1. **RNAseq用什么软件做比对？Bowtie2? STAR? HISAT? BWA?**

RNA比对工具和DNA比对工具的主要区别在于**是否会考虑跨外显子的比对**，RNA比对工具会将没有比对上的reads劈开，然后再次去比对。以上提到的四种比对工具里，BWA、Bowtie为DNA测序常用工具，**STAR、HISAT2**（HISAT2不仅支持RNA-seq的比对还支持DNA-seq比对，唯一需要做的就是加上一个参数--no-spliced-alignment）以及一些帖子提到的**Tophat**（不过有看到说HISAT是他的非正式升级版本，克服了他速度慢的缺点）为RNA比对的工具。

1. **STAR**

比对完成后，会输出许多文件，包含4个类别：

**log文件:**

Log.out文件记录了程序运行时的信息，可以用来回溯错误信息。

Log.progress.out文件报告比对进程情况，1分钟记录一次。

Log.final.out文件包含了比对结束后比对统计的信息。

**sam文件:**

存储测序数据和参考基因组比对结果的文件

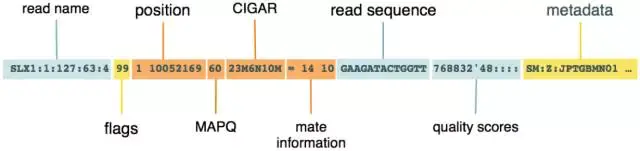
**bam文件:**

sam的二进制格式文件，由于sam文件是纯文本内容，因此内容巨大，比如一个人30x全基因组测序的sam大小超过600G，非常容易造成存储爆满。因此开发者想到了压缩sam 文件这一方法。bwa作者还为bam 专门开发了samtools，更便于处理bam文件。

Bam文件格式如下



（对照例子看可能更直观…）



**剪切位点文件**:

剪切位点文件实际上是根据mapping情况，估算出来的intron区间的信息，默认的文件名称为SJ.out.tab

1. **STAR和HISAT软件**

HISAT2因为索引的优势(**看不太懂这个局部索引和全局索引的原理**)，可以相对轻松比对跨区域的reads（可变剪切)，而**Tophat2耗时久，STAR耗内存，HISAT2克服了两个的缺点**。

至于**HISAT2**，他的**junction**（junction双端测序，中间会留有200bp测不到的东西，我们叫junction）**正确率最高**，但是**灵敏度相对较低**。而**STAR灵敏度更高**，它对于双端测序的reads，要么全部比对上，要么全部抛弃，不会像像TopHat和HISAT2一样只比对上某一个reads（也就是说STAR会将没有paired mapping上的reads都剔除，避免single reads比对到基因组上）。

1. **Cellranger用的比对软件是什么**

用的是**STAR**

Cellranger包含有与单细胞基因表达分析相关的四个pipelines，分别是

cellranger mkfastq：将 Illumina 测序仪产生的 raw base call (BCL) 文件解析成 FASTQ 文件

cellranger count ：将 cellranger mkfastq 产生的或其他来源的 FASTQ 文件进行比对、过滤、barcode 计数以及 UMI 计数，并可以生成 feature-barcode 定量矩阵，随后确定细胞群并进行基因表达分析

cellranger aggr ：将多个 cellranger count 产生的数据进行整合、标准化，并可以对整合后的数据进行分析

cellranger reanalyze ：使用 cellranger count 或 cellranger aggr 产生的表达矩阵重新进行降维、聚类等后续分析（**那这样一来后面本来用R做的事情都可以用cellranger做…嘛？**）

以上四个pipeline 均将转录组常用**比对软件 STAR** 封装其中。

1. **RPKM、TPM、CPM、FPKM的区别**

RPKM、FPKM 和 TPM这三个指标都用于标准化测序深度和基因长度。CPM/RPM➗exon length(KB))=RPKM

1. **RPKM**

**RPKM = total exon reads/ (mapped reads (Millions) \* exon length(KB))**

total exon reads：某个样本mapping到 特定基因 的外显子上的所有的reads；

mapped reads (Millions) :某个样本的 所有 reads总和；

exon length(KB)：某个基因的长度（外显子的长度的总和，以KB为单位）

1. **FPKM**

FPKM 与 RPKM 非常相似。针对Single-end RPKM与FPKM基本没有差异。针**对Paired-end**，如果一对paired-read都比对上,**那么FPKM计算方法中认为这一对read为一个fragment**（RPKM则计为2），如果一对中仅有一个比对上，则将比对上的计为一个fragment。

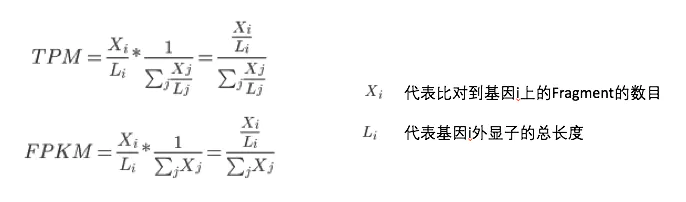
1. **TPM**

以下为TPM计算过程：

标准化基因长度：将所有基因的Read数除以基因长度（基因长度单位为kb）；

计算总Read数：计算每一个样本的总Read数，然后将其换算为以百万位单位（M）；

标准化总Read数：将所有基因的Read数除以总Read数。



因此比对TPM和FPKM的公式可以发现，FPKM的分母没有考虑基因长度的影响，所以TPM更加符合我们对相对表达量的定义。（有个很多帖子都在用的例子害挺直观，这儿就不放了）

1. **CPM**

RPM/CPM = total exon reads/ (mapped reads (Millions))

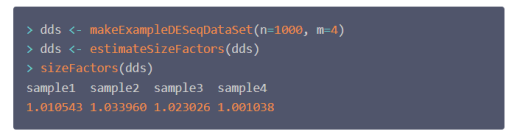
多进行样本间比较，无法进行样本内差异表达分析

1. **edgeR和DEseq**

**感觉下面两个R包跟上面几个的算法不一样哎**

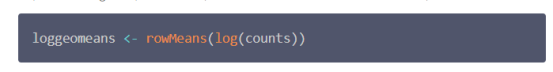
1. DEseq

为了在样本间进行差异分析，首先就需要**对原始的raw count 表达量数据进行归一化**。在DESeq2中，通过**estimateSizeFactors函数**为每个样本计算一个系数，称之为**sizefactor**, 示意如下



具体的计算过程如下：

先进行**log转换**，转换之后，**计算每个基因在所有样本中的均值**，代码如下



计算单个样本的sizafactor时，将该样本中**每个基因的表达量减去对应的所有样本中的均值**，然后**取中位数**。由于开始进行了log转换，最后在转换回来。 假定一个样本中所有基因的表达量为cnts, 代码如下



需要注意的时，在计算中位数时，对基因进行了过滤，需要满足以下两个条件

1.在该样本中该基因的表达量大于0

2.在所有样本中该基因的表达量都大于0，而且取log之后的和不为0

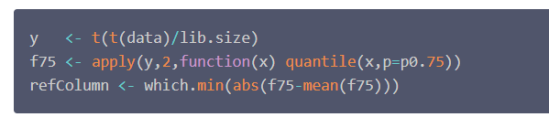
实际上第二个条件已经包含第一个条件了，在原始的表达量矩阵中，肯定会有基因在部分样本表达量为0的情况，所以最终计算中位数时，只会用到部分基因。

计算出每个样本的sizefactor之后，将**该样本原始的表达量除以该样本的sizefactor, 就得到了归一化之后的表达量**。

1. edgeR

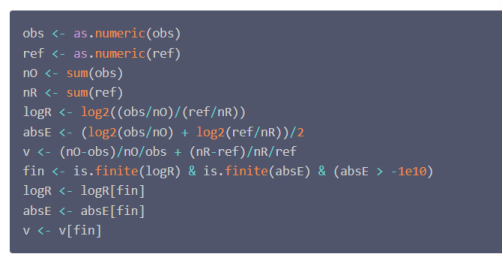
DESeq2的归一化算法只考虑在所有样本中表达量都大于零的基因。edgeR采取了参照样本的策略，首先从所有样本中挑选一个样本作为参照，在对其他样本进行归一化时，只考虑那些在参照样本和待归一化的样本间共有的RNA。

选取参照样本的代码如下



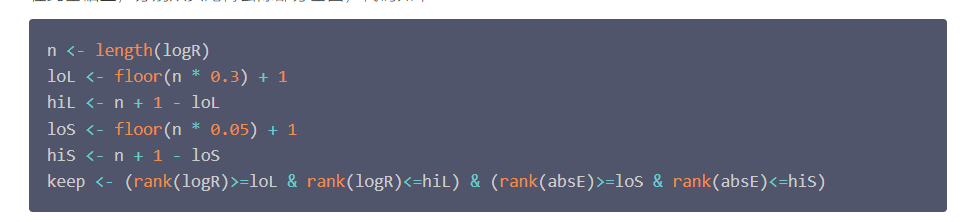
根据相对丰度，**计算每个样本的第三四分位数**，采用该值代表每个样本的表达水平，选取与**所有样本第三四分位数均值相差最小的样本作为参照样本**。

参照样本选好之后，采用循环对每个样本进行归一化。在归一化时，**重点关注基因的选取**。代码如下



ref代表参照样本的表达量，obs代表待归一化样本的表达量。通过构建的3个统计量对基因进行初步过滤，logR 其实就是样本间的log2FD值，absE是表达量的均值，通过fin指定的过滤措施(**不太懂这个过滤的意思**)，过滤掉在任意样本中表达量为的基因，通过absE过滤掉一部分表达量很低的基因。

在此基础上，**分别从头尾再去除部分基因**，代码如下



这样做的目的是**保留在样本间表达量没有差异的基因**。最后在利用下列公式计算每个样本的sizefactor。

