XX大学XX基因组测序 组装分析报告

项目名称：XX大学XX基因组测序

项目编号：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

XX大学XX基因组测序 组装分析报告

1 技术背景
2 实验流程
2.1 DNA质检
2.2 文库构建及质量检测
2.3 DNA测序
3 分析流程
3.1 数据质控
3.1.1 HiFi数据质控
3.1.1.1 HiFi数据质控方法
3.1.1.2 HiFi数据统计
3.1.2 ONT数据质控
3.1.2.1 ONT数据质控方法
3.1.2.2 ONT数据统计
3.1.3 HiC数据质控
3.1.3.1 HiC数据质控方法
3.1.3.2 HiC数据统计
3.2 长序列组装
3.3 基因组组装结果评估
3.3.1 BUSCO评估
3.3.2 GC-Depth分析
4 分析软件
5 分析方法
5.1 数据质控
5.2 基因组组装
5.3 基因组组装结果评估
5.3.1 BUSCO评估
5.3.2 GC-Depth分析
6 参考文献
7 联系我们
联系方式

1 技术背景

从早期的Sanger测序到第二代高通量测序，再到当前的单分子测序技术（第三代高通量测序，Third-generation Sequencing, TGS），DNA测序技术持续推动着生命科学研究的快速发展。第二代测序虽然通量高、成本低，但其读长较短、PCR扩增引入偏好性等局限在复杂基因组、重复区域、大片段结构变异等场景中难以胜任。第三代测序技术以PacBio单分子实时测序（SMRT）和纳米孔测序为代表，具备无需PCR扩增、读长更长、可直接检测表观修饰等优势，有效解决了二代测序在组装完整性和变异检测方面的不足[1]。

PacBio Revio是PacBio公司于2022年推出的最新一代SMRT测序平台，相较于Sequel II平台，在通量、准确度、运行成本和自动化水平上实现了显著提升。Revio系统每个SMRT Cell拥有25M个ZMW孔，单芯片产出可达120-160Gb，支持同时运行4个SMRT Cell，通量高、成本低，适合多样本并行测序与高精度全基因组研究，适用于高质量动植物基因组组装、全长转录组分析、结构变异检测等多种高通量测序需求[2]。Revio平台的推出大大加快了高质量基因组学研究的进程，已广泛应用于农业育种、物种演化、医学研究及微生物组学等多个领域[3]。

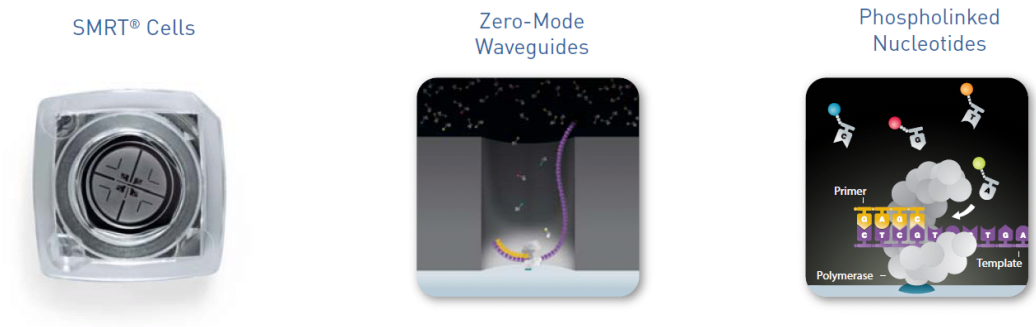

图1 PacBio SMRT 测序原理示意图

2 实验流程

Pacbio测序技术对样品DNA的要求非常高，浩瑞基因对客户提供的DNA进行严格的样品检测，从源头上保证质量。对检测合格的样品建库、上机测序，每个环节都严格把控，保证测序数据的准确性和PacBio测序长读长的特性。

首先提取高质量的DNA，之后进行构建文库，文库构建完成后利用Pacbio Revio系列测序仪对DNA进行单分子实时荧光测序，获得原始测序数据。实验流程如下图所示：

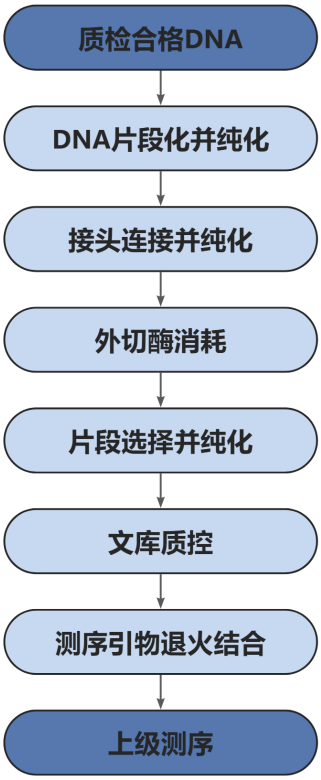

图2 实验流程图

2.1 DNA质检

采用以下四种方法检测DNA是否合格：

1. 样品的外观性状是否含有异物或颜色；
2. 0.75%琼脂糖电泳：检测样品是否有降解以及DNA片段大小；
3. Nanodrop：检测DNA纯度（OD260/280在1.8-2.0之间；OD 260/230在2.0-2.2之间）；
4. Qubit：对DNA进行精确定量。

2.2 文库构建及质量检测

样本质检合格后，根据建库的片段大小对基因组DNA进行打断；将片段化的DNA进行损伤修复和末端修复；在DNA片段两端连接茎环状测序接头，然后利用外切酶去除未连接接头的DNA片段；再用核酸片段仪筛选回收目的片段，纯化后即文库。然后检测文库片段大小。文库大小合格后继续进行后续上机测序。

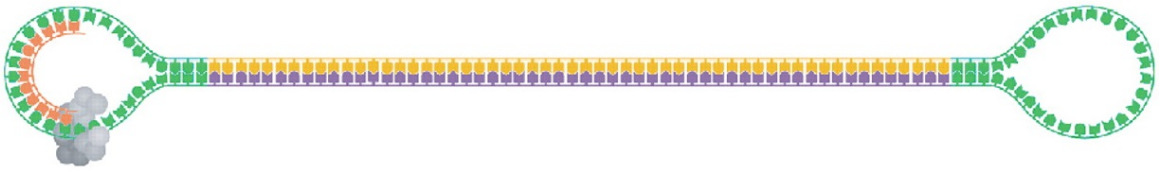

图3 SMRTbell模板结构示意图

> 注：发夹状接头（绿色）连接到双链DNA分子末端（黄色和紫色），构成闭环。锚定于ZMW纳米孔底部的聚合酶（灰色）与测序引物序列（橘黄色）结合，开启测序。

2.3 DNA测序

建库完成后需对文库进行上机前制备，将DNA模版与测序聚合酶进行结合，之后将一定浓度和体积的DNA模板和酶复合物加入到测序试剂板中，并和测序Cell及相关耗材转移到Revio测序仪内开始实时单分子测序。

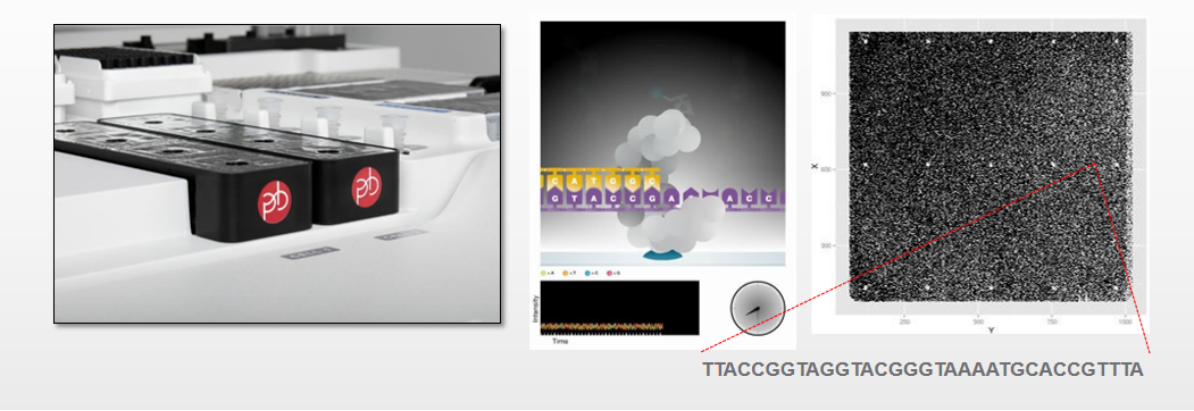

图4 PacBio Revio单分子实时测序

3 分析流程

测序获得物种的组学数据后，对数据质量进行评估，利用高质量的数据进行组装得到基因组，进而对组装得到的基因组进行相关质量评估。具体分析流程如下图：

图5 数据分析流程

3.1 数据质控

3.1.1 HiFi数据质控

3.1.1.1 HiFi数据质控方法

在PacBio的测序平台中，将通过零模波导孔的DNA产生的荧光信号记录成movie进而转化为相应的碱基序列的过程，称为basecalling。使用官方提供的工具SMRTLink进行basecalling获得含接头的测序序列，即酶读（polymerase reads），其长度由反应酶的活性和上机时间决定。酶读去除低质量序列和接头序列后得到subreads。环化共有序列（Circular Consensus Sequencing，CCS）测序模式获得subreads，通过校准同一序列模板多次测序的subreads的随机错误，可将测序准确率提升至99%以上，通过Q20阈值过滤获得高准确性的HiFi reads。

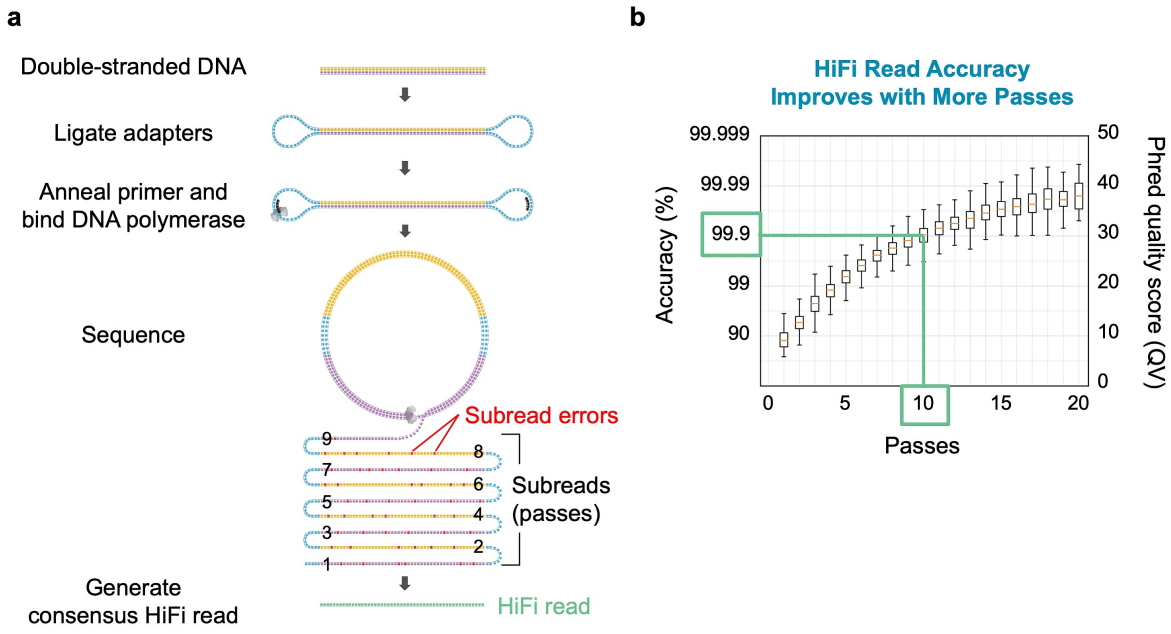

图6 HiFi read数据质控流程图

3.1.1.2 HiFi数据统计

对HiFi reads数据进行统计，总数据量为89.21Gb。各文库及总数据量统计信息如下表所示:

表1 HiFi下机数据统计

Library id	Total bases(nt)	Total reads	Mean length(nt)	Max length(nt)	N50 length(nt)	>10kb rate(%)	>20kb rate(%)	>40kb rate(%)
bam-1	72,574,858,984	4,705,870	15,422	44,073	15,348	98.71	5.05	0
bam-2	16,637,238,013	1,064,501	15,629	41,955	15,543	99.11	6.13	0

> 注：下机数据的质量会根据前期实验提取建库质量的差别而有所不同。表格各列说明如下表：

列名	说明
Library id	测序文库编号；
Total bases(nt)	有效数据的总碱基数；
Total reads	有效数据的总reads数；
Mean length(nt)	有效数据的平均长度；
Max length(nt)	有效数据的最长reads长度；
N50 length(nt)	有效数据的N50长度；
>10kb rate(%)	有效数据中长度大于10kb的reads比例；
>20kb rate(%)	有效数据中长度大于20kb的reads比例；
>40kb rate(%)	有效数据中长度大于40kb的reads比例。

文库r84071_230928_001的read长度分布如下图所示:

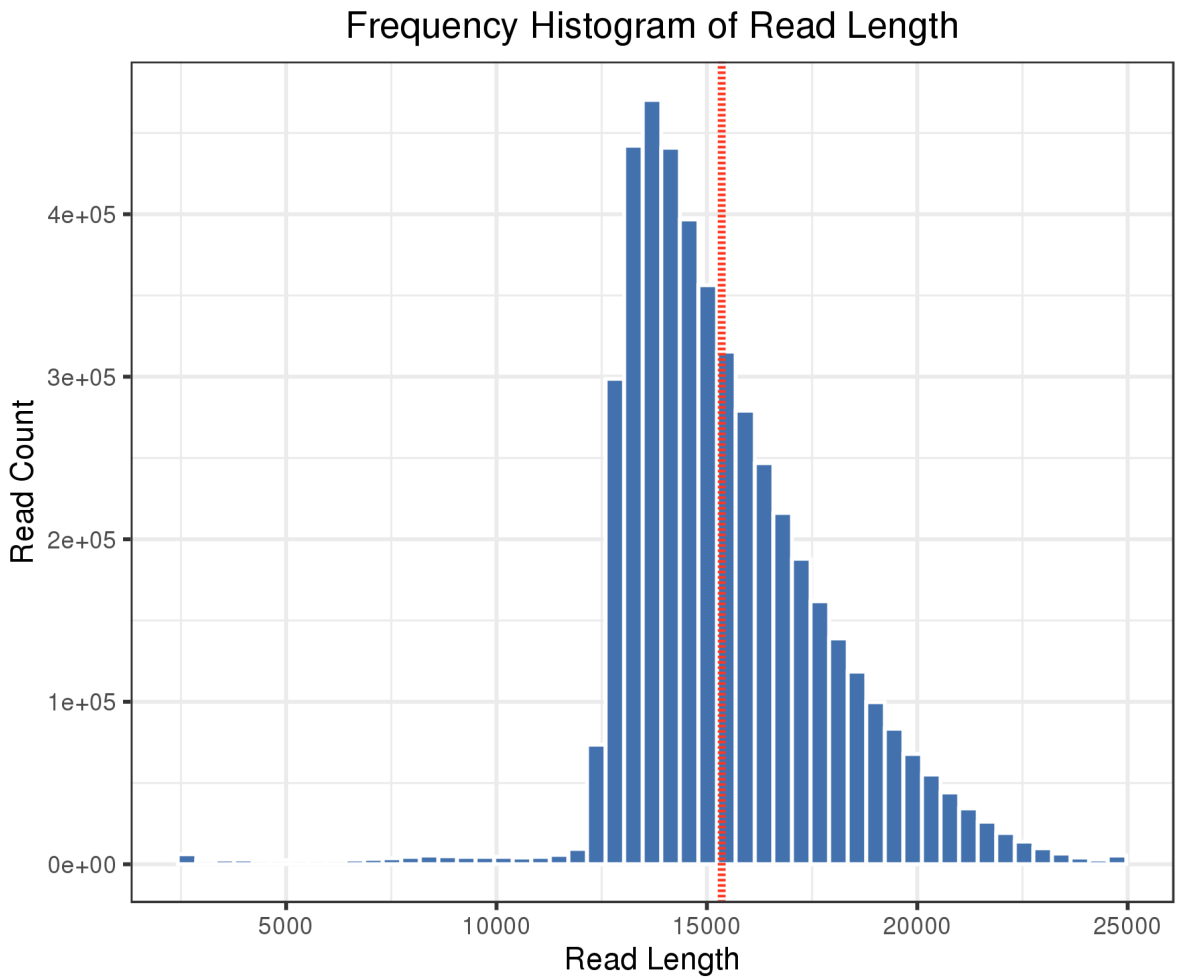

图7 文库r84071_230928_001的读长分布图

其他HiFi测序数据统计结果文件见目录：src/summary/1_data/HiFi/。

根据下机数据过滤及质控的统计分析，该项目的HiFi reads总数据量为89.21Gb，reads数为5,770,371条，reads平均长度为15.46Kb，其中最长reads长度为44.07Kb。相对其他测序平台，HiFi reads兼顾长读长和高准确性，是目前推荐用于基因组组装的最佳选择。

3.1.2 ONT数据质控

3.1.2.1 ONT数据质控方法

在ONT的测序平台中，将通过纳米孔的DNA或RNA链产生的电位信号转化为相应的碱基序列的过程，称为basecalling。使用官方提供的工具Guppy进行basecalling[4]，以mean_qscore_template的数值大于等于7为标准获得pass reads[5]，pass reads即可直接用于后续的组装。

3.1.2.2 ONT数据统计

对ONT下机数据进行统计，各文库及总数据量统计信息如下表所示:

表2 ONT下机数据统计

Library id	Total bases(nt)	Total reads	Mean length(nt)	Max length(nt)	N50 length(nt)	>10kb rate(%)	>20kb rate(%)	>40kb rate(%)
ONT-1	8,888,984,351	204,112	43,550	1,084,846	50,000	100	92.97	35.45
ONT-2	10,770,673,531	247,910	43,446	685,957	50,000	100	92.89	36.16

> 注：下机数据的质量会根据前期实验提取建库质量的差别而有所不同。表格各列说明见HiFi数据说明列表。

文库ONT-1的read长度分布如下图所示：

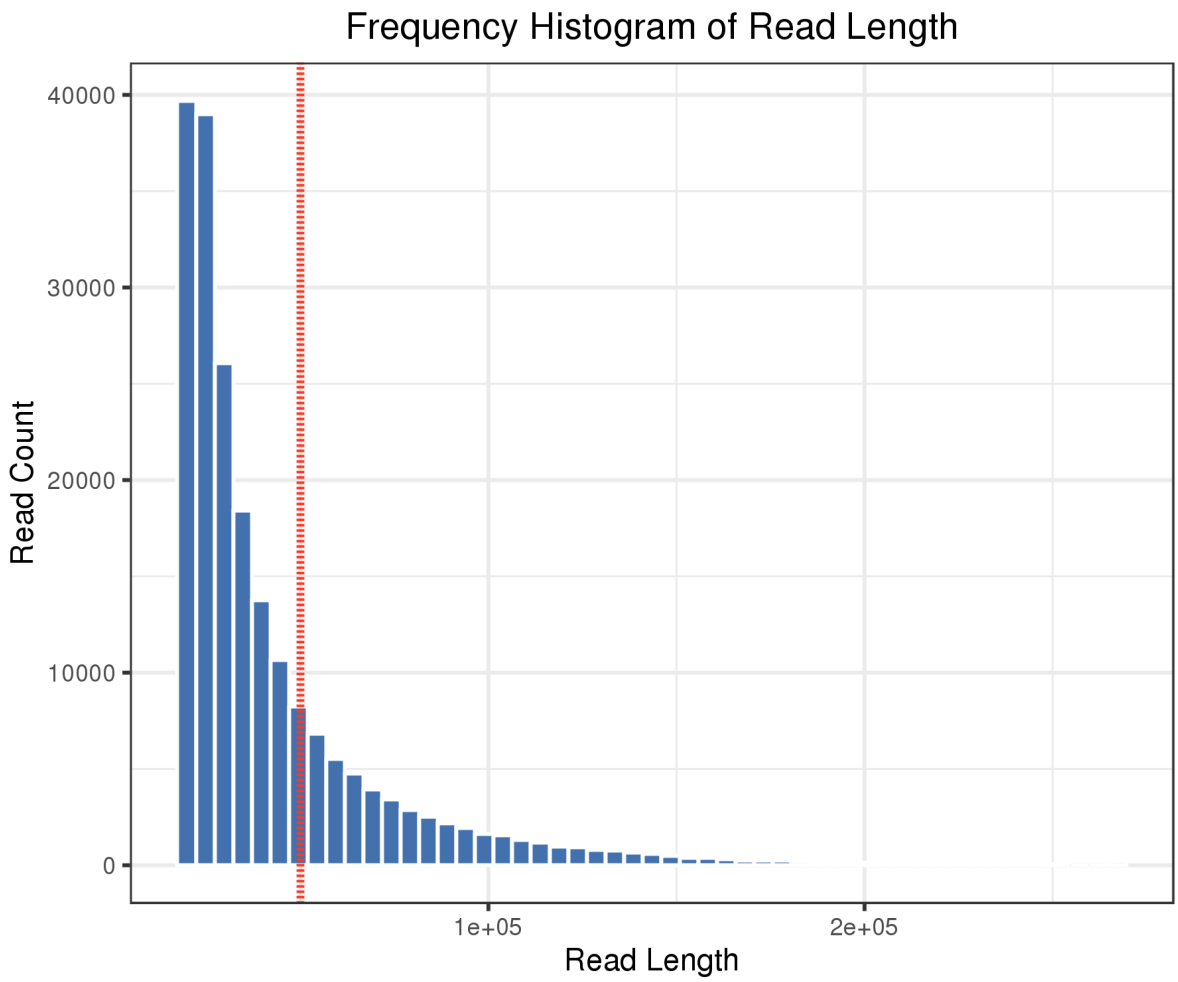

图8 文库ONT-1的读长分布图

其他ONT测序数据统计结果文件见目录：src/summary/1_data/ONT/。

根据ONT数据过滤及质控的统计分析，该项目的质控后ONT测序总数据量为19.66Gb，reads数为452,022条，下机reads平均长度为43.50Kb，其中最长reads长度为1084.85Kb。

3.1.3 HiC数据质控

3.1.3.1 HiC数据质控方法

利用fastp[6]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. 去除当中还含有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。

3.1.3.2 HiC数据统计

原始数据及质控后的高质量HiC数据统计结果见下表:

表3 HiC数据统计

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：表格各列说明如下表：

列名	说明
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。

其他HiC测序数据统计结果文件见目录：src/summary/1_data/HiC/。

根据下机数据过滤及质控的统计分析，该项目的HiC测序总数据量为118.88Gb，reads数为792,533,734条，质控后的总数据量为104.22Gb，reads数为696,691,224条。

3.2 长序列组装

质控过后，基于高质量的HiFi reads、ONT reads和HiC reads，使用hifiasm[7]软件对基因组进行组装。

hifiasm的分析流程如下，主要分为3个步骤：

1. 通过所有序列的相互比对，对潜在测序错误进行纠正。如果一个位置只存在两种碱基类型，且每个碱基类型至少有3条read支持，那么这个位置会被当作杂合位点，否则，视作测序错误，将被纠正；
2. 根据序列之间的重叠关系，构建分型的字符串图（phased string graph）。其中调整朝向的序列作为顶点（vertex），一致重叠作为边（edge）。字符串图中的气泡（bubble）则是杂合位点；
3. 如果没有额外的信息，hifiasm会随机选择气泡的一边构建primary assembly，另一边则是alternate assembly。该策略和HiCanu，Falcon-Unzip一样。对于杂合基因组而言，由于存在一个以上的纯合haplotype，因此primary assembly可能还会包含haplotigs。HiCanu依赖于第三方的purge_dups, 而hifiasm内部实现了purge_dups算法的变种，简化了流程。如果有额外的信息，那么hifiasm就可以正确的对单倍体进行分型。hifiasm组装流程图如图9所示：

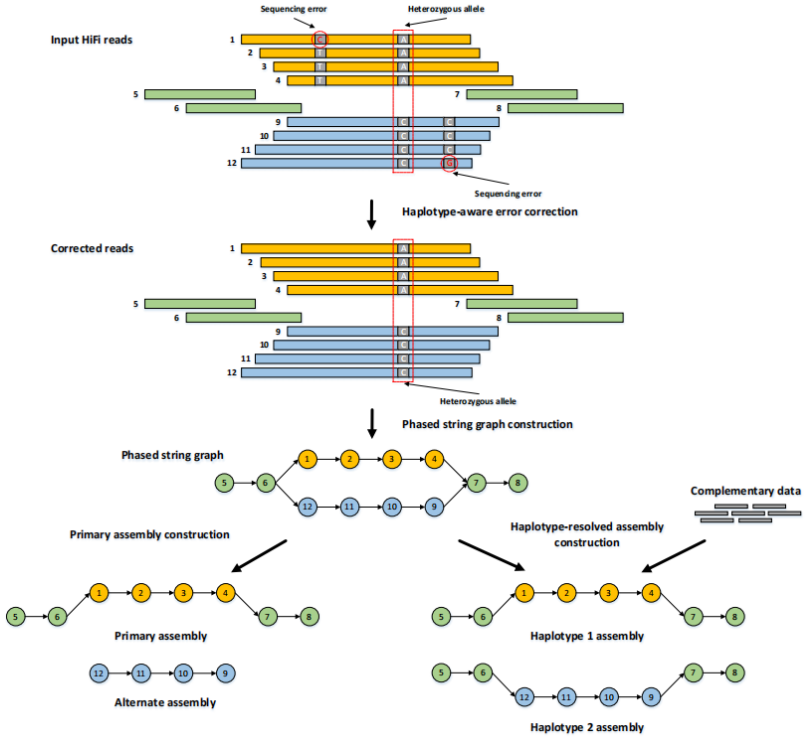

图9 hifiasm组装流程图

最终得到基因组大小为930,067,953bp，N50为101,985,394bp，详细的统计结果如下表：

表4 组装结果统计

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：表格第一列为Contig统计指标，第二列为Contig长度，第三列为Contig序号或数目。表格各行统计指标说明如下表：

行名	说明
Longest	基因组组装序列中最长的一条Contig的长度；

行名	说明
N50	按从长到短排序后累加，累加长度达到总长度的50%时，最后加上的一条序列的长度和序号；
N60	按从长到短排序后累加，累加长度达到总长度的60%时，最后加上的一条序列的长度和序号；
N70	按从长到短排序后累加，累加长度达到总长度的70%时，最后加上的一条序列的长度和序号；
N80	按从长到短排序后累加，累加长度达到总长度的80%时，最后加上的一条序列的长度和序号；
N90	按从长到短排序后累加，累加长度达到总长度的90%时，最后加上的一条序列的长度和序号；
Length>=50kb	序列长度大于等于50Kb的序列总长度和总数；
Length>=10kb	序列长度大于等于10Kb的序列总长度和总数；
Length>=5kb	序列长度大于等于5Kb的序列总长度和总数；
Length>=1kb	序列长度大于等于1Kb的序列总长度和总数；
Total	基因组组装所有序列的总长度和总数。

基于组装校正后的基因组进行序列长度排序绘图，发现组装基因组的连续性良好。具体序列长度分布信息如下图：

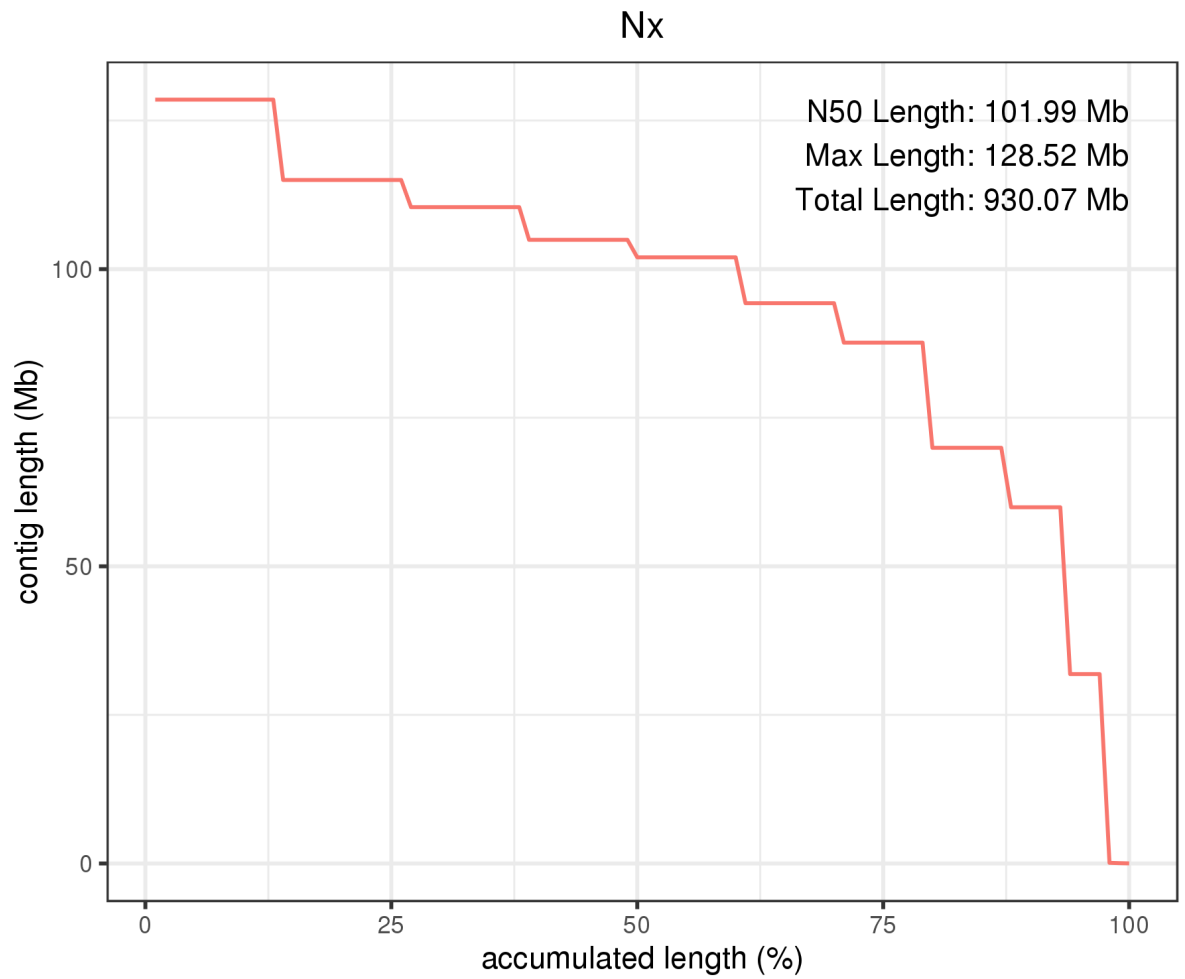

图10 Contigs长度累计图

基因组的碱基含量统计如下表所示：

表5 碱基含量结果统计

Content	Number(bp)	Rate(%)
A	284,092,130	30.55
T	284,294,788	30.57
C	180,814,042	19.44
G	180,866,993	19.45
N	0	0
GC	361,681,035	38.89
Total	930,067,953	100

注：表格第一列为碱基统计指标，第二列为碱基数目，第三列为碱基占比。表格各行统计指标说明如下表：

行名	说明
A	腺嘌呤（A）的数量及其占总长的比例；
T	胸腺嘧啶（T）的数量及其占总长的比例；
C	胞嘧啶（C）的数量及其占总长的比例；
G	鸟嘌呤（G）的数量及其占总长的比例；
N	未知碱基的数目及其占总长的比例；contig是无gap的序列，其未知碱基长度为0；
GC	胞嘧啶（C）与鸟嘌呤（G）的数量之和及其占总长的比例；
Total	组装序列的总长度（单位为bp）。

其他组装统计结果文件见目录：src/summary/2_assembly/。

3.3 基因组组装结果评估

3.3.1 BUSCO评估

根据OrthoDB数据库中进化分支eudicots的**通用单拷贝直系同源基因集**（Benchmarking Universal Single-Copy Orthologs，**BUSCOs**[8]）预测基因组现有序列的基因情况，进而评估组装基因组的完整性，详细评估结果如下表所示：

表6 BUSCO预测统计

Type	Number	Percent(%)
Complete BUSCOs (C)	2,270	97.59

Type	Number	Percent(%)
Complete and single-copy BUSCOs (S)	2,218	95.36
Complete and duplicated BUSCOs (D)	52	2.24
Fragmented BUSCOs (F)	11	0.47
Missing BUSCOs (M)	45	1.93
Total BUSCO groups searched	2,326	100

> 注：表格第一列为BUSCO统计指标，第二列为BUSCO数目，第三列为BUSCO占比。表格各行统计指标说明如下表：

行名	说明
Complete BUSCOs (C)	序列完全比对上BUSCO；
Complete and single-copy BUSCOs (S)	一个BUSCO比对上一个基因；
Complete and duplicated BUSCOs (D)	一个BUSCO比对上多个基因；
Fragmented BUSCOs (F)	部分序列比对上BUSCO；
Missing BUSCOs (M)	未比对上BUSCO；
Total BUSCO groups searched	总BUSCO集。

基于BUSCO预估结果进行绘图得下图：

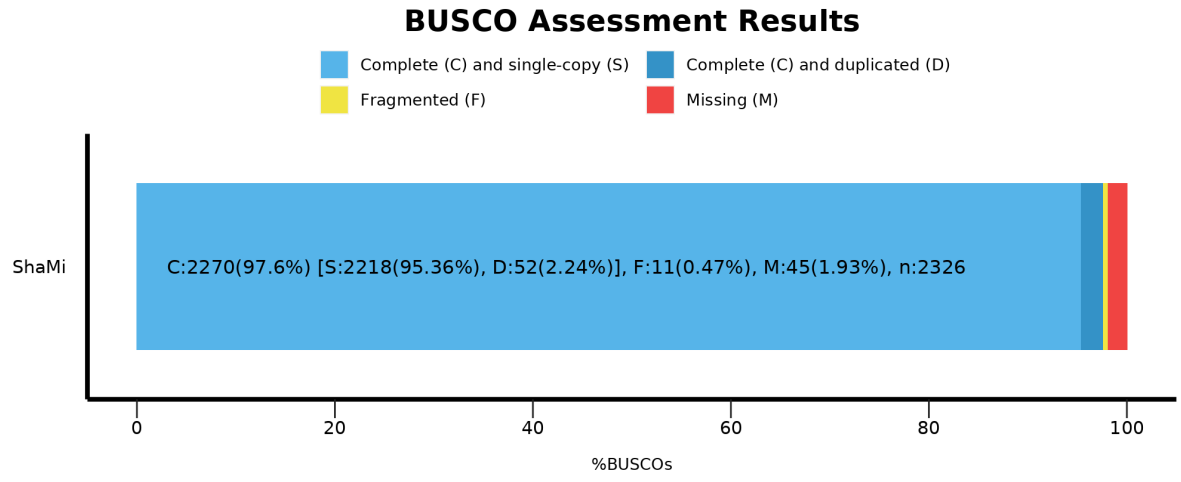

图11 BUSCO预测统计图

该基因组中可以找到约97.59%的完整基因元件，说明绝大部分保守基因组装得比较完整，从侧面反映组装结果可信度较高。

3.3.2 GC-Depth分析

利用 minimap2[9] 将HiFi测序reads回比到基因组，统计下机reads比对率、深度、GC含量等信息。

基于HiFi数据统计结果显示HiFi数据的比对率为99.79%。具体结果见下表：

表7 HiFi reads比对结果统计

Total Reads	Mapped Reads	Mapped Rate(%)
5,770,371	5,758,158	99.79

> 注：此表格为samtools基于bam文件统计得到，输出虽为reads数但实际统计为 alignments数，当read存在多次比对时统计得到的Total Reads会比真实数据多，此表格统计可反应真实比对情况，为得到准确的比对率可用`1-（Total Reads-Map Reads）/ Reads num`计算。表格各列说明如下：

列名	说明
Total Reads	Reads总数；
Mapped Reads	比对上的Reads数目；
Mapped Rate(%)	与HiFi数据的比对率。

基于HiFi数据回比到基因组深度统计结果显示HiFi数据的平均深度为95.62X。统计覆盖深度为1X时，整个基因组覆盖度为99.8171%。具体结果见表8和表9。

表8 基因组覆盖度统计

Depth(X)	Base number	Coverage ratio(%)
1	928,367,110	99.8171
5	920,358,235	98.956
10	918,770,357	98.7853
20	917,382,066	98.636

> 注：表格各列说明如下：

列名	说明
Depth(X)	基因组覆盖深度；
Base number	碱基数；
Coverage ratio(%)	覆盖度；

表9 基因组覆盖度深度统计

Assemly Bases	Coverage Bases	Coverage depth(X)
930,067,953	88,929,451,842	95.62

> 注：表格各列说明如下：

列名	说明
Assemly Bases	基因组碱基总数；

列名	说明
Coverage Bases	比对上的HiFi数据碱基数；
Coverage depth(X)	HiFi数据基因组评估覆盖深度。

基于HiFi数据回比基因组结果，统计不同bin的平均GC及平均深度信息，进行二维散点图绘制，结果显示如图12。若存在分离聚团现象，说明基因组可能存在外源污染。

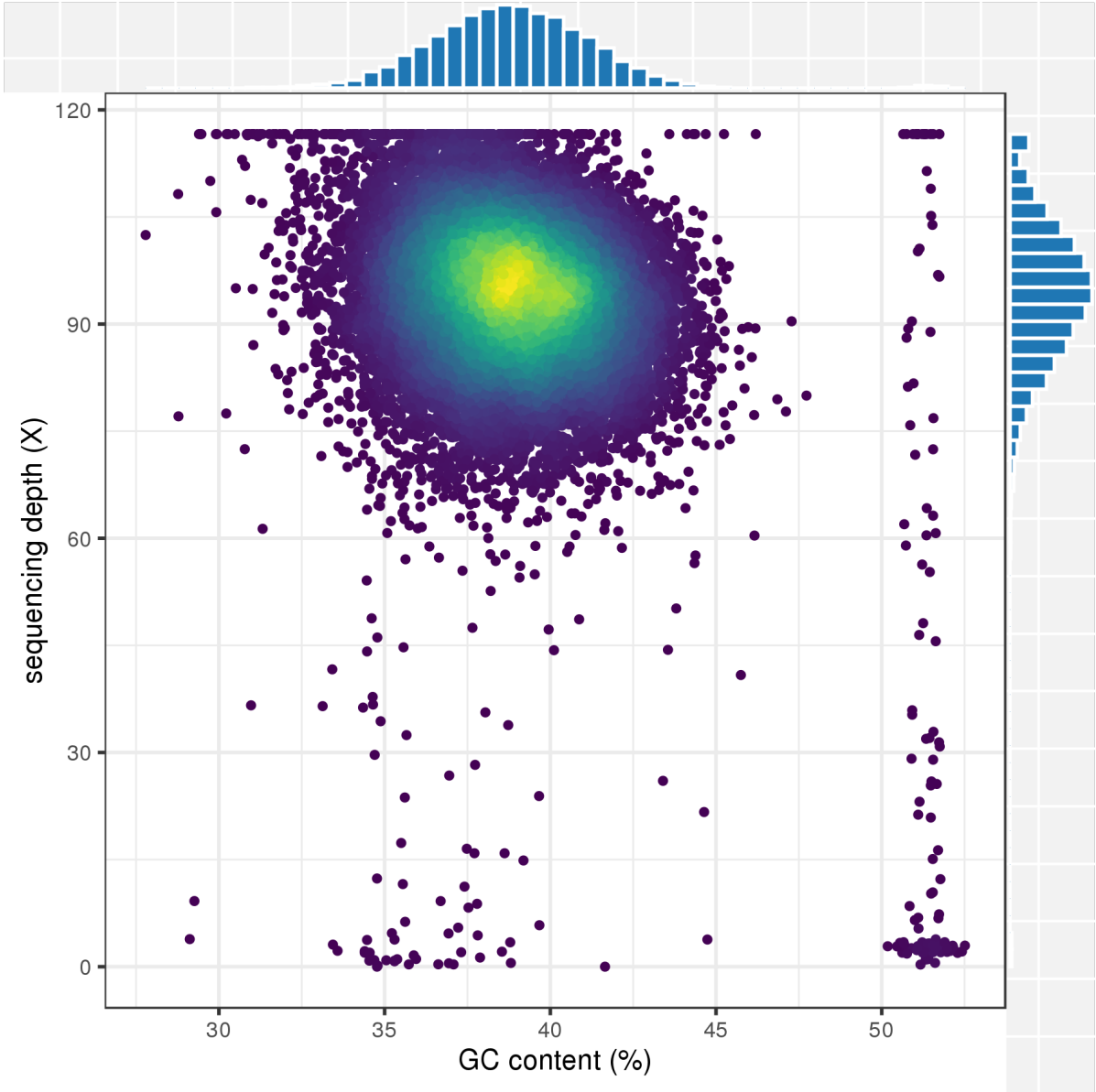

图12 GC Depth分布

> 注：上图为基因组GC Depth分布，横坐标为GC含量，纵坐标为Depth，此二值是以50Kb窗口依次统计。

其他评估结果文件见目录：src/summary/2_assembly/HiFi/。

4 分析软件

表10 分析所用软件信息表

分析模块	分析内容	工具名称	版本
数据质控	HiFi数据评估	SMRTLink	v10.1.0

分析模块	分析内容	工具名称	版本
数据质控	HiC数据评估	fastp	v0.22.0
基因组组装	基因组组装	hifiasm	v0.25.0
基因组评估	完整性评估	BUSCO	v5.0.0
基因组评估	完整性评估	BLAST	v2.12.0
基因组评估	完整性评估	HMMER3	v3.3.2
基因组评估	完整性评估	Augustus	v3.4.0
基因组评估	HiFi数据比对	minimap2	v2.17
基因组评估	比对文件处理	samtools	v1.13
基因组评估	比对文件处理	sambamba	v0.8.2
基因组评估	一致性评估	bcftools	v1.13
基因组评估	GC-Depth分析	mosdepth	v0.3.3

5 分析方法

5.1 数据质控

在PacBio的测序平台中，将通过零模波导孔的DNA产生的荧光信号记录成movie进而转化为相应的碱基序列的过程，称为basecalling。使用官方提供的工具SMRTLink进行basecalling获得含接头的测序序列，即酶读（polymerase reads），其长度由反应酶的活性和上机时间决定。酶读去除低质量序列和接头序列后得到subreads。环化共有序列（Circular Consensus Sequencing，CCS）测序模式获得subreads，通过校准同一序列模板多次测序的subreads的随机错误，可将测序准确率提升至99%以上，通过Q20阈值过滤获得高准确性的HiFi reads。

碱基识别qscore与Phred分值之间的计算公式为：$Q\text{-}score=-10\times\log\ 10\ P$

碱基识别qscore与Phred分值之间的简明对应关系见下表：

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	99.9%	Q30
40	1/10000	99.99%	Q40

5.2 基因组组装

由于测序的随机性及技术局限，测序获得的序列其长度通常会根据使用的测序平台和方法局限在某一范围内，特别是对大型基因组，单条测序所得的序列只占整个基因组十分微小的部分。因此，为了构建完整的物种基因组，需要使用软件对零散的'短'测序序列进行拼接，使其按照一定算法进行排列和延伸，得到较'长'的一致性序列。一般来说，基因组装软件主要有三大类，分别基于不同的计算策略：DBG（de Bruijn Graph）、OLC（Overlap Layout Consensus）[10]、string graph[11]。

基于此样本质控后的测序数据，使用hifiasm软件对基因组进行组装。hifiasm分析流程主要分为3个步骤：（1）通过所有序列的相互比对，对前在测序错误进行纠正。如果一个位置只存在两种碱基类型，且每个碱基类型至少有3条read支持，那么这个位置会被当作杂合位点，否则，视作测序错误，将被纠正；（2）根据序列之间的重叠关系，构建分型的字符串图（phased string graph）。其中调整朝向的序列作为顶点（vertex），一致重叠作为边（edge）。字符串图中的气泡（bubble）则是杂合位点；（3）如果没有额外的信息，hifiasm会随机选择气泡的一边构建primary assembly，另一边则是alternate assembly. 该策略和HiCanu，Falcon-Unzip一样。对于杂合基因组而言，由于存在一个以上的纯合haplotype，因此primary assembly可能还会包含haplotigs。HiCanu依赖于第三方的purge_dups, 而hifiasm内部实现了purge_dups算法的变种，简化了流程。如果有额外的信息，那么hifiasm就可以正确的对单倍体进行分型。

5.3 基因组组装结果评估

5.3.1 BUSCO评估

BUSCO评估是指通过鉴定物种进化过程中保守单拷贝基因来评估组装结果完整性。首先，构建好进化中各分支BUSCO数据库（例如真核生物数据库，真菌数据库，鸟类数据库等等）。其中，运行BUSCO需要blast[12]、HMMER3[13]和Augustus[14]等软件的支持。其次，对于基因组组装评估，利用tBLASTn与对应的BUSCO数据库进行比对从而确定其候选区域，然后使用 Augustus 软件进行基因结构预测，并使用 HMMER3 进行比对，从而评估其完整性。如果匹配长度在BUSCO配置文件匹配长度的预期范围内，则将其归类为“完成”。如果不止一次发现它们，则它们被归类为“重复的”。仅部分匹配被分类为“碎片”，没有匹配上的被分类为“缺失”。

利用BUSCO通过BUSCO单拷贝基因数据库基于默认参数对基因组进行评估。

5.3.2 GC-Depth分析

核苷酸序列中鸟嘌呤(G)和胞嘧啶(C)所占的比例称为GC含量，GC含量在物种间存在一定特异性，不同的基因组具有不同的GC含量，通过对组装的基因组进行GC depth分析可以用于评估基因组是否存在外源污染。首先使用minimap2将组装的基因组和HiFi测序数据进行比对得到比对的bam文件，然后使用samtools基于比对结果计算组装基因组每个位点的测序深度，然后以50Kb为滑动窗口分别统计每条contigs序列各个区段的平均GC含量和平均测序深度，最后以平均GC含量为横坐标，平均测序深度为纵坐标绘制GC-depth散点图，及平均GC含量和平均测序深度的分布图。GC-depth图中点的密度越高，点的颜色越深，表明基因组序列的平均GC含量和平均测序深度集中富集在此区域。由于物种间的GC含量存在一定特异性，因此若GC depth集中富集在同一区域则说明基因组不存在污染，反之可能存在外源污染。

6 参考文献

1. Logsdon GA, Vollger MR, Eichler EE. **Long-read human genome sequencing and its applications**. *Nat Rev Genet*. 2020 Oct;21(10):597-614. doi: 10.1038/s41576-020-0236-x. Epub 2020 Jun 5. PMID: 32504078; PMCID: PMC7877196.
2. Du H, Yu Y, .etc. **Sequencing and de novo assembly of a near complete indica rice genome**. *Nat Commun*. 2017 May 4;8:15324. doi: 10.1038/ncomms15324. PMID: 28469237; PMCID: PMC5418594.
3. Rhoads A, Au KF. **PacBio Sequencing and Its Applications**. *Genomics Proteomics Bioinformatics*. 2015 Oct;13(5):278-89. doi: 10.1016/j.gpb.2015.08.002. Epub 2015 Nov 2. PMID: 26542840; PMCID: PMC4678779.
4. Wick RR, Judd LM, Holt KE. **Performance of neural network basecalling tools for Oxford Nanopore sequencing**. *Genome Biol*. 2019 Jun 24;20(1):129. doi: 10.1186/s13059-019-1727-y. PMID: 31234903; PMCID: PMC6591954.

5. Senol Cali D, Kim JS, Ghose S, Alkan C, Mutlu O. **Nanopore sequencing technology and tools for genome assembly: computational analysis of the current state, bottlenecks and future directions**. *Brief Bioinform*. 2019 Jul 19;20(4):1542-1559. doi: 10.1093/bib/bby017. PMID: 29617724; PMCID: PMC6781587.
6. 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.
7. Cheng, H., Concepcion, G.T., Feng, X., Zhang, H., Li H. (2021) **Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm**. *Nat Methods*, 18:170-175. https://doi.org/10.1038/s41592-020-01056-5
8. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. **BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs**. *Bioinformatics*. 2015 Oct 1;31(19):3210-2. doi: 10.1093/bioinformatics/btv351.
9. Li,H. (2018). **Minimap2: pairwise alignment for nucleotide sequences**. *Bioinformatics*, 34:3094-3100. doi:10.1093/bioinformatics/bty191
10. Li Z, Chen Y, Mu D, Yuan J, Shi Y, Zhang H, Gan J, Li N, Hu X, Liu B, Yang B, Fan W. **Comparison of the two major classes of assembly algorithms: overlap-layout-consensus and de-bruijn-graph**. *Brief Funct Genomics*. 2012 Jan;11(1):25-37. doi: 10.1093/bfgp/elr035. Epub 2011 Dec 19. PMID: 22184334.
11. Myers EW. **The fragment assembly string graph**. *Bioinformatics*. 2005 Sep 1;21 Suppl 2:ii79-85. doi: 10.1093/bioinformatics/bti1114. PMID: 16204131.
12. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. **Basic local alignment search tool**. *J Mol Biol*. 1990;215(3):403-410. doi:10.1016/S0022-2836(05)80360-2
13. Eddy SR. **Accelerated Profile HMM Searches**. *PLoS Comput Biol*. 2011;7(10):e1002195. doi:10.1371/journal.pcbi.1002195
14. Stanke M, Diekhans M, Baertsch R, Haussler D. **Using native and syntenically mapped cDNA alignments to improve de novo gene finding**. *Bioinformatics*. 2008;24(5):637-644. doi:10.1093/bioinformatics/btn013

7 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Sequell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


浩瑞基因