RPKM, FPKM and TPM,迷糊？

2016-12-27 struggle 生信人

RPKM (Reads Per Kilobase Million) 、 FPKM (Fragments Per Kilobase Million)、TPM (Transcripts Per Kilobase Million)这三个名词常常作为转录组数据定量的表示方法，而好多人也不清楚这三个到底有啥区别，下面小编就讲解一下。

首先这三种方法都是对表达量进行标准化的方法。为啥标准化呢，不直接用read数表示呢？因为基因越长read数目越多，测序深度越高，一个基因对应的read数目越多，所以必须要标准化，而标准化的对象就是基因长度与测序深度了。

1、 RPKM

主要用来对单端测序（single-end RNA-seq）进行定量的方法。英文释义： mapped reads per kilobase of exon per million mapped reads ,

$$RPKM_g = \frac{r_g \times 10^9}{fl_g \times R}$$

rg：map到该基因的read数目

flg：该基因外显子长度，以kb为单位

R：read总数目

主要意思就是将map到某个基因外显子上的read数除以map到所有基因的外显子上的read数(以million为单位)与该基因外显子的总长度(以KB为单位)

“per million” 主要为了标准化测序深度
"per kilobase of exon"主要标准化基因外显子长度

作为最广泛使用的归一化算法，RPKM却有很多弊端：

如果不考虑 RNA-seq 实验测序深度的影响，RPKM 方法实际上就是用读段的覆盖度来刻画

基因的表达水平。假设读段在某基因的各外显子区域内都是均匀分 布的，且该基因不含有选择性剪接事件，用 RPKM 自然能较准确的刻画基因表达水平。因此，我们认为 RPKM 方法暗含了读段在所研究区域内均匀分布这一 假设条件。然而在选择性剪接基因上，由于非组成性外显子并不包含于所有剪接异构体，则非组成性外显子的读段覆盖度可能比组成性外显子的读段覆盖度低。RPKM 方 法实际上是对各外显子读段覆盖度的加权平均，从而在选择性剪接基因上不能真 实反映基因的整体表达水平，导致低估。

其次基因表达平衡问题。个别表达量很高的基因，会引起其他低表达量的差异假阳性。假设2个样品A、B，二者差异表达的只有一个基因，差异量为rDE.由于数据量R相同，B的平均测序深度必将降低。则对于某个相同表达的基因g：

则A的RPKMg=(rg*109)/(flg*R);

且B的RPKMg =((rg*R/(R+ rDE)*109)/(flg*R);

二者显然不同（一般认为rDE<<R情况下可以无视)

再次基因数的影响。二个样品检测到的基因数不同，会影响RNA-seq结果。如样品A表达12000个基因，样品B则表达10000个基因。则只在A中表达的基因2000个必定是差异表达基因（相对B中表达量为0）。但样品总reads不一定A>B，因为其他基因的表达量差异未知。

2、FPKM

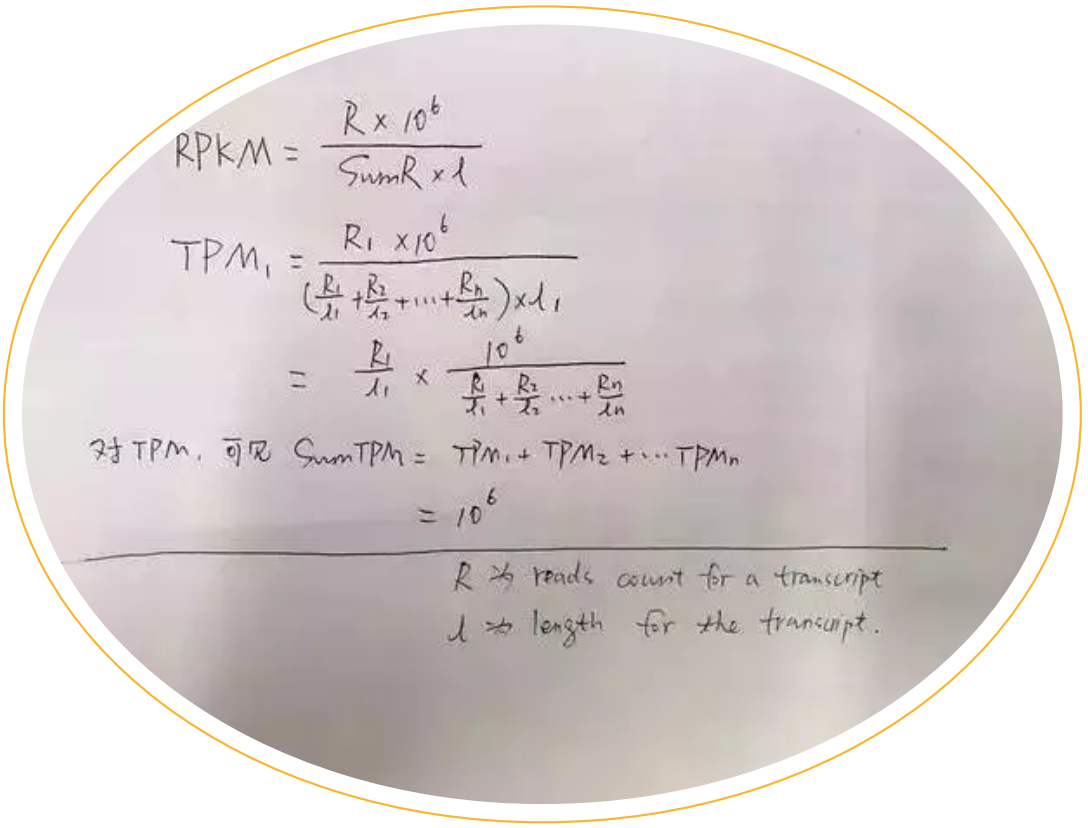
主要针对pair-end测序表达量计算。
F是fragments，R是reads，每个fragments会有两个reads，FPKM只计算两个reads能比对到同一个转录本的fragments数量。使用fragment作为统计单位，能够有效克服重复序列对比对的影响，提高unique map的比例，定量会更加准确。

3、TPM

Transcripts per million (TPM):优化的RPKM计算方法，TPM可以用于同一物种不同组织间的比较。

那么什么是TMP呢？

公式是下面这样的



R1：map到该基因的read数目

l1：该基因外显子长度

R2 R3.. Rn代表其他所有基因

由此可知，TPM概括了基因的长度、表达量和基因数目。TPM可以用于同一物种不同组织间的比较，因为Sum(TMPs)的值总是唯一的。

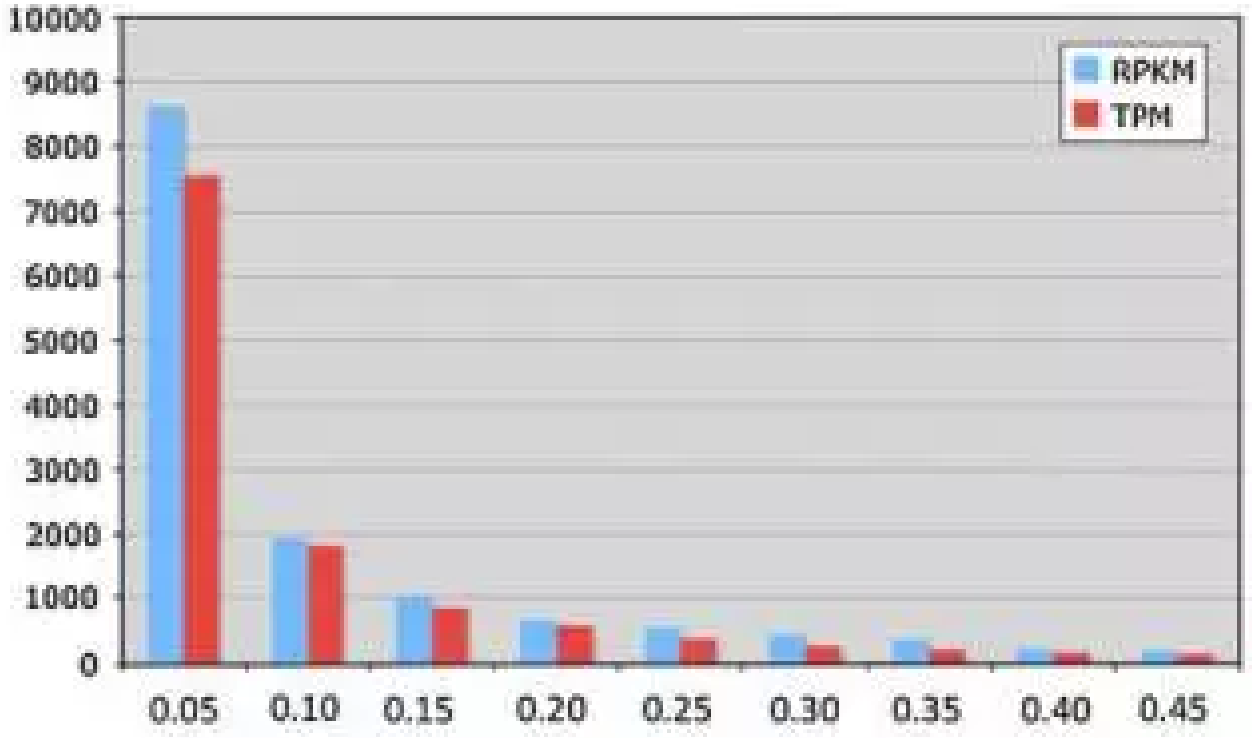


Fig. 2 *p* value distribution of RNA abundance data from human chondrocytes and myometrial cells for data expressed in RPKM and TPM. The *p* values were calculating from two-tailed *t* test assuming different variances. The *p* values are binned in 0.05 bins. Note that *t* tests using RPKM lead to higher number of low *p* values as expected if RPKM introduces artifactual differences in RNA abundance measures between samples

上图描述的是用TPM与RPKM分析两种人体细胞样品的表达量差异，所得结果进行t检验后得到的p值的分布。可以看到，RPKM相对TPM，明显较高P值的差异结果较多。说明RPKM可能引入了人为的表达量差异。具体文献参见：

Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples

近期热帖：

生信人春联作品展

生信人春联作品展

转录组分析常用软件汇总--精华版

miRNA-LncRNA-单细胞RNA分析软件汇总


ShengXin Ren