**GaeaFastqQC产品介绍**

**深圳华大基因研究院**

**大数据平台**

**2014年9月**

目录

[1. 简介 3](#_Toc399693463)

[2. 软件架构及组件说明 3](#_Toc399693464)

[2.1 质控过滤 4](#_Toc399693465)

[2.2质控报告 4](#_Toc399693466)

[2.3 接头污染算法 6](#_Toc399693467)

[3.炎黄数据测试 6](#_Toc399693468)

[3.1精度测试 6](#_Toc399693469)

[3.2性能测试 6](#_Toc399693470)

[4总结 6](#_Toc399693471)

# 简介

GaeaFastqQC是一款基于Hadoop并行计算框架的对Fastq测序下机数据进行质控的软件。 GaeaFastqQC在BGI生产需求的基础上开发完成具有很强的实用性。GaeaFastqQC主要在接头、低质量和含N比例三个方面对Fastq数据进行质控，并且输出详细的质控图表信息。GaeaFastqQC支持动态去除reads末端接头污染功能，对于Nextera建库的数据有良好的支持。GaeaFastqQC支持多样品混合质控，解决了WES及TRS等小数据在Hadoop平台上的性能问题。

# 软件架构及组件说明

GaeaFastqQC读入fastq文件以及adapter readsID列表文件（此处为华大特有）；然后将属于同一readID的数据聚到一起，包括一对PE reads和其对应的adapter信息，对其进行质控操作；最后输出过滤后的Fastq以及质控报告。如图1所示。

**fq1**

**fq2**

**adp1**

**adp2**

**…**

**(rID1, read\_infom)**

**(rID1, adp\_infon)**

**(rID2, read\_infop)**

**(rID2, adp\_infoq)**

**…**

**(rID1, [read\_infom, adp\_infon, …])**

**(rID2, [read\_infop, adp\_infoq, …])**

**…**

**(filter\_read1)**

**read2**

**…**

***Input***

***Map***

***Reduce***

***Output***

**图1.GaeaFastqQC流程图**

## **质控过滤**

下机数据的质控是做好后面一切分析的基础，许多噪音在后续分析中会逐渐放大影响，对于最后的变异检测产生不好的影响。另外测序数据的质控也是我们把握测序质量的，避免实验等方面出问题的一个必选分析。

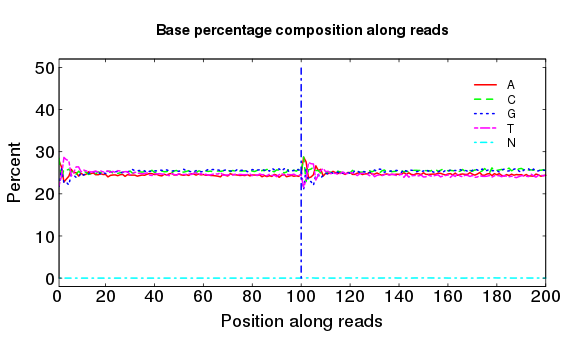
对于Fastq数据的质控主要分为三个部分：接头、低质量、含N比例，这三个部分是主要影响测序质量的三个因素。对于接头数据，GaeaFastqQC支持提供含有接头的read ID文件以及接头序列，对于提供接头序列，GaeaFastqQC在用户设置的允许mismatch数下对read头部进行比对，如果发现匹配则判定为含接头序列，这些序列会被过滤掉。对于低质量，我们按照含有低于某个质量值的碱基比例大于某个值来进行质控，比如我们分别设置低质量的值为10，低质量碱基比例为50%，则如果一条read含有质量值低于10的碱基的比例大于50%，则会被过滤掉。对于含N比例，GaeaFastqQC默认过滤掉含N碱基超过10%的reads，当然这个也是参数可设。另外对于PE reads，如果其中一条read不通过上面三项质控，则这对reads也将会被去除。

另外针对如nextera等建库会出现read 尾端的接头部分污染的问题，我们开发了基于BM方法的seed-extend查找算法去除接头污染部分的序列，保留剩下序列。但是针对污染过于严重的reads（一般为剩余序列短于30bp），考虑到对于下游分析特别是比对中引入错误比对的概率较大，我们会直接过滤掉。

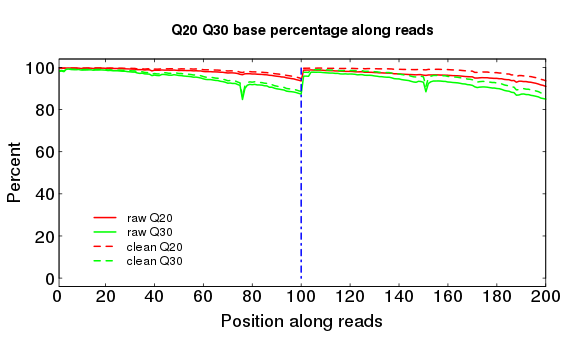
## **2.2质控报告**

对于质控标准中的三项指控标准的详细过滤信息，以及reads的统计信息我们最后会生成一份报告，具体如下表和图所示

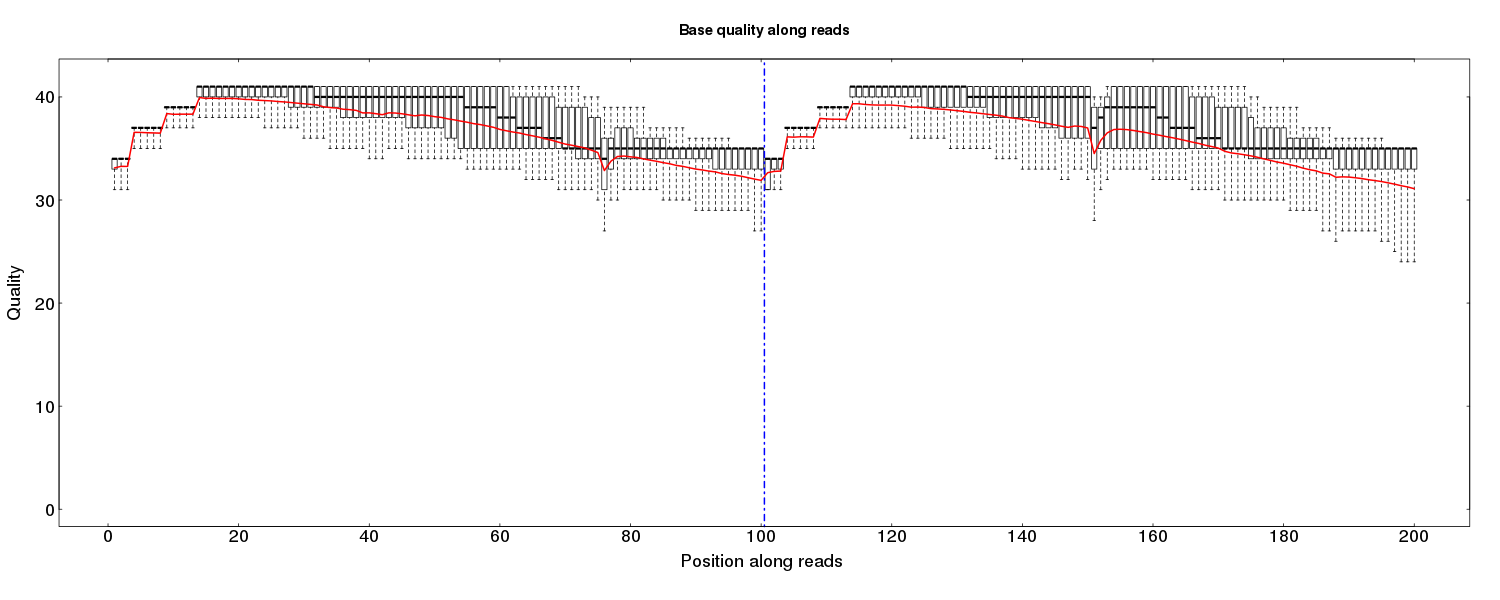
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 编号 | 名称 | 值 | 是否通过 | 未通过条件 |
| 1.1.1 | Read总数 |  | X | X |
| 1.1.2 | 过滤后read总数 |  | X | X |
| 1.1.3 | Read过滤比例 |  |  |  |
| 1.1.4 | 含接头read总数 |  | X | X |
| 1.1.5 | 低质量read总数 |  | X | X |
| 1.1.6 | 含过量N的read总数 |  | X | X |
| 1.1.7 | 由于成PE关系的两个read中有1个含有接头，而被过滤掉的另一个正常read的总数 |  | X | X |
| 1.1.8 | 由于成PE关系的两个read中有1个含有过多低质量碱基，而被过滤掉的另一个正常read的总数 |  | X | X |
| 1.1.9 | 由于成PE关系的两个read中有1个含有过多N，而被过滤掉的另一个正常read的总数 |  | X | X |
| 原始数据 | | | | |
| 1.1.10 | 原始数据总量 |  | X | X |
| 1.1.11 | 原始数据GC比例 |  |  |  |
| 1.1.12 | 原始数据含N比例 |  |  |  |
| 1.1.13 | 原始数据Q20比例 |  |  |  |
| 1.1.14 | 原始数据Q30比例 |  |  |  |
| 过滤后数据 | | | | |
| 1.1.15 | 过滤后数据总量 |  | X | X |
| 1.1.16 | 过滤后数据GC比例 |  |  |  |
| 1.1.17 | 过滤后数据含N比例 |  |  |  |
| 1.1.18 | 过滤后数据Q20比例 |  |  |  |
| 1.1.19 | 过滤后数据Q30比例 |  |  |  |



**图1.1.1 碱基分布图**



**图1.1.2 Q20/Q30分布图x**



**图1.1.3 质量值分布图**

## **2.3 接头污染算法（石泉）**

# 3.炎黄数据测试

## **3.1精度测试**

##### 3.1.1

## **3.2性能测试**

# 4总结

GaeaFastqQC有着丰富的质控指标以及合理的过滤准则，同时由于其基于Hadoop并行平台，具有优秀的扩展性，能够在几分钟内完成成百个区域测序数据的质控，并输出过滤后结果。