小RNA过滤Soapnuke与生产不一致的原因（蒋帅）

1. 需要过滤的项目主要包括以下几点
   1. 低质量
   2. 3’接头缺失及空载
   3. 5’污染
   4. 小片段
   5. polyA片段
2. soapnuke与生产条件一致的几个方面
   1. **低质量与生产条件完全一致**
      1. 整条read中，N碱基的个数>2

或者

* + 1. 第1-30号碱基中，N碱基的个数>0。

或者

* + 1. 严格条件：

第1-30号碱基中

* + - 1. 质量值q<10的个数>2

或者

* + - 1. 质量值q<13的个数>3
    1. 宽松条件：

第1-30号碱基中

* + - 1. 质量值q<10的个数>4

或者

* + - 1. 质量值q<13的个数>6
  1. **小片段与生产条件完全一致**
     1. 片段长度小于18(参数可调),依赖于3’接头检测
  2. **polyA片段与生产条件完全一致**
     1. Tag中含有A占70%(参数可调),依赖于3’接头检测

1. Soapnuke与生产不太一致的方面
   1. 3’接头检测
      1. 相同条件: 3’接头与Sequence比对, 错配数<=4, 最小比对长度>=5, 错配数/正配数<=0.4, N碱基不参与统计
      2. 造成不同的原因, 有个别的sequence按上述条件,可能会得到多个结果.(若我们把条件设置更严格的话,结果将会更一致)

对于这种个别的sequence生产中的策略是:

* + - 1. 如果3’接头比对的起始位置定位到了第1,2个碱基,并且错配数<=3, 则不管后边是不是会出现更好的比对结果,都选则这个位置.
      2. 如果3’接头比对的起始位置定位到第22±5个碱基,如果后边出现了比对结果更好的位置,也还是用22±5的这个情况
      3. 排除1,2情况后,以错配数/正配数做优先考虑

Soapnuke的测略是: 在所有的比对结果中,选择最佳的比对结果

1. 正配数更大,错配数更少.

将会造成的不同主要如下

一

CTCGTATGCTGTCTTCTGCTTG CATCGTATGCCGTCTTCTGCTTGCGAT

TCGTATGCCGTCTTCTGCTTG

二

CTCGTATGCTGTCTTCTGCTTG CATCGTATGCCGTCTTCTGCTTGCGAT

TCGTATGCCGTCTTCTGCTTG

**在生产中,序列将会按第一种方法,被过滤**

**在Soapnuke中,将会以第二种方法被过滤**

***个人觉得,应该Soapnuke的选择更好吧***

* 1. 5’接头检测
     1. 共同条件,错配数<=4, 最小连续正确比对数>=6或者Tag长度小于12, 正配数/tag长度>=0.8

***其实,对于tag长度小于18的情况,都不会对最终结果有影响,因为会被当作小片段处理掉***

* + 1. 不同原因, 在做更深层比对的时候

生产中选择了正常比对,和开gap(Smith-Waterman algorithm)的方式.

其中正常比对的时候, 生产中默认Tag长度比5’接头短的时候才会进行,并且比对范围是从接头左端比对到右端,如下

起始

GTTCAGAGTTCTACAGTCCGACGATC(接头)

ACGATTGGAAGG →

→

终址

GTTCAGAGTTCTACAGTCCGACGATC(接头)

ACGATTGGAAGG

开gap比对的时候, 要求gap不能多于1

Soapnuke选择的方案是, 从tag的左起第第6个,比对到Tag的右起第6个,如下:

起始:

GTTCAGAGTTCTACAGTCCGACGATC (接头)→

GAGTTCTACAGTCCGACGATC

→

终止

(接头)GTTCAGAGTTCTACAGTCCGACGATC

GAGTTCTACAGTCCGACGATC

这样会造成的不同主要如下:

**如下情况,在生产中,会被当作5’污染,而soapnuke是检测不出来的**

GTTCAGAGTTCTACAGTCCGACGATC (5’接头)

GTTCAGAGTT-----------TCCGACGATC (tag: GTTCAGAGTTTCCGACGATC)

GTTCAGAGTTCT--------- ACAGTCCGACGATC(5’接头)

GTTCAGAGTTCTAAAAAACAGTCCGACGATC(tag)

**如下情况,在生产中是检测不出来的,而在soapnuke中会被当作5’污染.**

GTTCAGAGTTCTACAGTCCGACGATC(5’接头)

AGAGTTCTACAGTCCGACGATCATCC(tag)

***个人觉得,开gap的算法不太符合测序原理. 而且我们是不是应该加一个条件,就是正配数/5’接头>=0.8时,也算作是5’污染.***

***总结: 小RNA在过滤的时候, 由于生产中条件过于宽松,所以经常会有很大一部分tag没法比对到基因组上, 而两种过滤方案,所产生的这些误差,与那些比对不到基因上的tag相比,基本影响不到结果.***

RNA-seq过滤 GPT与生产上差异分析：(陈浩森)

BC：

1. 去接头的时候，用两条adapter sequence以及两条adapter sequence的反向互补链，去匹配read，如果有连续>=10bp完全匹配到，则认为是adapter。

2. 在去N比例的时候，是要一条read的N值的比例大于所设定的Nrate。

3. 去低质量值时，去除低质量比例大于所设定的值Qrate。

GPT：

1. 去接头的时候，只用3'端adapter sequence，进行匹配。最小匹配长度为参数设定值。

(

程序这样实现的依据：

RNA-seq中也有adapter.list的情况，而adapter.list都是通过fqcheck\_adapter.cpp 这个程序生成的。

从RNA-seq的原始数据目录中，可以知道adapter.list只是使用3‘端接头序列

）

2. 在去N比例的时候，是要一条read的N值的比例大于等于所设定的Nrate。

3. 去低质量值时，去除低质量比例大于等于所设定的值Qrate。

另外：

由于GPT是对单个fq文件进行统计，如read1是低质量值而read2不是低质量，则fq1的统计信息中低质量值+1， 而fq2的统计信息中不改变。

这样，如果参数设置一样的话，得出的统计信息，总过滤条数肯能跟生产上的一致， 但对于pe，具体到某项（adapter，低质量值...)的时候，就有有较大差异了。