## 1 二代测序数据质量评估

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3.2.1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**

表1 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：

* 过滤掉含有接头序列的 reads；
* 截掉reads头部和尾部低质量、不稳定碱基；
* 当单端测序 read 中含有的 N 的含量超过该条 read 长度比例的10% 时，需要去除此对双末端 reads；
* 由于PCR扩增等原因，需过滤掉duplication的reads，尤其大片段文库；
* 当单端测序 read 中含有的低质量(<=5)碱基数超过该条 read 长度比例50% 时，需要去除此对双末端 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。

* 数据质控结果：

我们对clean data产量做了统计，结果如下：

表2 Clean data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}} |



图1 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。

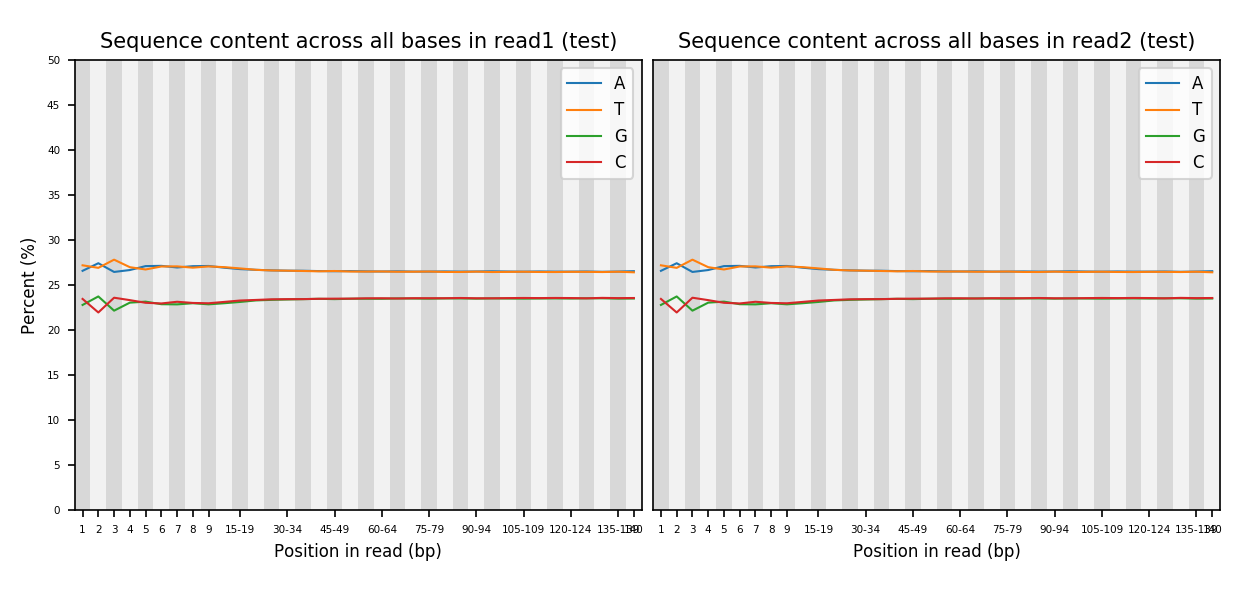


图2 碱基含量分布图

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

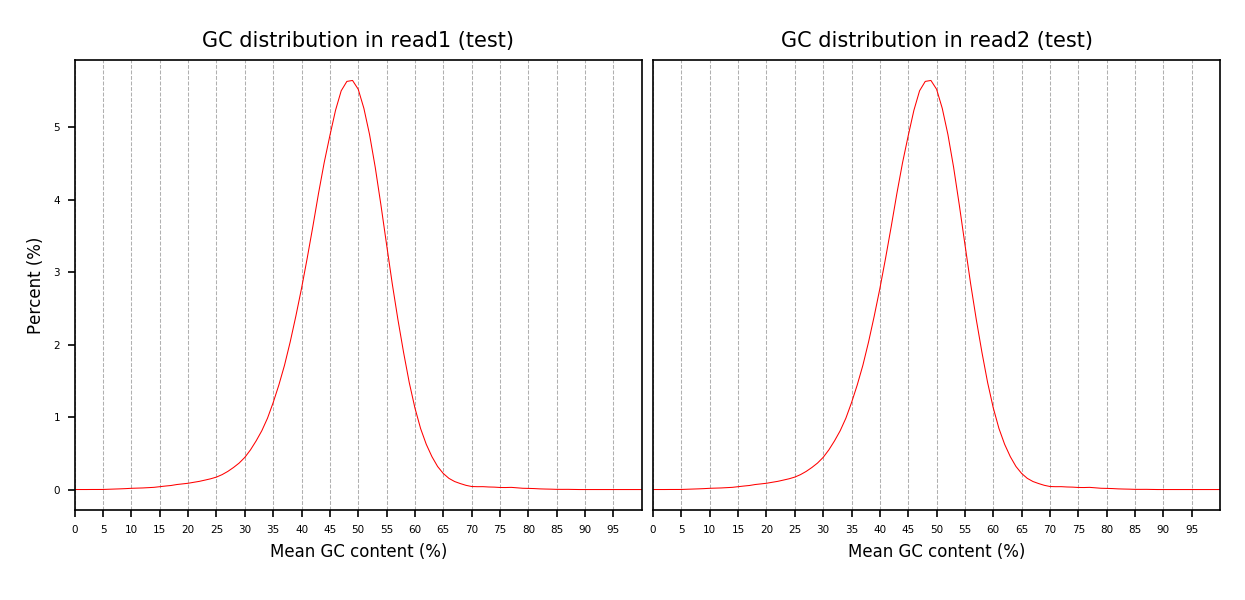


图3 平均GC分布图

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或是部分reads构成的子集有偏差。

### 数据评价：

原始数据{{raw\_data}} bp （见表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}} pb 的Clean data，数据质量良好，可用于后续基因组大小估计和contigs序列构建。