1、数据

使用samtools view 取 经过PCR验证的11000左右个位点（包括得证和未得证的位点）前后200bp的reads，生成fastq文件，共1.3Gb。
其中去除的reads绝大部分都是不配对的，所以后续质控及比对都是按照SE 的方法处理的。

已验证的位点去除重复后统计，共6269个位点得证，345个位点未得证。

2、使用初始参数处理（5.1之前流程中的参数）

已得证的位点共检出5189个，未得证的位点共检出87个。

```
PS：这一步的检出率都比较低，可能是下面几个方面的影响：
a、取出来的reads 不是配对的，在比对、去重复、GATK处理的时候有影响。
b、同一个位点附近的reads 可能来自不同的样本，相互之间有影响。（后针对未得证的位点，每个位点附近500bp内只取一个样本的数据，取出来的reads再处理，结果没有变化，所以可以排除这个因素）
c、未得证位点附近结构比较复杂，改变参数后，报出的位点再验证的位点附近。这方面的影响较大，核查了十几个位点，有一半是这种情况。
d、样本来自于不同时期，可能是不同的测序平台，对结果可能有影响。
```

3、GATK base quality 调整

添加--min_base_quality_score 15：
已得证的位点共检出5187个，未得证的位点共检出84个。
添加--min_base_quality_score 25：
已得证的位点共检出5180个，未得证的位点共检出77个。

4、使用fastp软件对reads进行过滤

fastp 参数：

```
-q 20 -u 20 \ # 质量值低于20的标记为低质量，低质量占比大于20%的reads去掉
-n 3 \ #去掉模糊碱基N数目大于3的序列
-f 2 \ #切除reads5’ 端 2个碱基
-l 35 \ #去除质控后长度小于35的序列
-W 8 \ #滑动窗口长度设为8bp
-M 25 \ #滑动窗口内平均碱基质量低于25的从窗口起始位置剪切掉
-3 \ #从reads末端进行滑动窗口的剪切
-y \ #开启低复杂度过滤
-x \ #trim 3’ 端ploy结构
-A #关闭去接头的操作，，，因为取出来的reads是经过质控的，且是不配对的序列，对fastp过滤接头有影响，故关闭
```

在设置--min_base_quality_score 25 的情况下：
已得证的位点共检出5169个，未得证的位点共检出72个。

5、参考序列补充未组装上染色体的序列

新参考序列路径：

```
/share/ofs1a/prod/bin/heh_tmp/hg19Un/hg19_addUn.fa
```

在设置--min_base_quality_score 15 的情况下：
已得证的位点共检出5166个，未得证的位点共检出73个。

在设置--min_base_quality_score 25 的情况下：
已得证的位点共检出5159个，未得证的位点共检出67个。

PS：曾尝试将hg38中未组装的序列添加进参考序列，但是已得证的位点损失太严重，故舍弃。

6、调整bwa参数

新的bwa参数

```
-M \ #mark shorter split hits as secondary 对后续去 PCR Duplicates 有帮助
-B 8 \ #碱基错配罚分 默认4
-O 12 \ #gap open 罚分 默认6
-L 15 \ #penalty for 5'- and 3'-end clipping 默认 5
-U 20 \ # penalty for an unpaired read pair [17]
-T 50 \ #minimum score to output 默认30
-d 75 \ #off-diagonal X-dropoff [100]
-r 1.3 \ #look for internal seeds inside a seed longer than {-k} * FLOAT [1.5]
-E 4 \ #gap延伸罚分 默认1
-k 32\ #seed 长度
```

总体而言比对罚分更严格，为尽可能减少错误比对而设
测试结果（使用新的参考序列，及质控后的序列）：
在设置--min_base_quality_score 15 的情况下：
已得证的位点共检出5173个，未得证的位点共检出73个。

在设置--min_base_quality_score 25 的情况下：
已得证的位点共检出5165个，未得证的位点共检出67个。

详细数据统计：

GATK未更新	q15	q25	fastp	hg19_addUn&q15	hg19_addUn&q25	bwa&hg19_addUn&q15	bwa&hg19_addUn&q25
5189	5187	5180	5169	5166	5159	5173	5165
87	84	77	72	73	67	73	67

结论：

1、gatk 添加--min_base_quality_score 25 参数 、参考序列添加未组装上基因组的序列 可以显著降低假阳性，并且对检出率影响不大
2、fastp 进一步质控对减低假阳性率有帮助，使用PEreads 的时候效果应该更明显
3、调整bwa参数后虽然假阳性率改善不明显，但是检出率有提升。

进一步调整GATK参数

GATK提供了很多对bam文件中reads进行过滤的方法，将这些方法添加进来进行测试。由于很多方法是基于PEreads 进行的。这次选用了未得证位点最多的10个样本的原始数据进行测试。
样本信息：

样本名称	57745	62528	65014	71648	53609	54610	67480	67483	71399	59387
不得证位点数目	4	4	4	4	5	5	5	5	5	6

测试参数：

fastp

```
-f 5 -F 5 \                    #切除read1，read2 5’ 端各5bp
-g --poly_g_min_len 6 \        #切除reads末端的ployG结构（大于6bp）
-x \                           #切除reads末端的ployX结构（10bp）
-3 \                           #从reads末端进行滑动窗口的剪切
-W 8 -M 25 \                   #滑动窗口长度设为8bp ,滑动窗口内平均碱基质量低于25的从窗口起始位置剪切掉
-q 20 -u 20 \                  # 质量值低于20的标记为低质量，低质量占比大于20%的reads去掉
-n 3 \                         #去掉模糊碱基N数目大于3的序列
-y -Y 20 \                     #过滤低复杂度序列
-c                             #根据overlap 修正低质量的碱基
```

bwa

```
-M \        #mark shorter split hits as secondary  对后续去 PCR  Duplicates  有帮助
-B 8 \      #碱基错配罚分  默认4
-O 12 \     #gap  open 罚分  默认6
-L 15 \     #penalty for 5'- and 3'-end clipping  默认 5
-U 20 \     # penalty for an unpaired read pair [17]
-T 50 \     #minimum score to output  默认30
-d 75 \     #off-diagonal X-dropoff [100]
-r 1.3 \    #look for internal seeds inside a seed longer than {-k} * FLOAT [1.5]
-E 4 \      #gap延伸罚分 默认1
-k 32\      #seed 长度
```

gatk添加的参数

```
-rf NotPrimaryAlignment \                  #
-rf MaxInsertSize -maxInsert 1000 \        #默认值为10000
-rf BadCigar \                             #
-rf BadMate \                              #
--min_mapping_quality_score 30 \           # 默认值为20 ，
-rf MateSameStrand                         #
```

主要是添加了不同的reads filter ，后来发现HaplotypeCaller 已默认添加了大部分reads filter ：

Additional Information

Read filters

```
* HCMappingQualityFilter
* MalformedReadFilter
* BadCigarFilter
* UnmappedReadFilter
* NotPrimaryAlignmentFilter
* FailsVendorQualityCheckFilter
* DuplicateReadFilter
* MappingQualityUnavailableFilter
```

These Read Filters are automatically applied to the data by the Engine before processing by HaplotypeCaller.

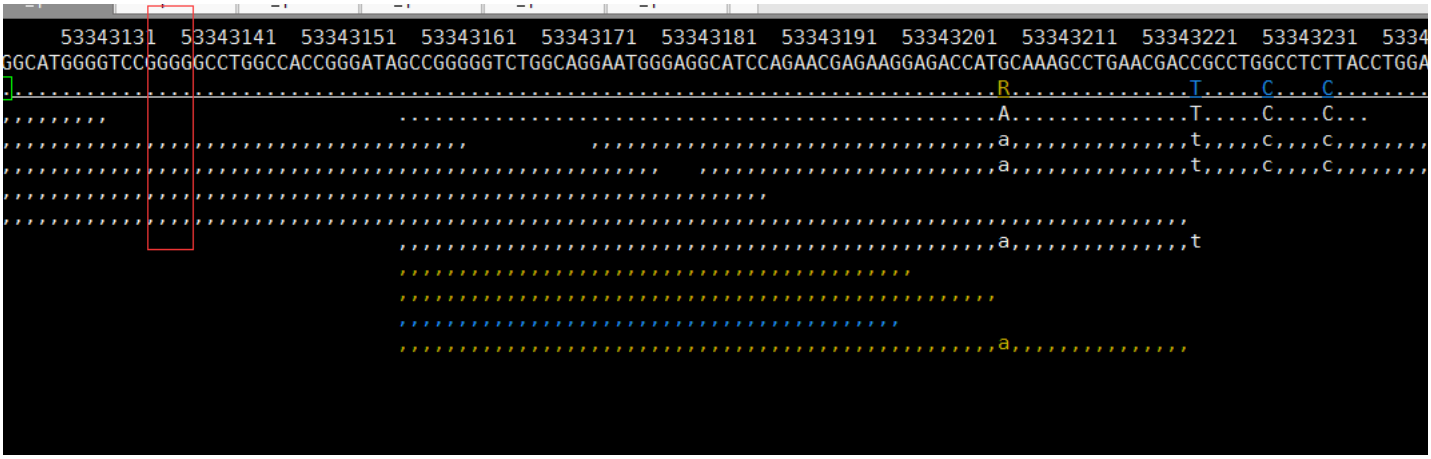
测试详细数据见excel文档。

测试结果

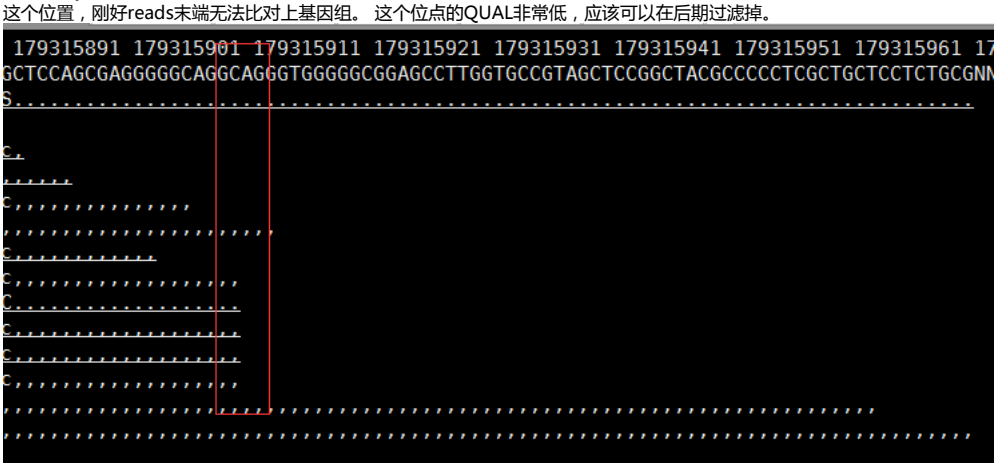
10个样本共47个未得证位点，新参数处理过后，可排除35个位点。即可排除74.4%的假阳性位点。
47个位点有部分比较集中，按100bp区域统计，共16个区域，可排除其中的12个 。

其中67483xbycn1 和 59387xbycn1A 两个样本中的位点都未能排除，查看结果如下：

67483xbycn1 从比对结果上未见明显变异情况，不知为何会报出变异



59387xbycn1A call出来的变异与之前的差别较大，新参数显示这个位点插入 CCGAGAAGGGGGTTTT ，分析发现，主要是reads比对错误，将本不应该在这个位置的reads比对到这个位置，刚好reads末端无法比对上基因组。 这个位点的QUAL非常低，应该可以在后期过滤掉。



针对上述情况，尝试添加GATK的过滤参数解决，但是没有效果，添加的参数：

```
-NoRequireSCBothEnds \
-filterTooShort 80 \
--dontUseSoftClippedBases
```

这个结果中没有对检出率进行分析，主要是数据中没有得证的检出位点，后续可以根据已知的高频位点进行估计。
这次分析的样本量还是比较少，所以涉及的位点数目也不多，后续可以加大样本量进行测试。

变异可靠性研发数据统计.xlsx
2018/5/6 17:37, 44.9 KB