统计、切换错误率和子代基因组相位一致性的可视化图形。
- Merqury的基本原理是通过比较基因组装得到的序列与高质量测序reads所产生的k-mer之间的差异来评估基因组装的准确性。通过这种基于k-mer的分析方法，Merqury能够检测出基因组装结果中的缺失序列、重复序列、序列变异等问题，并生成相应的评估指标。
- Merqury的优点在于可以不依赖已知的真实基因组序列进行评估，而是直接使用无组装的高质量reads产生的k-mer信息进行比较。这种方法可以避免对参考序列的依赖，特别适用于那些没有经过完整验证的参考序列或者没有参考序列的物种。此外，Merqury还具有评估亲代基因组序列的能力，可以评估长序列的相位一致性和区块统计等指标。
- 总体而言，Merqury是一种比较先进的基因组装评估工具，通过基于k-mer的分析方法，能够提供更准确、全面的基因组装质量评估结果。

第三章 基因组注释

主要内容

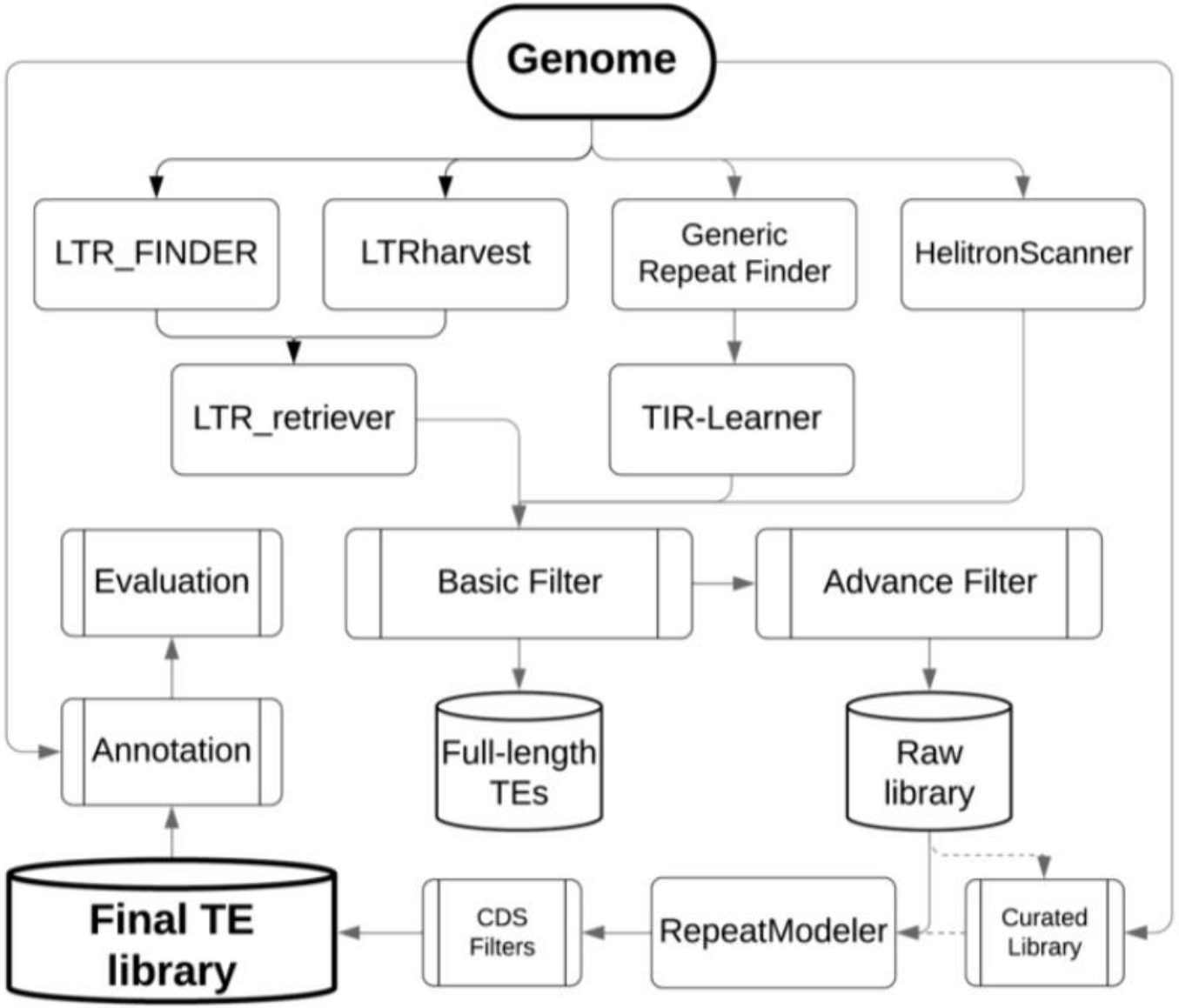
- 重复序列预测
- 编码基因预测
- 基因功能注释
- 假基因预测分析
- 非编码RNA预测分析

1、重复序列注释-分类

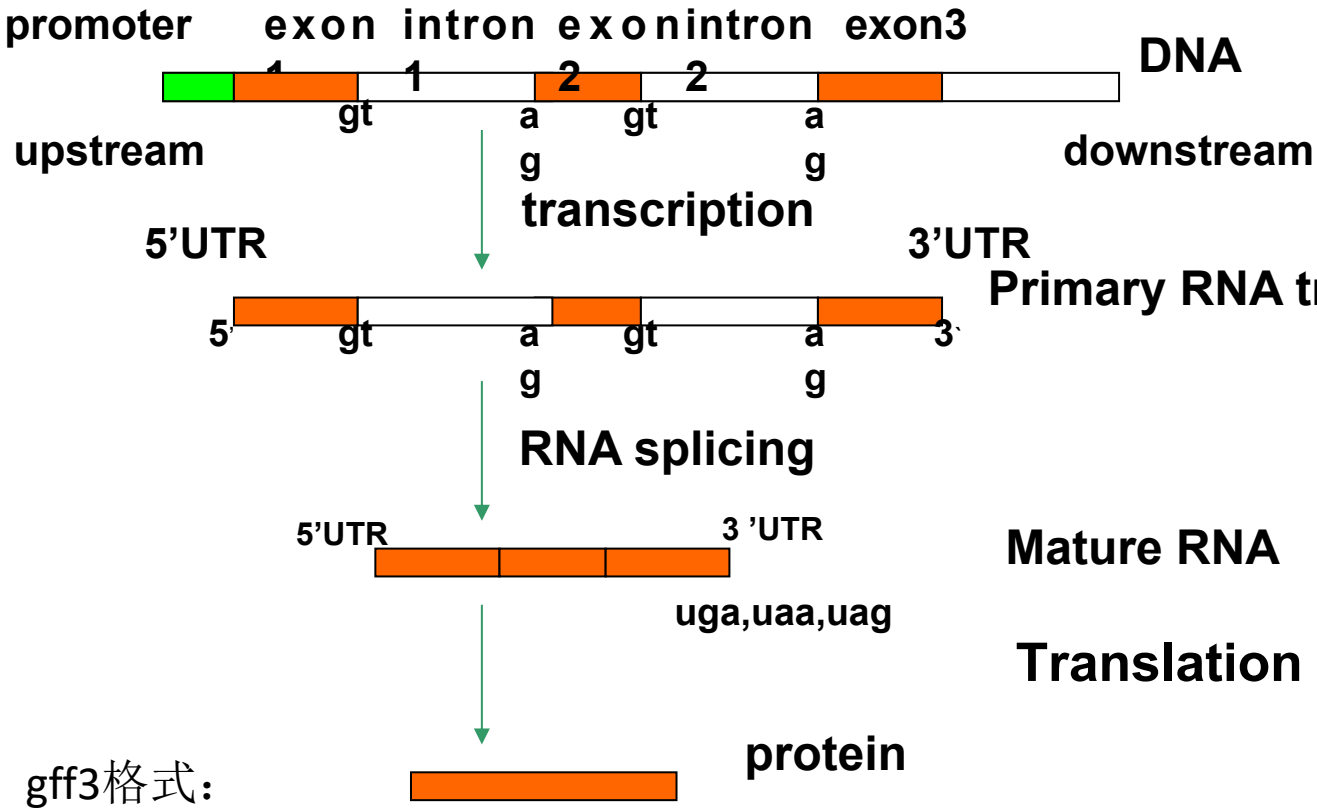
分成2类：串联重复(tandem repeats) 和 散在重复序列(interspersed repeats)，其中第二类主要是TEs（Transposon Element）.

- 第一类串联重复包含：microsatellites 或 simple sequence repeats（1-6个碱基为一个重复单元） 和 minisatellites（10-60个碱基的长序列为一个重复单元）.
- TEs包含2种类型：**class-I TEs通过RNA介导的(copy and paste）机制进行转座；class-II TEs通过DNA介导的(cut and paste）机制来转座. 前者称为retroposon，后者称为DNA transposons**。
- class-I TEs中主要由LTR(long terminal repeat)构成。LTR的部分序列可能具有编码功能。另外Class I还具有non-LTR，包含2个子类：LINEs(long interspersed nuclear elements)和SINEs(short interspersed elements），其中前者可能具有编码功能，后者则没有。
- class-II TEs中加入了一个子类 MITEs(miniature inverted repeat transposable elements),基于DNA的转座因子，但是确通过"copy and paste"的机制来转座(Wicker et al., 2007)。

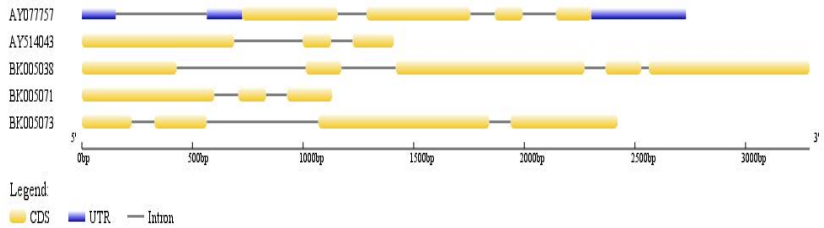
1、重复序列注释-EDTA



2、基因预测-基础知识



- gene:最长
- mRNA：转录本，不含内含子
- exon：外显子组合形成mRNA
- UTR:外显子末端部分
- CDS：属于外显子中编码的部分
- Intron：剪切掉的基因的一部分。
- GFF文件里可以有多个UTR吗?
- 可以。UTR之间也可能有内含子。



gff3格式：

```
Mt      protein_coding  gene    220471  220773  .       -       .       ID=ATMG00740;
Mt      protein_coding  mRNA    220471  220773  .       -       .       ID=ATMG00740.1;Parent=ATMG00740;
Mt      protein_coding  start_codon     220771  220773  .       -       0       ID=start_codon:ATMG00740.1:1;Parent=ATMG00740.1;
Mt      protein_coding  exon    220471  220773  .       -       .       ID=exon:ATMG00740.1:1;Parent=ATMG00740.1;
Mt      protein_coding  CDS     220471  220773  .       -       0       ID=CDS:ATMG00740.1:1;Parent=ATMG00740.1;
Mt      protein_coding  stop_codon      220471  220473  .       -       0       ID=stop_codon:ATMG00740.1:1;Parent=ATMG00740.1;
Mt      protein_coding  gene    57774   58331   .       +       .       ID=ATMG00210;
Mt      protein_coding  mRNA    57774   58331   .       +       .       ID=ATMG00210.1;Parent=ATMG00210;
Mt      protein_coding  start_codon     57774   57776   .       +       0       ID=start_codon:ATMG00210.1:1;Parent=ATMG00210.1;
Mt      protein_coding  CDS     57774   58331   .       +       0       ID=CDS:ATMG00210.1:1;Parent=ATMG00210.1;
```

2、基因预测-原理

主要采用**从头预测、基于同源蛋白预测、基于转录组（EST，全长转录组）预测3种不同的策略来预测编码基因，再利用整合软件（Maker，EVM等）对预测结果进行整合**。

1、同源预测软件：**GeMoMa**。主要借助同源基因的结构来预测的。同源物种（除模式物种外尽量一个科的）的基因组序列及注释文件(**>=4个**)。

2、转录预测：

1）二代转录组：混样转录组或多个单独测得多个时空条件的转录组样本（建议尽可能多的数据，以便获得更多的可变剪接）进行有参组装和无参genome-guide组装，软件可以用hisat2-stringtie,Trinity,transdecode等。

2）三代全长转录组。Pacbio CCS和ONT(需要二代数据联合使用进行纠错），软件可以采用gmap,transdecode。

3、从头预测

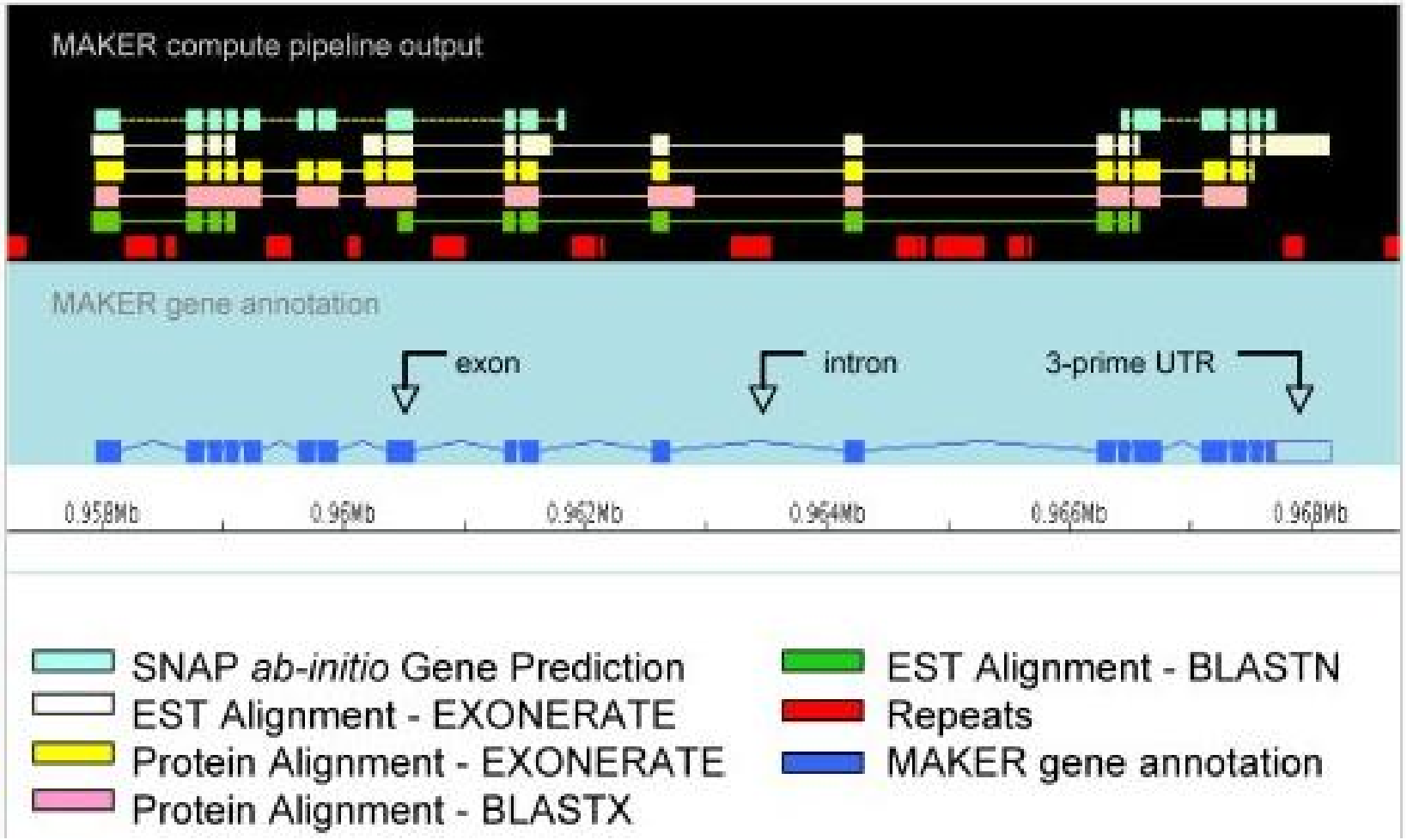
1）原理：基于基本特征序列：剪接位点序列：gt...ag；位于内含子中辅助剪接序列；翻译起始前的序列；编码序列；非编码序列；单外显子，起始外显子、中间外显子、终止外显子，基因间区长度分布；每个基因外显子数目分布与内含子长度分布

2）使用方式：自带训练集，在相应软件目录里；自己制作本物种训练集：可以使用同源预测/转录组预测单拷贝基因。

3）软件:Augustus,SNAP,GlimmerHMM,GeneScan,GeneID等。

2、基因预测-原理

基因预测整合原理-MAKER



2、基因预测-原理

基因预测整合-EVM

对于同源和转录组的预测的结果，偏好的认为其基因座位置是对的，但是基因结构不做要求，整合的时候只考虑internal外显子。而对于从头预测的软件，其结构识别是很准确的，但是整体结构不一定对，因此会将从头预测的结果按照四种外显子（initial,internal,terminal.single exon）类型划分，不利用整体的基因结构进行整合。最后对一段区域内的所有外显子（根据制定的**weight和长度**等特征）进行打分，然后从头到尾利用**动态规划算法**找到一条分数最高的路径，这个路径就是最佳的基因结构。

EVM中，基因注释分为3种，并将所有的注释信息（无论使用那种方法进行注释的）合并到到3个文件中：

1）gene_predictions.gff3:所有从头的预测产生的gff文件

2）protein_alignments.gff3：各物种同源预测gff文件

3）transcript_alignments.gff3： PASA结果和Hisat-Stringtie结果

权重文件：

```
ABINITIO_PREDICTION     augustus                    1
ABINITIO_PREDICTION     twinscan                    1
ABINITIO_PREDICTION     glimmerHMM                  1
PROTEIN                 spliced_protein_alignments  1
PROTEIN                 genewise_protein_alignments 2
TRANSCRIPT              spliced_transcript_alignments 1
TRANSCRIPT              PASA_transcript_assemblies  10
OTHER_PREDICTION        PASA_transdecoder           5
```

3、基因预测注意事项

- 最大内含子长度

粗略大小：植物一般设置在20,000 bp，哺乳动物（>=3Gb 的一般设置在500,000）；

- 结合模式物种或者近缘物种选择单软件最佳预测结果，确保整合后不丢失!

4、基因预测质控

- 基因数量：与近缘物种相差不大，相差大的话，过滤来源于从头预测的结果或者AED评分过滤
- BUSCO：在基因水平上的BUSCO要高于基因组水平上的BUSCO
- 功能注释：一般要在90%以上。

5、非编码RNA预测

- 非编码RNA即不编码蛋白质的RNA，包括**microRNA、rRNA和tRNA等**多种已知功能的RNA，针对不同非编码RNA的结构特点，采用了不同的策略来预测不同的非编码RNA。
- 分别基于**Rfam数据库和miRBase数据库**并利用 Infenal 1.1 进行rRNA和microRNA预测；
- 利用**tRNAscan-SE** 识别tRNA。

6、基因功能注释

➢基因功能注释的研究背景和意义通过功能注释，可以了解基因在细胞生理、代谢过程、发育过程等方面的功能。这有助于揭示基因在生物体内的作用，并推断其可能参与的生物学过程。链接基因到生物学过程和疾病：通过功能注释，可以将基因与已知的生物学过程、通路以及疾病关联起来。这有助于理解基因在正常生理和病理过程中的作用，为疾病的发生机制的研究提供线索。

➢注释比对工具：diamond(超级快)

➢功能注释数据库简介

NR：所有已知的基因

KEGG：代谢通路

KOG：（蛋白质直系同源簇）数据库是基于具有完整基因组的细菌、藻类、真核生物的编码蛋白的系统进化关系构建的。利用KOG数据库可以对基因产物进行直系同源分类，在功能层面上对基因进行分类。

GO：生物学过程、构成细胞的成分和实现的分子功能

TrEMBL：包含所有已知功能信息的基因序列

6、功能注释-GO、KEGG

- KEGG（Kyoto Encyclopedia of Genes and Genomes）是一个广泛使用的生物信息学数据库和分析工具。KEGG注释指的是将基因或蛋白质序列与KEGG数据库中的功能注释进行关联的过程。KEGG注释提供了有关基因或蛋白质序列的功能信息、代谢通路和相关基因组学数据。通过将基因或蛋白质序列与KEGG数据库中的已知信息进行比对，可以为这些序列提供功能注释和通路关联信息。注释数据库:KOBAS。
- GO（Gene Ontology）是一种用于描述基因和蛋白质功能的分类系统，提供了一套标准化的词汇和概念。GO注释是将基因或蛋白质序列与GO分类进行关联的过程，帮助研究人员理解基因或蛋白质的功能和相互作用。

- GO分类系统包括三个主要方面的功能注释:

  - 分子功能（Molecular Function）：描述蛋白质在分子层面上的活性、反应和结合特性，比如催化酶活性、配体结合等。
  - 细胞组分（Cellular Component）：描述蛋白质在细胞中的局部化位置和组织结构，比如核内、细胞膜等。
  - 生物过程（Biological Process）：描述蛋白质参与的生物学过程，比如代谢途径、信号传导等。

第四章 Hi-C三维基因组

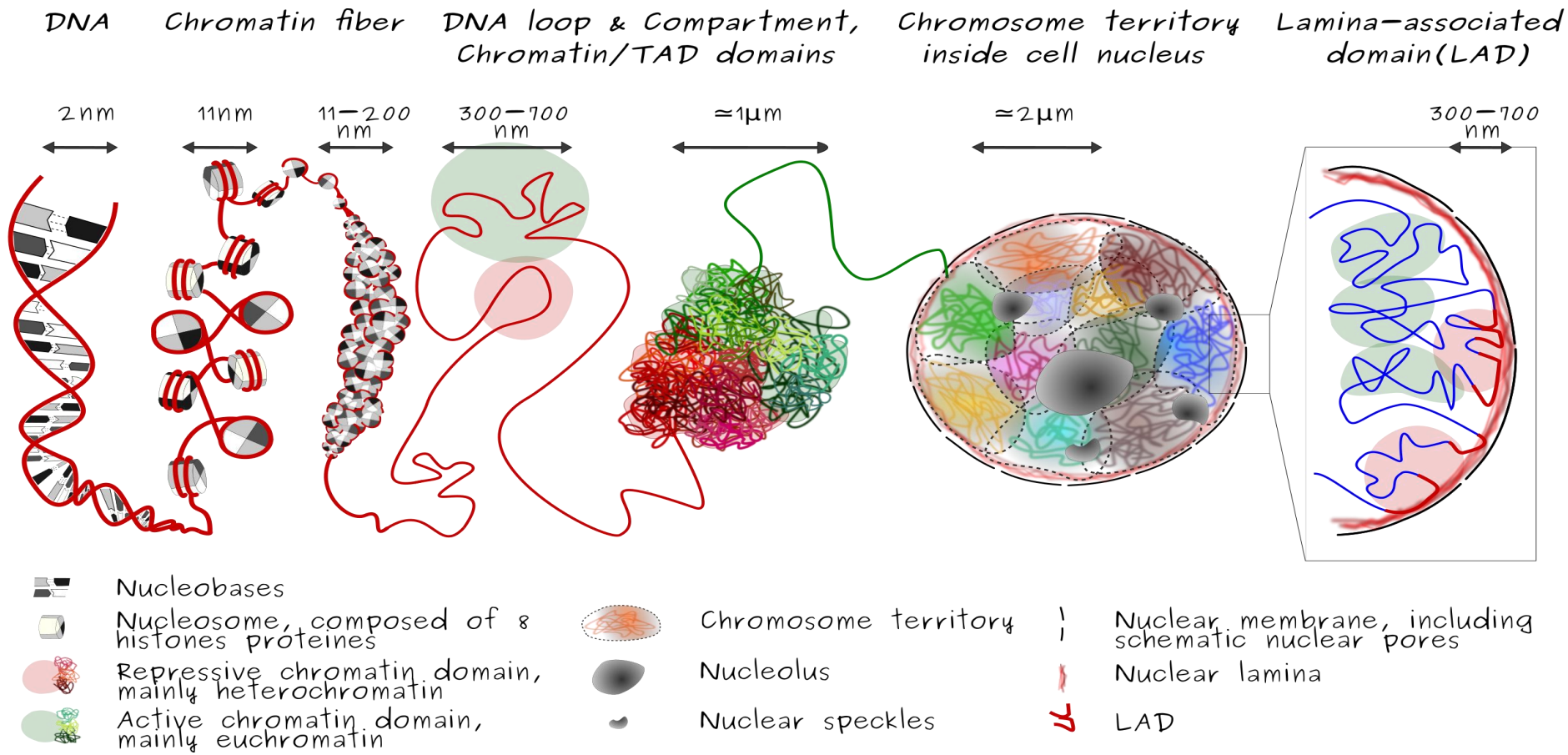
主要内容

- Hi-C简介
- Hi-C文库评估
- Hi-C辅助基因组组装
- Hi-C三维基因组互作分析

三维基因组简介

云南省生物大数据重点实验室
生信帮

基因组三维空间结构与功能研究的简称，旨在考虑基因组序列、基因结构、调控元件的同时，对基因组序列在细胞核内的三维空间结构，及其在基因转录、调控、复制和修复等生物过程中的功能进行研究。



三维基因组简介

云南省生物大数据重点实验室

生信帮

- 3C 通常是用启动子或者某一个基因或者基因组某一个短的片段在邻近的几十kb或者几百kb基因组扫描可以获得相互作用区域。由于实验需要特异性引物，因而实验室相当费力的，且检测范围小。
- 4C同3C一样做单位点的检测，但其检测扩展到了整个基因组上。主要是引入了反向PCR，因而只需要对这一单一位点设计引物即可。
- 5C 做两个大片段之间相互作用点的检测，可以达到10Mb水平。其仍需使用引物，且引物设计是其技术的难点。
- Hi-C 可以实现基因组对基因组水平的检测，一般采用**四碱基的Dpn II（六碱基只能做到TAD）**。目前分为两种交联方式：Dilution proximity ligation：裂解细胞核后进行交联；In situ proximity ligation：裂解细胞之前进行交联，**In situ Hi-C**的有效数据更多。
- 其他衍生技术:HiChIP,Micro-C,Ocean-C

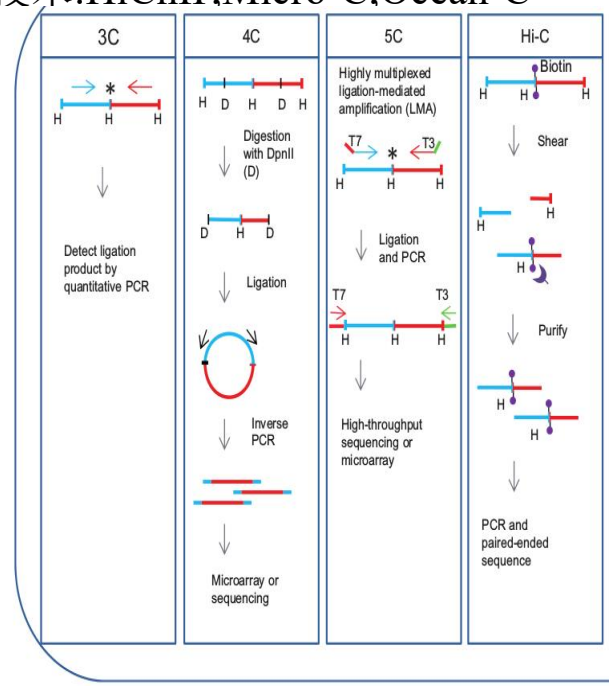


图1-1各种C技术

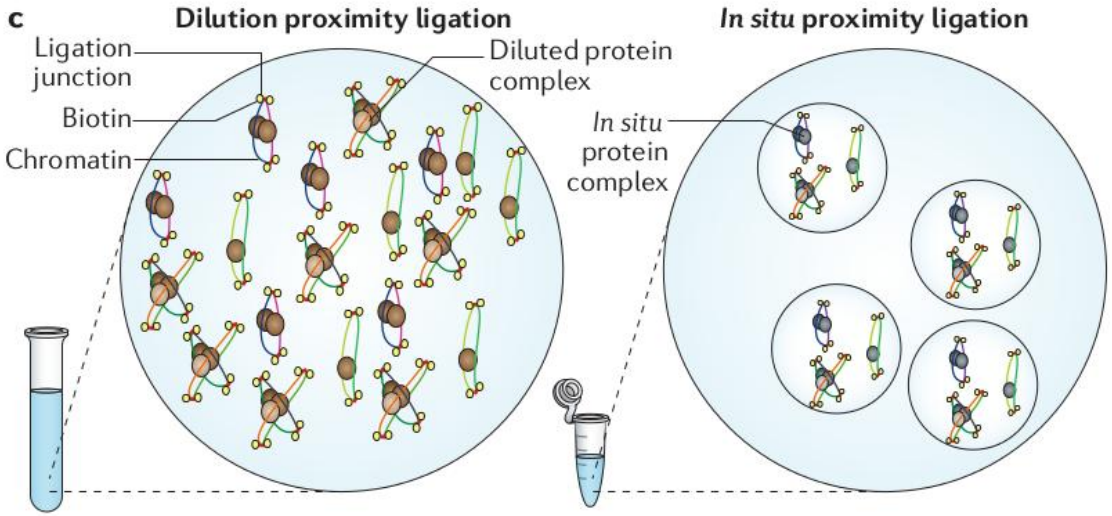
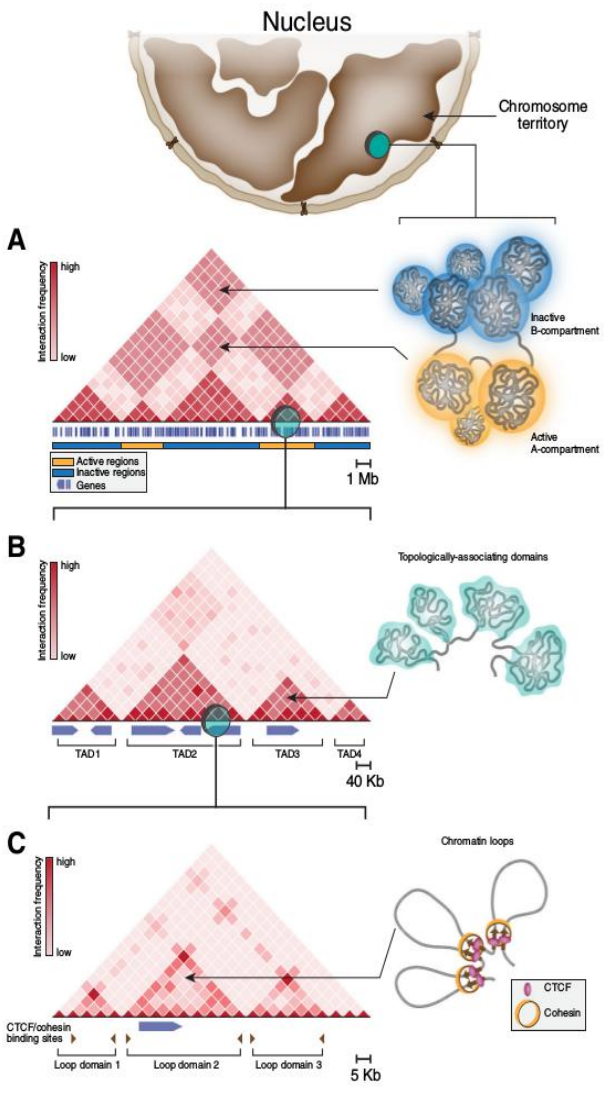


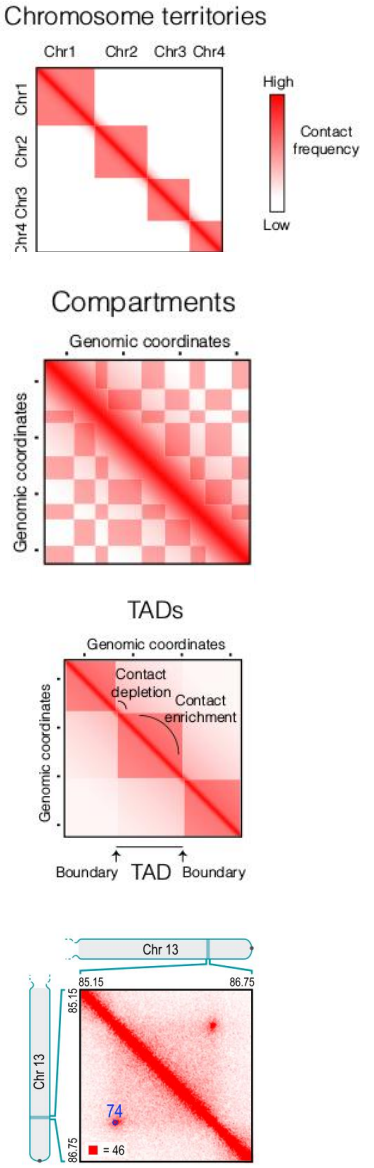
图1-2两种交联方式

三维基因组简介

三维基因组基本结构特征

- 从隔室（A/B Compartments）到拓扑相关结构域（动物中TAD，植物中TAD-like）,最后再到环（loop）的染色质的这种层级结构 .
- 隔室为常染色质与异染色质划分，热图呈现为棋盘格。
- TAD作为一个调控单位，其内部的基因拥有共同的调控元件，因而存在其内部各基因的协同表达特征。在热图对角线上呈现一个方形的结构。
- 环作为远端调控的一种形式。增强子－启动子环。

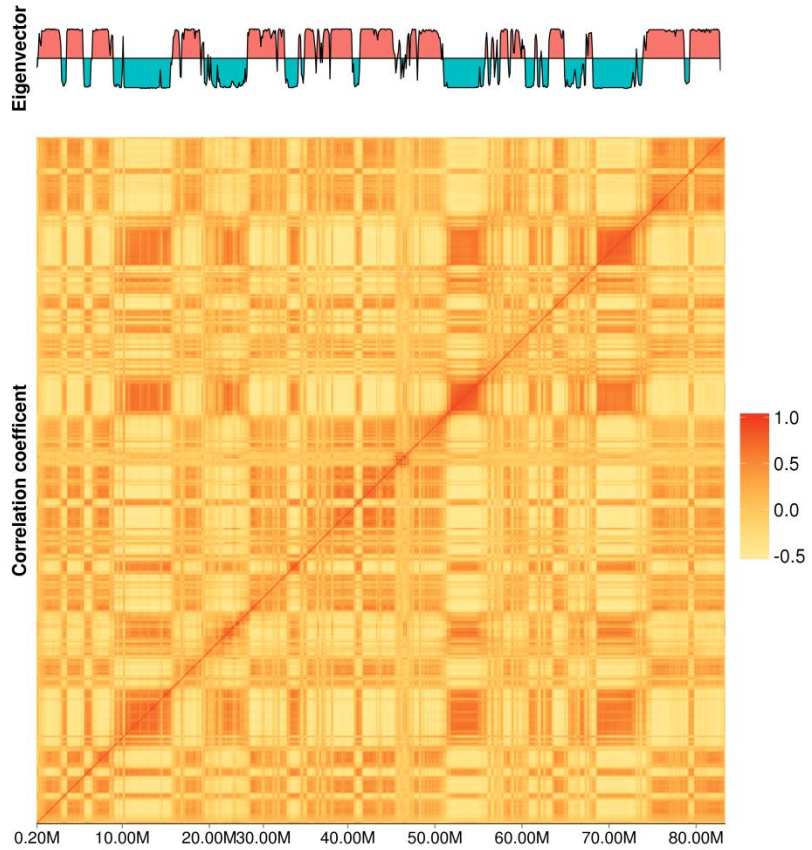




三维基因组简介

AB Compartment

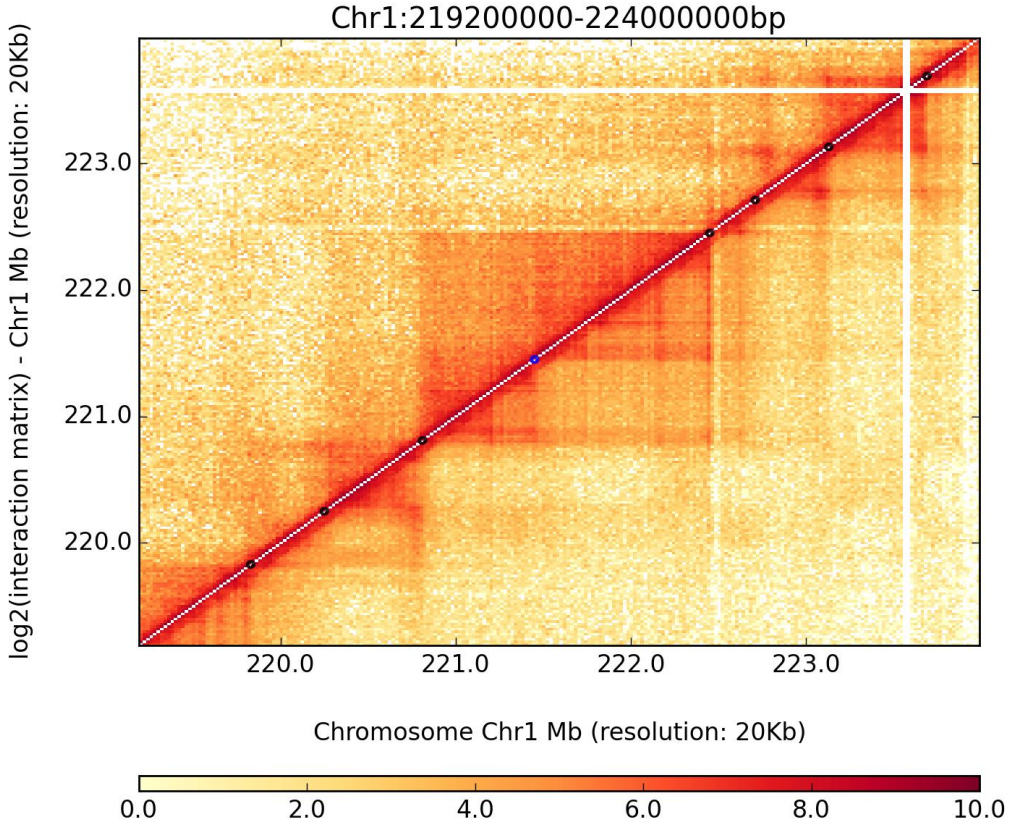
- 第一主成分可视化基因组区域的差异行为。正的特征值与松散染色质状态相关联，称为A Compartment;负的特征值与压缩染色质状态相关，称为B Compartment。热图呈现为棋盘格。
- A区域染色质是具有高基因密度，开放的，易接近的（DNAase I敏感区），转录活性区域的常染色质；B区域一般为基因少的异染色质区域。
- 染色体结构与表观遗传修饰。染色体上松弛的结构域的多富集与转录活性相关的表观遗传标记，例如H3K4me3,H3K27ac，而紧密的结构域则富集无活性的表观遗传标记H3K27me3;
- 数据量：>30X；分辨率：100 kb；分析软件：HiTC



三维基因组简介

TAD

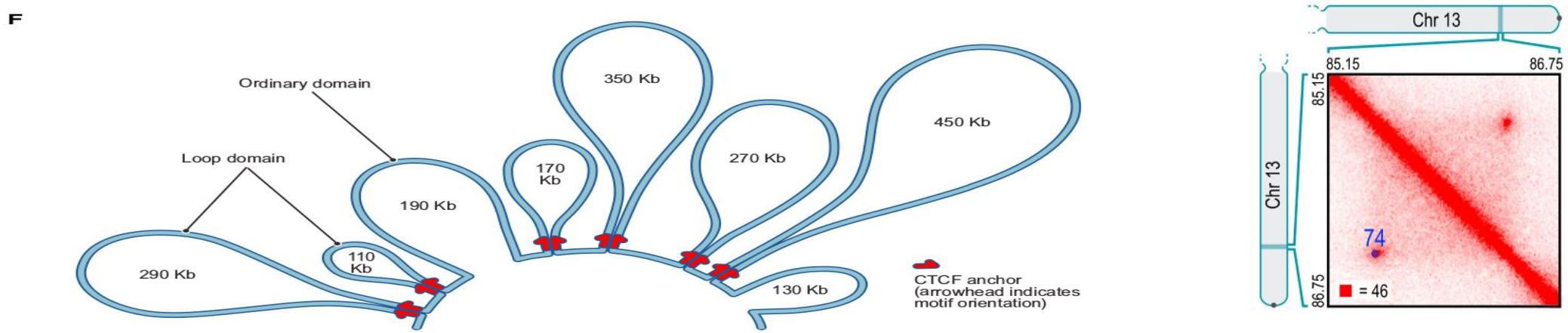
- 拓扑相关结构域（TAD）是一个高度自关联的连续的区域，并且其通过明显的边界与其相邻区域分离开来，形成一个独立的调控单位。
- 在该结构域边界通常存在大量的绝缘子，管家基因，tRNA，短片段重复（SINE）。其重要功能是形成基因调控的独立区域，同时，将它们与邻近区域隔离开来。
- 哺乳动物中大部分TAD由CTCF形成，CTCF占据TAD边界的大部分
- 数据量：>=50X
- 分辨率：40kb
- 分析软件：TadLib，HOMER
- 长度: 哺乳动物 ~ 1 Mb，水稻~60 kb
- 数目:哺乳动物2000~5000个，水稻>2000个



三维基因组简介

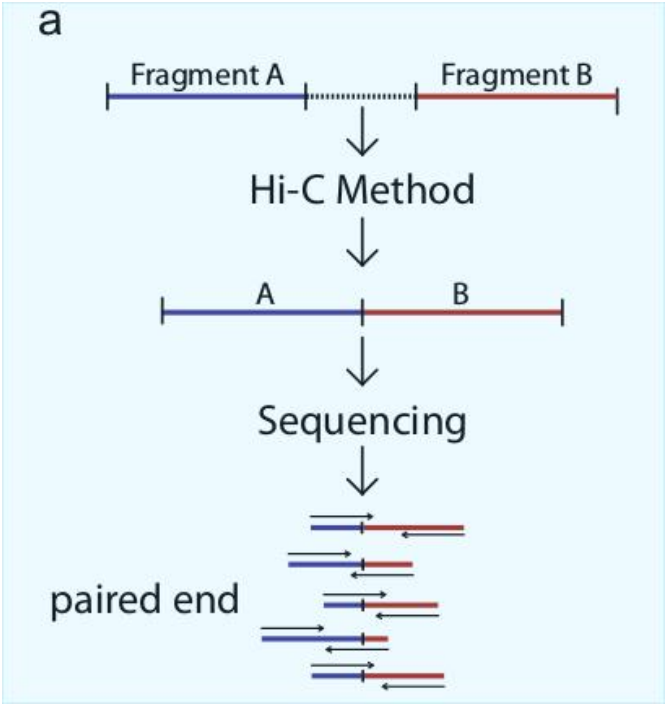
Loop

- 染色质环的位置,寻找一个片段比线性上相邻的片段在空间上更加显著接近（交互）的一个片段，这样的一对片段的交互频率往往高于线性上相邻的片段。这个高交互的位点称为Peak 位点，这反映了染色质环的存在。Peak位点为染色质环锚定的位点。
- 生物学意义：增强子/沉默子 – 启动子环通常需要与DNAase-seq、CUT&Tag，ATAC-seq结合使用。
- 数据量：>150X
- 分辨率：10 kb(<2 Gb:5 kb ;<0.5Gb :2 kb)
- 分析软件：HICCUP，FitHi-C
- 长度:>20kb(2 个bin长度)
- 数目:哺乳动物10000~20000个；植物：不确定

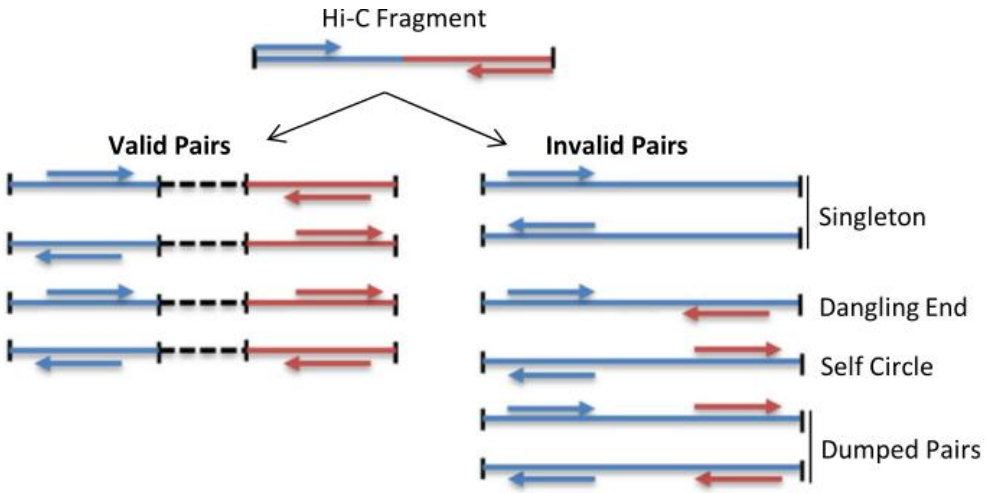


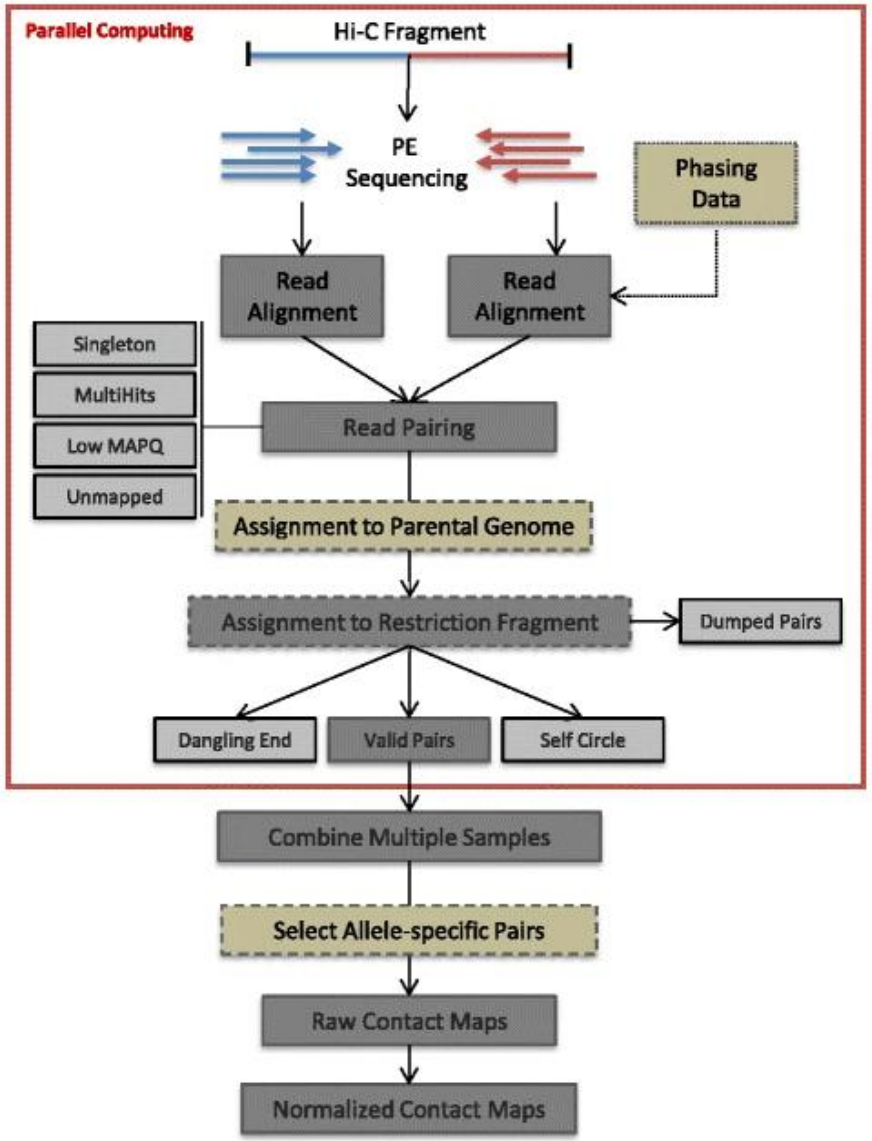
Rao et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. ***Cell.*** 2014 Dec 18

Hi-C文库评估



- 无参评估：truncter;hicup
- 有参评估：酶切位点;HiC-Pro
- 有效数据>30%
- trnuter>5%

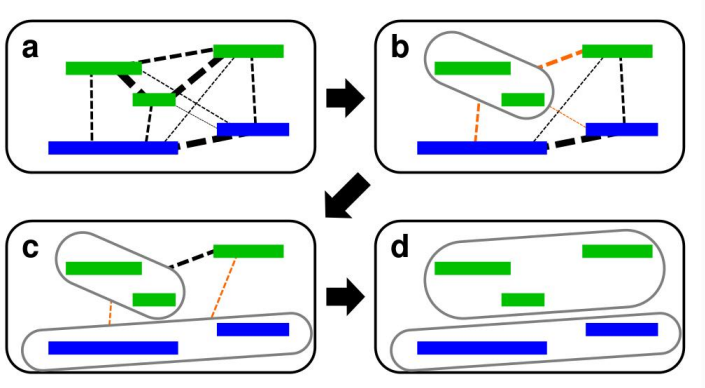




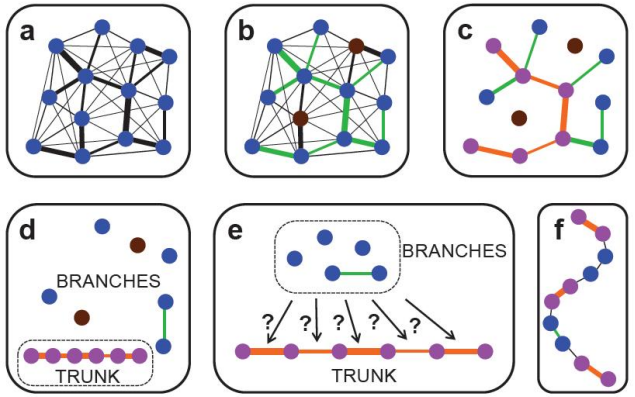
基因组染色体组装

云南农業大學 Yunnan Agricultural University
云南省生物大数据重点实验室
生信帮

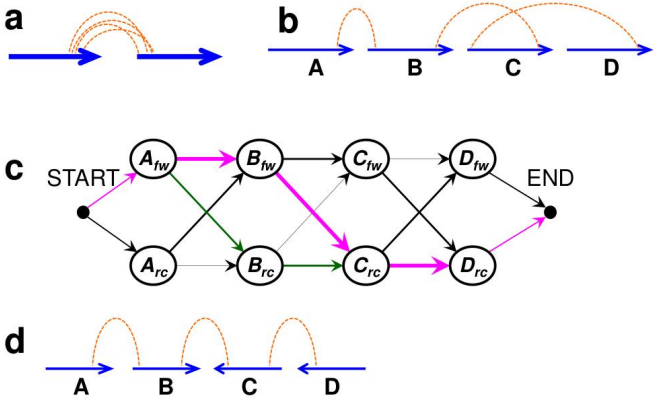
- Hi-C辅助基因组组装是指在已有二代或三代或光学图谱辅助组装的的contig/scaffold序列和已知染色体数目（**未知染色体也可以**）的前提下，利用Hi-C测序数据将这些序列进行染色体群组的划分，并确定各序列在染色体上的顺序和方向，使基因组组装组装水平提升到**染色体水平**的技术。
- 划分染色体群组原理：细胞核内包含同源染色体在内每条染色体的都占据着一个独特区域，称为**染色体疆域**，导致基因组各区段（ locus ）在同一染色体上的交互频率高于不同染色体的交互频率。因而实现了将初步组装的Contigs或Scaffolds分配到各染色体群组中。
- 染色体内部不同locus的交互频率与locus之间的线性距离一般近似服从幂次定律（power law），因而可以通过交互频率的高低确定每个染色体群组中的不同Contigs或Scaffolds顺序与方向。
- 组装软件:LACHESIS,3D-DNA,ALLHiC等。其中LACHESIS组装算法如下:



染色体聚类算法



Ordering算法

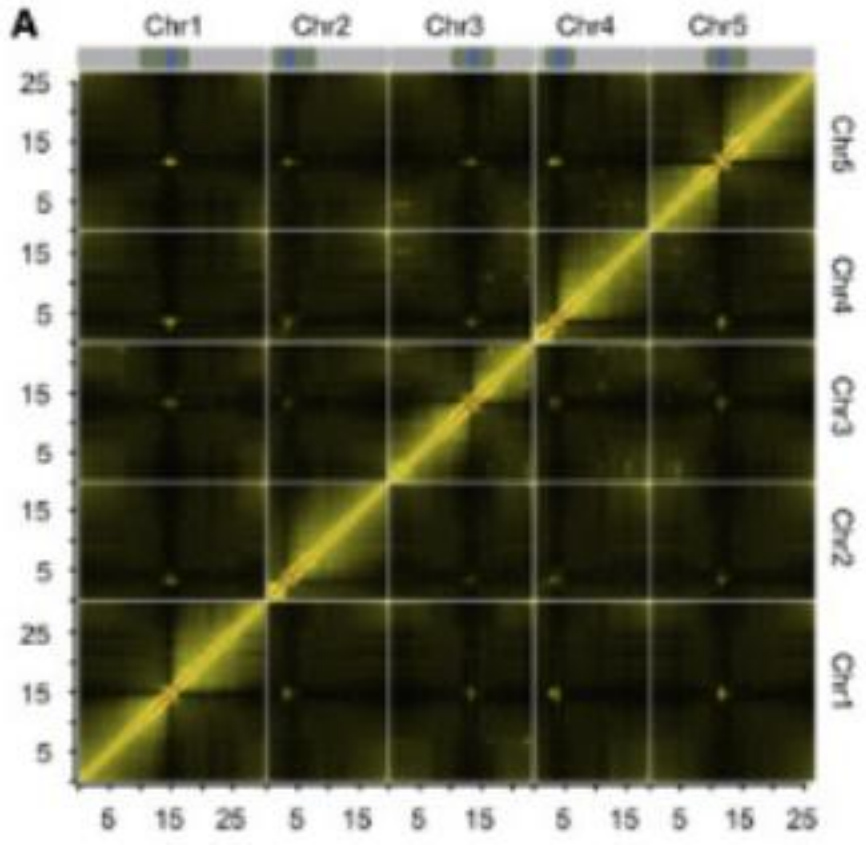


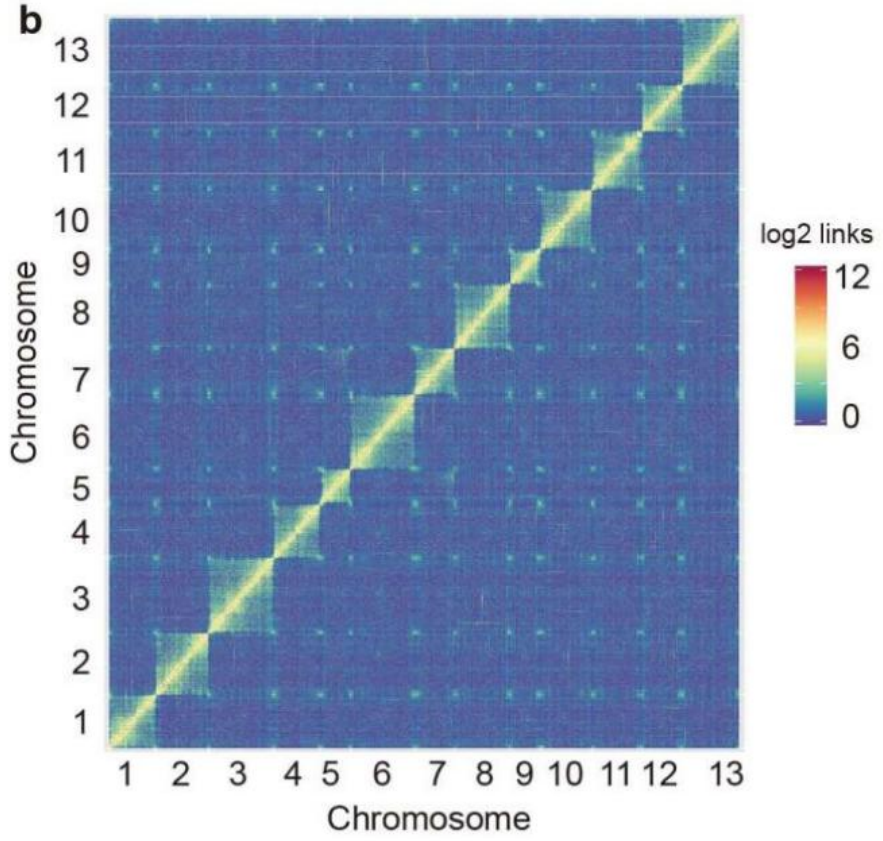
Orientation算法

Burton et al.. Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. ***Nat Biotechnol***. 2013 Dec

基因组染色体组装

组装热图如下：



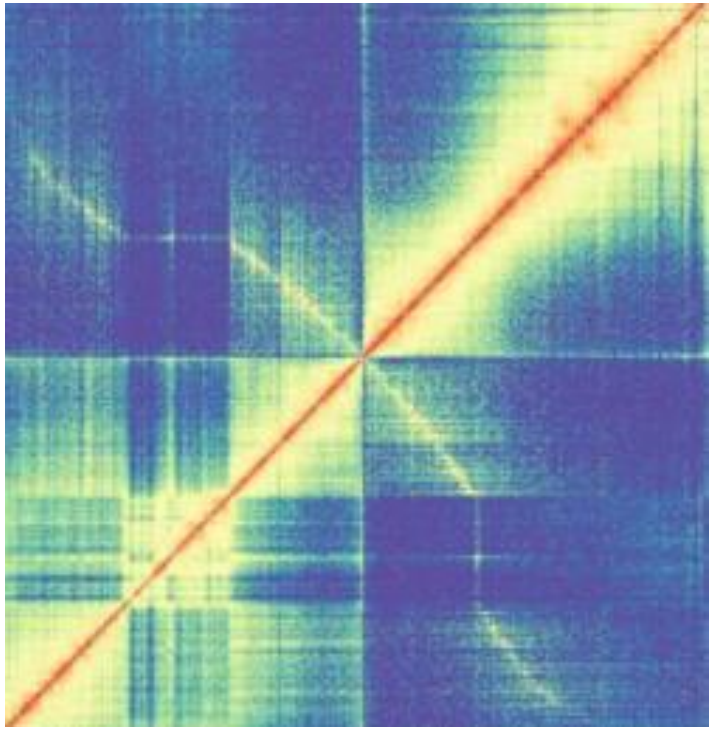


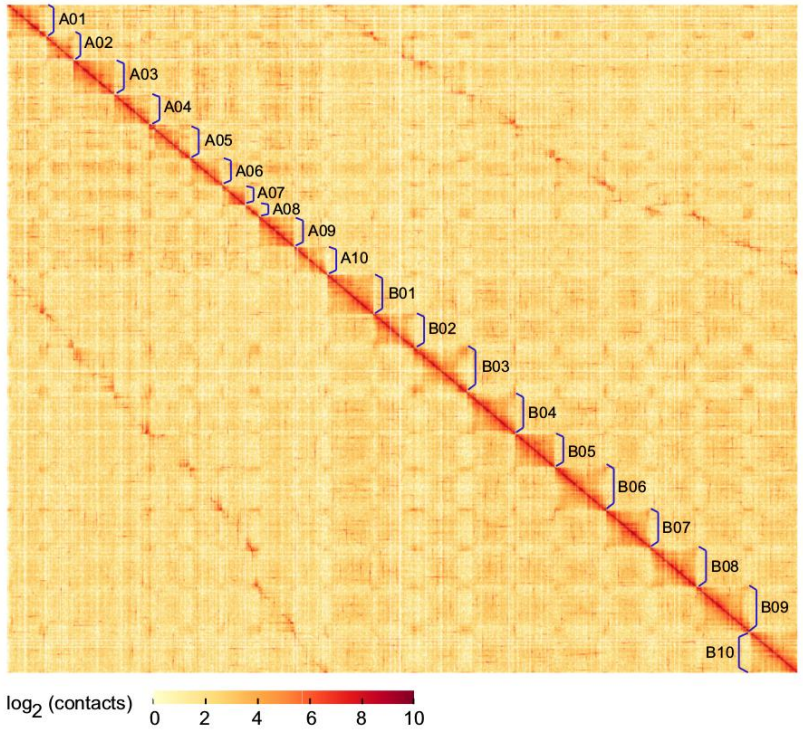
Grob S, Schmid MW, Grossniklaus U. Hi-C analysis in Arabidopsis identifies the KNOT, a structure with similarities to the flamenco locus of Drosophila. ***Mol Cell.*** 2014 Sep 4
Du, et al. (2018). Resequencing of 243 diploid cotton accessions based on an updated a genome identifies the genetic basis of key agronomic traits. ***Nature Genetics***.

基因组染色体组装

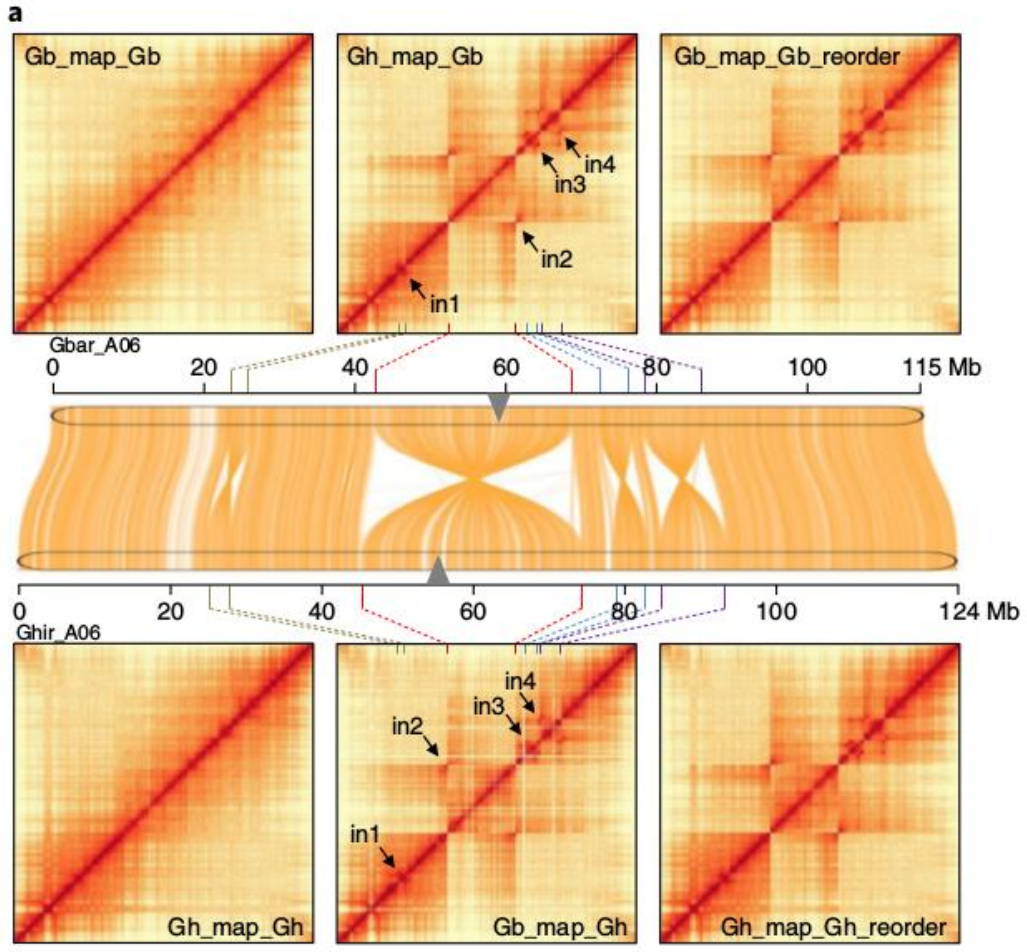
基因组染色体组装面临的挑战:

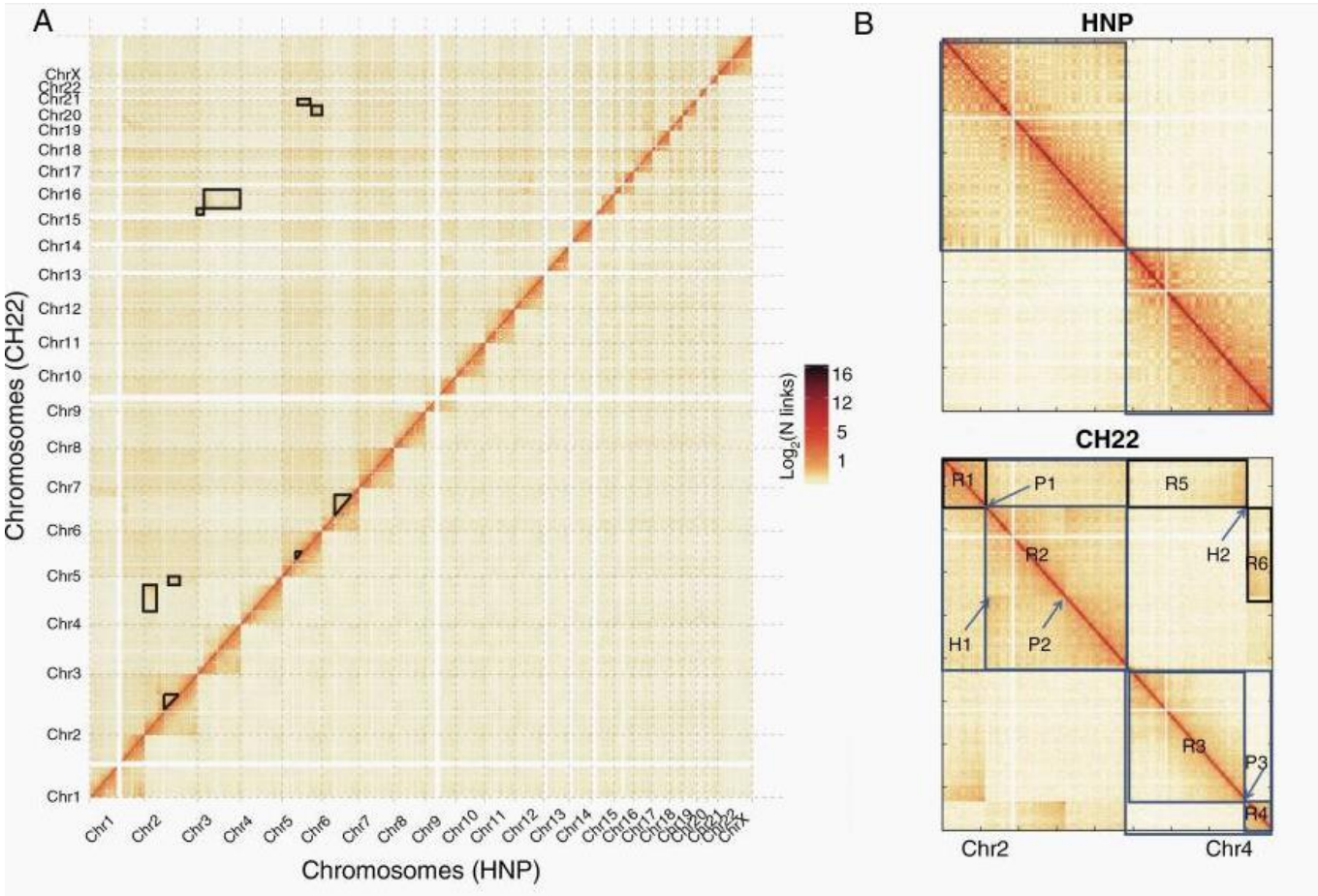
- 分型组装：受制于杂合率
- 多倍体





Hi-C技术大型结构变异鉴定



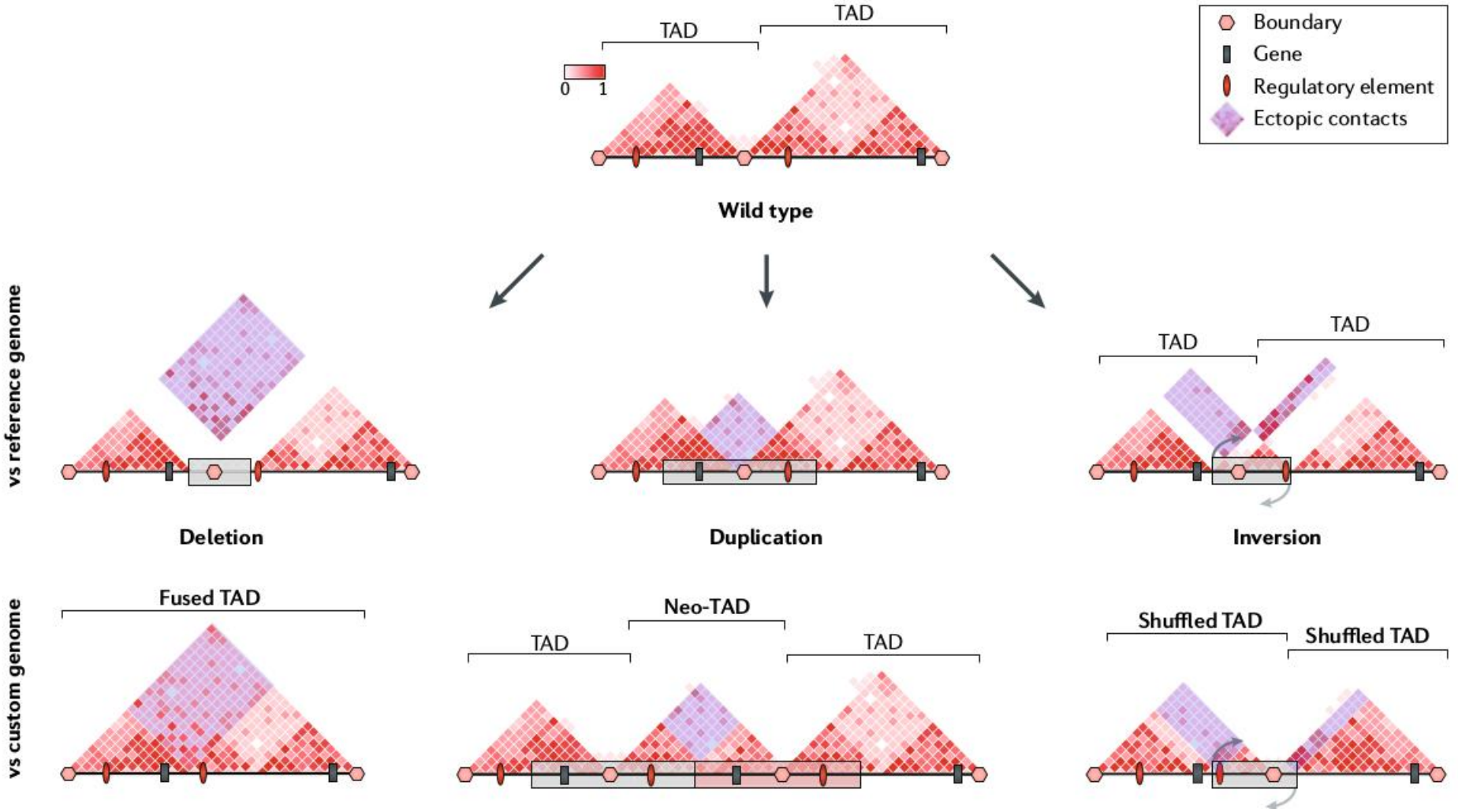


Wang et al. (2019). Reference genome sequences of two cultivated allotetraploid cottons, gossypium hirsutum and gossypium barbadense. ***Nature genetics***, 51(2), 224.
Meng, et al. (2021). The comparative integrated multi-omics analysis identifies ca2 as a novel target for chordoma. ***Neuro-Oncology***

Hi-C技术在表观调控中应用

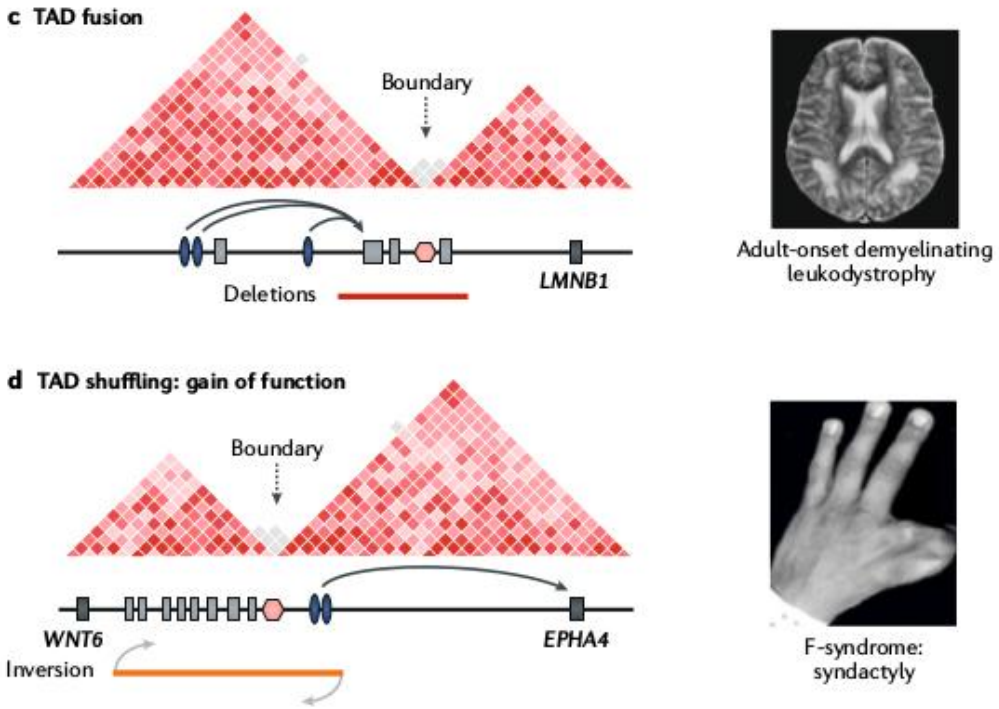
云南农業大學 Yunnan Agricultural University
云南省生物大数据重点实验室
生信帮

1、非编码区域SV影响TAD实现基因表达调控



Hi-C技术在表观调控中应用

1、非编码区域SV影响TAD实现基因表达调控



c、删除LMNB1基因座附近的TAD边界导致另一个TAD内部增强子调控LMNB1基因，进而导致成人发作性脱髓鞘性脑白质营养不良。

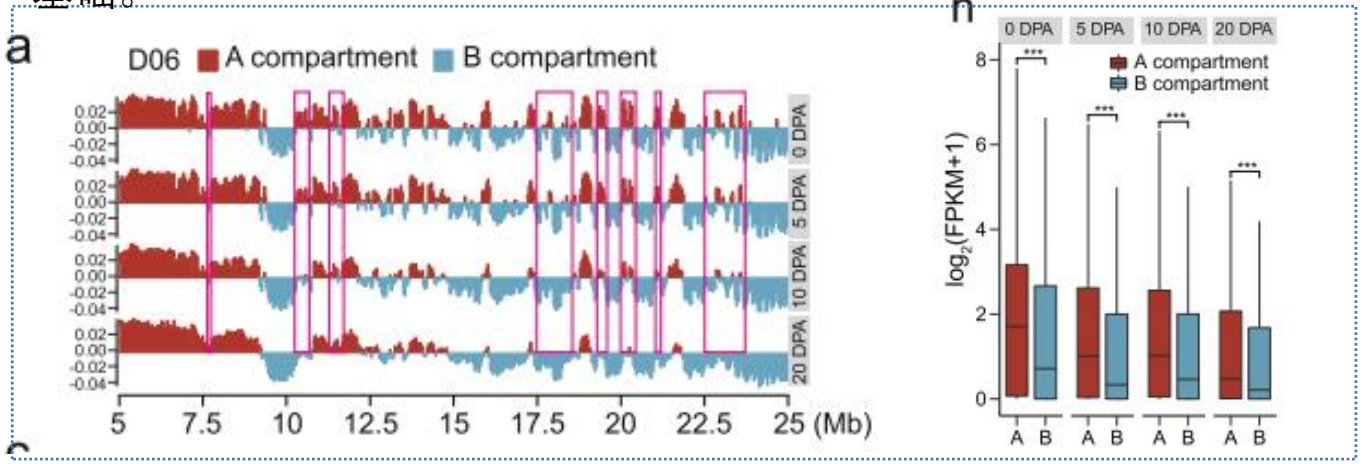
d、翻转EPHA4基因座增强子簇，导致增强子错误调控WNT6，导致拇指和食指并指畸形。

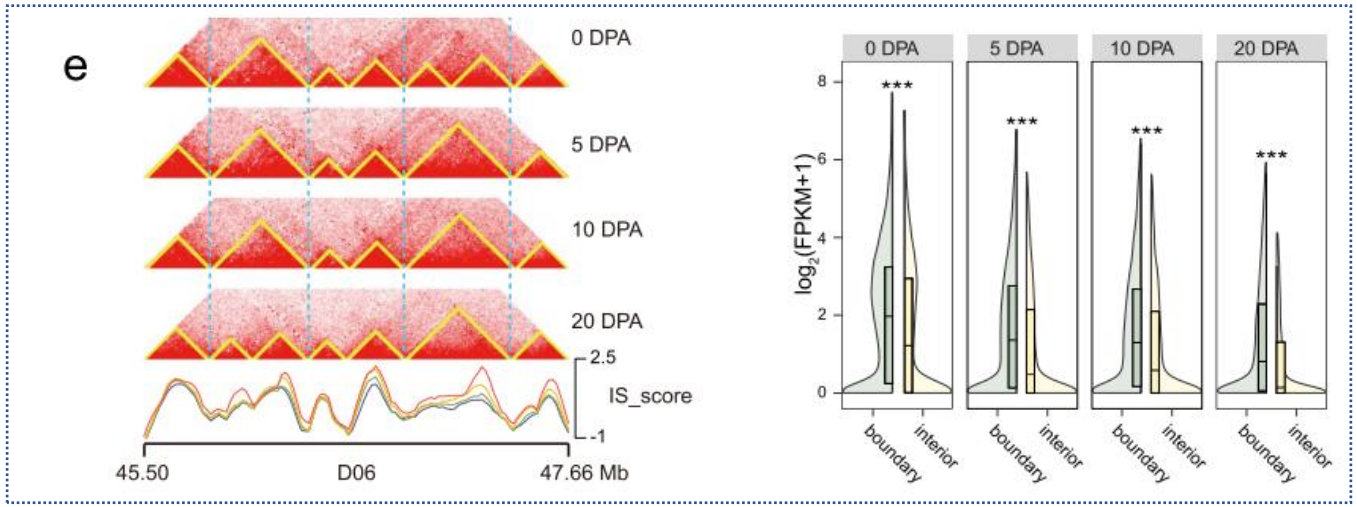
Hi-C技术在表观调控中应用

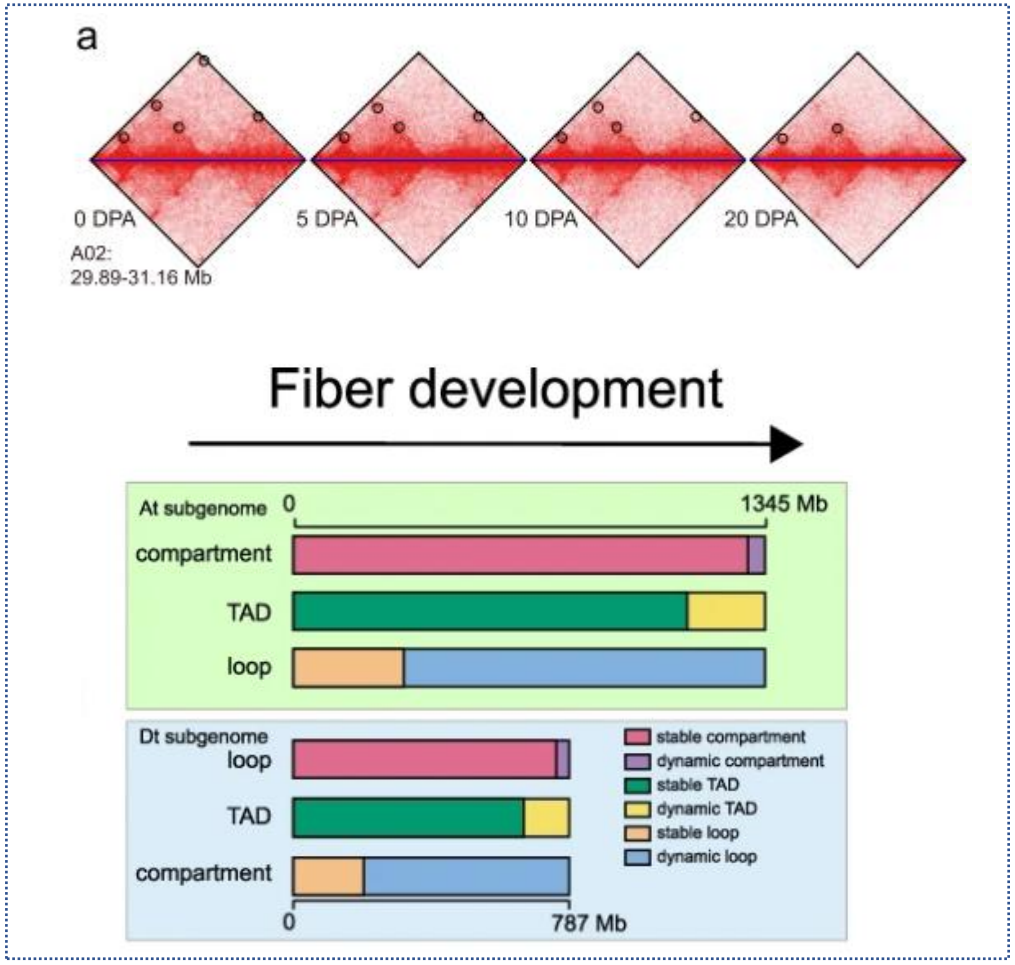
云南农業大學 Yunnan Agricultural University
云南省生物大数据重点实验室
生信帮

2、构建棉花纤维高分辨率动态3D基因组结构图

研究首次构建了棉花纤维的高分辨率动态三维基因组结构图谱，揭示了亚基因组协作调控异源四倍体棉花纤维发育的拓扑结构基础。







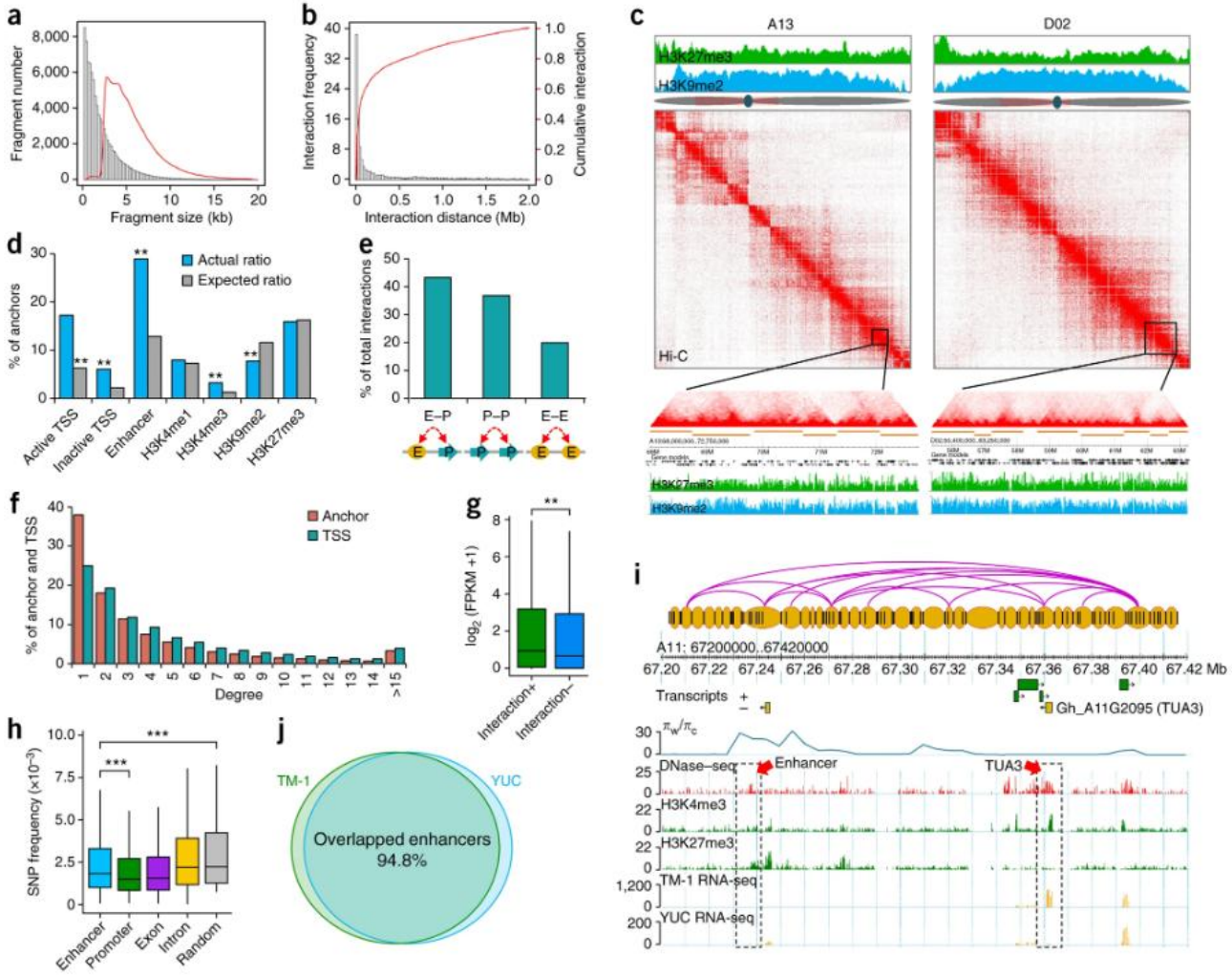
Pei L,etal. Dynamic 3D genome architecture of cotton fiber reveals subgenome-coordinated chromatin topology for 4-staged single-cell differentiation. ***Genome Biol.*** 2022 Feb 3;

Hi-C技术在表观调控中应用

云南农業大學 Yunnan Agricultural University
云南省生物大数据重点实验室
生信帮

3、棉花驯化过程中的不对称亚基因组选择和顺式调控分歧

- 该项研究首先利用267份材料对纤维品质相关性状进行全基因组关联分析，一共鉴定了19个显著位点，其中有4个位点位于驯化选择区间中。
- 对作物的人工驯化常常会改变大量基因的表达水平。
- 为了分析基因差异性表达的原因，研究者巧妙地将DNA酶I酶切测序和三维基因组技术结合起来，鉴定了大量启动子上的顺式调控元件和远距离作用的增强子元件。
- 这些转录调控元件受到了强烈的驯化选择，与基因的差异表达相关。
- 首次在植物中对非编码区的调控变异进行分析，为在其他物种中挖掘功能变异提供了重要参考



Wang et al. Asymmetric subgenome selection and cis-regulatory divergence during cotton domestication. ***Nat Genet.*** 2017 Apr

生信帮，答你所问

雲南農業大學

Yunnan Agricultural University

云南省生物大数据重点实验室