碱基质量分数校正(base quality score recalibration, BQSR)是数据预处理的其中一个步骤，其主要功能是检测测序仪在估计每个碱基的质量分数时出现的系统性误差并对其校正。

图一为BQSR的一般流程，主要分为三个步骤，根据输入的比对文件(sam/bam)为各个用于BQSR的变量构建tables，之后把这些table组合起来形成一个group(.grp)文件，最后一步根据grp文件更改原比对文件的碱基质量分数并输出校正后的比对文件(sam/bam)。整个流程中，第一步和第三步是用spark实现，第二步在单机上实现。

***Combine tables***

***Print Reads***

***Input***

***Create tables for different covarites***

Figure 1碱基质量分数校正流程

## Create tables for covariates

这一步主要由两个阶段组成：对输入文件的转换和创建表格。转换阶段主要是对reads进行过滤，初步解析以及窗口划分，划分窗口根据pos/winSize来得到winId，其中需要注意的是横跨两个窗口的reads需要在两个窗口都保存。格式转换后，通过二次排序把同一条染色体的相同窗口的reads分配到同一个窗口中，并且窗口里的reads按照pos排序。具体见图2。



Figure 2比对文件转换过程

PairWritable主要记录reads的染色体，所在窗口和起始位置，SAMRecordWritable主要记录read的信息。在创建表格阶段主要进行质量分数校正功能。在这个阶段我们先把处于同一个窗口的reads收集起来，然后根据reads的每个位点的比对情况判断是否需要进行校正（对于mismatch位点，若这个位点不在提供的dbsnp文件中出现，则对其进行校正）。如图3所示：



Figure 3创建表格流程

Map base主要对窗口里的碱基逐一检查，并根据该碱基位点是否包含在dbSNP和与碱基是否与ref匹配更新相关变量的table。具体实现步骤见图4。



Figure 4更新tables流程

更新tables主要是根据变量把碱基划分成不同的组，然后根据每个组中错误的碱基，即不在dbSNP中的碱基和该组中的碱基数计算该组碱基的质量分数，具体算法会在下面讲述。

## Combine tables

由于第一步是通过把一个大文件切割为很多个小文件进行统计，所以每个小文件都有自己的table。所以这一步需要把各个小文件的tables的信息整合在一起形成一个包含大文件信息的总table，然后根据总table的错误碱基数和观察碱基数计算经验分数（empirical quality score），最后输出格式化的table。具体过程见图5.



Figure 5Combine table过程

## Print Reads

这一步主要是根据上面两步得到的各个变量的report quality score和empirical quality score根据公式校正质量分数，并且输出校正后的bam文件。具体过程见图6.

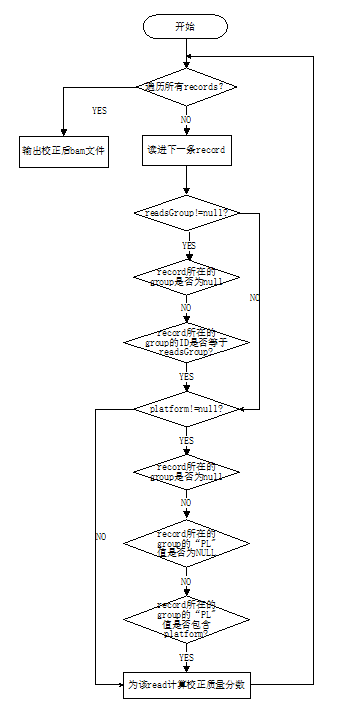


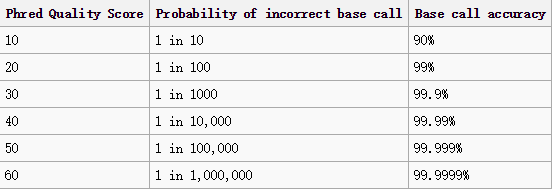
Figure 6Print Reads过程

## 碱基质量分数校正算法

DNA测序仪会为读到的每一个碱基提供质量估计，以此来评估错误概率，这个值称为Phred Quality Score.Phred Quality Score的计算公式如下：

（公式1）

其中P为错误概率。

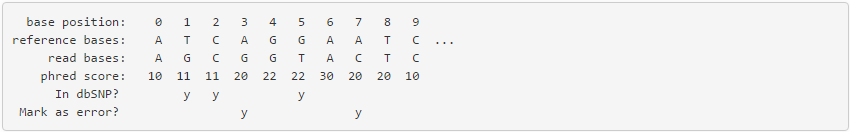


BQSR通过观察文件中的所有碱基调整Phred Quality Score使其更为精确。

执行BQSR的前提是输入一个包含所有已知SNP位点的文件，例如dbSNP，这是因为BQSR算法只会把比对文件中任何不在dbSNP的SNP认为是一个测序错误(error)。

### 寻找errors

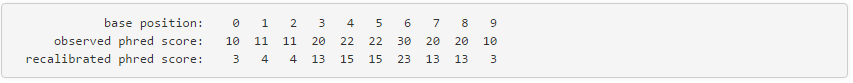
下面的例子可以帮助我们了解BQSR校正一个碱基的逻辑。假设我们有一条10个碱基长的序列，在染色体上的起始位置为0。



我们把第3位和第7位的碱基标记为error，因为它们不在dbSNP中出现，1,5位碱基虽然与ref不匹配，但因为在dbSNP中有记录，所以它们不是errors。

### 计算经验（Empirical）值

以上我们可以看到10个碱基中有2个error（第3位和第7位），因此错误概率P=0.2，根据公式1得phred quality score值为7。

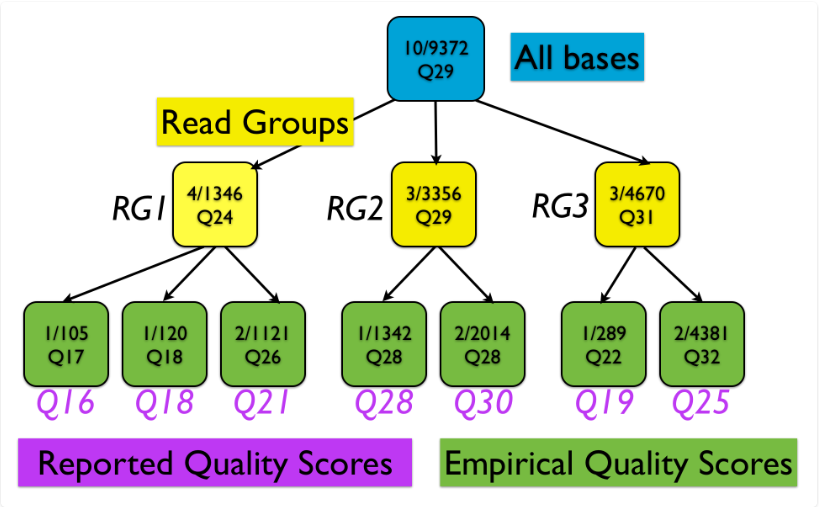
所以要校正这条read的分数，只需要将每个碱基的分数减去7. 

但是对于单一的一条序列执行BQSR是无效而且不精确的，下面观察一些更加实际的例子。

### 根据协变量(covariate)计算

BQSR包含四个变量，read group，quality group，cycle和context，其中read group和quality group变量是必须的，其余两个是可以选择的变量。

这是因为不同的测序机器的校正不同，我们把位于相同group的碱基收集起来，然后我们进一步把同一个group的碱基根据质量分数切分，相同分数的碱基将处于同一个区间里。 因为BQSR的目标是比较测序仪提供的分数和我们从经验错误数得出的经验分数。



这个例子里，我们有三个group：RG1，RG2和RG3，每个read group根据碱基的分数又切分成不同的sub group（实际中subgroup的数目会多的多）。每个节点中包含error数目和节点中总碱基数的比例和经验质量分数。

经验质量分数的计算公式如下：

（公式2）

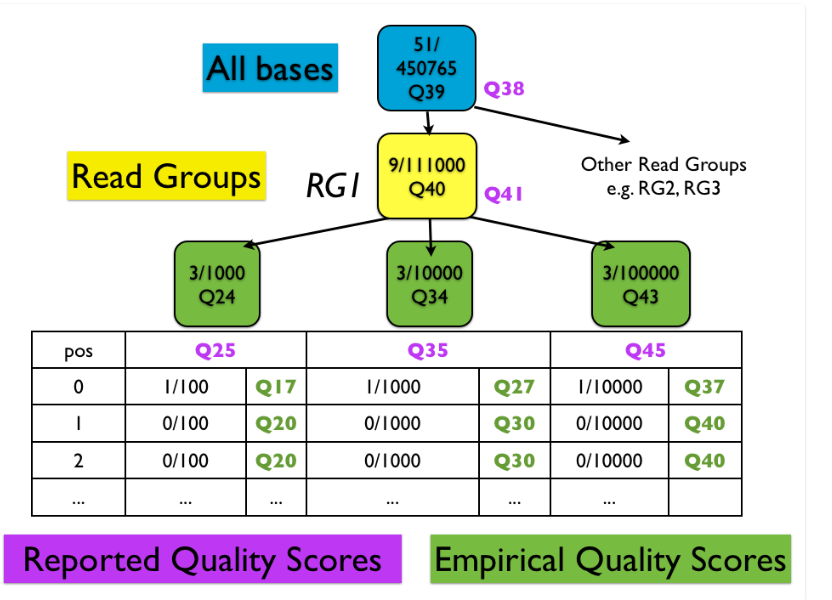
这个公式对碱基数目（obervations）较少的区间添加罚分。

BQSR的执行需要两个步骤。第一步是构建一个嵌套结构（见上图），通过嵌套结构来保存read group和质量分数组合中的总的error数目和观察到的碱基数目。当这个结构完成后，就会在第二步中被用来对碱基进行校正。例如，我们从RG1得到一条序列，我们把这条序列中质量值为16的碱基校正为17，质量值为18的保持不变，质量值为21的校正为26.

如果计算中还包括其他变量，可以按照以下公式计算：

其中Qrg表示read group，Qqs表示quality score，Qfr表示from read，即该碱基在序列中的分数，Qco表示其他的变量，包括cycle,context。

举个例子：



图中的协变量为cycle，于是RG1，Q24，Cycle为0的碱基校正为：