分析结果

1 项目信息

该项目基本信息展示如下:

- 项目编号：F22FTSEUHT0562_MUSomrhR
- 产品名称：只过滤
- 样本数：3
- 建库类型：DNBSEQ真核链特异性mRNA建库
- 测序平台：DNBseq
- 测序长度：PE100
- Clean fastq 质量得分体系：Phred+33

2 测序数据产出

测序完成后，对原始数据进行数据过滤，数据过滤包括去污染，去接头及去除低质量数据。

表1 有效数据产量统计表

Sample Name	Clean Reads	Clean Base	Read Length	Q20(%)	GC(%)
N21	36,235,360	7,247,072,000	PE100	98.21	51.06
N22	36,169,752	7,233,950,400	PE100	98.21	50.98
N23	36,351,830	7,270,366,000	PE100	98.14	51.29

- Sample Name：样本名
- Clean Reads：测序reads总量
- Clean Base：测序reads数与reads长度的乘积，代表测序总体碱基数
- Read Length：测序reads长度
- Q20(%)：测序reads中质量值≥20的碱基占总体碱基的百分比
- GC(%)：测序reads中碱基G、C占总体碱基的百分比

3 数据质控

原始数据进行数据过滤后，对数据质量进行监控。

3.1 数据过滤后各样碱基分布和质量分布图

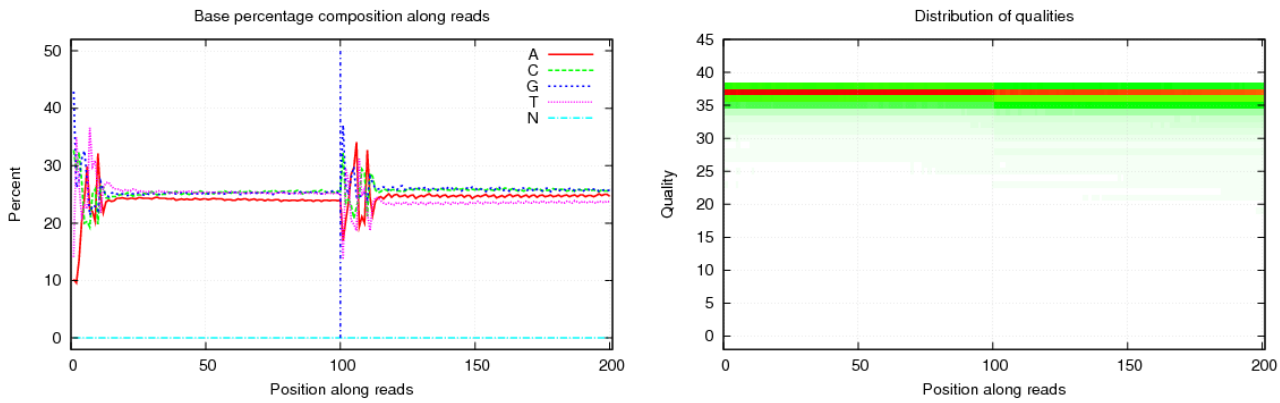

图1 N23单碱基组分分布图和单碱基质量分布图

左图中:横坐标为reads中的碱基位置，纵坐标为单碱基在该位点所占的比例；不同颜色代表不同的碱基类型。正常情况下样本无AT和CG分离，前几bp出现抖动情况是由于随机引物、测序反应开始酶和底物结合不太稳定导致，属于测序本身所带来的正常抖动。右图中:X轴表示各碱基在reads上的位置信息；Y轴表示单碱基的质量值；图中每个点表示reads上相应碱基的质量值；颜色深浅代表碱基数量的多少，某个质量值对应位置颜色越深表示测序数据中这个质量值的比例越高。

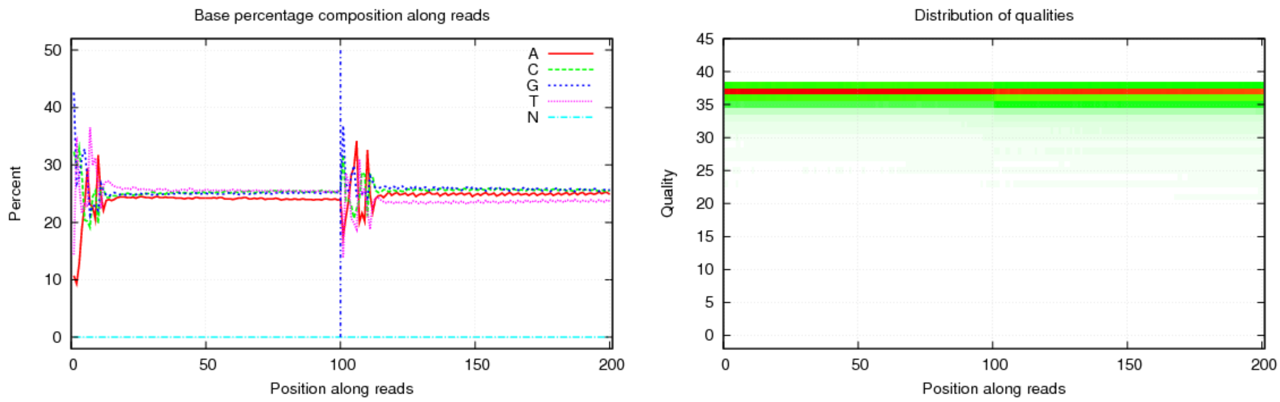

图2 N21单碱基组分分布图和单碱基质量分布图

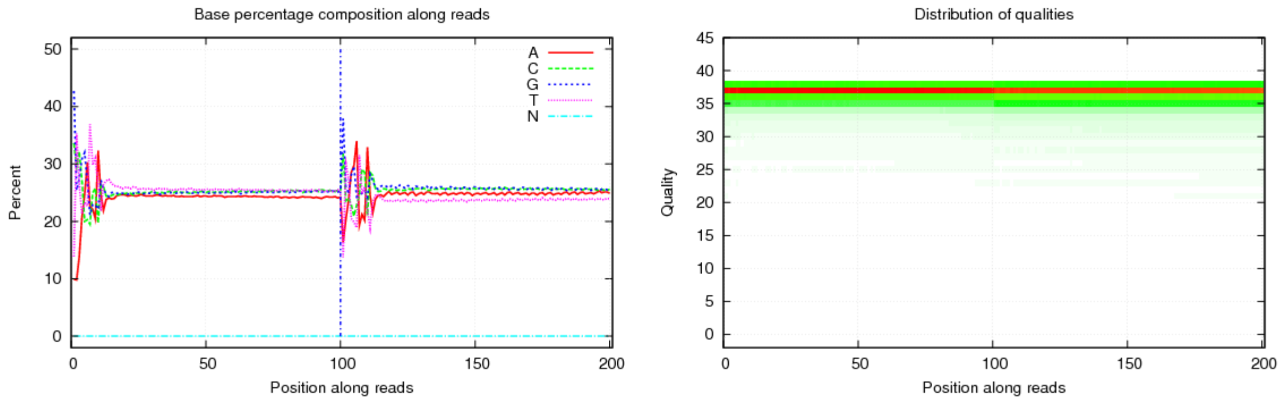

图3 N22单碱基组分分布图和单碱基质量分布图

左图中:横坐标为reads中的碱基位置，纵坐标为单碱基在该位点所占的比例；不同颜色代表不同的碱基类型。正常情况下样本无AT和CG分离，前几bp出现抖动情况是由于随机引物、测序反应开始酶和底物结合不太稳定导致，属于测序本身所带来的正常抖动。右图中:X轴表示各碱基在reads上的位置信息；Y轴表示单碱基的质量值；图中每个点表示reads上相应碱基的质量值；颜色深浅代表碱基数量的多少，某个质量值对应位置颜色越深表示测序数据中这个质量值的比例越高。

4 参考文献

[1] SOAPnuke: a MapReduce acceleration-supported software for integrated quality control and preprocessing of high-throughput sequencing data.

Chen Y, Chen Y, Shi C, et al.

PMID: 29220494 PMCID: PMC5788068 DOI: 10.1093/gigascience/gix120

分析方法

1 实验流程

文库构建和测序过程按以下步骤进行：

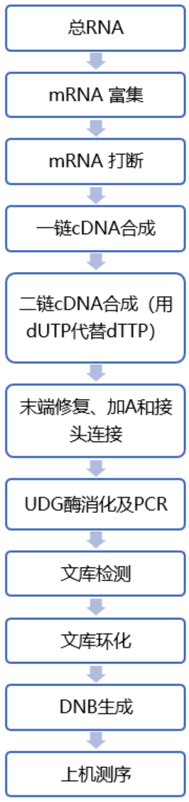

图1 实验流程图

1）取一定量Total RNA 样品，使用 oligodT从总RNA中获取mRNA； 2）mRNA打断；

3）片段化的mRNA加入随机引物进行cDNA一链的合成； 4）cDNA二链合成，用dUTP代替dTTP；

5）对双链cDNA进行末端修复、加“A”和接头连接；

6）用UDG酶消化掉带U标记的第二链模板后进行PCR以及PCR产物回收； 7）文库质量检测；

8）文库产物环化； 9）环状 DNA 分子通过滚环复制，形成DNA 纳米球（DNB）；

10）DNBSEQ平台上测序

10）DNBSEQ平台上测序。

2 生物信息分析流程

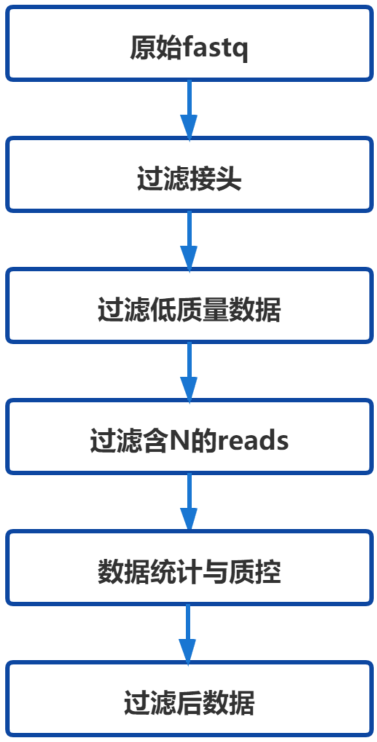

图2 生物信息分析流程

3 数据过滤

数据处理某些原始序列带有adaptor序列，或含有少量低质量序列。我们首先经过一系列数据处理以去除杂质数据，得到有效数据。我们利用华大自主开发的过滤软件SOAPnuke完成此步分析。

SOAPnuke软件过滤参数："-n 0.001 -l 20 -q 0.4 --adaMR 0.25 --ada_trim --minReadLen 100 "，过滤步骤：

1. 过滤接头：测序read匹配上adapter序列的25.0%或者以上（最多允许2个碱基错配）则切除adapter；
2. 过滤读长：如果测序read的长度小于100bp，则删除整条read；
3. 去N：如果测序read中N含量占整条read的0.1%或者以上，则删除整条read；
4. 过滤低质量数据：如果测序read中质量值低于20的碱基占整条read的40.0%或者以上则删除整条read；
5. 获得Clean reads：输出的read质量值体系设定为Phred+33。

4 参考文献

[1] SOAPnuke: a MapReduce acceleration-supported software for integrated quality control and preprocessing of high-throughput sequencing data.

Chen Y, Chen Y, Shi C, et al.

PMID: 29220494 PMCID: PMC5788068 DOI: 10.1093/gigascience/gix120

帮助

1 FASTQ 格式说明

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 （简称为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有四行描述：

```
@V350016857L4C001R0010000078/1
TTTTTCTGCTCCTTTTGATGCTATTAACAATTGCTTCAAGTTCAAGGGCACCTGCCTCAAAGTCCCTTTCTTCCAGACAAAATCTC
+
=,DDE@EFFF=DFDEFCCFDEFEGFEEAFDFFE=FFCFFEEEDFDEEEFDF8FFEFFEFF:FFEDF=EFDGE<1FDCEFFFFFDFE
```

第一行为序列标识以及相关的描述信息，以‘@’开头；第二行是碱基序列信息；第三行以‘+’开头，后面是序列标示符、描述信息，或者什么也没有；第四行是质量信息，和第二行的序列相对应，每一个碱基都有一个质量评分，根据碱基质量评分体系的不同，每个字符的代表的质量值也不相同。

下图为测序错误率与测序质量值简明对应关系。具体地，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：

$$SQ = -10 \times (\log\frac{E}{1-E})/(\log 10)$$

$$E = \frac{Y}{1+Y}$$

$$Y = \frac{SQ}{e^{-10 \times \log 10}}$$

1）对于测序质量值为 Phred+33 质量体系数据：碱基的测序质量值=质量信息字符对应的 ASCII 值-33，比如A对应的 ASCII 值为65，那么其对应的碱基质量值是65-33=32。DNBSEQ测序平台碱基质量值范围为2到42。

2）对于测序质量值为 Phred+64 质量体系数据：碱基的测序质量值=质量信息字符对应的 ASCII 值-64，比如c对应的 ASCII 值为99，那么其对应的碱基质量值是99-64=35。DNBSEQ测序平台碱基质量值范围为2到42

量值范围为2到43。

表1 测序错误率和测序质量关系简表

Sequencing error rate	Sequencing quality value	Character(Phred64)	Character(Phred33)
5%	13	M	.
1%	20	T	5
0.1%	30	^	?