PIBB 生物化学与生物物理进展
Progress in Biochemistry and Biophysics
2019,46(12):1202~1208
www.pibb.ac.cn

测序深度对RNA编辑识别算法的影响*

赵诚辉 洪 浩 李宛莹 李睿江 江 帅 李 昊 陈河兵** 伯晓晨**
（军事科学院军事医学研究院辐射医学研究所，北京 100850）

摘要 RNA编辑是重要的转录后修饰过程，目前已有多种算法用于识别RNA编辑，本文主要研究小鼠中测序深度对RNA编辑识别算法的影响，从而为RNA编辑的研究给出建议的方法. 本文使用STAR比对软件将小鼠的RNA-seq数据进行序列比对，然后使用GATK识别SNV，并用Separate Method、GIREMI、RNAEditor 3种方法识别出RNA编辑位点. 最后对3种方法识别RNA编辑位点的共同部分、识别效率、识别稳定性、识别与测序深度的关系进行分析. 结果发现3种方法识别的编辑位点数目差异大，共有位点较少，随着测序深度的增加，识别的RNA编辑位点数也在增加. 结果表明RNA编辑识别算法在小鼠中的识别性能与测序深度呈正相关.

关键词 RNA编辑，测序深度
中图分类号 Q-31
DOI：10.16476/j.pibb.2019.0036

RNA编辑是一个发生在转录后、翻译前，通过插入、删除或替换碱基从而改变从DNA模板，转录出RNA的生物现象[1]. 它在分子过程中发挥着作用，包括基因表达的调节以及非编码RNA的加工[2]. 同时它也被发现与人类的大脑发育以及神经系统疾病相关[3]，在很多癌症的发生中也发挥着作用[4]. RNA编辑还极大地增加了转录组和蛋白质组的多样性，从而在生物对环境的适应中发挥作用[5]. RNA编辑广泛地发生于原生生物中[6]，而由ADAR（adenosine deaminase acting on RNA）酶介导的A（adenosine）-to-I（inosine）编辑是最重要且发生最广泛的RNA编辑，也是目前RNA编辑主要的研究对象[7]. 肌嘌呤（I）通常在测序中被识别为鸟嘌呤（G）. 因此，A-to-G（Guanine）编辑也就是A-to-I编辑.

近年来，高通量RNA测序（RNA-Seq）技术发展，使得编辑位点的识别有了巨大的进步. 迄今为止，已在各种人体组织和实验条件下检测到数百万个RNA编辑事件[8]. 尽管有了这些进步，RNA编辑位点的全面识别仍然具有挑战性，编辑事件的组织特异性基因表达模式、相对高的测序错误率以及对潜在的基因组变异混淆分析等都是RNA编辑位点识别的难点[9]. 现今在各个物种中已经有十数种RNA编辑位点识别算法，常用的算法有Separate[10]、GIREMI[11]、JACUSA[12]、REDItools[13]、RES-Scanner[14]、RNAEditor[15]等. 这些算法原理不同，优缺点也各不相同. 但是在识别过程中，其识别效果均与RNA-seq的测序深度存在着明显的关联[16].

由于目前小鼠的RNA编辑识别缺乏数据金标集，因此评估算法时的敏感性和特异性等指标无法计算. 故我们选择对可靠性高且应用广泛的3种RNA编辑识别方法：Separate、GIREMI以及RNAEditor，在不同测序深度下的识别结果进行比较. 给出了3种方法在小鼠中识别的数目差异，以及不同测序深度数据下的算法稳定性.

1 材料与方法

1.1 小鼠RNA编辑数据获取

我们从Mouse ENCODE（encyclopedia of DNA elements）计划中下载了成熟小鼠4个细胞系

* 国家自然科学基金重点项目(U1435222)资助.
** 通讯联系人. Tel：010-66932251
伯晓晨. E-mail：boxiaoc@163.com
陈河兵. E-mail：chb-1012@163.com
收稿日期：2019-07-31，接受日期：2019-11-01

（Limb、Liver、Placenta、Whole Brain）的RNA-seq高通量测序数据. 然后使用3种方法对小鼠的RNA编辑位点进行了识别.

我们从UCSC Genome Browser上获取了小鼠的基因组参考序列文件（GRCm38/mm10），并从GENCODE上获取了小鼠的基因组注释文件（vM17）.

1.2　RNA编辑位点识别方法

Separate方法通过对从单个个体获得的基因组DNA和RNA序列进行细致分析，可以稳健地鉴定RNA编辑位点. 我们首先使用BWA（Burrows-Wheeler algorithm）将RNA-Seq数据与mm10的参考基因组以及剪接点区域的外显子序列对齐. 选择的剪接点区域长度短于Reads长度以防止多余的识别. 我们使用Picard的MarkDuplicates工具删除映射到相同位置的Reads（PCR重复）. 然后使用GATK工具围绕插入或缺失多态性进行局部重排，并重新校准基础质量分数，用GATK的UnifiedGenotyper工具（stand_call为0，stand_emit为0）求出变异位点，再从这些变异位点中删除dbSNP中已知的SNP，并且丢弃了每个reads中前6个碱基中的变异，避免人为的错配. 最后，排除了基因组高变区的变异位点，剩下的变异位点就是Separate方法求得的RNA编辑位点.

GIREMI方法使用SNV之间的等位基因连锁来检测候选编辑位点，并使用广义线性模型提高预测能力. 它结合RNA-seq读数中SNV对之间的互信息，然后通过统计推断与机器学习来预测RNA编辑位点. 对于GIREMI方法，需要使用BWA算法将RNA-seq数据比对到基因组上，同样的，我们使用GATK求出变异位点. 在这些变异位点中，对于每个错配位置，需要≥5的总读取覆盖率，并且要求变体等位基因存在于至少3个读数中. 然后丢弃了以下类型的错配：位于简单重复区域的或≥5 nt的均聚物；与偏向于单链的读数；具有极端变异等位基因频率（> 95%或<10%）和位于已知拼接交界处4 nt内的读数. 最后，将过滤后的结果输入GIREMI工具来识别RNA编辑.

RNAEditor方法可以在没有编程知识的情况下使用. RNAEditor接受FASTQ文件作为输入，并完全自动化识别RNA编辑事件. RNAEditor支持命令行以及图形界面两种模式，在图形界面中只需将感兴趣的FASTQ文件放入RNAEditor即可进行分析. 此外，RNAEditor实现了聚类算法来检测高度编辑的位点区域，并将其命名为“编辑岛”. 与单个编辑的站点相比，编辑岛表示的潜在ADAR结合位点，具有更高的可信度，更具有生物学意义的可能性. 3种方法的具体比较见表1.

Table 1　Comparing methods for RNA A−to−I editing site identifying

Method	Separate	GIREMI	RNAEditor
Required dependencies	STAR、GATK、SAMTOOLS	HTSlib、SAMtools、R	Pysam、pyqt4、matplotlib、numpy、BWA、Picard Tools、GATK、BLAT、BEDtools
Input	Fastq	Bam	Fastq
Strand orient	single end、pair end	single end、pair end	single end、pair end
DNA-RNA comparison	No	No	No
Multiple types of RNA editing	Yes	Yes	Yes
Sequence alignment	STAR	No	BWA
Annotation of RNA editing sites	No	No	Yes
SNV calling	Yes	No	Yes
De novo detection of RNA editing sites	Yes	Yes	Yes

1.3　小鼠RNA编辑识别流程

为了能够检验3种方法在不同测序深度下的识别效果，我们将得到的RNA-seq进行了非放回采样，从而可以得到不同测序深度的数据. 然后使用3种方法对得到的采样数据进行了RNA编辑位点的识别，最后再对识别的结果进行了详细分析，流程如图1所示.

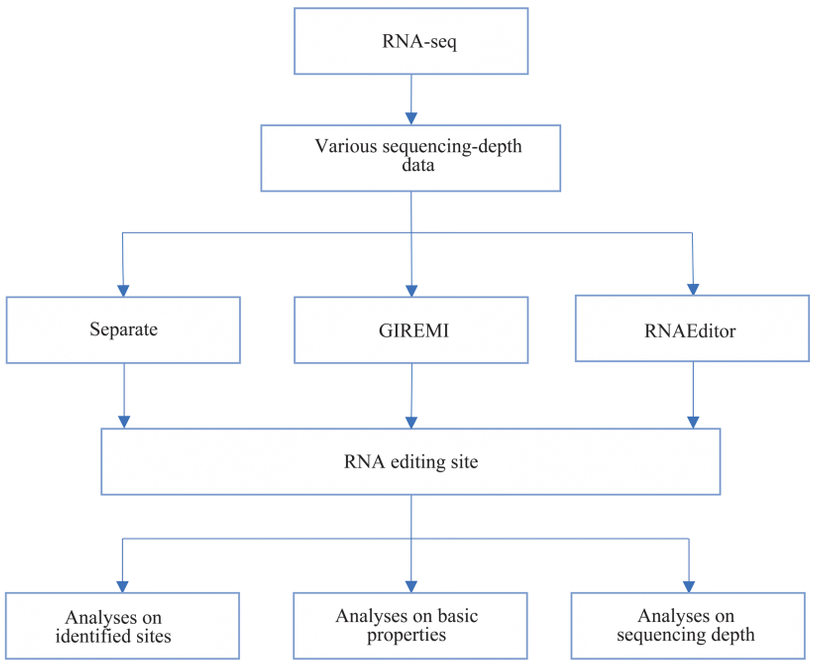

Fig. 1　Pipeline of data processing

2　结　　果

2.1　三种方法RNA编辑识别结果比较

我们使用Separate、GIREMI、RNAEditor 3种方法识别了来自成熟小鼠的4个细胞系Limb、Liver、Whole Brian以及placenta的RNA-seq数据，识别结果见表2和图2. 从表中可以看出3种方法在

Table 2　Comparing of identified RNA editing sites by three methods across four cell line

Cell line	Limb	Liver	Whole Brain	Placenta
Separate	2 096	1 239	3 937	1 741
GIREMI	435	300		492
RNAEditor	15 406	10 558	22 295	12 223

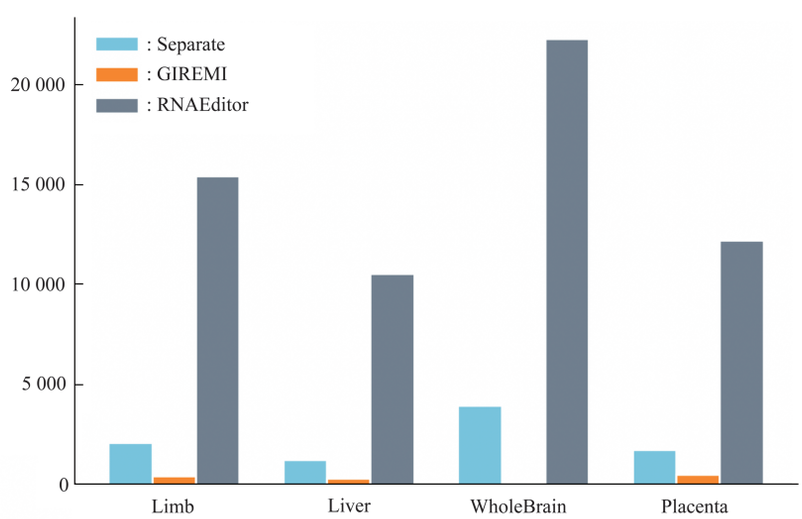

Fig. 2　Bar-chart of identified RNA editing sites using the three methods

识别RNA编辑时的个数差异很大. Separate方法识别出的RNA编辑位点数量级在10^3左右；GIREMI算法识别出的位点数量级在10^2，且有一个细胞系未能识别；RNAEditor算法识别出的RNA编辑位点数量最多，数量级在10^4左右.

从细胞系的角度来说，在Whole Brain细胞系中识别出的RNA编辑位点最多，Limb次之，Placenta第三，Liver细胞系最少.

2.2　小鼠RNA编辑识别结果的基本性质

在使用3种方法识别得到4个细胞系的RNA编辑位点后，我们对3种方法识别得到的位点进行了统计分析. 发现3种方法的共有位点较少（图3）. 其中Separate和GIREMI方法的共有位点数目约是其识别位点的50%. RNAEditor方法识别出的位点较多，但是与GIREMI、Separate重合的结果只有一小部分.

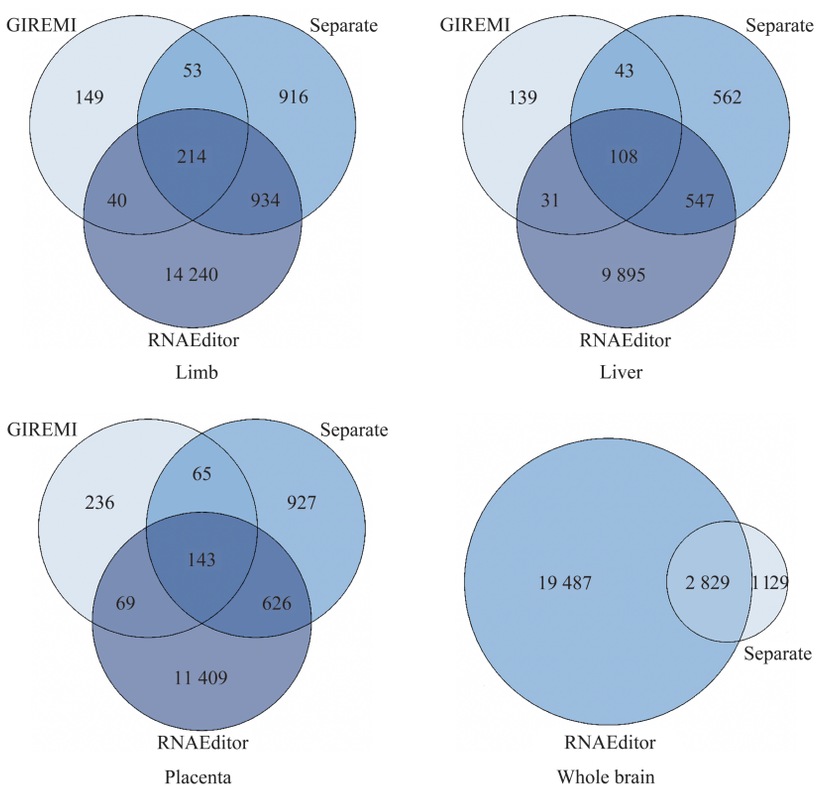

Fig. 3　Venn diagrams of identified editing sites by the three methods

这说明在识别结果上，3种方法各有优势，并不能够互相取代. 这是由算法特性所导致的. Separate算法采用了严格的生物信息学过滤流程进行RNA编辑的识别；GIREMI是利用机器学习中的互信息方法进行识别；而RNAEditor则在过滤后使用聚类进行识别. 算法原理的不同导致了其结果并不能够相互覆盖. 因此在使用算法识别RNA编辑时，最好能够使用多种方法，这样可以保证结果的准确性.

我们还统计了同一种方法，在不同细胞系之间识别RNA编辑的差异. 通过维恩图（图4）可以看出，3种方法识别的结果在不同细胞系之间也存在很大差异. 各个细胞系之间，只有10%左右的RNA编辑位点是重合的，剩下的编辑位点都是细胞特异的RNA编辑位点. 这说明了RNA编辑作为转录后调控因素，在不同的细胞中发挥作用的RNA编辑是不同的. 以后的研究过程中，也许可以将RNA编辑作为不同细胞的标志.

2.3　小鼠RNA编辑识别与测序深度

识别RNA编辑位点的算法与测序深度之间存在着明显的关联，因此我们特地通过实验来验证3种识别算法与测序深度之间的关系. 结果显示（图5），RNAEditor和Separate算法识别出的RNA编辑位点个数与测序深度呈明显的线性正相关关系，在4个细胞系中均是如此. 而GIREMI算法的识别结果与测序深度则完全没有关系，这与GIREMI算法

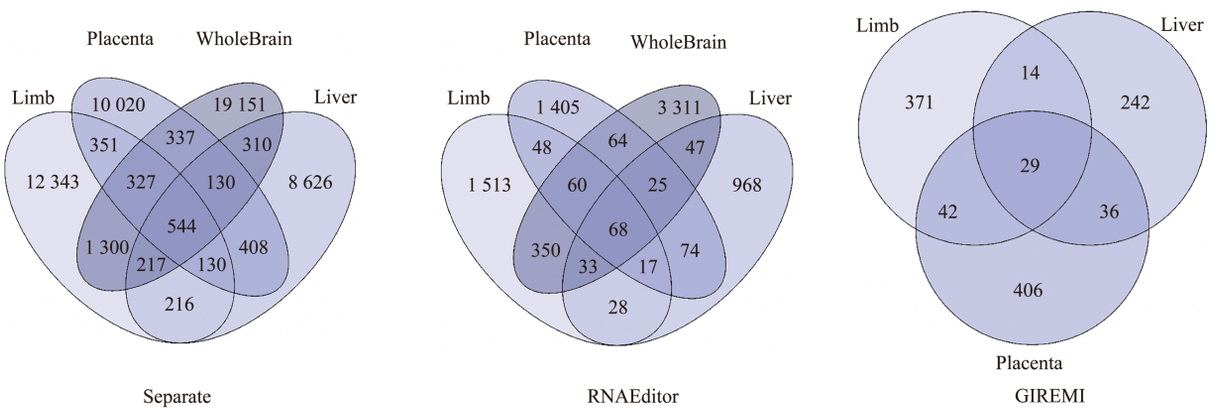

Fig.4　Venn diagrams of identified editing sites across four cell lines data

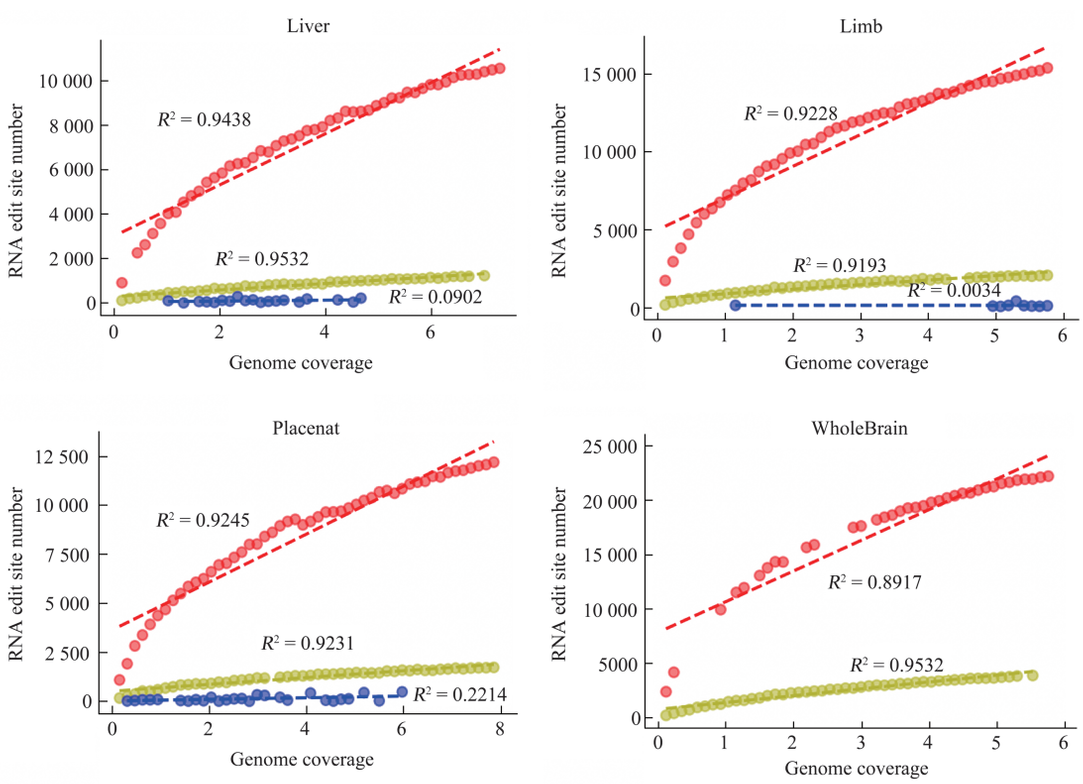

Fig. 5　The relationship between identification results and sequencing depth

Red: RNAEditor; Yellow: Separate; Blue: GIREMI.

在小鼠RNA编辑位点的识别不稳定有关. 很多样本GIREMI算法都没有办法计算出结果. 相较而言，RNAEditor和Separate则要稳定得多，在各种测序深度下都能够稳定地识别出结果，只有个别的样本会出现无法识别的情况.

我们的结果显示，在当前数据的测序深度（6~8倍基因组覆盖）下，RNA编辑的识别与测序深度仍呈明显的线性正相关关系，因此在进行RNA编辑研究时，一定要考虑数据测序深度的影响.

3 讨　论

随着高通量RNA测序（RNA-Seq）技术发展，人们已经能够获得多种样本的RNA测序数据. 然而如何从这些数据发现有意义的信息，破解生物的谜题是一个关键的问题. RNA编辑正是解决这一问题的一把钥匙，作为转录后的关键修饰过程，人们对于RNA编辑的认识还存在很大的研究空间. 我们已经知道在癌症以及神经系统疾病中RNA编辑的水平会显著升高，但这中间的机制仍然有待进一步研究.

在进一步研究RNA编辑的作用机制前，我们必须保证所识别RNA编辑位点的准确性. 本文对如今常用的3种RNA编辑识别算法Separate、GIREMI、RNAEditor在小鼠不同测序深度RNA-seq数据中的识别结果进行了对比，从而可以给出后续研究工作建议选择的算法. 研究结果显示，3种方法的识别结果并不存在完全重合的情况，因此不能做相互替代. GIREMI在小鼠的RNA编辑识别中表现很不稳定，识别出的编辑位点最少. Separate和RNAEditor方法稳定，识别结果与测序深度呈明显的线性相关，因此在小鼠的RNA编辑识别中较为推荐. 研究过程中，建议综合几个算法的结果作为最终的识别结果.

目前RNA编辑的识别与测序深度密切相关，但是相关数据远远未达到使算法饱和的深度. 希望随着技术进步，我们可以获得测序深度足够深的数据，从而识别出该细胞系中所有的RNA编辑事件，以便对RNA编辑的作用机制开展更深入的研究.

参　考　文　献

[1] Licatalosi D D, Darnell R B. RNA processing and its regulation: global insights into biological networks. Nat Rev Genet, 2010, **11**(1): 75-87

[2] Nishikura K. A-to-I editing of coding and non-coding RNAs by ADARs. Nat Rev Mol Cell Biol, 2016, **17**(2): 83-96

[3] Hwang T, Park C K, Leung A K, *et al*. Dynamic regulation of RNA editing in human brain development and disease. Nat Neurosci, 2016, **19**(8): 1093-1099

[4] Baysal B E, Sharma S, Hashemikhabir S, *et al*. RNA editing in pathogenesis of cancer. Cancer Res, 2017, **77**(14): 3733-3739

[5] Liscovitch-Brauer N, Alon S, Porath H T, *et al*. Trade-off between transcriptome plasticity and genome evolution in cephalopods. Cell, 2017, **169**(2): 191-202. e111

[6] Danecek P, Nellaker C, McIntyre R E, *et al*. High levels of RNA-editing site conservation amongst 15 laboratory mouse strains. Genome Biol, 2012, **13**(4): 26

[7] Nishikura K. Functions and regulation of RNA editing by ADAR deaminases. Annu Rev Biochem, 2010, **79**: 321-349

[8] Picardi E, D'Erchia A M, Lo Giudice C, *et al*. REDIportal: a comprehensive database of A-to-I RNA editing events in humans. Nucleic Acids Res, 2017, **45**(D1): D750-D757

[9] Diroma M A, Ciaccia L, Pesole G, *et al*. Elucidating the editome: bioinformatics approaches for RNA editing detection. Brief Bioinform. 2019, **20**(2): 436-447

[10] Ramaswami G, Zhang R, Piskol R, *et al*. Identifying RNA editing sites using RNA sequencing data alone. Nat Methods, 2013, **10**(2): 128-132

[11] Zhang Q, Xiao X. Genome sequence-independent identification of RNA editing sites. Nat Methods, 2015, **12**(4): 347-350

[12] Piechotta M, Wyler E, Ohler U, *et al*. JACUSA: site-specific identification of RNA editing events from replicate sequencing data. BMC Bioinformatics, 2017, **18**(1): 7

[13] Picardi E, Pesole G. REDItools: high-throughput RNA editing detection made easy. Bioinformatics, 2013, **29**(14): 1813-1814

[14] Wang Z, Lian J, Li Q, *et al*. RES-Scanner: a software package for genome-wide identification of RNA-editing sites. Gigascience, 2016, **5**(1): 37

[15] John D, Weirick T, Dimmeler S, *et al*. RNAEditor: easy detection of RNA editing events and the introduction of editing islands. Brief Bioinform, 2017, **18**(6): 993-1001

[16] Picardi E, Horner D S, Chiara M, *et al*. Large-scale detection and analysis of RNA editing in grape mtDNA by RNA deep-sequencing. Nucleic Acids Res, 2010, **38**(14): 4755-4767

The Impact of Sequencing Depth on RNA Editing Identification*

ZHAO Cheng-Hui, HONG Hao, LI Wan-Ying, LI Rui-Jiang,
JIANG Shuai, LI Hao, CHEN He-Bing**, BO Xiao-Chen**
(*Institute of Radiation Medicine, Academy of Military Medical Sciences, Beijing* 100850, China)

Abstract RNA editing is an important post-transcriptional modification process. There are many algorithms used to identify RNA editing. This study of the sequencing depth's effect on RNA editing will provide a suggested method for RNA editing research. Mouse reference genome was mapped to the RNA-seq by STAR. Then identified SNV by GATK, finally RNA Editing sites were filtered by Separate method, GIREMI and RNAEditor. The results showed the identification of RNA editing varied in three methods. There was little overlap in results. As the sequencing depth increases, the identified number of RNA editing sites also increases. In conclusion, the identification of RNA editing sites is positive correlated with the sequencing depth.

Key words RNA editing, sequencing depth
DOI：10.16476/j.pibb.2019.0036

* This work was supported by a grant from The National Natural Science Foundation of China（U1435222）.
** Corresponding author. Tel：86-10-66932251
BO Xiao-Chen. E-mail：boxiaoc@163.com
CHEN He-Bing. E-mail：chb-1012@163.com
Received：July 31，2019 Accepted：November 1，2019