基迪奥生物
GENE DENOVO

# RNA-seq项目设计：生物学重复和单个样本测序量对结果的影响

基迪奥生物
GENE DENOVO

# 前言

在RNA-seq项目设计过程中，老师经常会问两个问题：
1）低丰度的基因是否能够被检测到（有或无）；
2）基因定量的结果是否准确（高或低）；

毫无疑问，提高生物学重复或提高单个样本测序量，都可以改善这些问题。但在研究经费有限的情况下，“提高生物学重复数而降低单个样本的测序量”或“提高单个样本测序量而降低生物学重复”，哪个更有效？

我们经常会建议老师：“3个生物学重复样本（2G/样本）的定量准确性 大于 单个样本6G数据量。即相同的总数据量拆分到更多的生物学重复中，实际上定量可靠性是提高了。”这个结论的出处是哪里？ 下面，我们通过一篇参考文献解答这个问题。

Robles et al. BMC Genomics 2012, 13:484
http://www.biomedcentral.com/1471-2164/13/484

BMC Genomics

RESEARCH ARTICLE Open Access

# Efficient experimental design and analysis strategies for the detection of differential expression using RNA-Sequencing

背景：这篇2012年 BMC genomiss的方法学文章，主要讨论了生物学或技术重复、单样本测序量、分析算法这3类因素对RNA表达差异分析的影响；

结论：1）软件DESeq的效果优于edgeR或NBPSeq。

2）差异分析对数据量并不敏感，甚至当单个样本测序量降低为原来的15%的时候，也不会大幅度降低差异基因检出率。

3）增加生物学重复对提高表达差异分析结果可靠性的效果要优于单样本测序量，

备注：第二、三部分的内容，我们将在下文重点解读。

基迪奥生物
GENE DENOVO

# 判定差异分析结果可靠性的指标

**假阳性与真阳性**

直观一些说，如果某个基因在RNA-seq结果显示差异表达，但Qpcr结果表明这个基因表达差异不显著，可以认为这个基因RNA-seq结果为假阳性；反之，这个结果就是真阳性。

而老师往往会关心某些低表达基因的表达差异变化能否被正确检测，那么这就要求我们提高实验的**真阳性率**。

**假阳性率（FPR）**：真实非差异表达中的基因中，被错误判定为差异表达的比例，FDR越低越好；

**真阳性率（TPR）**：真实差异表达的基因中，能够正确判定为差异表达的比例，TPR越高越好；真阳性率这个概念，如果换用为“**差异基因的检出率**”更容易理解，下文我们会并用这两个概念。

基迪奥生物
GENE DENOVO

# 问题的重点

下文，我们将重点摘抄文章中三个方面的问题，并做总结：

1. 生物学重复对差异表达分析的影响；
2. 单样本测序量对差异表达分析的影响；
3. 总数据量不变，生物学重复数与单样本测序量最佳组合；

# 1. 生物学重复对差异表达分析的影响

如表1所示，在单样本测序量保持不变的情况下，随着生物学重复数（n）的提高，差异分析的假阳性率（FPR）基本稳定，但真阳性率（TPR）在不断提高。也就是说提高生物学重复数，实验对差异表达基因的检测更加敏感，那些差异倍数较小或表达量较低的差异表达基因（此类基因的差异检测难度较大）能够更容易被检测到。

**Table 1 Effects of biological replication on power to detect DE using DESeq**

| % | $n = 2$ | $n = 3$ | $n = 4$ | $n = 6$ | $n = 8$ | $n = 12$ |
|---|---|---|---|---|---|---|
| call rate % | 0.44 | 1.15 | 1.76 | 3.03 | 4.08 | 5.12 |
| FPR % | 0.04 | 0.06 | 0.06 | 0.06 | 0.05 | 0.04 |
| TPR % | 3.26 | 8.95 | 13.95 | 24.30 | 32.72 | 41.57 |

Effects of biological replication on power to detect DE using DESeq. FPR and TPR are defined in Eqs. 5 & 6 respectively at 1%. "call rate" is the total number of reported positives / the total number of transcripts. These values are also represented in Figure 3 at 100% sequencing depth.

# 2.单样本测序量对差异表达分析的影响

**Table 2 Effects of sequencing depth on FPR at different *n* and depths**

| Depth | *n* = 2 | *n* = 3 | *n* = 4 | *n* = 6 | *n* = 8 | *n* = 12 |
|---|---|---|---|---|---|---|
| 25% | 0.02 | 0.02 | 0.04 | 0.03 | 0.03 | 0.03 |
| 50% | 0.03 | 0.03 | 0.04 | 0.05 | 0.04 | 0.03 |
| 75% | 0.04 | 0.06 | 0.05 | 0.07 | 0.04 | 0.04 |
| 100% | 0.04 | 0.06 | 0.06 | 0.06 | 0.05 | 0.04 |

Effects of sequencing depth on FPR values for a subset of our tested depths = 25%, 50%, 75% & 100%.

**Table 3 Effects of sequencing depth on TPR at different *n* and depths**

| Depth | *n* = 2 | *n* = 3 | *n* = 4 | *n* = 6 | *n* = 8 | *n* = 12 |
|---|---|---|---|---|---|---|
| 25% | 1.57 | 6.24 | 10.40 | 19.18 | 26.08 | 35.41 |
| 50% | 2.58 | 7.63 | 12.40 | 22.34 | 29.66 | 39.16 |
| 75% | 3.01 | 8.47 | 13.16 | 23.44 | 31.57 | 40.65 |
| 100% | 3.26 | 8.95 | 13.95 | 24.30 | 32.72 | 41.57 |

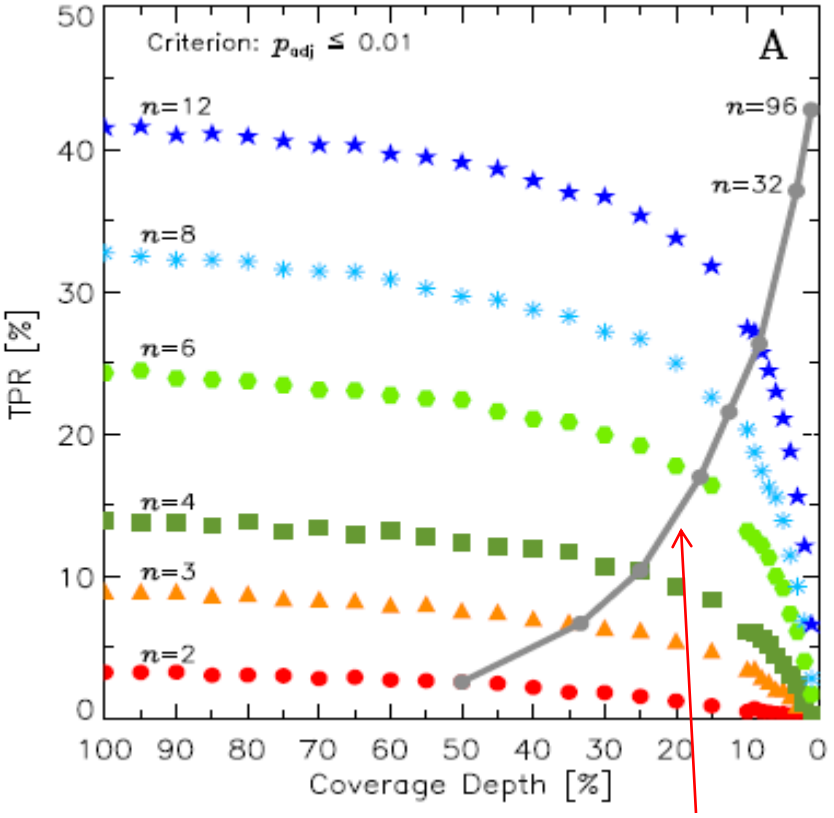
Effects of sequencing depth on TPR values for a subset of our tested depths = 25%, 50%, 75% & 100%.

如表2、表3所示，在一定的生物学重复数（n）的情况下，随着单样本测序量（Depth）的提高（25%→100%），假阳性率（FDR）和真阳性率（TPR）都只有有限的提高。例如在n=3的情况下，单个样本的测序量从25%提高到100%，FDR仅仅从0.02%提高到0.04%，TPR仅仅从6.24%提高到8.95%。

在表3中，如果Depth等于25%不变，当n从2提高到12，TPR的提高则是非常明显的。因此测序深度对结果改善效果并不如增加生物学重复。在下文，我们将详细比较。

# 3.总数据量不变，生物学重复数与单样本测序量最佳组合

Figure1（a）

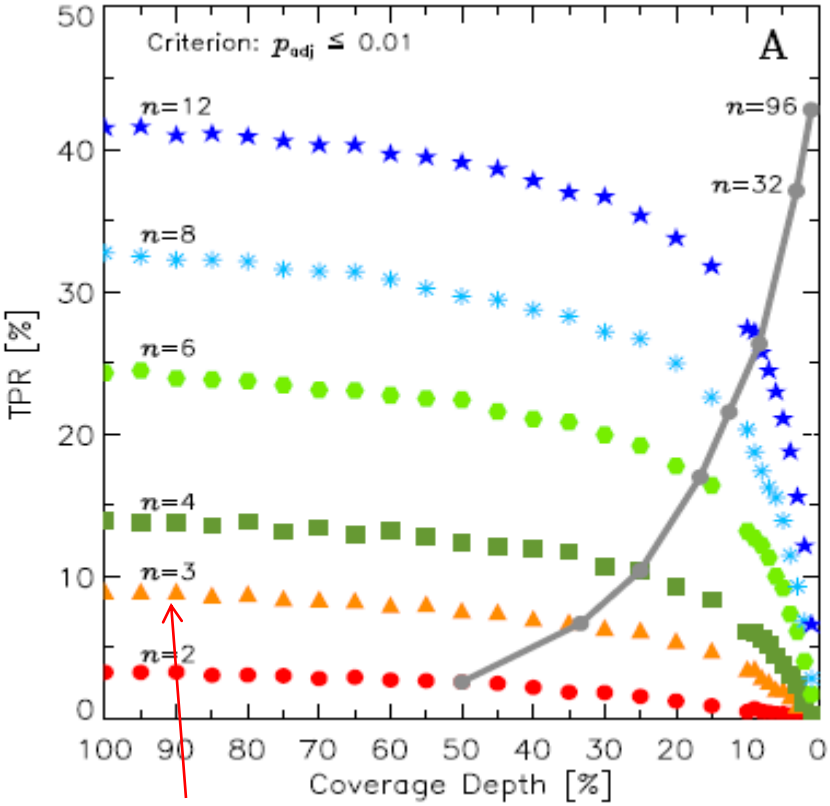



不同单样本测序量与生物学重复数组合，对应的TPR变化

如果保持总测序量不变（即如果生物学重复数为n，则单个样本的测序量降低为1/n，总数据量为n×1/n=1，保持不变）。
如图1（a），灰色实线代表不同生物学重复数（n）和单样本数据量（1/n）组合的情况下，真阳性率（TPR）的变化。结果表明，随着n的提高，TPR率不断提高。例如，如果n=2，TPR约为3%，如果n=6，TPR则提高到22%

# 3.总数据量不变，生物学重复数与测序量最佳组合

Figure1（a）



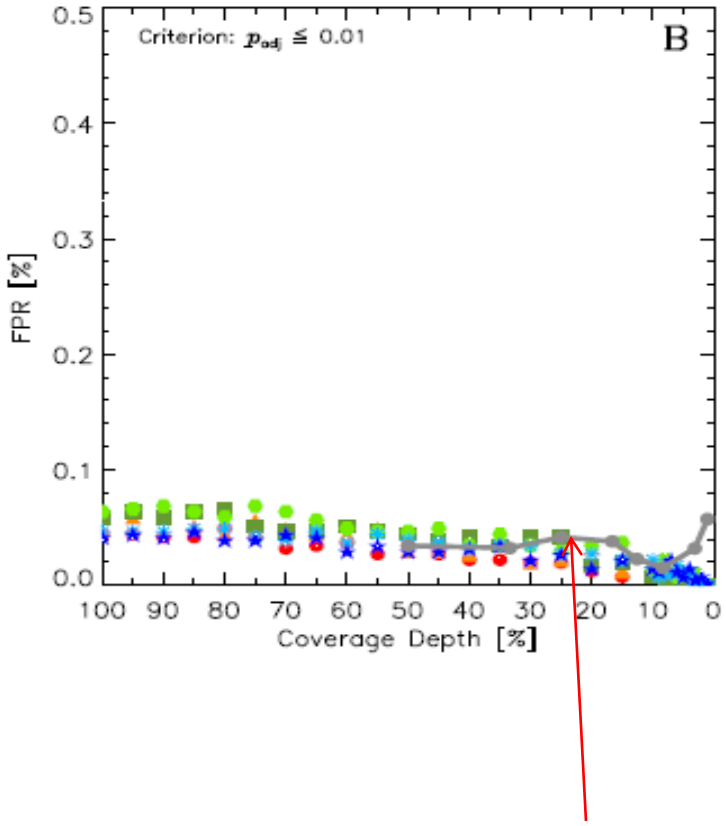

如果n=3固定不变，单个样本数据量降低，TPR的变化；

从图1（a），我们也可以对上文的第二点内容“单样本测序量对差异表达分析的影响”在进行深入观察。

我们也可以观察到如果n保持不变，但单个样本的数据量不断降低，TPR的降低十分缓慢。例如，当n=3，单个样本的数据量从100%降低到15%，TPR的值一直处于平台期，仅仅从9%降低到5%

# 3.总数据量不变，生物学重复数与测序量最佳组合

Figure1（b）




不同测序量与生物学重复数组合，对应的FPR变化

但是不同的生物学重复数和单样本测序量的组合，对假阳性率（FPR）的影响却较小。如图1（b），灰色实线代表不同生物学重复数（n）和单样本数据量（1/n）组合的情况下，真阳性率（FPR）的变化。虽然n从2变化到96，FPR基本没有太大变化。

从图中我们很容易发现，基于负二项分布的差异分析检验（P value），FPR对生物学重复数和单个样本数据量均不敏感，始终保持低于0.1%水平。或者说，这个算法对FPR的控制还是非常理想的。

# 讨论

## 1. RNA-seq老师关心的问题

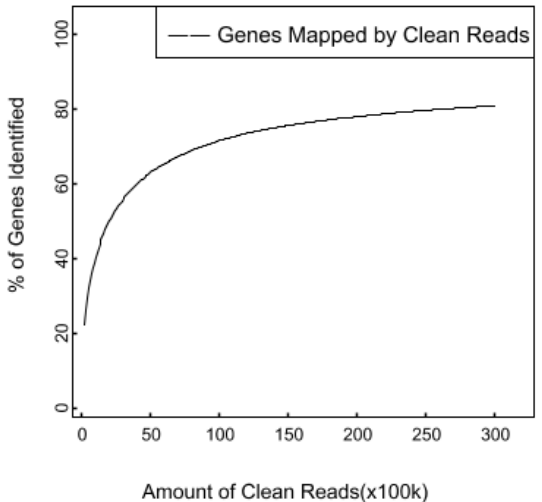
上文我们提到的老师关心两个问题：

1）低丰度的基因是否能够被检测到（有或无）；

2）基因定量的结果是否准确（高或低）；

大部分老师对第一个问题的关心程度要大于第二个，第二个问题常常被忽略。

但实际上，随着测序单价的下降，目前市场上RNA-seq类项目的单样本测序量正在不断提高。以2G，PE100测序的表达谱项目为例，其对应的测序量为20M条reads。如果一条长度为1kbp的低表达基因的表达量为RPKM=0.5，其理论上可以检测到的reads数为20×0.5=10。所以低丰度基因的检测，对RNA-seq这个技术来说并非最大问题。




如左图，大部分RNA-seq类项目，老师都会看到测序的饱和曲线达到平台期。也就是说再增加测序量，新检测出的基因数并不会有明显增加。

基迪奥生物
GENE DENOVO

# 讨论

第二个问题“转录本表达量的高低变化”比“转录本的有无”更具有普遍的生物学意义。虽然个别基因的表达量变化程度，可以使用Qpcr来验证。但我们往往也使用所有差异基因来统计某些规律。例如使用差异基因的pathway富集分析来寻找与性状相关的pathway。如果在全局水平的差异基因集并不可靠，那么pathway富集分析得出的结论的可靠性自然也受到影响。而全局水平的差异基因数量巨大，是难以使用Qpcr验证的。

因此，定量以及差异分析的准确性是在RNA-seq中更值得老师关心的问题。在讨论的第二部分，我们重点展开叙述。

基迪奥生物
GENE DENOVO

# 讨论

## 2. 重复数、单样本测序量的取舍

我们将前文提到的三个问题在进行总结：

1）生物学重复对差异表达的影响；

目前，主流期刊对生物学重复慢慢会有一定的要求。从本文，我们可以看到，设定生物学重复对差异基因的检出率（真阳性率，TPR）的提高具有明显效果。所以，设定生物学重复对提高结果的可靠性，是非常有意义的。

2）单个样本的测序量

老师对测序量比较关心，主要还是由于担心低丰度基因无法检测的问题。讨论的第一部分，我们也解释过，目前RNA-seq 的数据量（一般不低于2G，对于lncRNA测序，数据量一般更大）已经足以保证大部分低丰度基因的检测。而且，从本文我们可以看到，在其他条件不变的情况下，单样本数据量从100%降低到15%，差异基因的检出率（真阳性率，TPR) 降低较为平缓。所以，单样本数据量对RNA-seq定量和差异分析的影响实际上是十分有限的。

基迪奥生物
GENE DENOVO

# 讨论

3）总数据量不变，生物学重复数与单样本测序量最佳组合

由于大部分老师科研经费有限，无法无限制地增加样本数或数据量。所以在生物学重复数和单个样本测序量上必须找到平衡点。从本文我们可以看出，在总数据量不变的情况下，将总数据量分配到更多的生物学重复样本中，差异分析结果的可靠性在不断提升。这也与前两点得出的结论一致——对于RNA-seq，生物学重复数的价值要大于单个样本测序量。

但增加生物学重复的样本数，意味着要增加建库费用。因此，即使总数据不变，设置过多的生物学重复也是不合理的。一般而言，设定3个生物学重复，依然是最高性价比的选择。

基迪奥生物
GENE DENOVO

# 讨论

## 3. 其他

增加单样本数据量对定量的改良是有限的。但对于低丰度转录本*de novo*拼接（无参考基因组）或低丰度新转录本检测（有参考基因组），更高的数据量的确可以潜在改善拼接效果。

那么对于此类情况，我们可以采取以下策略：1）在拼接的步骤，我们可以将所有数据合并（例如每个生物学重复2G数据量，3个重复，全部合并），足够大的数据量来保证拼接效果；2）完成拼接后，在定量这个步骤，每个生物学重复样本独立定量。从而，可以在控制整个项目测序量的情况下，兼顾转录本拼接和定量这两个方面的问题。

这个策略也可以解释，对于lncRNA测序，如果不设置重复，我们建议老师单样本测序量为8~10G。如果设置了重复，而老师经费有限，那么可以将单个样本的数据量降低（例如5~6G），其效果依然要优于不设置重复的实验设计。

基迪奥生物
GENE DENOVO

# 用我们的服务，加速您的研究！

扫一扫

**关注基迪奥微信公众号**

- **订阅高通量测序前沿资讯**
- **有机会免费获得生物信息学培训**