

**基因组survey分析报告**

撰写人：{{author}}

审核人：{{reviewer}}

{{year}}年{{month}}月{{day}}日

武汉未来组生物科技有限公司

**目 录**

[一 项目基本信息 1](#_Toc16779157)

[二 项目流程 2](#_Toc16779158)

[2.1 实验流程 2](#_Toc16779159)

[2.2 生物信息分析流程 3](#_Toc16779160)

[三 分析结果 4](#_Toc16779161)

[3.1 数据质控 4](#_Toc16779162)

[3.2 数据污染评估 5](#_Toc16779163)

[3.3 基因组大小和杂合度 6](#_Toc16779164)

[3.4 复杂基因组大小和杂合度 7](#_Toc16779165)

[3.5 基因组组装结果 8](#_Toc16779166)

[3.6 GC-Depth分析及污染评估结果 9](#_Toc16779167)

[四 分析结论及建议 10](#_Toc16779168)

[五 分析方法 11](#_Toc16779169)

[5.1 分析软件 11](#_Toc16779170)

[5.2 二代数据介绍及质控方法 11](#_Toc16779171)

[5.3 *K*-mer分析方法 13](#_Toc16779172)

[5.4 基因组组装 14](#_Toc16779173)

[5.5 GC-Dpeth分析 15](#_Toc16779174)

[六 参考文献 16](#_Toc16779175)

[附件一 相关技术介绍 17](#_Toc16779176)

[附件二 联系我们 18](#_Toc16779177)

# 项目基本信息

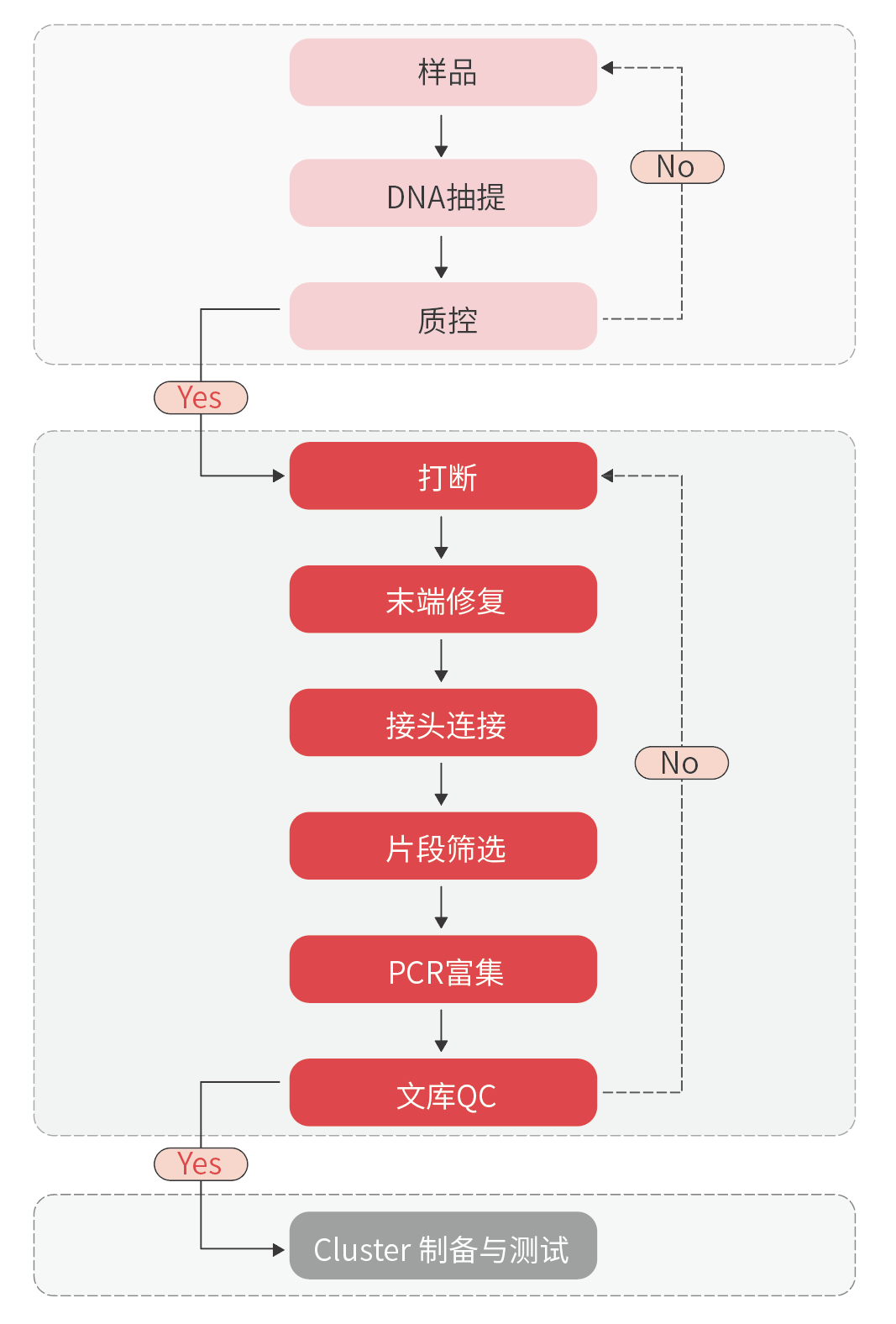
|  |  |
| --- | --- |
| 样本编号 | **{{name}}** |
| 物种名 | ***{{species}}* {{strain}}** |
| 测序平台 | **{{sequencer}}** |
| 测序数据量 | **{{raw\_data }} bp** |
| 预估基因组大小 | **{{estimate\_size}} bp** |
| 预估基因组杂合度 | **{{heterozygosity}}** |

备注：基因组大小和杂合度是依据软件{{software}}评估所得，仅作为参考用于初步判断基因组复杂程度。

# 项目流程

## 实验流程

利用给定的样品进行DNA抽提，随后对抽提DNA进行质量检测评估。对质量检测合格的DNA样品进行高通量测序（High Throughput Sequencing, HTS）的文库制备和上机测序，最终获得相应核苷酸序列数据。对于不合格的DNA样品，如果有样品剩余则重新进行DNA抽提，没有样品剩余则需要合作方重新安排样品寄送。实验流程参见图2.1。



##### 基因组survey测序流程

## 生物信息分析流程

第一：原始下机序列文件（raw data）的质量控制，过滤低质量序列，质控包括：

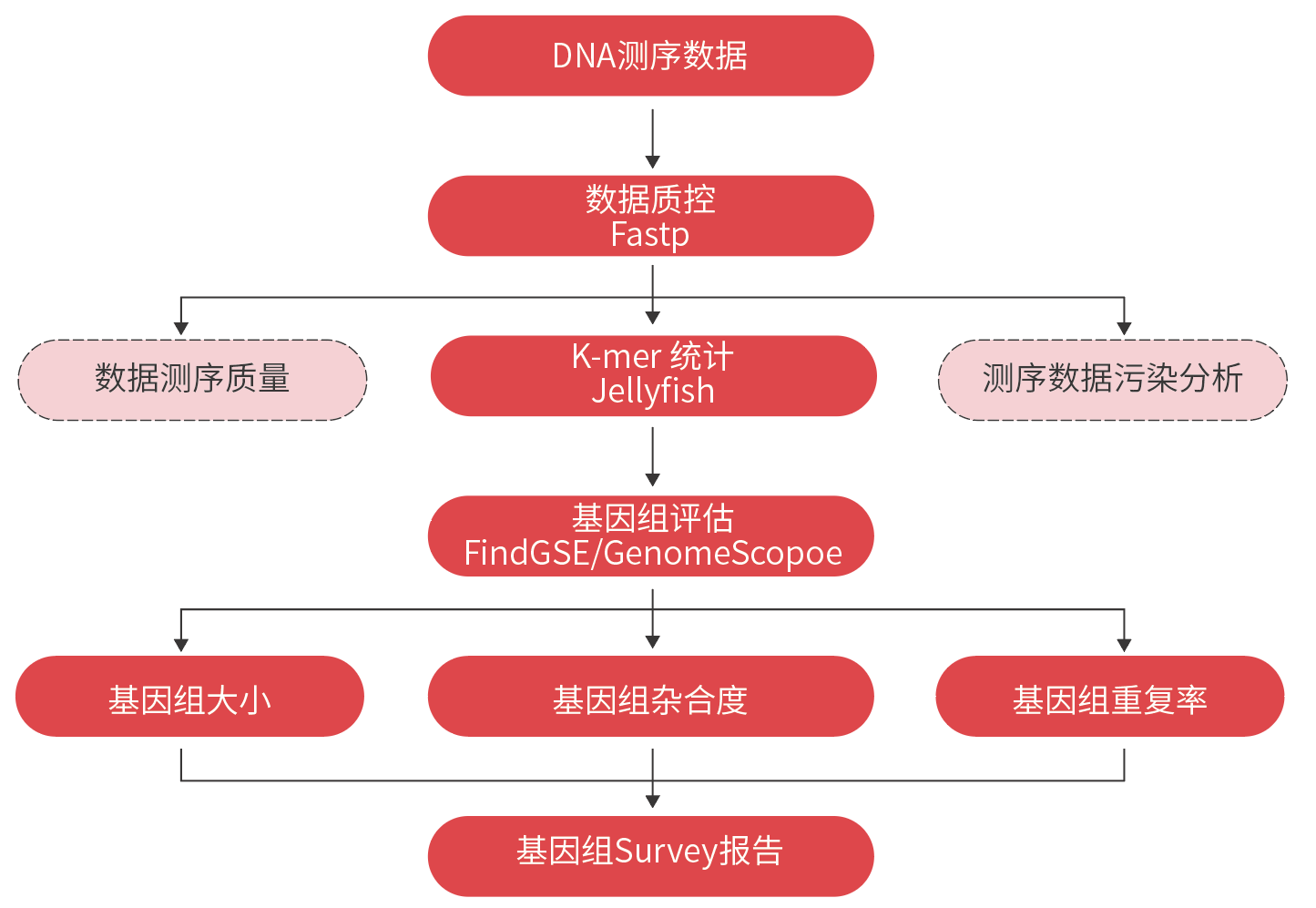
1) 去掉reads的接头序列；

2) 截掉reads两端测序不准确的碱基，直接截取左端5个碱基，右端5个碱基；

3) 去除当中还含有N的reads；

4) 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。

第二：基于质控后序列文件（clean data）进行包括基因组大小和基因组杂合度的评估。首先，根据合作方提供的基因组大小预估值（基于流式细胞仪评估或近源物种的先验知识）随机获取30-50X的序列数据，对于不足30X的数据集则利用全部数据信息进行后续分析；然后，通过jellyfish [1] 对序列文件进行K-mer的计数和统计；随后，分别利用偏正态分布模型（skew normal distribution model），负二项式模型 (negative binomial model) 对K-mer 数据进行拟合分析，通过两种算法对应的软件FindGSE[2] 与 GenomeScope[3] 进行基因组大小以及杂合度的评估，并生成最终基因组评估结果。由于FindGSE与GenomeScope适用于二倍体简单基因组，因此当基因组较为复杂（高杂合、多倍体等）时，利用Kmerfreq[4]进行评估，并利用拟南芥基因组模拟对应深度的短片段数据，在杂合率不同梯度组合情况下进行K-mer曲线拟合来估计样本的杂合率。具体分析流程见图2.2。



##### 基因组survey流程

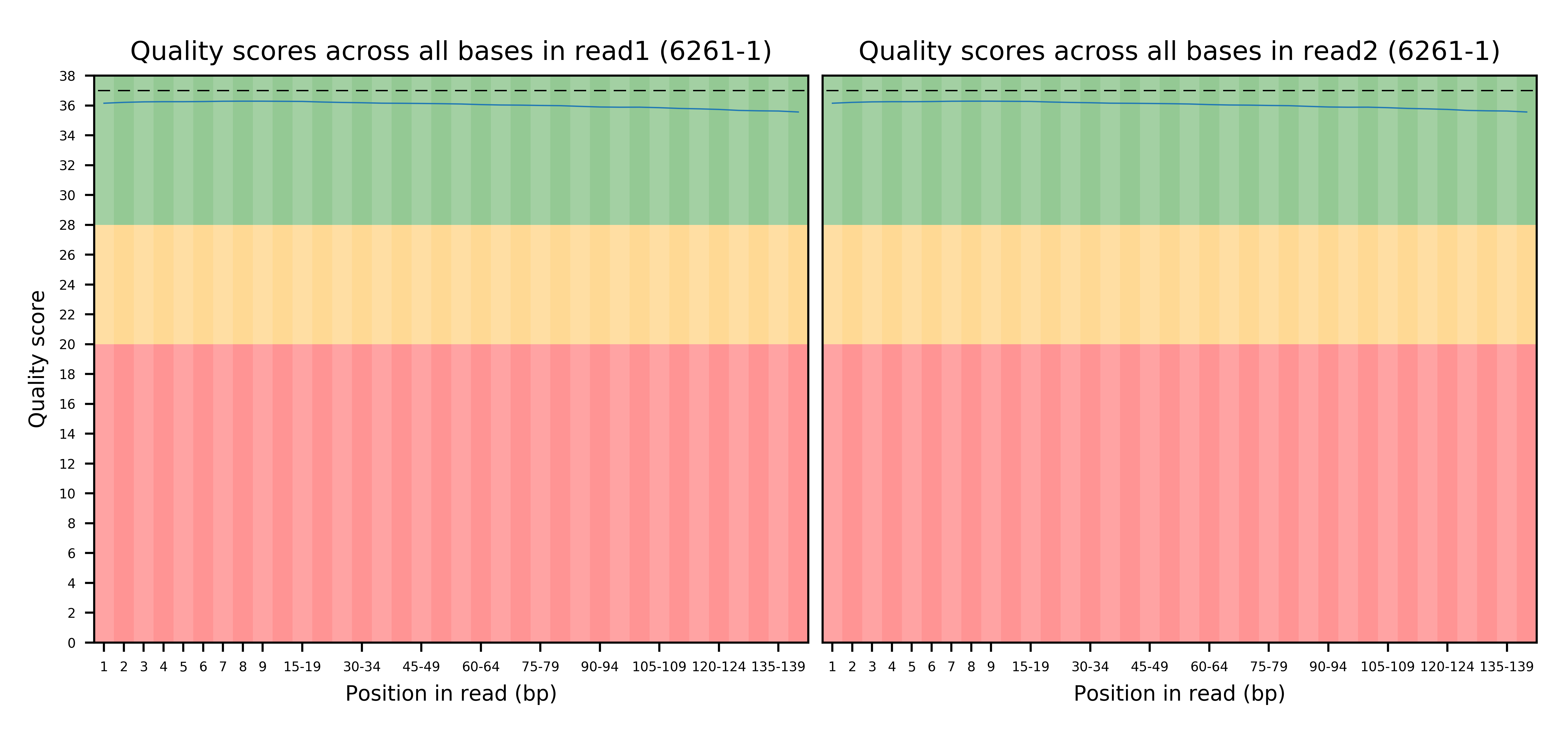
# 分析结果

## 数据质控

利用fastp对原始数据进行过滤，然后对原始数据（Raw data）及过滤后数据（Clean data）进行质量统计。具体结果见表3.1，及图3.1、图3.2和图3.3。

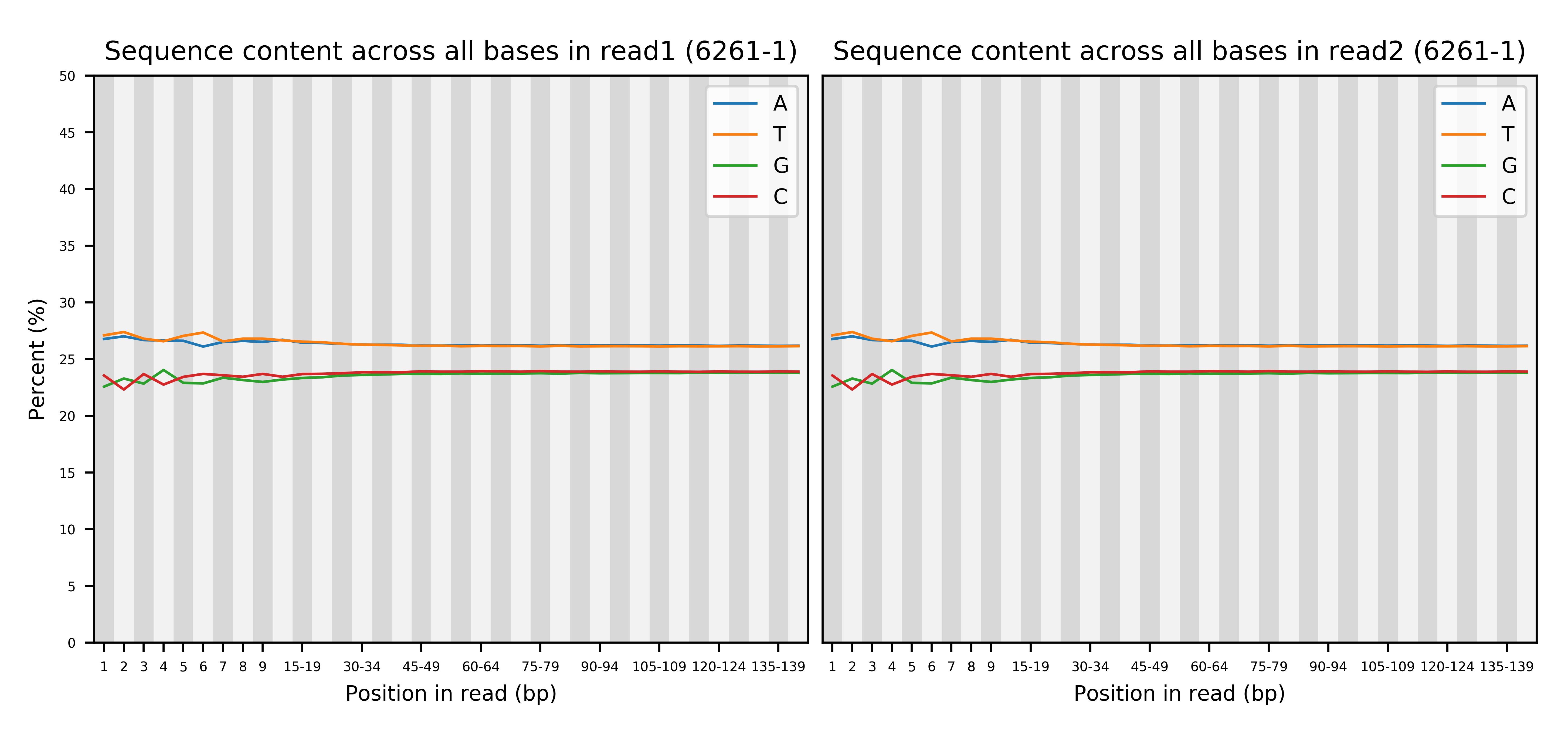
###### 数据产量统计

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **sample** | **total reads** | **total bases** | **clean reads** | **clean bases** | **Q20 rate (%)** | **Q30 rate (%)** | **GC (%)** |
| {{table\_data.0}} | {{table\_data.1}} | {{table\_data.2}} | {{table\_data.3}} | {{table\_data.4}} | {{table\_data.5}} | {{table\_data.6}} | {{table\_data.7}} |



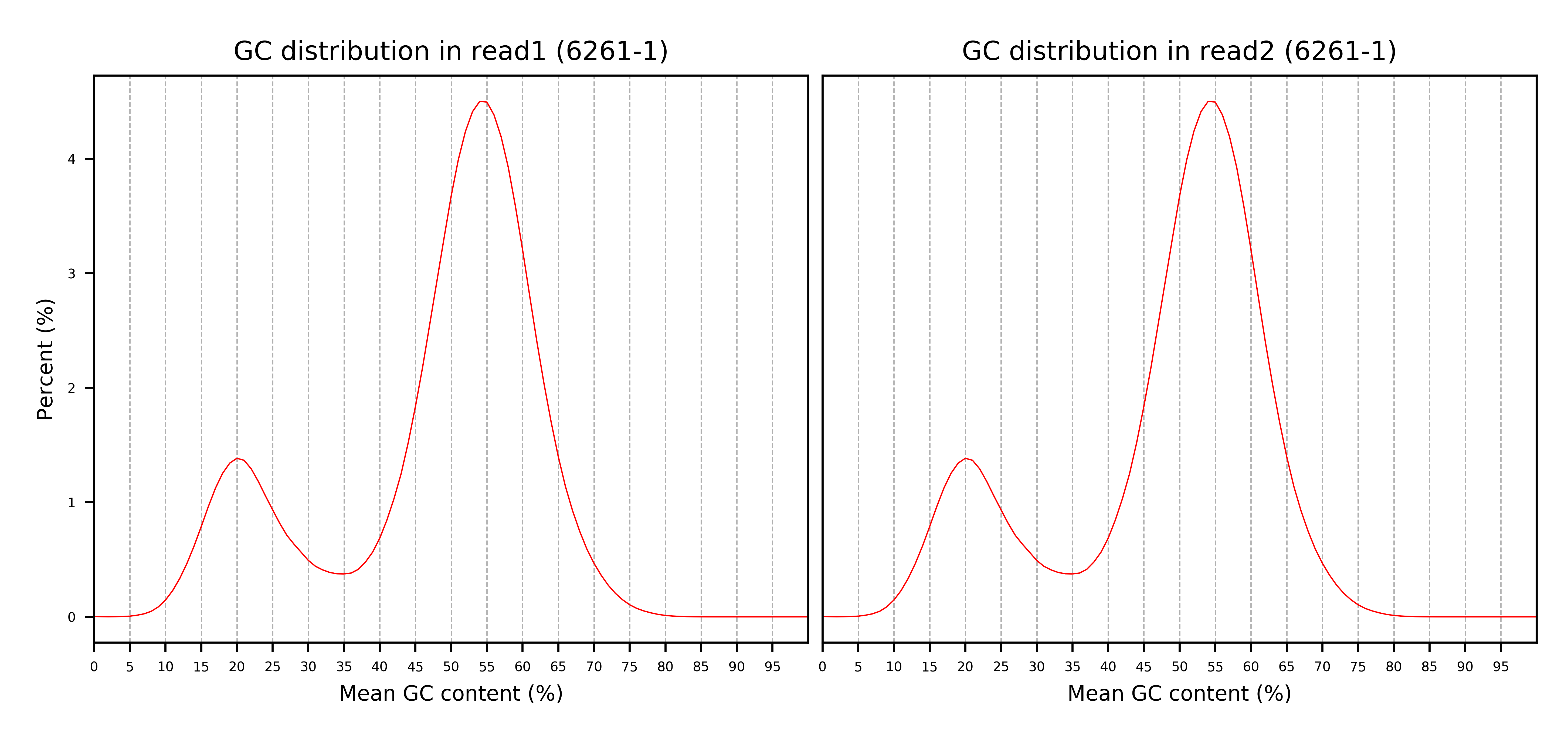
##### 碱基质量分布图

**说明：**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，其中红色线条为质量值中位数，蓝色线条是质量值平均值，箱线图的上下线分别是质量值的最大值和最小值。



##### 碱基含量分布图

**说明：**上图为碱基含量分布图：其中，蓝色线条，橙色线条，绿色线条，红色线条分别是A,T,G,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。

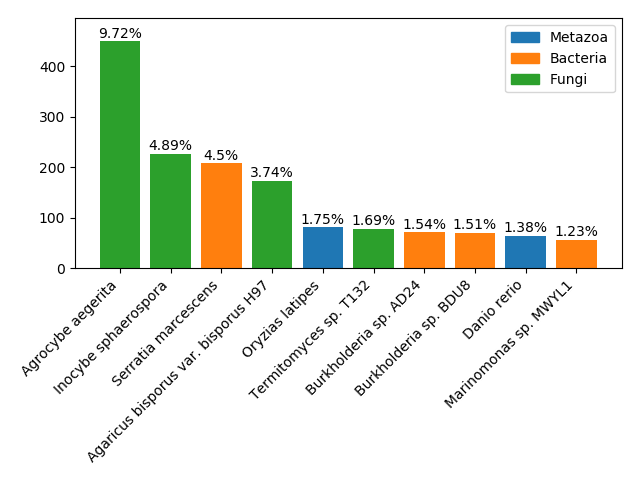


##### 平均GC分布图

**说明**：上图为平均GC含量分布图；横轴表示GC含量，纵轴表示不同GC含量对应的read的占比，曲线形状接近正态，出现偏差往往是由于文库的污染、部分reads构成的子集有偏差或物种特异性。

## 数据污染评估

随机取100,000条质控后的reads，利用blastn将其与nt数据进行比对，统计reads在nt库中的分布情况及比对上的物种分布，结果信息见表3.2及图3.4。



##### 主要物种分布图

###### 物种分布统计

|  |  |  |
| --- | --- | --- |
| **Species** | **Reads Number** | **Taxonomy** |
| **{%tr for table in table\_species %}** | | |
| **{{table.0}}** | {{table.1}} | {{table.2}} |
| **{%tr endfor %}** | | |

注意：这里只展示部分结果，Unmap表示没有比对上NT库的reads数目，其他的为物种名字，按照比对上的reads数目从高到低依次往下排列。

{% if homogeneous %}

## 基因组大小和杂合度

基于K-mer分析的结果，利用FindGSE预估基因组大小为{{findgse\_genome}} bp (图3.5a)，而利用GenomeScope预估基因组大小为{{scope\_genome}} bp （图3.5b）。

两种分析方法得到的结果相近（差异≤20%），因此推断目标物种基因组大小在 {{average\_genome}} bp 左右（{{min\_genome}} bp - {{max\_genome}} bp）。表3.3中总结了软件分析结果中目标物种的基因组大小、杂合度和重复度。

|  |  |
| --- | --- |
|  |  |
| a | b |

##### 样本{{name}} K-mer分布曲线

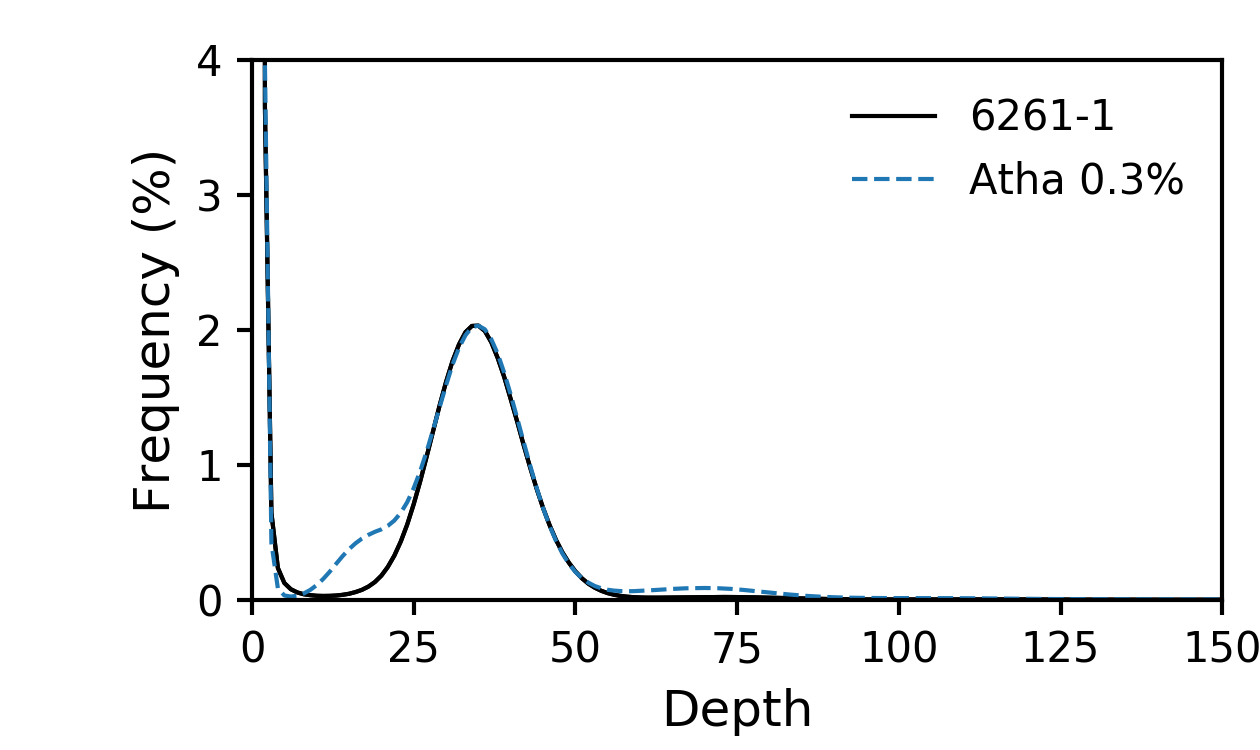
###### 样本{{name}}基因组预估结果

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Software** | **K-mer** | **Genome\_size (bp)** | **Heterozygous ratio (%)** | **Repeat ratio (%)** |
| **FindGSE** | {{kmer}} | {{table\_findgse.0}} | {{table\_findgse.1}} | {{table\_findgse.2}} |
| **GenomeScope** | {{kmer}} | {{table\_scope.0}} | {{table\_scope.1}} | {{table\_scope.2}} |

{% else %}

## 复杂基因组大小和杂合度

样本基因组较为复杂（可能为高杂合或多倍体等），需利用Kmerfreq[4]进行评估，并利用拟南芥基因组模拟对应深度的短片段数据，在杂合率不同梯度组合情况下进行K-mer曲线拟合来估计样本{{name}} 的杂合率。结果（图3.5和表3.3）显示基因组大小为{{table\_kmer.4}} bp ，杂合度为{{table\_kmer.5}}。



##### 样本K-mer分布曲线及杂合率模拟曲线

**注**：横轴为 K-mer 深度，纵轴为 K-mer 深度频率， 如AthaH0.010\_X31表示（Atha）拟南芥深度杂合度(H) 1.0%，（X）为深度31，其他依次类推。

###### 样本{{name}}的K-mer数据统计

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Name** | **K-mer** | **K-mer num** | **K-mer depth** | **Genome size (bp)** | **Heterozygous ratio (%)** |
| {{table\_kmer.0}} | {{table\_kmer.1}} | {{table\_kmer.2}} | {{table\_kmer.3}} | {{table\_kmer.4}} | {{table\_kmer.5}} |

{% endif %}

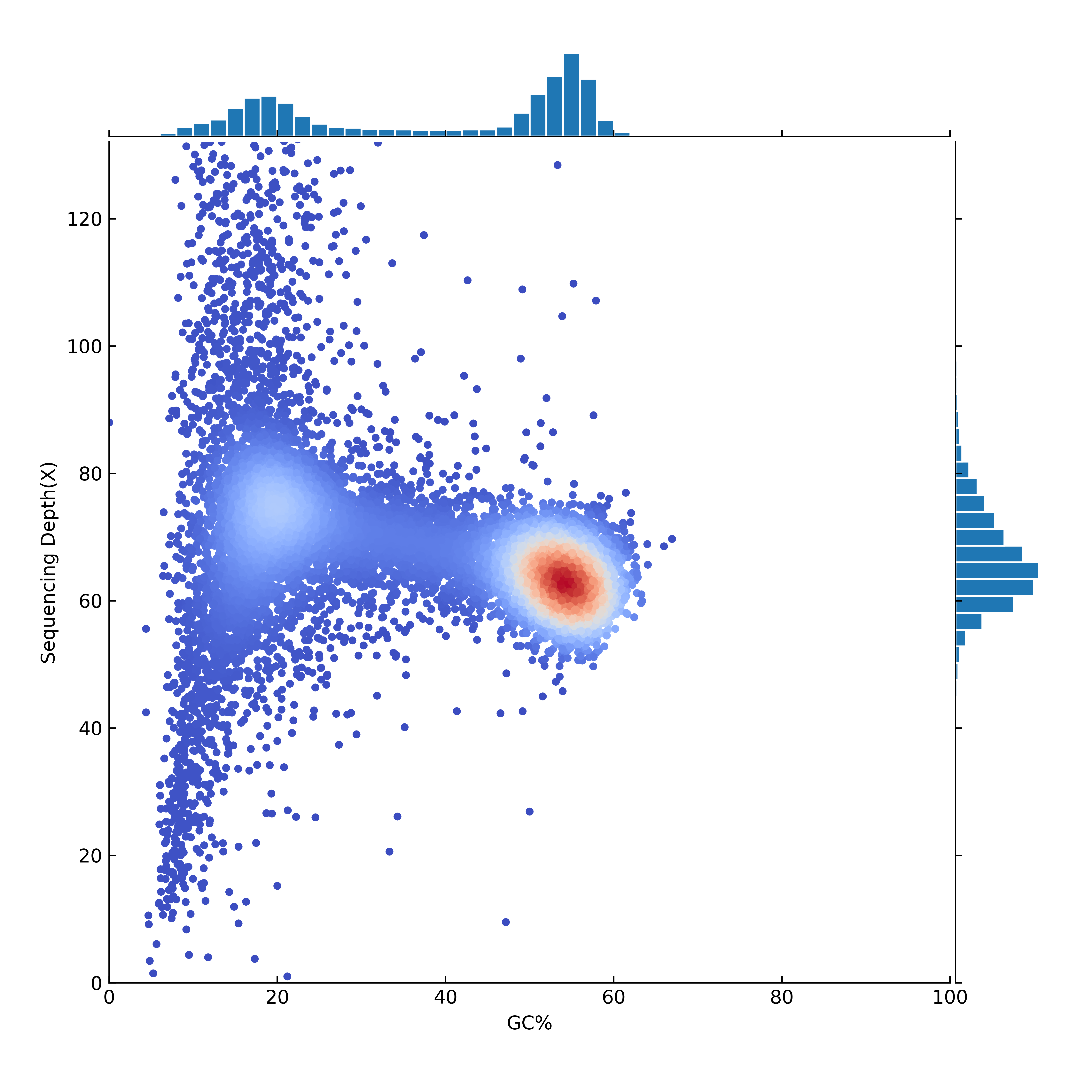
## 基因组组装结果

利用样本{{name}}过滤后的Pair-end reads进行基因组组装，最终组装后的基因组{{assembly\_size}} bp，详细组装统计结果见组装结果统计（表3.4）。

###### 样本{{name}}组装结果统计表

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| StatType | ContigLength | ContigNumber | ScaffoldLength | ScaffoldNumber | GapLength | GapNumber |
| {%tr for table in table\_assembly %} | | | | | | |
| {{table.0}} | {{table.1}} | {{table.2}} | {{table.3}} | {{table.4}} | {{table.5}} | {{table.6}} |
| {%tr endfor %} | | | | | | |

## GC-Depth分析及污染评估结果



##### 样本{{name}}的GC深度分布图

由GC深度图（图3.6）结果可以看出，{{depth\_description}}。

# 分析结论及建议

1. 原始数据{{raw\_data}} bp （见表3.1）；经过质控，去除不成对匹配的reads以及低质量reads、接头污染，然后过滤掉duplication的reads，处理后得到约{{clean\_data}} bp 的Clean data，数据质量良好，可用于后续分析。
2. 数据污染评估：基于污染分析分析结果显示，{{pollution\_description}}。

{% if homogeneous %}

1. K-mer分析：样本{{name}} 的k-mer分析预估基因组大小为{{estimate\_size}} bp左右，杂合率为{{heterozygosity }}左右。{% else %}
2. 样本{{name}} 的k-mer分析预估计基因组信息显示，该物种基因组为复杂基因组（多倍体、高杂合或数据存在污染）现有软件预估偏差较大。kmerFreq预估结果仅供参考，预估基因组大小为{{table\_kmer.4}} bp 左右，杂合度为{{table\_kmer.5}}左右，为更加精确预估基因组大小推荐利用流式细胞仪或近缘物种信息预估基因组大小。{% endif %}
3. 样本{{name}}采用SOAPdenovo进行预组装，组装Scaffold Length：{{scaffold\_length}} bp，Scaffold N50：{{scaffold\_n50}} bp，Scaffold Number:{{scaffold\_number}}；Contig Length：{{contig\_length}} bp，Contig N50：{{contig\_n50}} bp，Contig Number：{{contig\_number}}，详细组装结果见表3.4。

# 分析方法

## 分析软件

###### 分析所用软件信息

|  |  |  |
| --- | --- | --- |
| **Software** | **Version** | **Website** |
| ***Fastqc*** |  | https://github.com/s-andrews/FastQC |
| ***Fastp*** | *v0.20.0* | https://github.com/OpenGene/fastp |
| ***Jellyfish*** |  | https://github.com/gmarcais/Jellyfish |
| ***FindGSE*** |  | https://github.com/schneebergerlab/findGSE |
| ***GenomeScope*** |  | https://github.com/schatzlab/genomescope |
| ***SOAPdenovo*** | *2.04-r241* | https://github.com/aquaskyline/SOAPdenovo2 |
| ***bwa*** | *0.7.15-r1140* | https://github.com/lh3/bwa |
| ***samtools*** | *1.4* | https://github.com/samtools/samtools |

## 二代数据介绍及质控方法

测序得到的原始图像数据经碱基识别（base calling）转化为序列数据，我们称之为raw data或raw reads，结果以fastq（简称fq）文件格式存储。fastq文件为用户得到的最原始文件，里面存储reads的序列以及reads的测序质量。在fastq格式文件中每个read由四行描述：

@HWI-ST966:160:C28PYACXX:8:1101:1474:2178 1:N:0:GCCAAT

CAAAAACTGAAAGCATTTGGTTTCTACGGAACATACATATCCAGCAACCAGGCCAACTTAATTAACTCGGGGTCTAACGAAAGCTGCGTTCTTCTCTTAG

+

BBBFFFFFFFFFFIIIIIIIFIIBFIIIIFIIIIIIIFFIIIIIBFFIIIIIIFIIIFBFIBFFFIFFFFFFFBBBBBFBBBBBFBFF7BBBBBBBBBBF

其中第一行以“@”开头，随后为Illumina测序标识别符(Sequence Identifiers)和描述文字(选择性部分)；第二行是碱基序列；第三行以“+”开头，随后为Illumina测序标识别符和描述性文字(有的fq文件为了节省存储空间会省略第三行“＋”后面的内容)；第四行是对应序列的测序质量(Cock et al.,2010)。

第四行每个字符对应的ASCII值减去33，即为该碱基的测序质量值，比如B对应的ASCII值为66，那么其对应的碱基质量值是33。表5.2为测序错误率与测序质量值简明对应关系。如果测序错误率用E表示，碱基质量值用sQ表示，则有下列关系：

**sQ = -10log10E**

###### 测序错误率与测序质量值简明对应关系

|  |  |  |
| --- | --- | --- |
| **测序错误率** | **测序质量值** | **对应字符** |
| 5% | Q13 | . |
| 1% | Q20 | 5 |
| 0.10% | Q30 | ? |

基因组采用全基因组鸟枪法（WGS）策略，利用第二代测序技术，根据基因组的特点，构建二代小片段 DNA 文库，在 Illumina Hiseq Xten 上对这些片段两端进行双末端（Paired-End）测序，获得全基因组测序数据。

测序数据的产生是经过了 DNA 提取、建库、测序多个步骤的，这些步骤中产生的无效数据会对后续信息分析带来严重干扰，如测序接头序列，建库长度的偏差，以及测序错误、低质量碱基、未测出的碱基（以 N 表示）等情况，须通过一些手段将上述无效数据过滤掉，以保证分析的正常进行。

我们使用fastp (<https://github.com/OpenGene/fastp>)软件对原始数据进行过滤，过滤标准为：

* 去掉reads的接头序列；
* 截掉reads两端测序不准确的碱基，直接截取左端5个碱基，右端5个碱基；
* 去除当中还含有N的reads；
* 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。

然后使用FastQC ([http://www.bioinformatics.babraham.ac.uk/projects/fastqc](http://www.bioinformatics.babraham.ac.uk/projects/fastqc/)) 软件对clean data进行质控，若质控合格，则进行后续分析。

## *K*-mer分析方法

在基因组组装前，为了用测序所得的read信息估计基因组特征，我们采用基于*k*-mer的分析方法来估计基因组大小和杂合率等，即从一段连续序列中迭代地选取长度为*K*个碱基的序列，若read长度为*L*，*k*-mer长度为*K*，那么可以得到L-K+1个*k*-mer[5]。

定义：

*k*-mer深度分布曲线：*k*-mer深度~*k*-mer深度频率。

假设：

1. 从read中逐碱基取出的所有*K*-mer能够遍历整个基因组。根据Lander Waterman算法，基因组大小（G）满足如下公式：



*Knum*为*K*-mer个数， *Kdepth*为*k*-mer期望深度， *bnum*为碱基个数， *bdepth*为碱基期望深度。

1. *k*-mer深度频率分布服从泊松分布。



λ为期望值，取整后得到众数，即峰值为对应的测序*k*-mer深度，作为*k*-mer深度的估计值*k*-mer。

根据上述两个假设，即可从所有测序read中逐碱基取*k*-mer，统计*k*-mer频数分布，计算获得*k*-mer深度分布曲线和深度乘积曲线。根据曲线获得*k*-mer深度估计值，用于估计基因组大小。

利用Jellyfish对reads进行处理，获得不同频率的*k*-mer信息，然后利用相应的软件（FindGSE、GenomeScope）进行基因组大小预估。而Kmerfreq利用自身算法进行K-mer剪切（K-mer<=19）统计，利用K-mer频率分布进行基因组大小预估。

## 基因组组装

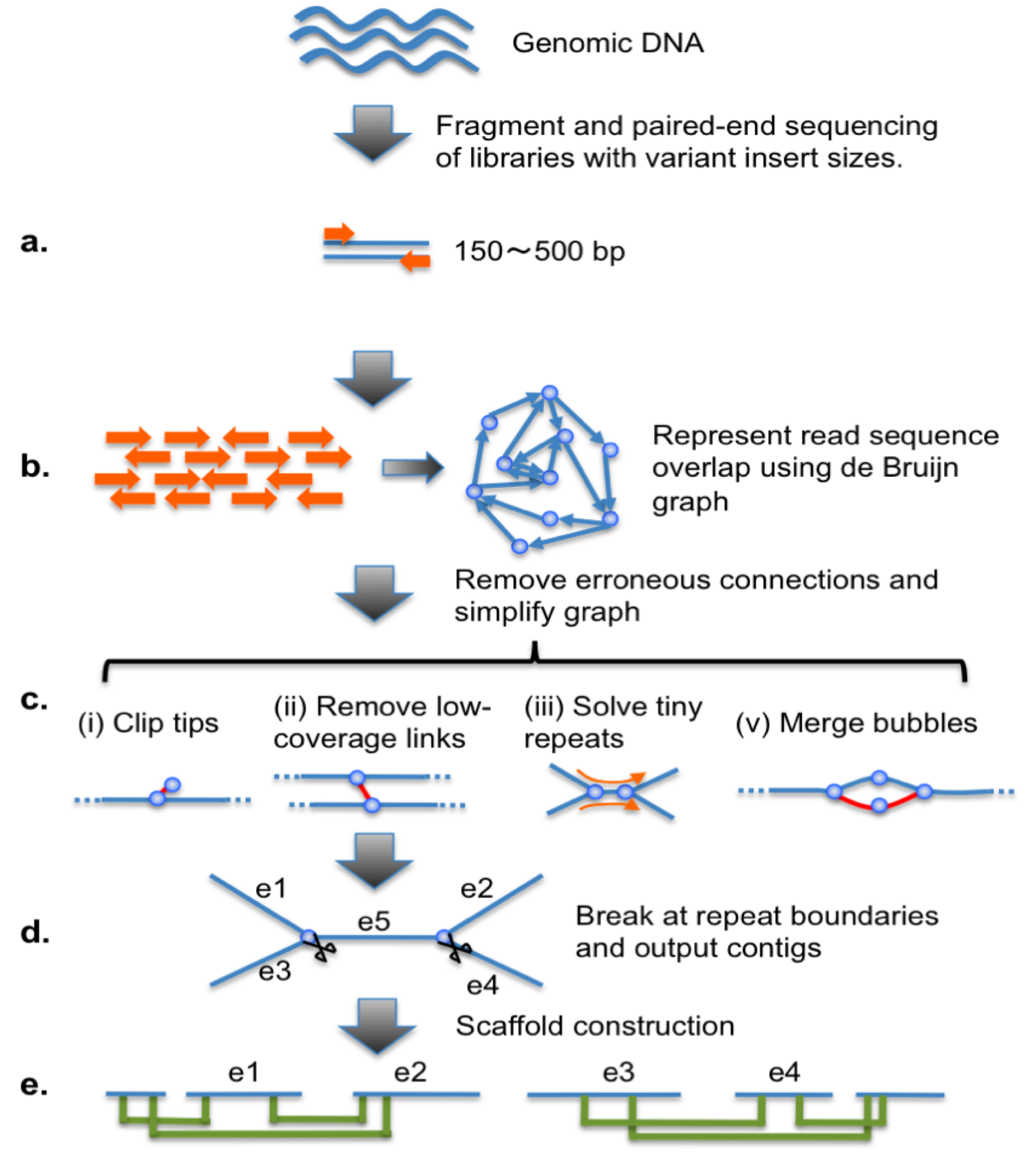


图6 组装流程图

1) Contigs 组装

利用Reads间Overlap和覆盖度信息，拼接出Contigs序列（连续的无 “N”序列）；Contigs组装方法较多，具体细节比较多，但从算法思想上来看，都是先利用Reads间Overlap关系构一个图，然后具体来简化这个图，简化图的过程就会涉及到具体的算法细节。

对短Reads（<=200bp）一般采用K-mer构建De-Brujin-Graph（DBG，也是我们采用的组装策略。

2) Scaffolds 构建

Scaffolds构建需要将所有文库测序得到的Reads比对回初步得到的 Contigs 序列，利用Reads间的Paired-End关系（短片段文库）,将Contigs 序列构图形成更长的Scaffolds序列，并利用不同插入片段估计出Contigs间的距离，用“N”填起来。

**{{name}}**基因组的组装流程如图6所示[6]，DNA序列先打断建库，构建二代小片段 DNA 文库数据，质控过滤处理得到高质量数据，将短Reads打断构建De-Brujin-Graph，然后利用一定数学方法解图，得到Contigs序列，并利用PE文库初步构建Scaffolds序列。

## GC-Dpeth分析

对组装的基因组序列以5kb为windows，无重复计算片段的平均GC含量和平均深度并作图。基于每一个windows对应的平均GC和平均深度进行绘图。可以根据此图分析测序数据是否存在GC偏向性以及样本是否存在污染。

# 六 参考文献

[1] Marcais, G. and C. Kingsford, A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics, 2011. **27**(6): p. 764-70.

[2] Sun, H., DIng, J., Piednoël, M. & Schneeberger, K. FindGSE: Estimating genome size

variation within human and Arabidopsis using k -mer frequencies. Bioinformatics 34, 550–

557 (2018).

[3] Vurture, G. W. et al. GenomeScope: Fast reference-free genome profiling from short reads.

Bioinformatics 33, 2202–2204 (2017).

[4] Luo, R., et al., SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. Gigascience, 2012. 1(1): p. 18.

[5] Liu, B., et al., Estimation of genomic characteristics by analyzing k-mer frequency in de novo genome projects. Quantitative Biology, 2013.

[6] Li, R., et al., De novo assembly of human genomes with massively parallel short read sequencing. Genome Research, 2010. 20(2): p. 265-272.

# 附件一 相关技术介绍

从早期的Sanger测序到第二代高通量测序、再到现在的单分子测序技术（第三代高通量测序），DNA测序技术一直推动着生命科学研究的发展。

基于Sanger法原理的测序仪经历了多次技术升级，但低通量、高成本的缺点仍然制约了人类对各种大基因组物种的组学研究。1990年开始，新的测序方法（Next generation sequencing）开始初具雏形，并在随后的若干年中不断发展，最终形成了以Illumina NovaSeq6000和MGISEQ-2000为主要平台类型的第二代高通量测序技术（以及被称为2.5代测序技术的Complete Genomics），它突破了Sanger法在通量和成本上的瓶颈，从此人类全基因组重测序技术开始被大量应用于医学研究；也带动了基因组、转录组学和表观基因组学等领域的高通量测序研究，同时强化了多组学相结合的生物信息分析思路。

基因组Survey基于小片段文库的低深度测序数据，通过生物信息学方法，评估基因组的大小及复杂程度等情况，以确定适合该物种的全基因组测序研究策略。是初步了解某一物种基因组特征的有效方法。

附件二 联系我们

武汉未来组生物科技有限公司（Nextomics Biosciences Co., Ltd）于2011年8月8日成立于武汉光谷生物城，目前在北京生命科学园成立了兄弟公司希望组且在美国费城设立有分支机构。

作为一家专注于第三代测序技术服务的公司，未来组早在2013年就成功开发了基于PacBio第三代单分子实时测序平台的生物信息分析流程，于2016年创建了Sequel基因组学中心并于2017年9月成功搭建Oxford Nanopore测序平台。2018年初，未来组又获得了Nanopore认证服务供应商资质，并正式面向全球提供GridION/PromethION等平台的测序及生物信息学分析服务。

2018年7月，未来组迁至武汉软件新城智慧健康园，占地4500m2，成立了基因组编辑中心和武汉希望组医学检验实验室，且拥有PacBio、Oxford Nanopore、MGISEQ、BioNano光学图谱及Hi-C染色体构象捕获等技术和平台，同时在华为云、天河二号及阿里云上搭建高性能集群用于生信分析，为高等院校和科研院所提供动植物基因组、全长转录组、表观基因组、微生物基因组、宏基因组和结构变异检测等领域的科研服务。

通过三代测序生物信息学工具和流程的开发，未来组解决了复杂动植物基因组、微生物基因组、全长转录组、微生物群体研究及人类基因组变异检测等领域的技术瓶颈，并推出了“TOP1000昆虫基因组计划”、“个人参考基因组服务计划”、“华夏万人结构变异计划”等，推动了基因组学研究的升级换代。因为专注于三代测序技术开发和应用推广，未来组成为众多科研机构优先合作的三代测序基因组中心，已完成了数百个三代测序科研项目，合作发表了一系列高质量的三代技术文章。

未来组致力于打造世界领先的三代测序基因组中心，在承诺高标准交付指标的同时将进一步大幅压缩项目服务周期，为合作伙伴提供完善的基因测序分析服务。

武汉未来组生物科技有限公司

地址：武汉市花城大道软件新城2期C11栋

电话：173 6290 2840

邮箱：support@grandomics.com

网址：www.nextomics.cn