XX大学XX基因组分析 Hi-C挂载分析报告

项目名称：XX大学XX基因组分析

项目编号：XX

样本编号：XX

分析人员：韩露

审核人员：李威、刘潇潇

报告日期：2025年XX月XX日

报告单位：西安浩瑞基因技术有限公司

1 背景介绍

基因组DNA在细胞核中并不是呈线性的一字排列，而是以三维结构高度折叠并浓缩成染色体的方式储存于核内，具有特定的高级空间结构和构象[1]。高通量染色体构象捕获(high-througnput chromosome conformation capture, Hi-C)技术于2009 年首次被提出[2]，目前已得到大规模运用，使得人们对于三维基因组学有了更深刻的认识。Hi-C技术是以整个细胞核为研究对象，利用高通量测序技术，结合生物信息学方法，研究全基因组范围内整个染色体DNA的互作关系[3]。染色体三维构象的维持对于细胞行使正常的功能起着至关重要的作用，染色体构象的变化对基因转录、DNA复制等生物学过程具有调节作用[4-7]。

利用Hi-C 技术能够对染色体内部或所有染色体之间的相互作用进行精细分析，从而把基因表达调控引入到空间的、全局性的研究层面，为全面解析与DNA 有关的生物学过程的机理开启新的契机[8]。Hi-C数据与ChIP-Seq，转录组数据联合分析，可以从基因调控网络和表观遗传网络来阐述生物体性状形成的相关机制。

1.1 实验流程

样品交联完成后，进行细胞裂解，并取样抽提检测样本质量。

检测合格后进入“Hi-C片段”制备流程。使用限制性内切酶（HindIII/DpnII，报告中后续提到的‘酶切位点’或者‘酶切片段’均指的是此酶，本报告使用的是DpnII）进行染色质消化，并取样检测酶切效果。之后经生物素标记、平末端连接及DNA纯化提取，制备Hi-C样本，并取样进行DNA质量检测。检测合格后进入标准文库构建流程。Hi-C片段经去除末端标记的生物素，超声打断，末端修复，加碱基A，钓取含有生物素的片段，加测序接头形成加接头产物。最后进行PCR条件的筛选并扩增获得文库产物。

理想的文库是由来自两个不同酶切片段的DNA片段连接构成，片段连接位点会形成一种新的内切酶切割位点（例如：HindIII酶切，平末端连接后形成的新酶切位点为NheI所识别切割）。文库扩增产物取样进行“Hi-C片段连接点质控检测”，检测合格后完成整个文库制备。构建好的文库经过文库质控合格后，用Illumina HiSeq/MGI进行测序，测序策略PE150。

以下是实验流程示意图:

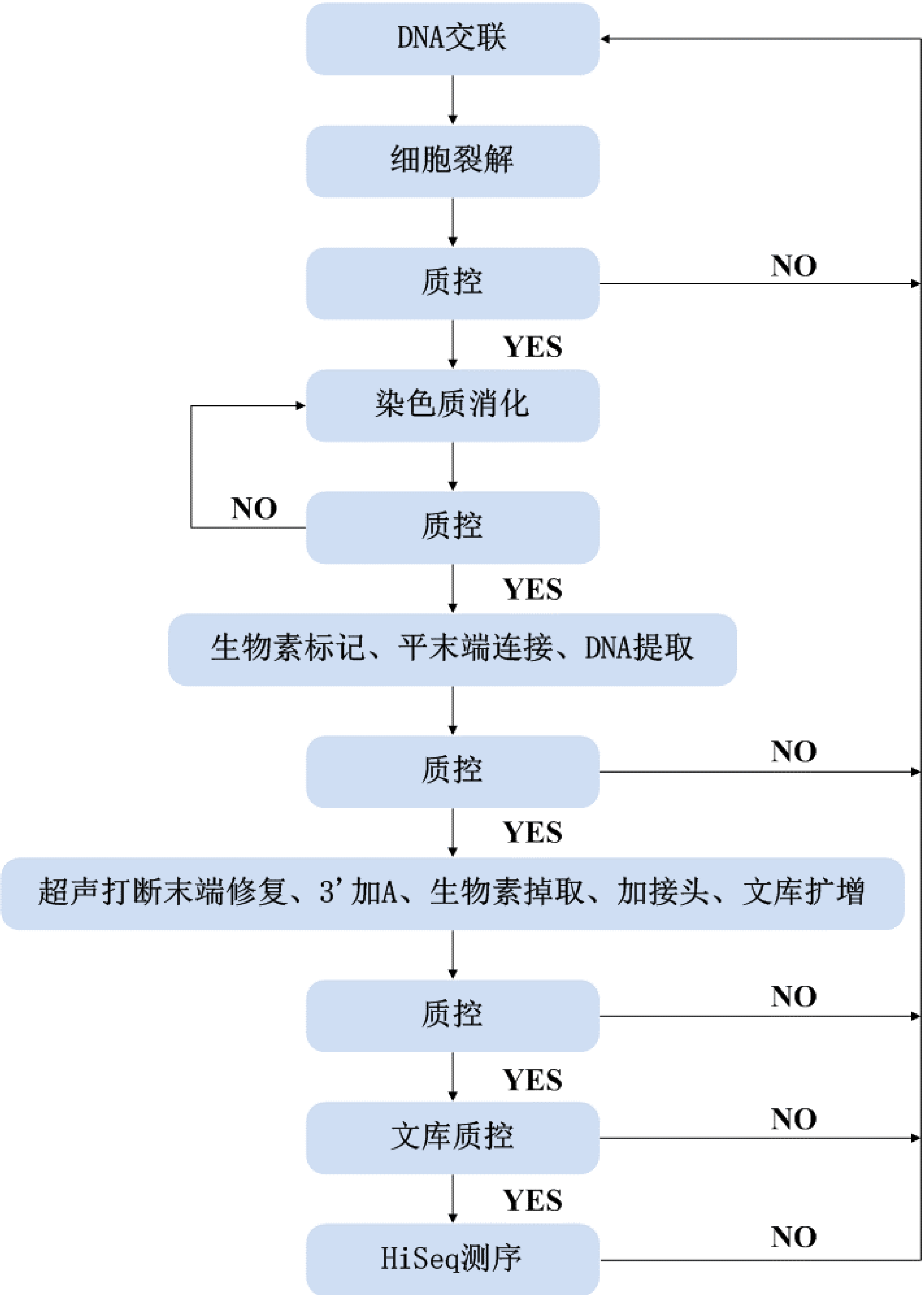

图1 实验流程图

以下是实验流程示意图[1]:

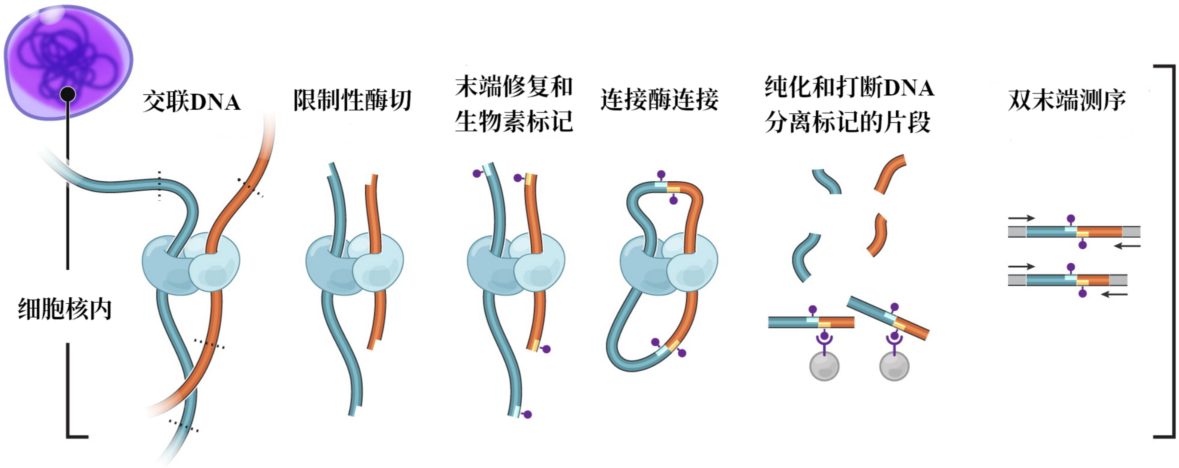

图2 实验流程示意图

1.2 生信分析流程

对原始下机数据过滤，获得高质量的Reads，比对到基因组，提取PE两端均比对到基因组唯一位置的Reads Pair用于后续的分析。利用参考基因组酶切片段（实验所用内切酶）信息，过滤得到Valid reads pair（PE 落在不同的酶切片段），最后分析酶切片段相互作用并将草图组装到染色体级别。

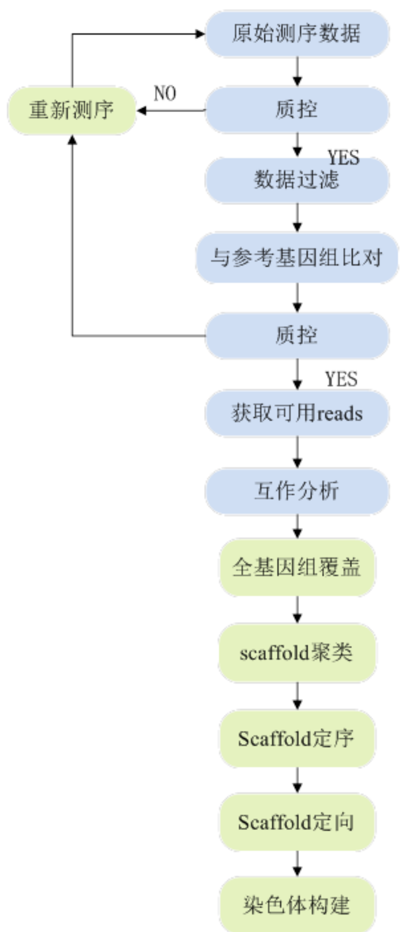

图3 信息分析流程图

2 数据过滤

2.1 原始数据

Illumina/MGI高通量测序结果最初以原始图像数据文件存在，经CASAVA软件进行碱基识别（Base Calling）后转化为原始测序序列（Sequenced Reads），其结果以FASTQ文件格式存储。FASTQ文件中每条测序序列（Read）包含名称、碱基序列以及其对应的测序质量信息。FASTQ格式示例如下：

```
@HWI-7001268:911:H2MVGBCXX:1:1101:1180:2085 1:N:0
TCCAGCTCCAATAGCGTATATTTAAGTTGTTGCAGTTAAAAAGCTCGTAGTTGGACCTTGG
+
DDDDDIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIHIIIIIIIIIIIIIII
```

图4 FASTQ文件格式示例图

> 注：FASTQ文件中每4行对应一条Read的信息。其中第一行以“@”开头，随后为测序标识别符和描述文字；第二行是碱基序列；第三行是以“+”开头的可选信息；第四行中每个字符对应的ASCII值减去33或64，即为第二行对应碱基的测序质量值。

2.2 测序数据质量评估

利用fastp[9]对原始数据（Rawdata）进行过滤，质控指标一般包括：

1. 去掉Reads的接头序列；
2. 去除序列中含有N的Reads；
3. 当一条Reads中超过20%的碱基质量分数小于20，则舍弃该Reads所对应的一对Reads。

然后对原始数据及过滤后的高质量数据（clean data）进行质量统计，具体结果见表1。

表1 数据产量统计

Sample id	Total reads	Total bases	Clean reads	Clean bases	Q20 rate(%)	Q30 rate(%)	GC(%)
XXX	792,533,734	118,880,060,100	696,691,224	104,217,305,752	99.49	98.16	39.80

> 注：Sample id：测序样本编号；Total reads：总Reads数；Total bases：总碱基数；Clean reads：质控过滤后的Reads数；Clean bases：质控过滤后的碱基数；Q20 rate(%)：质控过滤后测序质量值大于Q20的碱基百分比；Q30 rate(%)：质控过滤后测序质量值大于Q30的碱基百分比；GC(%)：质控过滤后数据的GC含量百分比。

2.3 测序数据质量值分布

测序错误率与碱基质量有关，受测序仪本身、测序试剂、样品等多个因素共同影响。每个碱基测序错误率是通过测序Phred数值（Phred Score，Qphred）通过公式转化得到：

$$\mathrm{Phred} = -10\log_{10} e$$

图5 Phred计算公式

其中e为碱基测序错误率，Phred数值是在碱基识别过程中，通过一种预测碱基判别发生错误概率模型计算得到的，对应关系如下表所示：

表2 碱基识别与Phred分值之间的简明对应关系

Phred分值	不正确的碱基识别	碱基正确识别率	Q-score
10	1/10	90%	Q10
20	1/100	99%	Q20
30	1/1000	99.90%	Q30
40	1/10000	99.99%	Q40

为了粗略反映测序过程中测序质量的稳定性，以过滤后序列的碱基位置作为横坐标，每个位置的平均质量值作为纵坐标，得到下面的测序质量分布图：

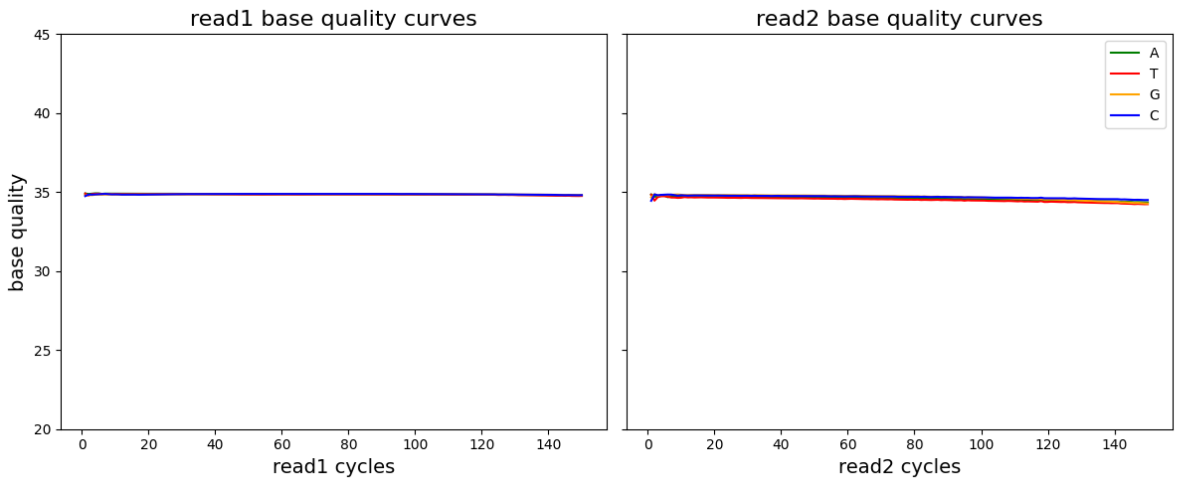

图6 样品测序质量分布图

> 注：上图为碱基质量分布图，横坐标为Reads的位置，纵坐标为测序质量值，不同颜色的线条分别代表4种碱基质量值的平均值。

2.4 测序数据碱基分布

以过滤后Reads的碱基位置作为横坐标，以每个位置的ATGC碱基的比例作为纵坐标，得到Reads的碱基分布图。在测序中Reads的每个位置上A碱基和T碱基比例相等，G碱基和C碱基比例相等，表明了测序没有偏好性。样品碱基含量分布图如下所示：

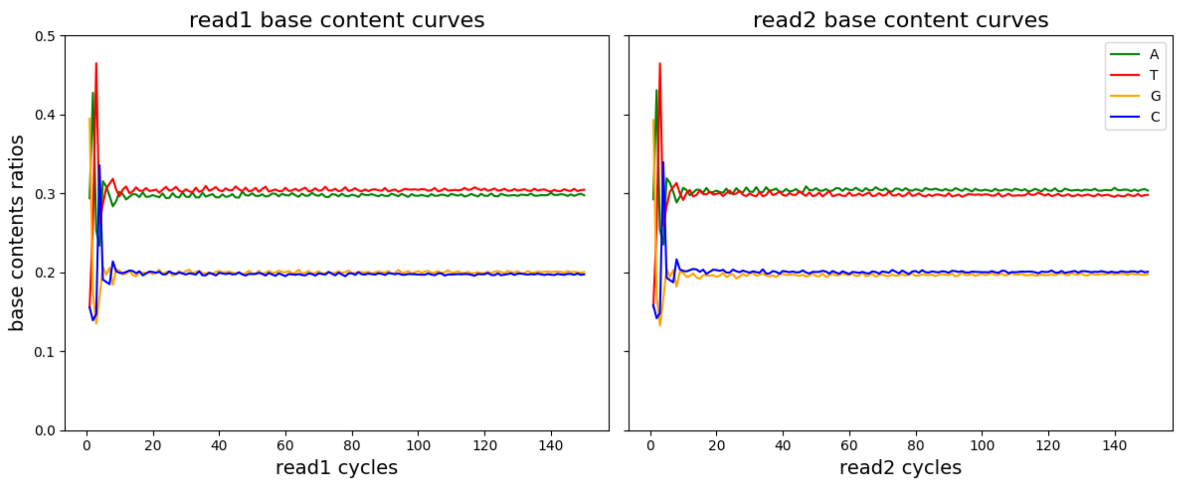

图7 样品碱基含量分布图

> 注：上图为碱基含量分布图。其中，绿色、红色、橙色和蓝色线条分别是A、T、G和C四种碱基的含量，由于二代测序的本身特性，前十几个bp碱基含量会有波动。从碱基含量分布图中可以看出，在十几个bp以后，A与T、G与C含量基本一致，数据碱基含量合格。

3 比对及片段分析

3.1 比对分析

由于Hi-C建库的特殊性，使用HiC-Pro[10]的比对策略，调用bowtie2进行比对[11]。比对及过滤策略如下图所示：

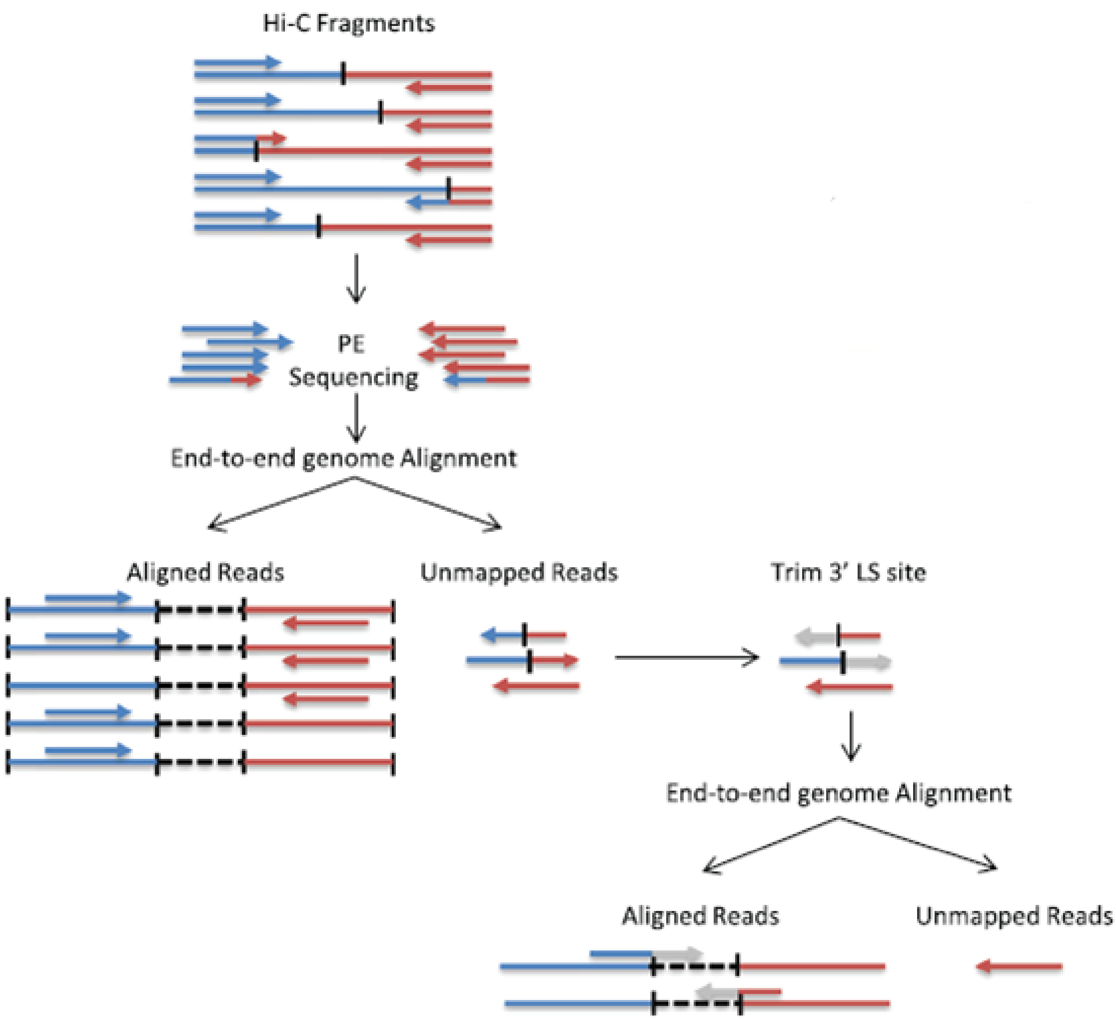

图8 PE Reads比对示意图

分别将Reads1和Reads2与参考基因组进行单端比对，得到比对结果。对于未比对到参考基因组（unmap）的Reads，如果能找到酶切之后文库的连接位点（ligation-site），那就截掉连接位点之后的部分再次进行比对，合并两次比对的结果，挑选PE两端都比对到基因组唯一位置的Reads Pair（Unique Mapped Paired-end Reads），进行后续的分析。

比对结果如下表所示：

表3 样品的比对结果统计

Type	Count	Percent(%)
Clean Paired-end Reads	348,345,612	100.00
Unmapped Paired-end Reads	3,413,984	0.98
Paired-end Reads with Singleton	41,585,742	11.94
Multi Mapped Paired-end Reads	0	0.00

Type	Count	Percent(%)
Low Quality Paired-end Reads	157,306,426	45.16
Unique Mapped Paired-end Reads	146,039,460	41.92

> 注：Clean Paired-end Reads：过滤后得到的高质量的Reads对数；Unmapped Paired-end Reads：两端均未比对到基因组的Reads对数；Paired-end Reads with Singleton：有一端比对到基因组的Reads对数；Multi Mapped Paired-end Reads：比对到基因组多个位置的Reads对数；Low Quality Paired-end Reads：低质量比对的Reads对数；Unique Mapped Paired-end Reads：比对到基因组唯一位置上的Reads对数。

3.2 基于唯一比对Reads搜索片段

Hi-C建库会产生两种情况的数据，分别为有效对（Valid Interaction Pairs）和无效对（Invalid Interaction Pairs）[12]。其中Valid Interaction Pairs指测序得到的双端Reads分别来源于空间上相邻但线性上不相邻的两个酶切后的DNA片段，其能够提供有效的交互信息。而不能够提供片段间交互信息的称为Invalid Interaction Pairs。其中Invalid Interaction Pairs主要包含自环（Self Circle）、边缘悬挂（Dangling End）和其他丢弃的类型（Dumped Pairs）三种。如下图所示：

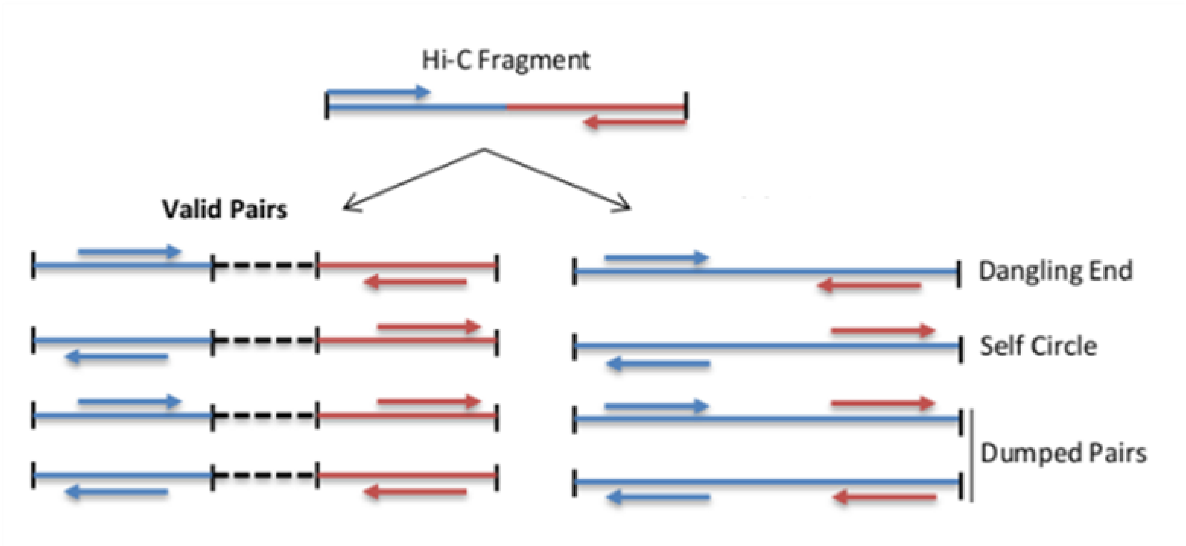

图9 建库会产生不同的分子类型

Invalid Interaction Pairs都是来自相同的酶切片段，根据试验原理，不同的Invalid PE Reads在比对方向上有其不同的特点，具体表现为：若比对方向为`<-->`识别为自环；若比对到基因组的方向为`-><-`，则识别为生物素落在Reads的一端，即边缘悬挂；若比对方向为`->->`或者`<-<-`识别为Dumped Pairs。我们会对插入片段（Insert Size）进行预测（Reads1比对起始位置到Reads1比对方向下游的第一个酶切位点的距离 + Reads2比对起始位置到Reads2比对方向下游的第一个酶切位点的距离），如果预测的插入片段不符合试验测定的插入片段范围也会归为Dumped Pairs。

表4 统计唯一比对的Paired-end Reads

Type	Count	Percent(%)
Unique Mapped Paired-end Reads	146,039,460	100.00
Dangling End Paired-end Reads	41,098,229	28.14
Self Circle Paired-end Reads	66,808	0.05

Type	Count	Percent(%)
Dumped Paired-end Reads	23,646	0.02
Religation Paired-end Reads	2,864,522	1.96
Valid Paired-end Reads	101,986,255	69.83

> 注：Unique Mapped Paired-end Reads：唯一比对到基因组上的Reads对数；Dangling End Paired-end Reads：PE比对方向为Dangling End的Reads对数；Self Circle Paired-end Reads：PE比对方向为Self Circle的Reads 对数；Dumped Paired-end Reads：PE两端的Reads落在同一酶切片段内部且比对方向相同的Reads对数与预测插入片段不符合预期的Reads对数之和；Religation Paired-end Reads：PE两端的Reads发生重新连接的Reads对数；Valid Paired-end Reads：去除PCR过程中产生的重复序列的Reads对后PE两端的Reads落在不同酶切片段的Reads对数。

4 辅助组装

4.1 Hi-C辅助组装原理

基于顺势相互作用（同一染色体内的互作）远大于反式相互作用（不同染色体间的互作），且顺势相互作用中线性距离越近则互作越强的原理，将Contigs或Scaffolds进行聚类、排序、定向，得到染色体水平基因组。原理如下图：

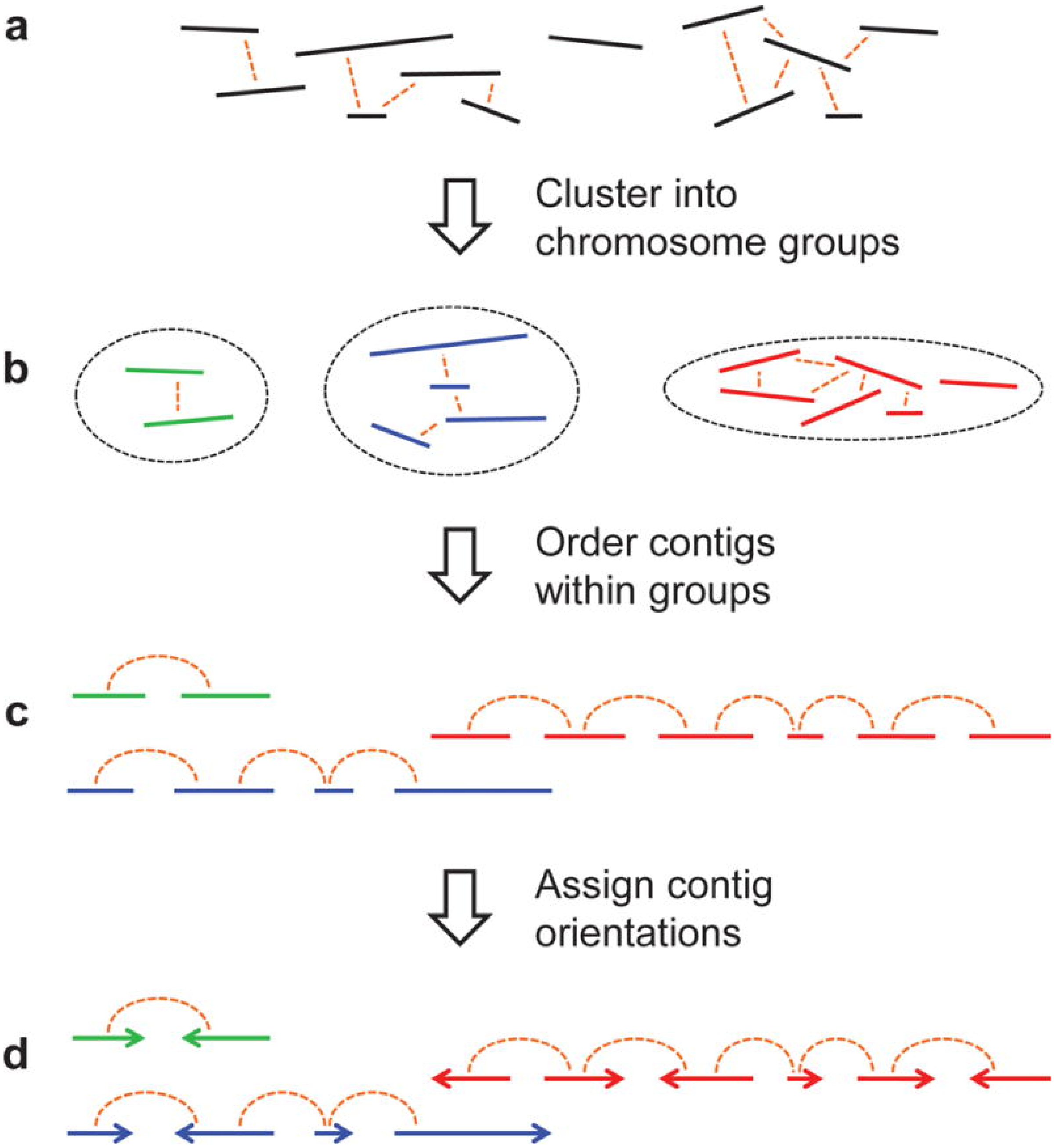

图10 Hi-C辅助组装原理

> 注：红色虚线代表两条Contigs或Scaffolds间有互作；根据染色体具有'chromosome territory'的特性，将Contigs 或者Scaffolds进行聚类，理想情况每条染色体会单独聚为1类，黑色虚线代表1类；Contigs或者Scaffolds间的相对顺序，根据一维距离越近其互作数越多的规律进行排序；排序后的序列根据与排序一样的原理进行定向。

4.2 Hi-C辅助组装及纠错

使用ALLHiC[13]或者Juicer[14]和3D-DNA[15]等软件[16]对比对到草图基因组上的有效数据进行自动聚类、排序和定向，并生成最初的组装挂载文件。

使用JuiceBox[17]软件对最初的组装挂载文件进行可视化纠错。通过3D-DNA等构建全基因组互作图谱，基于互作热图来发现自动聚类中存在的Contigs顺序、方向或其内部的组装错误，并进行校正以及染色体疆域划分。使校正后的全基因组互作图谱呈现染色体内部互作强于染色体间互作，并且线性距离越近互作越强、局部互作平滑的特点。

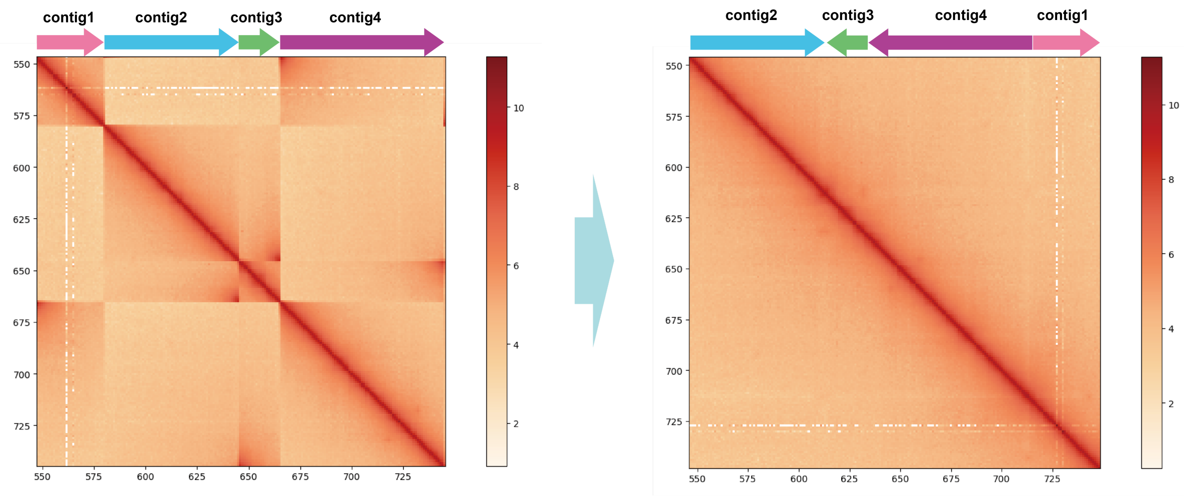

图11 JuiceBox纠错原理

> 注：左图中的四个Contigs存在位置或方向错误，依据Contigs之间互作强度关系进行调整得到右图。

5 结果统计

5.1 组装结果统计

对草图组装指标进行统计，结果见下表：

表5 草图的组装结果数据统计表

Type	Length(bp)	Number
Longest	128,515,687	1
N50	101,985,394	5
N60	101,985,394	5
N70	94,269,498	6
N80	69,937,382	8
N90	59,929,633	9
Length>=50kb	922,521,230	198
Length>=10kb	930,067,953	405
Length>=5kb	930,067,953	405
Length>=1kb	930,067,953	405
Total	930,067,953	405

> 注：Longest: 组装得到的最长序列的长度，单位bp；N(5-9)0: 将组装好的序列按照从长到短进行排列并进行累加，当累加长度达到总长的(5-9)0%的时候，那一条序列的长度，以bp为单位。比如N50 表示将组装好的序列按照从长到短进行排列并进行累加，当累加长度达到总长的50%的时候，那一条序列的长度；Length>=50kb: 组装得到的序列长度长于50Kb的条数；Total: 组装得到的序列总长，以bp为单位。

通过Hi-C辅助组装后，将草图Contigs进行了染色体挂载，挂载后指标统计结果见下表。

表6 Hi-C辅助组装的组装结果数据统计表

Type	Length(bp)	Number
Longest	128,515,687	1
N50	104,599,850	5
N60	104,599,850	5
N70	101,985,394	6
N80	94,269,498	7
N90	87,629,897	8
Length>=50kb	922,521,530	195

Type	Length(bp)	Number
Length>=10kb	930,068,253	402
Length>=5kb	930,068,253	402
Length>=1kb	930,068,253	402
Total	930,068,253	402

> 注：Longest: 染色体挂载得到的最长序列的长度，单位bp；N(5-9)0: 将挂载后的序列按照从长到短进行排列并进行累加，当累加长度达到总长的(5-9)0%的时候，那一条序列的长度，以bp为单位。比如N50 表示将挂载得到的序列按照从长到短进行排列并进行累加，当累加长度达到总长的50%的时候，那一条序列的长度；Length>=50kb: 挂载得到的序列长度长于50Kb的条数；Total: 挂载得到的序列总长，以bp为单位。

草图Contigs挂载得到Pseudomolecule，对所得Pseudomolecule进行统计，结果见下表（由于在使用JuiceBox对基因组进行进行可视化纠错时可能会根据Hi-C互作信号对Contigs进行切分，因此此处统计的Contigs数量可能会与组装草图的Contigs数量有所差异）。

表7 Hi-C辅助组装的组装结果Pseudomolecule长度统计表

Pseudomolecule	Contig Num	Length(bp)	Percent(%)
Chr01	1	128,515,687	13.82
Chr02	1	114,998,005	12.36
Chr03	2	110,551,454	11.89
Chr04	1	104,929,016	11.28
Chr05	3	104,599,850	11.25
Chr06	1	101,985,394	10.97
Chr07	1	94,269,498	10.14
Chr08	1	87,629,897	9.42
Chr09	1	59,929,633	6.44
Total anchored	12	907,408,434	97.56
Unanchored	393	22,659,819	2.44
Total	405	930,068,253	100.00

> 注：Pseudomolecule：组装得到的Contigs经过聚类、排序和定向得到最终的序列；Contig Num: 每个Pseudomolecule 挂载的Contigs个数；Length(bp): Pseudomolecule长度，单位bp，Contigs之间以100 bp的'N'连接；Percent(%): 每条Pseudomolecule占基因组总长度的百分比；Total anchored: 全部染色体挂载的Contigs情况；Unanchored: 基因组草图中未挂载Contigs情况；Total: 基因组总体情况。

5.2 组装结果的热图验证

对组装的染色体，进行热图验证，是辅助组装准确性的最优标准。

热图构建的方法：对于辅助组装的染色体，切割成等长Bin（常见的1Mb、500Kb、150Kb等），以两两Bin之间支持的有效比对Read对，作为两两Bin之间交互的强度信号，构建热图[18]。热图坐标表示各染色体的所有Bins，每个点的颜色代表了相应基因组Bin pair交互强度的log值，交互强度从黄到红依次增强。

基于Hi-C辅助组装的两个基本原理：（1）染色体内部的互作强度强于染色体之间的互作关系；（2）同一个染色体上，线性距离近的互作强度强于线性距离远的互作关系。对应的，染色体热图验证中，存在两个基本规律：（1）对角线附近的互作关系，明显强于远离对角线的互作关系；（2）在距离近的两两Bin之间的互作关系，强于距离远的Bin之间的互作关系。如果出现了远距离的两两Bin强于近距离的两两Bin的互作关系，就可能是组装或挂载错误导致。

下图展示了全基因组Hi-C交互热图，从全局的热图来看，组装效果比较理想。

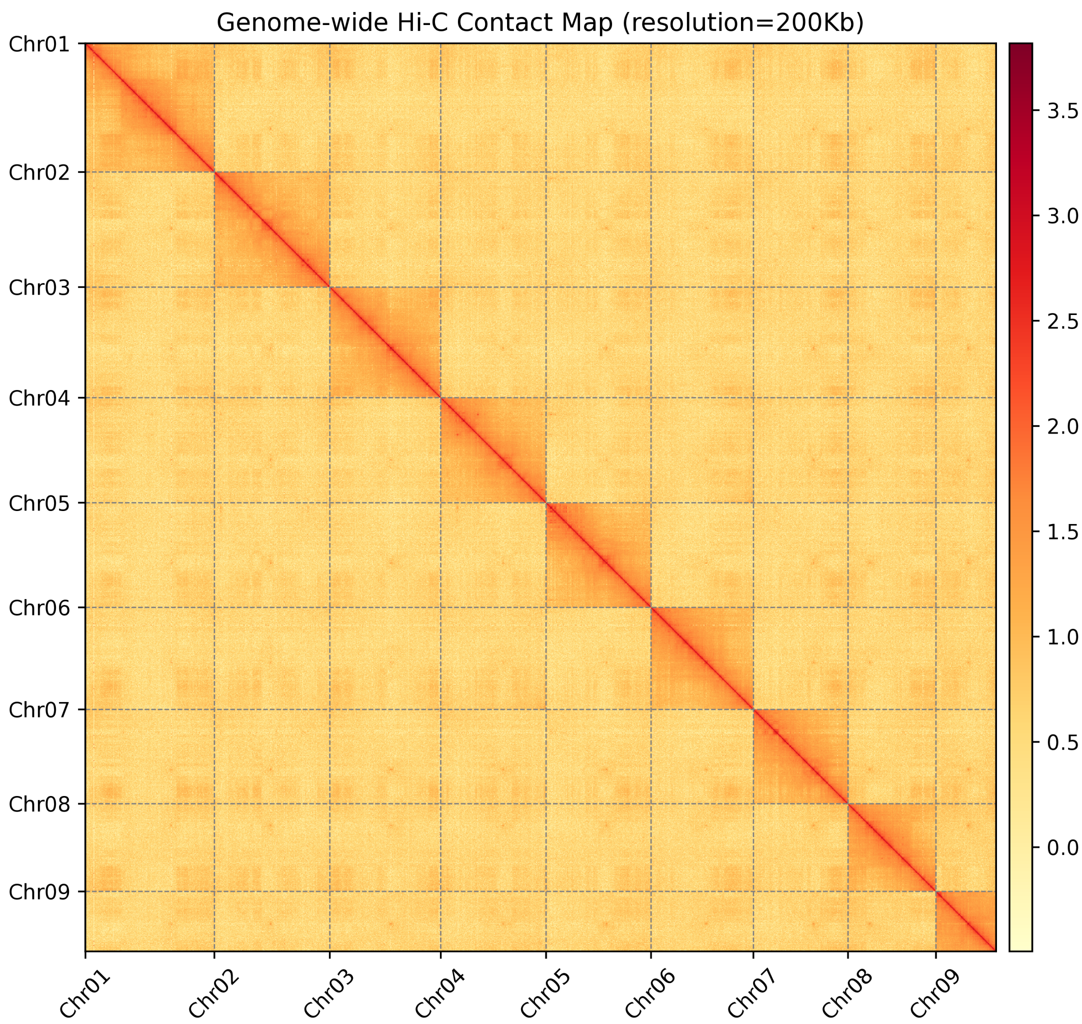

图12 全基因组Hi-C交互热图

染色体内部互作热图见目录：src/summary/4_Heatmap/Heatmap_Chr/。

6 使用软件列表

该项目中使用的软件，如下表所示：

表8 软件列表

分析内容	软件	版本
Hi-C数据质控	fastp	v0.21.0
Hi-C序列比对	Bowtie2	v2.4.4
Hi-C文库质控	HiC-Pro	v3.0.0
Hi-C数据过滤	juicer	v1.6
Hi-C辅助组装	3D-DNA	v180922
Hi-C辅助组装	ALLHiC	v0.9.13
Hi-C辅助组装	YaHS	v1.1
可视化纠错	JuiceBox	v2.13.07
Hi-C互作热图	HicExplorer	v3.7.2

7 参考文献

1. Lanctôt C, Cheutin T, Cremer M, Cavalli G, Cremer T. **Dynamic genome architecture in the nuclear space: regulation of gene expression in three dimensions**. *Nat Rev Genet*. 2007 Feb;8(2):104-15. doi: 10.1038/nrg2041. PMID: 17230197. ↩
2. Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO, Sandstrom R, Bernstein B, Bender MA, Groudine M, Gnirke A, Stamatoyannopoulos J, Mirny LA, Lander ES, Dekker J. **Comprehensive mapping of long-range interactions reveals folding principles of the human genome**. *Science*. 2009 Oct 9;326(5950):289-93. doi: 10.1126/science.1181369. PMID: 19815776; PMCID: PMC2858594. ↩
3. van Berkum NL, Lieberman-Aiden E, Williams L, Imakaev M, Gnirke A, Mirny LA, Dekker J, Lander ES. **Hi-C: a method to study the three-dimensional architecture of genomes**. *J Vis Exp*. 2010 May 6;(39):1869. doi: 10.3791/1869. PMID: 20461051; PMCID: PMC3149993. ↩
4. Misteli T. **Spatial positioning; a new dimension in genome function**. *Cell*. 2004 Oct 15;119(2):153-6. doi: 10.1016/j.cell.2004.09.035. PMID: 15479633. ↩
5. Misteli T. **Beyond the sequence: cellular organization of genome function**. *Cell*. 2007 Feb 23;128(4):787-800. doi: 10.1016/j.cell.2007.01.028. PMID: 17320514. ↩
6. Dekker J. **Gene regulation in the third dimension**. *Science*. 2008 Mar 28;319(5871):1793-4. doi: 10.1126/science.1152850. PMID: 18369139; PMCID: PMC2666883. ↩
7. Miele A, Dekker J. **Long-range chromosomal interactions and gene regulation**. *Mol Biosyst*. 2008 Nov;4(11):1046-57. doi: 10.1039/b803580f. Epub 2008 Aug 13. PMID: 18931780; PMCID: PMC2653627. ↩
8. Fortin JP, Hansen KD. **Reconstructing A/B compartments as revealed by Hi-C using long-range correlations in epigenetic data**. *Genome Biol*. 2015 Aug 28;16(1):180. doi: 10.1186/s13059-015-0741-y. PMID: 26316348; PMCID: PMC4574526. ↩

9. Chen S, Zhou Y, Chen Y, Gu J. **fastp: an ultra-fast all-in-one FASTQ preprocessor**. *Bioinformatics*. 2018 Sep 1;34(17):i884-i890. doi: 10.1093/bioinformatics/bty560. PMID: 30423086; PMCID: PMC6129281. ↩
10. Servant N, Varoquaux N, Lajoie BR, Viara E, Chen CJ, Vert JP, Heard E, Dekker J, Barillot E. **HiC-Pro: an optimized and flexible pipeline for Hi-C data processing**. *Genome Biol*. 2015 Dec 1;16:259. doi: 10.1186/s13059-015-0831-x. PMID: 26619908; PMCID: PMC4665391. ↩
11. Langmead B, Wilks C, Antonescu V, Charles R. **Scaling read aligners to hundreds of threads on general-purpose processors**. *Bioinformatics*. 2019 Feb 1;35(3):421-432. doi: 10.1093/bioinformatics/bty648. PMID: 30020410; PMCID: PMC6361242. ↩
12. Belton JM, McCord RP, Gibcus JH, Naumova N, Zhan Y, Dekker J. **Hi-C: a comprehensive technique to capture the conformation of genomes**. *Methods*. 2012 Nov;58(3):268-76. doi: 10.1016/j.ymeth.2012.05.001. Epub 2012 May 29. PMID: 22652625; PMCID: PMC3874846. ↩
13. Zhang X, Zhang S, Zhao Q, Ming R, Tang H. **Assembly of allele-aware, chromosomal-scale autopolyploid genomes based on Hi-C data**. *Nat Plants*. 2019 Aug;5(8):833-845. doi: 10.1038/s41477-019-0487-8. Epub 2019 Aug 5. PMID: 31383970. ↩
14. Durand NC, Shamim MS, Machol I, Rao SS, Huntley MH, Lander ES, Aiden EL. **Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments**. *Cell Syst*. 2016 Jul;3(1):95-8. doi: 10.1016/j.cels.2016.07.002. PMID: 27467249; PMCID: PMC5846465. ↩
15. Dudchenko O, Batra SS, Omer AD, Nyquist SK, Hoeger M, Durand NC, Shamim MS, Machol I, Lander ES, Aiden AP, Aiden EL. **De novo assembly of the Aedes aegypti genome using Hi-C yields chromosome-length scaffolds**. *Science*. 2017 Apr 7;356(6333):92-95. doi: 10.1126/science.aal3327. Epub 2017 Mar 23. PMID: 28336562; PMCID: PMC5635820. ↩
16. Zhou C, McCarthy SA, Durbin R. **YaHS: yet another Hi-C scaffolding tool**. *Bioinformatics*. 2023 Jan 1;39(1):btac808. doi: 10.1093/bioinformatics/btac808. PMID: 36525368; PMCID: PMC9848053. ↩
17. Durand NC, Robinson JT, Shamim MS, Machol I, Mesirov JP, Lander ES, Aiden EL. **Juicebox Provides a Visualization System for Hi-C Contact Maps with Unlimited Zoom**. *Cell Syst*. 2016 Jul;3(1):99-101. doi: 10.1016/j.cels.2015.07.012. PMID: 27467250; PMCID: PMC5596920. ↩
18. Wolff J, Rabbani L, Gilsbach R, Richard G, Manke T, Backofen R, Grüning BA. **Galaxy HiCExplorer 3: a web server for reproducible Hi-C, capture Hi-C and single-cell Hi-C data analysis, quality control and visualization**. *Nucleic Acids Res*. 2020 Jul 2;48(W1):W177-W184. doi: 10.1093/nar/gkaa220. PMID: 32301980; PMCID: PMC7319437. ↩

8 联系我们

西安浩瑞基因成立于2019年，公司引入了三代测序平台--3台PacBio Revio和7台SequelI设备，致力于深耕动植物基因组学、转录组和微生物组学研究的科研技术服务。2024年，与华大智造携手共建西北首家DCSLab组学前沿实验室，引入DNBSEO-T7测序平台，开展基于二代测序的单细胞转录组、时空转录组等前沿技术服务。凭借专业的一站式多组学技术，为广大科研客户提供专业、高效、可靠的组学科研技术服务。

联系方式

热线电话：+86 029-89303503

官方网站：www.xahorizon.cn

邮　箱：project@xahorizon.cn

地　址：陕西省西安市沣东新城中兴深蓝科技产业园A区2号楼3层