PIBB 生物化学与生物物理进展
Progress in Biochemistry and Biophysics
2012, 39(3): 282~293
www.pibb.ac.cn

基于转录组测序数据识别黑猩猩 RNA 编辑位点 *

王端青[1, 2] 何 涛[2] 汪 莉[2]** 王玉民[2] 邵卫东[1]**
([1] 苏州大学电子信息学院，苏州 215006；[2] 军事医学科学院生物工程研究所，北京 100071)

摘要 使用转录组测序(RNA-Seq)数据识别黑猩猩 RNA 编辑位点，探索了 RNA 编辑的识别机制以及潜在的功能影响．基于黑猩猩 RNA-Seq 数据与基因组序列的比对信息发现 RNA-DNA 错配位点，并构建编辑位点候选集．从中滤除基因组或转录组测序质量低的位点，其他的过滤条件包括 3′端测不准、覆盖度、SNP 位点以及估算的编辑水平．构建二项分布统计模型和 Bonferroni 多重检验滤除候选集中的随机错误，得到 RNA 编辑位点．选取落在已知基因上的编辑位点进行功能分析，并用 Two Sample Logo 软件分析编辑位点上下游序列的特征．识别出黑猩猩 12 种碱基替换型 RNA 编辑位点 8 334 个，其中有 41 个编辑位点改变原有的氨基酸，另有 3 个编辑位点落在 microRNA(miRNA)潜在靶基因的种子结合区．统计学分析表明，分别有 640 和 872 个 RNA 编辑位点存在组织和性别差异．上下游碱基频率分析表明，多种类型的编辑位点紧邻碱基具有显著偏好．结果显示， RNA 编辑在黑猩猩体内大量存在，且潜在具有重要的生物学功能，为进一步深入研究灵长类 RNA 编辑的机制奠定了基础．

关键词 黑猩猩，RNA 编辑，转录组测序，计算识别
学科分类号 Q344+.14，Q752
DOI: 10.3724/SP.J.1206.2011.00328

RNA 编辑是一种重要的转录后修饰事件，通过核苷酸的替换、插入或删除而改变初始转录本，产生与 DNA 模板之间存在差异的 RNA 序列[1]．RNA 编辑可发生于细胞核、线粒体、叶绿体和质粒中[2]，在植物、动物、真菌、原生生物、细菌和病毒体内都有 RNA 编辑现象被报道[3]．A-to-I 和 C-to-U RNA 编辑是最常见的两种编辑类型，它们都是由碱基脱氨所致[4]．C-to-U RNA 编辑主要发生在高等植物线粒体和叶绿体中[5]，而 A-to-I RNA 编辑广泛存在于脊椎动物体内[6-8]．RNA 编辑能改变氨基酸序列、改变翻译起始或终止密码子、破坏或新建剪接信号、影响 miRNA 前体的加工及成熟体的靶向功能[2, 9]，异常的 RNA 编辑与色素异常症、癫痫、抑郁症、急性白血病等疾病密切相关[10-12]．

2003 年以前只有少量的编辑位点被报道，它们是在生物学实验中被偶然发现的．近年来，大规模计算识别 RNA 编辑位点(主要是 ADARs 蛋白家族介导的 A-to-I RNA 编辑)通常采用序列比对法，即从转录组与基因组序列比对的错配位点中发现 RNA 编辑位点．2003 年，Hoopengardner 等[13]采用比较基因组学方法，识别并验证了果蝇 16 个编辑位点和人的 1 个编辑位点．2004 年，多篇报道[14-17]在人 EST/cDNA 序列与基因组比对的基础上，计算预测了人类转录组中存在大量的 A-to-I RNA 编辑位点，且这些位点绝大多数落在 Alu 序列上．基于 EST/cDNA 序列与基因组比对识别编辑位点的计算策略的提出，标志着大规模识别编辑位点的开始和广泛存在的 RNA 编辑事件日益被科学研究者重视．但是 EST 序列来源庞杂且存在很大偏性(组织之间、疾病与正常状态和不同转录区域)，测序成本较高，一定程度上限制了新 RNA 编辑位点识

* 国家自然科学基金资助项目(30800644), 国家高技术研究发展计划(863) (2007AA022204)和国家科技重大新药创制专项(2008ZXJ09007-001)资助.
** 通讯联系人.
邵卫东. Tel: 18913146453, E-mail: shaowd@suda.edu.cn
汪 莉. Tel: 010-66948793, E-mail: liwang@tsinghua.edu.cn
收稿日期：2011-07-14，接受日期：2011-11-04

别．新一代测序技术[18-20]的出现，大大推动了基因组学和转录组学的发展．其中 RNA-Seq 技术以其高通量、高灵敏性以及可扩展性等优势[20]，目前已经被成功用于 RNA 编辑位点识别[21-24]、新基因识别[25]、基因表达分析[26-27]、SNP 识别[28]、可变剪接识别[29-31]等．2009 年，Li 等[21]最先使用新一代测序技术识别 A-to-I RNA 编辑位点．作者同时测序同一个体的基因组和 7 个组织的转录组，排除了单核苷酸多态性的干扰，识别出 710 个落在非 Alu 序列的 A-to-I 编辑位点．同年，Wahlstedt 等[22]使用小鼠脑组织的 RNA-Seq 数据，研究 A-to-I RNA 编辑水平在脑不同发育阶段的变化，结果显示 RNA 编辑水平随着小鼠的发育显著增长．2010 年，Picardi 等[23]使用两种测序平台对葡萄线粒体的 mRNA 进行测序，结合两组测序数据，识别出 401 个落在编码序列和 44 个落在非编码 RNA 上的 C-to-U RNA 编辑位点，该方法显著降低了编辑位点的假阳性率．2011 年，Rosenberg 等[24]通过比较野生型小鼠和 *Apobec1* 基因敲除小鼠小肠组织的 RNA-Seq 数据，识别出 33 个 C-to-U 编辑位点．

黑猩猩是灵长类中最接近人类的类人猿动物．黑猩猩基因组序列已测序完成，人和黑猩猩的基因组相似度高达 98.8%[32]．通过对黑猩猩的研究，有助于了解人类自身，以及人类进化过程中的未解之谜．RNA 编辑作为一种重要的转录后修饰事件，影响各种转录后调控事件和翻译水平事件，丰富物种的转录组和蛋白质组．Paz-Yaacov 等[33]采用生物信息学和 454 测序比较研究了 3 种灵长类动物(人、黑猩猩和猕猴)8 个基因 Alu 序列上成簇存在的 100 个 A-to-I 位点的编辑水平差异，发现人脑中的编辑水平显著高于其他 2 种灵长类动物．本研究在全转录组范围开展黑猩猩编辑位点的计算识别工作，基于黑猩猩 RNA-Seq 数据，识别出 8 334 个碱基替换型 RNA 编辑位点，其中 41 个编辑位点改变氨基酸序列，3 个编辑位点落在 microRNA 潜在靶基因的种子结合区．统计学分析显示，640 个编辑位点具有组织差异，872 个编辑位点具有性别差异．另外，通过序列分析，发现不同类型的编辑位点上下游序列存在着显著的碱基偏好性．

1 材料和方法

1.1 数据来源

黑猩猩基因组序列及其相应的测序分数(panTro2 版)、RefSeq 基因标注从 UCSC 网站获得(http://hgdownload.cse.ucsc.edu/goldenPath/panTro2/)、SNP 数据下载自 NCBI dbSNP(http://www.ncbi.nlm.nih.gov/projects/SNP/, 第 132 版)．黑猩猩转录组数据来源于 Wetterbom 等的工作[25]，他们分别提取了 1 只雌性和 1 只雄性黑猩猩的脑和肝组织 RNA，并利用 RNA-Seq 技术在 AB SOLiD 2.0 平台上对其进行测序，其中雌性黑猩猩的脑和肝组织按 35 nt 和 50 nt 2 种长度分别进行了测序．相应的 RNA-Seq 数据下载自 EMBL 网站 ERA 数据库(ERA000160).

1.2 序列比对及软件参数设置

在进行黑猩猩转录组序列与基因组序列比对时，我们采用 TopHat(v-1.3.2)软件，可将跨剪接位点的 read 同时比对上．TopHat[34]是一款成熟的可用于剪接位点识别的软件，在内部调用第三方比对软件 bowtie．TopHat 不需要已知的基因注释信息，将第一步无法比对到基因组上的序列进一步分割成小片段，同第一步比对建立的剪接结合区进行二次比对．在本研究中，我们对序列比对参数进行如下设置：a．每条序列最多允许出现 2 个 color 错配(-v 2)；b．比对到基因组上的序列位置唯一(-g 1).

1.3 编辑位点筛选方法

碱基替换型 RNA 编辑的直观表现是基因组序列与转录组序列在对应位点发生碱基错配，错配位点就是潜在的 RNA 编辑位点．将黑猩猩 6 组 RNA-Seq 数据比对到黑猩猩基因组上，滤掉一条 RNA-Seq 序列含有 2 个或 2 个以上碱基错配的比对结果，从剩余比对结果中提取转录组序列与基因组序列错配的位点，保留仅有一种替换类型的错配位点构建编辑位点候选集．

雌性黑猩猩的脑组织和肝组织各有 2 组 RNA-Seq 数据，测序长度分别为 35 nt和 50 nt．测序使用同一批样品，不存在实验条件的差异，属于独立重复实验．为尽可能降低编辑位点的假阳性，在过滤前，我们取 2 次测序数据中错配位点的交集作为候选编辑位点．编辑位点候选集有 4 组数据，分别为雄性脑组织(BrainM)、雌性脑组织(BrainF)、雄性肝组织(LiverM)和雌性肝组织(LiverF).

由于基因组与转录组测序来源于不同黑猩猩个体，编辑位点候选集潜在含有单核苷酸多态性(SNP)，同时还可能有测序错误等导致的随机背景噪声．本研究用图 1 所示的方法对 4 组编辑位点候选集进行过滤，以降低编辑位点的假阳性．

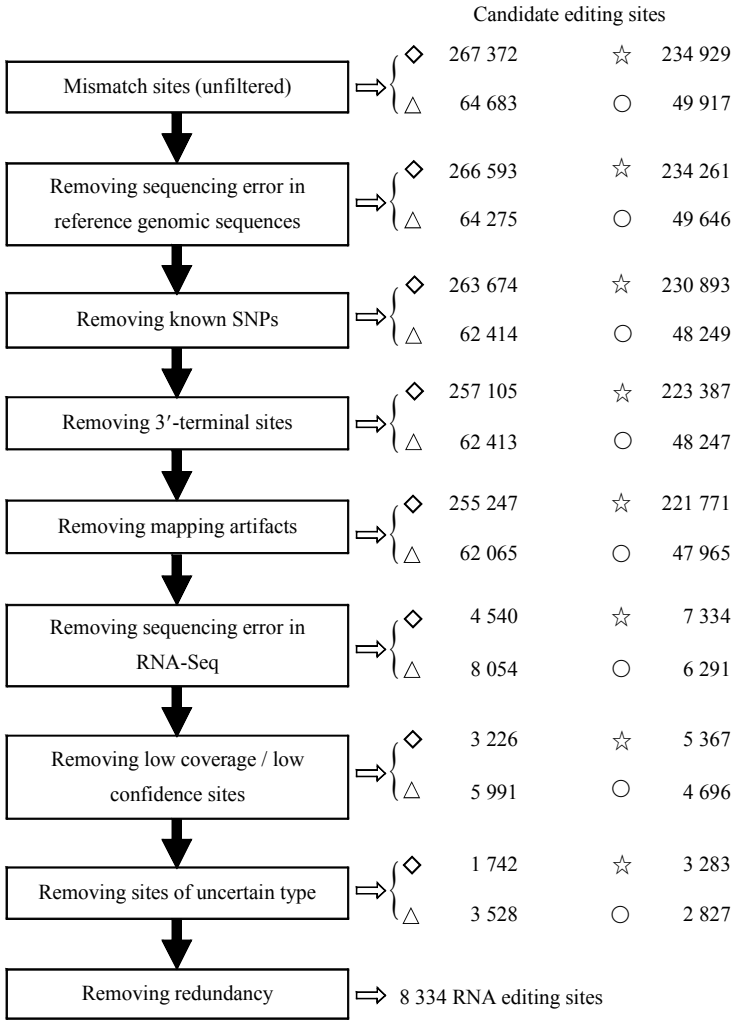

Fig. 1　Flows and results of RNA editing sites prediction

The left part of Figure 1 was a flowchart for identification of RNA editing sites. All RNA-DNA mismatch sites were regarded as a candidate set. Then several filters were performed on each candidate set in order to remove non-editing sites. The left sites by every filter step in each candidate set were shown on the right. ◇:BrainM; △:BrainF; ☆: LiverM; ○:LiverF.

a．滤除基因组测序错误．错配位点可能来自基因组测序错误，根据 UCSC 注释，测序分值低于 45 被认为测序质量不高，发生测序错误的概率较大．我们提取错配位点的基因组测序分值，滤除测序分值小于 45 的错配位点．

b．滤除已知 SNP 位点．从 NCBI 数据库下载到黑猩猩 SNP 序列，共 1 474 539 个 SNP 位点．用 BLAST 软件比对到黑猩猩基因组上确定 SNP 位点在 panTro2 的基因组位置，进而从编辑位点候选集中滤除已知的 SNP 位点．

c．滤除落在 RNA-Seq 序列 3′端的位点．由于新一代测序技术存在系统缺陷，序列 3′端碱基被测错的概率明显高于上游碱基．候选编辑位点如果落在覆盖该位点的转录序列的 3′端，则该条支持序列被删除，过滤后若序列比对不再支持错配，则该位点从候选集中被滤除．

d．滤除比对到内含子区域的 RNA-Seq 序列．由于转录组测序样品制备时采用 oligo(dT)合成模板链，故理论上测序得到的转录序列均来自于加尾后的成熟 mRNA，是不含内含子的．我们从比对结果中滤除与内含子有重叠的序列，同时修正了候选编辑位点的序列支持条数或删除了候选编辑位点．

e．滤除转录组测序错误．测序仪存在一定的测序错误率，尽管 SOLiD 测序仪精度理论上达到 99.94%，但是对于上百亿的碱基数据，产生的测序错误不容忽视．我们估算了平均测序错误率

(Pe=0.001 2)，每一个碱基是否被测错服从概率为 Pe 的二项分布．用二项分布统计检验和 Bonferroni 多重校正判断错配位点是否由测序错误引起，从编辑位点候选集里滤除测序错误位点(以修正后的 P-value 作为域值：雄性脑组织为 1.96×10^{-7}，雌性脑组织为 8.05×10^{-7}，雄性肝组织为 2.25×10^{-7}，雌性肝组织为 1.04×10^{-6}).

f．滤除可信度不高的编辑位点．为了降低识别的假阳性，我们保留至少有 5 条序列覆盖的编辑位点．另外，设置阈值为 5%，滤掉编辑水平低于该阈值的编辑位点．

g．滤除替换类型不定的编辑位点．本研究使用的 6 组 RNA-Seq 数据是无方向的，无法确定序列由基因组正义链或者反义链转录．我们保留落在已知基因上的编辑位点，根据基因的转录方向确定编辑位点的替换类型．由于黑猩猩已知基因(RefmRNA)不足 1 500 条，故我们依照 Tay 等[35]使用的方法，利用基因组版本转换工具 liftOver[36]将人的已知基因转换成黑猩猩基因组标注，有 27 126 个人类基因成功转换到黑猩猩基因组标注，看作黑猩猩参考基因．根据参考基因的转录方向，确定编辑位点的碱基替换类型．

1.4 编辑位点在已知基因上的分布

黑猩猩有 1 471 个已知基因，包括编码基因和非编码基因．从黑猩猩 RefSeq 基因标注提取编码基因的 5′UTR、CDS 和 3′UTR 在基因组上的坐标，对于非编码基因仅提取基因全长坐标．将编辑位点坐标同基因的不同区域比较，得到黑猩猩编辑位点在已知基因上的分布．

1.5 黑猩猩 miRNA 靶标位点识别

黑猩猩 3′UTR 区碱基序列从 NCBI 网站获得，黑猩猩 miRNA 成熟体下载自 miRBase[37]第 17 版，含 525 个 miRNA．使用 miranda(v3.3a)软件预测 miRNA 在 3′UTR 区的靶标位点，限定 miRNA 种子结合区与靶标位点必须完全互补配对，最小自由能在–15 以下，其他参数使用默认值．

1.6 编辑位点的组织差异性和性别差异性

我们以 RNA 编辑位点编辑前后碱基在脑和肝组织中的 read 支持数来估算不同组织的编辑效率，同样的策略也用于估算雄性和雌性的编辑效率．使用 Fisher 检验衡量编辑位点在组织和性别间编辑水平差异的统计显著性，再结合 Bonferroni 多重校正来识别组织或性别差异的 RNA 编辑位点(以修正后的 P-value 作为域值：组织差异为 7.53×10^{-6}，性别差异为 7.05×10^{-6}).

1.7 编辑位点上下游序列碱基特征分析

编辑位点按照碱基替换类型的不同可分成 12 类，使用 Two Sample Logo 软件[38]分析各类型编辑位点所在序列的特征，该软件能分析出阳性样本和阴性样本在相同位点上是否存在统计学差异．阳性样本集为黑猩猩编辑位点所在序列，序列以编辑位点为中心，长度为 11 nt．阴性样本集从黑猩猩 RefmRNA 序列中获取，mRNA 上任何 11nt 序列均可作为阴性样本．Two Sample Logo 软件参数设置：显著性水平设为 0.005，统计模型选用 t-test，选用 Bonferroni 校正．

2 结　果

2.1 转录组序列比对信息

黑猩猩 6 组样品经测序后分别得到 3 800 万至 1.7 亿条序列，序列总条数超过 5 亿条，如表 1 所示，接近 1.5 亿条序列成功比对到黑猩猩基因组上，占序列总数的 29.2%．两组测序长度为 50 nt 的 RNA-Seq 数据都有部分序列被比对到剪接结合区，其中 BrainF 有 493 202 条, LiverF 有 1 419 061 条，占各自总条数的 1.3%和 2.4%．比对结果显

Table 1 Mapping summary for all RNA-Seq data

Sample	Read length	Raw reads	Non-uniquely mapped reads	Unmapped reads	Uniquely mapped reads
BrainM	35	88 598 445	14 103 998(15.9%)	47 075 555(53.1%)	27 418 892(31.0%)
LiverM	35	78 533 657	8 681 175(11.0%)	48 114 479(61.3%)	21 738 003(27.7%)
BrainF	35	170 016 027	20 708 836(12.2%)	99 274 150(58.4%)	50 033 041(29.4%)
BrainF	50	38 733 951	2 348 370(6.1%)	23 712 052(61.2%)	12 673 529(32.7%)
LiverF	35	77 388 286	8 739 248(11.3%)	47 797 779(61.8%)	20 851 259(26.9%)
LiverF	50	58 610 173	2 836 402(4.8%)	39 212 311(66.9%)	16 561 460(28.3%)
Total		511 880 539	57 418 029(11.2%)	298 066 048(59.6%)	149 276 184(29.2%)

Chimpanzee RNA-Seq sequences were mapped to chimpanzee genome sequences by TopHat software with allowing up to two color mismatches in each alignment.

示，11.2%的序列由于在基因组上有多个位置而被过滤掉，6 组数据中该比例从 4.8%至 15.9%不等，4 组测序长度为 35 nt 的数据中这项比例都在 10%以上，明显高于测序长度为 50 nt 的数据，反映出序列越长在基因组上出现多个拷贝的概率越小．剩余 59.6%的序列没能比对到基因组上，部分由于测序质量不高而被滤掉，另外，由于比对参数设置仅允许 2 个 color 错配，致使部分序列无法比对到基因组．

2.2　编辑位点识别

我们获取转录组与基因组的错配位点构建编辑位点候选集，由于存在测序错误、比对错误以及 SNP 等干扰因素，候选集里含有大量非编辑位点．针对不同影响因素，设置了不同的过滤条件，从候选集里识别了非冗余的 8 334 个潜在的 RNA 编辑位点(BrainM、LiverM、BrainF 和 LiverF 中编辑位点数量分别为 1 742、3 283、3 528 和 2 827 个).

按照碱基替换类型的不同，编辑位点可分成 12 类．我们将 BrainF、BrainM、LiverF 和 LiverM 以及总的编辑位点按照不同的碱基替换类型统计，结果如图 2 所示．A->G、G->A、U->C 和 C->U 4 种类型的编辑位点数量明显多于其他 8 种替换类型，说明碱基转换型 RNA 编辑比碱基颠换型 RNA 编辑更容易发生，这与先前的研究结果一致[15].

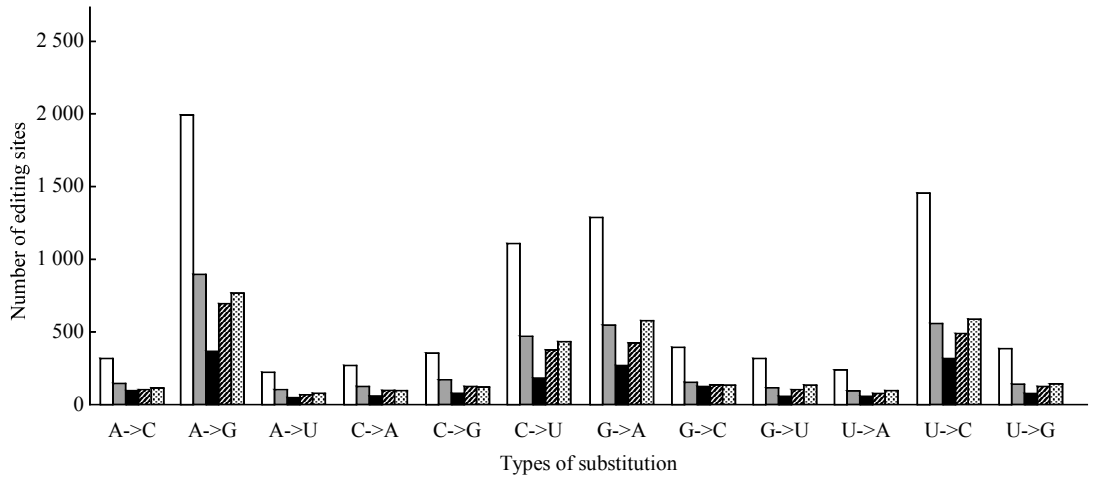

Fig. 2　The distribution of RNA editing sites in chimpanzee

With gene orientation information, we observed all 12 possible categories of base differences between RNA and its corresponding DNA. □: Total; ▩: BrainF; ■: BrainM; ▨: LiverF; ▦: LiverM.

2.3　编辑位点在黑猩猩 RefSeq 基因中的分布

落在黑猩猩已知基因上的编辑位点有 127 个，1 个落在非编码 RNA 上，其余 126 个编辑位点落在编码蛋白质的基因上．按照标注的基因结构，我们统计出落在 5′UTR 区、CDS 区和 3′UTR 区的编辑位点个数．如表 2 所示，落在 CDS 区的编辑位点数量最多(61 个)，其次为 3′UTR 区(59 个)，而落在 5′UTR 的编辑位点仅有 6 个．CDS 区的编辑位点潜在改变原有密码子，可能影响原蛋白质的功能或活性．3′UTR 区的编辑位点潜在影响 mRNA 的稳定性以及 miRNA 的靶结合位点．

Table 2　The distribution of editing sites in known gene

RefSeq gene		127 editing sites	Number of gene
Protein-coding gene	5′UTR	6	3
	CDS	61	50
	3′UTR	59	41
Non-protein-coding gene		1	1

The editing sites that resided in protein-coding genes were specified into 5′UTR, CDS or 3′UTR.

2.4　编辑位点改变原有的氨基酸

如表 2 所示，落在 CDS 区的编辑位点有 61 个，落在 50 个不同基因上．我们根据黑猩猩基因注释，获取 50 个基因的 CDS 区坐标范围，找出每个编辑位点在 CDS 区的相对坐标，从而得出编辑位点落在第几个密码子上以及密码子第几位．如表 3 所示，CDS 区的 61 个编辑位点有 41 个改变了原有的氨基酸，分布在 34 个基因上，其中 *LOC456645*、*AGT*、*EIF3L*、*UBC*、*SCG5* 和 *RHAG* 5 个基因各含至少 2 个编辑位点，RNA 编辑将使

Table 3 The alteration of amino acids by 41 RNA editing sites

Position	Type	Position in CDS	Position in Codon	Gene	Codon change	Amino acid alteration	*P*_value
chr1_+_25547158	A->G	457	1	*LOC456645*	AUG-GUG	Met-Val	<1.0×10^{-16}(LiverF)
chr1_+_25559021	A->G	916	1	*LOC456645*	AUC-GUC	Ile-Val	<1.0×10^{-16}(LiverF)
chr1_+_25588788	A->G	1240	1	*LOC456645*	AAU-GAU	Asn-Asp	<1.0×10^{-16}(LiverF)
chr1_-_88312395	A->G	409	1	*SEP15*	AAU-GAU	Asn-Asp	1.4×10^{-14}(LiverF)
chr1_+_150549854	G->A	895	1	*FMO3*	GUA-AUA	Val-Ile	1.7×10^{-8}(LiverF)
chr1_-_211313104	G->A	778	1	*AGT*	GUG-AUG	Val-Met	1.1×10^{-16}(LiverF)
							9.6×10^{-8}(LiverM)
chr1_-_211313353	A->G	529	1	*AGT*	ACA-GCA	Thr-Ala	<1.0×10^{-16}(LiverM)
chr1_+_220793035	C->G	1723	1	*CHRM3*	CAG-GAG	Gln-Glu	1.7×10^{-15}(BrainM)
chr10_+_17875532	U->A	931	1	*VIM*	UGG-AGG	Trp-Arg	<1.0×10^{-16}(BrainM)
							<1.0×10^{-16}(LiverM)
chr10_+_42720282	C->U	308	2	*EIF3L*	GCU-GUU	Ala-Val	<1.0×10^{-16}(BrainF)
							2.4×10^{-15}(BrainM)
							<1.0×10^{-16}(LiverF)
							2.4×10^{-15}(LiverM)
chr10_+_42720587	G->C	613	1	*EIF3L*	GGU-CGU	Gly-Arg	<1.0×10^{-16}(BrainF)
							<1.0×10^{-16}(BrainM)
							7.2×10^{-11}(LiverF)
chr11_+_48142916	A->C	650	2	*NDUFS3*	AAG-ACG	Lys-Thr	2.8×10^{-9}(LiverM)
chr12_+_67396538	A->G	923	2	*LDHB*	AAG-AGG	Lys-Arg	9.2×10^{-15}(BrainF)
chr12_-_126826456	U->C	694	1	*UBC*	UGU-CGU	Cys-Arg	3.3×10^{-9}(BrainF)
							3.0×10^{-9}(LiverF)
chr12_-_126826758	C->U	392	2	*UBC*	CCU-CUU	Pro-Leu	<1.0×10^{-16}(LiverM)
chr13_-_24478427	G->U	3922	1	*CENPJ*	GUA-UUA	Val-Leu	<1.0×10^{-16}(BrainM)
							<1.0×10^{-16}(LiverF)
chr15_+_29463145	G->A	323	2	*SCG5*	AGU-AAU	Ser-Asn	<1.0×10^{-16}(BrainF)
chr15_+_29467882	C->A	407	2	*SCG5*	ACC-AAC	Thr-Asn	<1.0×10^{-16}(BrainF)
chr15_+_76479318	U->C	691	1	*CHRNA5*	UGU-CGU	Cys-Arg	5.2×10^{-8}(BrainF)
chr17_+_16153682	C->U	790	1	*KRT31*	CAG-UAG	Gln-Stop	6.7×10^{-16}(BrainF)
chr17_+_50161033	U->C	428	2	*NME1*	AUG-ACG	Met-Thr	<1.0×10^{-16}(LiverM)
chr17_-_65416807	A->G	685	1	*APOH*	AAA-GAA	Arg-Glu	<1.0×10^{-16}(LiverM)
chr18_+_27492229	A->G	14	2	*TTR*	CAU-CGU	His-Arg	1.7×10^{-15}(LiverF)
chr22_-_25293818	U->C	2470	1	*TFIP11*	UGG-CGG	Trp-Arg	<1.0×10^{-16}(LiverM)
chr2b_-_211730316	G->A	1195	1	*NDUFS1*	GUU-AUU	Val-Ile	7.2×10^{-11}(LiverM)
chr3_+_138351241	A->G	1771	1	*TF*	AAG-GAG	Lys-Glu	<1.0×10^{-16}(LiverM)
chr3_+_185018683	A->G	401	2	*NDUFB5*	GAA-GGA	Glu-Gly	3.4×10^{-8}(LiverM)
chr3_+_192198094	G->U	787	1	*AHSG*	GGU-UGU	Gly-Cys	<1.0×10^{-16}(LiverF)
chr4_+_169759922	G->A	902	2	*CPE*	CGC-CAC	Arg-His	<1.0×10^{-16}(BrainF)
chr5_+_72747893	U->C	1102	1	*SEPP1*	UGC-CGC	Cys-Arg	<1.0×10^{-16}(LiverF)
							3.8×10^{-11}(LiverM)
chr6_-_31353959	G->U	1220	2	*FLOT1*	AGU-AUU	Ser-Ile	2.1×10^{-7}(LiverM)
chr6_-_50742166	G->A	1219	1	*RHAG*	GAG-AAG	Glu-Lys	<1.0×10^{-16}(LiverF)
chr6_-_50751977	A->G	634	1	*RHAG*	AUG-GUG	Met-Val	<1.0×10^{-16}(LiverF)
chr6_-_162618429	A->U	278	2	*SOD2*	GAU-GUU	Asp-Val	<1.0×10^{-16}(LiverM)
chr7_+_141356466	A->C	8	2	*NDUFB2*	GAU-GCU	Asp-Ala	1.4×10^{-7}(LiverM)
chr9_+_71972675	U->C	1027	1	*ANXA1*	UGU-CGU	Cys-Arg	7.9×10^{-9}(LiverF)
chr9_+_97829349	U->C	731	2	*SEPT7*	UUC-UCC	Phe-Ser	4.9×10^{-12}(BrainF)
							3.1×10^{-11}(LiverF)
chr9_-_120006979	C->U	5014	1	*CDK5RAP2*	CGC-UGC	Arg-Cys	1.1×10^{-12}(BrainF)
chrX_-_47479521	G->C	444	3	*NDUFB11*	CAG-CAC	Gln-His	9.6×10^{-8}(BrainF)
chrX_-_77021727	G->C	4090	1	*ATRX*	GCA-CCA	Ala-Pro	<1.0×10^{-16}(BrainF)
chrX_+_120423271	A->G	1568	2	*GLUD2*	CAC-CGC	His-Arg	<1.0×10^{-16}(LiverF)

Among the 61 RNA editing sites, 41 RNA editing sites resided in 34 genes had altered amino acid residues. Five genes had at least two RNA editing sites. The different isoforms could be produced with various combinations of editing sites. *P*_value: *P*_vaules calculated by Binomial test are corrected for Bonferroni multiple testing.

该基因产生多种蛋白质亚型．*KRT31* 是角蛋白基因，位于黑猩猩 17 号染色体上，角蛋白参与毛发和指甲的形成[39]，该基因上 chr17_+_16153682 位点发生 C->U RNA 编辑，使得密码子由 CAG 变成 UAG，将提前终止翻译过程．分析发现，4 个改变了原有的氨基酸的编辑位点导致该残基的亲疏水性发生明显变化，如 chr12_–_126826758 和 chr6_–_162618429 编辑位点，分别从亲水的脯氨酸和天冬氨酸变为疏水的亮氨酸和缬氨酸，chr5_+_72747893 和 chr9_+_97829349 编辑位点，分别从疏水的半胱氨酸和苯丙氨酸变为亲水的精氨酸和丝氨酸．氨基酸亲疏水性等理化性质的改变可能影响蛋白质的结构、溶解性和结合脂肪的能力等．chr10_+_42720282 的 C->U 位点影响真核翻译起始因子(EIF3L)第 308 位的氨基酸，落在保守的 Paf67 功能域内(Pfam ID: pfam10255)，潜在影响蛋白质行使该功能域的正常功能．

2.5 编辑位点对 miRNA 的潜在影响

miRNA 是一种大约 22 nt 的非编码 RNA，参与多细胞生物的转录后基因调控，通过与 mRNA 靶向结合而影响 mRNA 的翻译和稳定性[40]．mRNA 的 3′UTR 区是 miRNA 的主要靶标区域[41]，因此我们选择 3′UTR 区进行研究，分析编辑位点是否落在 miRNA 的靶标区．miranda 软件预测出 miRNA 在 3′UTR 区共有 992 个靶标位点，3′UTR 区的 59 个编辑位点有 11 个落在 miRNA 靶标区域(如表 4 所示)，其中 3 个编辑位点(chr12_+_934576 的 A->G、chr2a_+_26970053 的 G->C 和 chr6_+_3057563 的 G->C)落在 miRNA 靶标区的种子结合区，RNA 编辑的发生导致 miRNA 5′端 2～8 位与靶标序列不完全匹配，潜在破坏 miRNA与靶基因的靶向结合，导致功能丧失．

Table 4 Editing sites that are located in microRNA targets

Position	Type	miRNA	Region	miRNA targets	*P*_value
chr10_+_17878271	U->C	ptr-miR-630	non-seed	*VIM*	<1.0×10^{-16}(LiverM)
chr12_+_934576	A->G	ptr-miR-1208	seed	*WNK1*	<1.0×10^{-16}(BrainF)
					<1.0×10^{-16}(LiverF)
chr16_-_4904281	G->A	ptr-miR-329	non-seed	*ZNF500*	5.5×10^{-14}(BrainF)
chr16_-_4904281	G->A	ptr-miR-362	non-seed	*ZNF500*	5.5×10^{-14}(BrainF)
chr16_-_4904281	G->A	ptr-miR-612	non-seed	*ZNF500*	5.5×10^{-14}(BrainF)
chr16_-_10829488	A->G	ptr-miR-761	non-seed	*EMP2*	3.4×10^{-8}(LiverM)
chr20_+_4618535	G->A	ptr-miR-1272	non-seed	*PRNP*	1.5×10^{-14}(BrainF)
chr2a_+_26970053	G->C	ptr-miR-1288	seed	*SELI*	<1.0×10^{-16}(LiverM)
chr2a_+_26970630	A->C	ptr-miR-7	non-seed	*SELI*	<1.0×10^{-16}(LiverM)
chr4_-_125065463	C->U	ptr-miR-34a	non-seed	*ANXA5*	<1.0×10^{-16}(BrainF)
					<1.0×10^{-16}(LiverF)
chr4_-_125065463	C->U	ptr-miR-449b	non-seed	*ANXA5*	<1.0×10^{-16}(BrainF)
					<1.0×10^{-16}(LiverF)
chr6_+_30575637	G->C	ptr-miR-661	seed	*HLA-A*	1.0×10^{-9}(LiverM)
chr6_-_32004001	G->A	ptr-miR-148a	non-seed	*HLA-B*	6.7×10^{-16}(LiverM)
chr6_-_32004165	U->C	ptr-miR-23b	non-seed	*HLA-B*	1.4×10^{-13}(LiverM)
chr6_-_32004165	U->C	ptr-miR-23a	non-seed	*HLA-B*	1.4×10^{-13}(LiverM)

Region: The region where RNA editing sites resided in. *P*_value: *P*_vaules calculated by Binomial test are corrected for Bonferroni multiple testing.

2.6 编辑位点的组织差异性和性别差异性

4 组 RNA 编辑位点 BrainF、BrainM、LiverF 和 LiverM 按照组织划分可以分成脑组织与肝组织，按照性别划分可以分成雄性与雌性．其中，组织间共有 6 440 个 RNA 编辑位点所在的编辑底物在两个组织均表达(即都有 RNA-Seq 序列覆盖)，性别间共有 7 096 个．使用 Fisher 检验和 Bonferroni 多重校正分别识别到 640 和 872 个在组织和性别间存在显著差异的 RNA 编辑位点．如表 5 和表 6 所示，分别有 17 和 24 个组织和性别差异的编辑位点落在已知基因上．其中，*UBC* 基因上的编辑位点 chr12_-_126826758、*NME1* 基因上的编辑位点 chr17_+_50161033 以及 *SEPP1* 基因上的编辑位点 chr5_+_72747893 均导致氨基酸的改变．*UBC* 编码泛素蛋白，与蛋白质降解、DNA 修复、细胞周期调节等相关[42]．*NME1* 编码核苷二磷酸激酶，该基因是一个癌基因，在卵巢癌中与临床指标和病理组织显著相关[43]．*SEPP1* 编码硒蛋白，在胞

Table 5 17 RNA editing sites showed a significant difference between brain and liver

Position	Type	Gene	Region	Brain	Liver	*P*_value
chr1_+_80000601	G->C	*IFI44*	5′UTR	65/67	81/144	7.0×10^{-11}
chr10_-_100867296	U->C	*NDUFB8*	CDS	14/116	4/947	3.4×10^{-11}
chr10_+_17878271	U->C	*VIM*	3′UTR	1/103	7/10	2.4×10^{-8}
chr12_-_92043895	U->C	*DCN*	CDS	0/1499	58/167	3.9×10^{-63}
chr12_-_126826758	C->U	*UBC*	CDS	0/12	21/26	2.3×10^{-6}
chr14_-_102528569	C->U	*HSP90AA1*	CDS	0/53	12/31	1.2×10^{-6}
chr16_-_4903604	U->G	*ZNF500*	3′UTR	0/38	5/5	1.0×10^{-6}
chr17_+_50161033	U->C	*NME1*	CDS	0/27	20/22	8.2×10^{-12}
chr18_-_42307133	U->C	*ATP5A1*	CDS	4/202	29/119	2.8×10^{-10}
chr2a_+_26970053	G->C	*SELI*	3′UTR	0/366	42/111	2.1×10^{-30}
chr2a_+_26970630	A->C	*SELI*	3′UTR	0/295	24/63	1.1×10^{-20}
chr7_-_5518025	A->C	*ACTB*	3′UTR	0/1340	9/42	9.0×10^{-15}
chr7_-_10957341	A->G	*NDUFA4*	5′UTR	0/43	16/18	7.5×10^{-13}
chr8_-_14257861	U->C	*ASAH1*	3′UTR	196/213	527/670	3.5×10^{-6}
chr9_+_22996194	A->G	*MTRNR2L14*	5′UTR	0/1111	40/43	6.5×10^{-71}
chr9_+_22996366	C->U	*MTRNR2L14*	5′UTR	2/3287	922/1022	0
chr9_+_22996680	A->G	*MTRNR2L14*	5′UTR	0/37	17/18	2.6×10^{-13}

Brain: Edited sequences/total sequences. Liver: Edited sequences/total sequence. *P*_value: *P*_vaules calculated by Fisher's test are corrected for Bonferroni multiple testing.

Table 6 24 RNA editing sites showed a significant difference between male and female

Position	Type	Gene	Region	Male	Female	*P*_value
chr1_+_80000601	G->C	*IFI44*	5′UTR	74/134	72/77	1.2×10^{-9}
chr10_-_100867296	U->C	*NDUFB8*	CDS	14/116	5/1261	3.5×10^{-12}
chr10_-_99733008	C->G	*GOT1*	3′UTR	37/126	92/159	1.6×10^{-6}
chr12_-_92043895	U->C	*DCN*	CDS	58/167	3/3027	3.2×10^{-75}
chr12_-_126826758	C->U	*UBC*	CDS	21/26	0/30	4.9×10^{-11}
chr14_-_102528569	C->U	*HSP90AA1*	CDS	12/31	0/108	2.1×10^{-9}
chr15_+_29480205	G->A	*SCG5*	3′UTR	8/129	696/2508	2.3×10^{-9}
chr16_-_4903604	U->G	*ZNF500*	3′UTR	5/5	0/38	1.0×10^{-8}
chr17_+_50161033	U->C	*NME1*	CDS	20/22	0/28	4.9×10^{-12}
chr18_-_42307133	U->C	*ATP5A1*	CDS	29/119	3/274	1.4×10^{-13}
chr2a_+_26970053	G->C	*SELI*	3′UTR	42/111	0/411	4.0×10^{-32}
chr2a_+_26970630	A->C	*SELI*	3′UTR	24/63	0/335	7.9×10^{-22}
chr3_+_138342400	C->U	*TF*	CDS	18/93	0/410	1.5×10^{-14}
chr4_-_125065463	C->U	*ANXA5*	3′UTR	3/85	21/44	2.8×10^{-9}
chr5_+_72747893	U->C	*SEPP1*	CDS	5/20	71/78	9.5×10^{-9}
chr5_+_72748119	G->A	*SEPP1*	3′UTR	23/52	179/209	3.1×10^{-9}
chr5_-_116928428	A->G	*TMED7*	3′UTR	2/78	28/72	1.2×10^{-8}
chr6_+_173828052	C->U	*TBP*	3′UTR	9/14	0/45	1.6×10^{-7}
chr6_+_32517034	C->U	*CFB*	CDS	0/130	14/60	3.1×10^{-8}
chr7_-_10957341	A->G	*NDUFA4*	5′UTR	16/18	0/103	4.3×10^{-18}
chr7_-_99487539	C->U	*CYP3A7*	CDS	47/86	0/53	2.5×10^{-13}
chr9_+_22996194	A->G	*MTRNR2L14*	5′UTR	40/43	0/1658	9.5×10^{-78}
chr9_+_22996366	C->U	*MTRNR2L14*	5′UTR	922/1022	3/5274	0
chr9_+_22996680	A->G	*MTRNR2L14*	5′UTR	17/18	0/77	7.0×10^{-18}

Male: Edited sequences/total sequences. Female: Edited sequences/total sequence. *P*_value: *P*_values calculated by Fisher's test are corrected for Bonferroni multiple testing.

外起抗氧化剂作用[44]. chr12_-_126826758 和 chr17_+_50161033 既存在组织差异也有性别差异，我们的结果显示这两个位点仅在雄性黑猩猩的肝组织中分别发生 C-to-U 和 U-to-C RNA 编辑. chr5_+_72747893 位点存在性别差异，雌性黑猩猩的编辑水平显著高于雄性.

2.7　编辑位点上下游序列特征分析

使用 Two Sample Logo 软件分析编辑位点上下游序列的特征. 黑猩猩 8 334 个编辑位点按替换类型的不同分成 12 个阳性样本集，各样本集所含的样本量从 222 至 1 995 不等. 对应的阴性样本集也有 12 个，样本容量都是 30 000 个. 阳性样本集和阴性样本集共同作为 Two Sample Logo 软件的输入.

Two Sample Logo 输出具有统计学差异的位点特征，如图 3 所示，横坐标表示碱基序列，坐标为

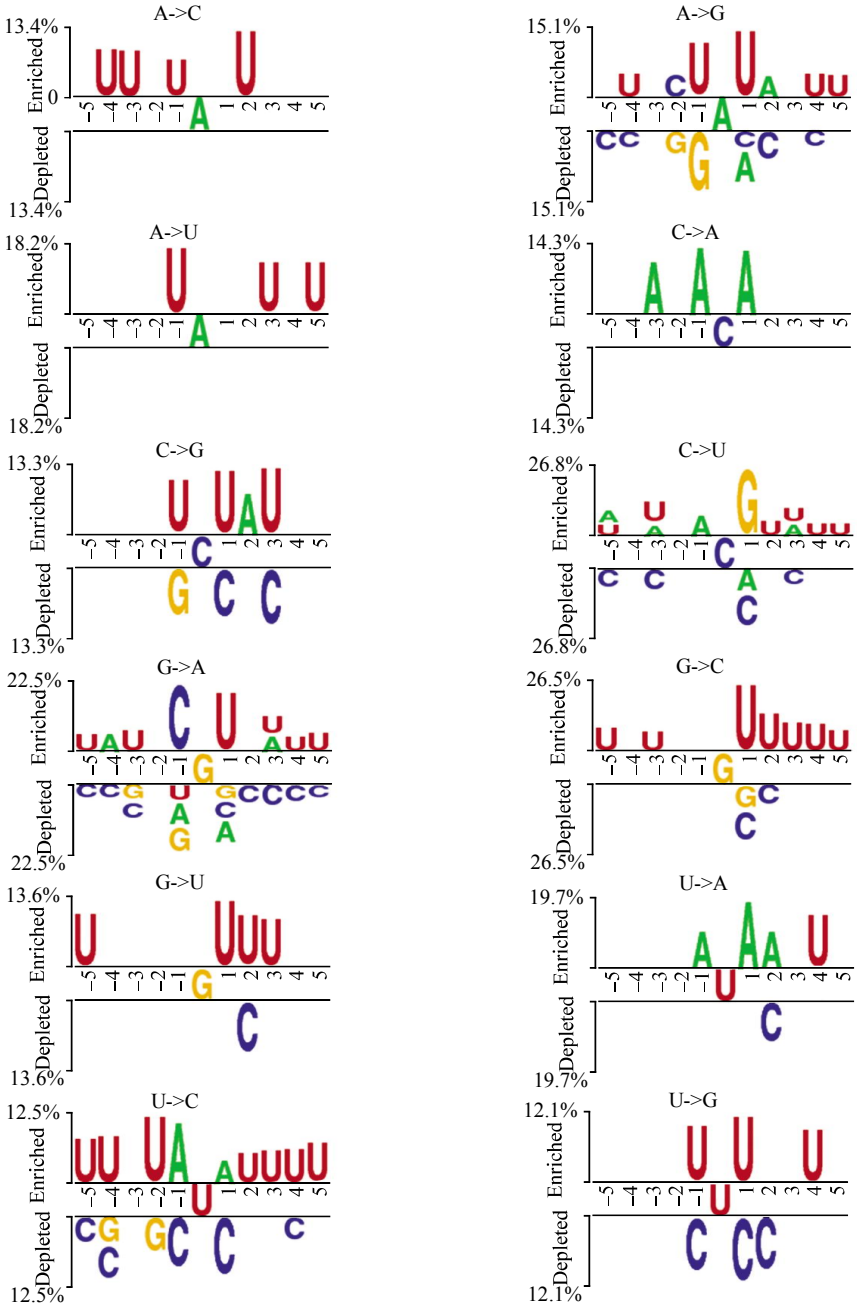

Fig. 3　Base preference of different types of RNA editing

Logo displays enriched bases and depleted bases in upstream/downstream region of the RNA editing site. The level of enrichment/depletion is shown by letter heights.

0 的位点是编辑位点，左侧序列来自编辑位点上游(–5 至–1 位)，右侧序列来自编辑位点下游(+1 至+5 位)．纵坐标分为 enriched 和 depleted，分别表示编辑位点上下游某碱基频率比背景集显著高或低．12 种编辑类型都有明显的序列特征．A->G 编辑在–1 位点倾向于出现 U 而拒绝 G，+1 位点 U 的出现频率显著高于背景中的对照位点，这些特征与先前的报道一致[2]．C->U 编辑在–1 位点 A 显著高频出现，+1 位置更倾向于 G，上下游序列整体呈现 AU 富集而 C 缺乏，与文章报道类似[24]．其余 10 种编辑类型的序列特征都没有文献报道．从图 3 中可以看出，A->C、A->U、C->A 3 种编辑类型只有偏好的碱基而无排斥碱基，其中 A->C 和 A->U 都偏好 U 而 C->A 偏好 A．C->G、G->A、G->C、G->U、U->A、U->C 和 U->G 7 种编辑类型既有碱基偏好又有碱基排斥，除 G->A 编辑在–1 位点显著倾向 C 而排斥 UA 以及在+1 位点排斥 A 外，其他编辑类型在特定位点都显著偏好 A 或 U，而排斥 C 或 G.

为了分析各样本集内部是否有显著的差异，我们对所有样本集做随机对照，随机抽取部分样本构建阳性样本集，用剩余样本构建阴性样本集，用 Two Sample Logo 分析两组数据之间是否有差异，结果表明同一个样本集的样本均无显著的碱基偏好，排除了序列特征由样本抽取不均匀所致的情况．

3 讨　论

RNA-Seq 以其高通量、高灵敏性以及扩展性强等优势，被广泛用于转录组研究．随着测序通量的增大、测序费用的降低以及测序长度的增加，该技术势必得到更广泛的应用．本课题利用黑猩猩 RNA-Seq 数据成功识别了 8 334 个碱基替换型 RNA 编辑位点．其中 A->G、G->A、C->U 和 U->C 4 种类型的 RNA 编辑位点数量明显高于其他 8 种类型，显示出转换型 RNA 编辑较颠换型 RNA 编辑更易发生．进一步的编辑位点功能分析表明，有 41 个编辑位点改变了氨基酸残基，如 *LOC456645* 基因上有 3 个编辑位点，理论上可生成 8 种不同的蛋白质．3 个编辑位点落在 miRNA 靶标区的种子结合区，编辑的发生潜在破坏 miRNA 与靶基因的结合致使功能丧失．编辑位点上下游碱基频率分析显示，12 种替换类型的 RNA 编辑位点紧邻碱基有显著的序列偏好，其中 A->G 和 C->U 编辑的序列特征与以往文献报道一致．

碱基替换型 RNA 编辑的发生使转录后序列同相应基因组序列存在碱基错配．生物信息学正是利用序列比对的方法，识别转录后序列和相应基因组序列之间的错配位点，这些位点是潜在的 RNA 编辑位点．当然，错配位点还可能由 SNP、PCR 扩增错误、测序错误、比对错误等原因所致，这些干扰必须考虑在内．我们先将黑猩猩 RNA-Seq 数据比对到基因组上，从比对结果中提取所有错配位点，构建编辑位点候选集，并从中滤除可能的干扰，再结合黑猩猩基因标注信息，确定落在已知基因上的编辑位点的替换类型，识别出潜在的 RNA 编辑位点．尽管经过多步严格过滤条件的筛选，我们报道的编辑位点中仍可能存在假阳性．例如，从编辑位点候选集里滤除了已知 SNP 位点，但是由于已知黑猩猩 SNP 位点仅有 150 万，而人的 SNP 位点已达 3 000 万(dbSNP 第 132 版)[45]，编辑位点中可能混杂有未报道的 SNP 位点．目前 PanMap 计划[46]正致力于研究黑猩猩个体之间的差异，该计划同时对 10 只黑猩猩的基因组进行测序，将大大增加 SNP 位点数据量，届时我们将重新滤除新的 SNP 位点以进一步降低识别的假阳性．

由于已知的黑猩猩基因不足 1 500 条，我们用 liftOver 工具和人、黑猩猩基因组坐标关联文件，将人的已知 37 630 个基因转换到黑猩猩基因组上，成功转换了 27 126 个基因作为黑猩猩参考基因．依据这些参考基因的转录方向来确定黑猩猩编辑位点的替换类型，落在未知转录本上的编辑位点，由于无法确定替换类型而被滤除，这样可能会使我们的识别结果中丢失部分 RNA 编辑位点．

本研究在全转录组范围系统识别黑猩猩各种替换类型的 RNA 编辑位点，表明多种 RNA 编辑事件在黑猩猩转录组的存在，为开发从头预测编辑位点的算法提供了原始数据，为进一步研究黑猩猩 RNA 编辑的功能以及 RNA 编辑发生的机理奠定了基础．

参 考 文 献

[1] Gott J M, Emeson R B. Functions and mechanisms of RNA editing. Annu Rev Genet, 2000, **34**: 499−531

[2] Bass B L. RNA editing by adenosine deaminases that act on RNA. Annu Rev Biochem, 2002, **71**: 817−846

[3] Knoop V. When you can't trust the DNA RNA editing changes transcript sequences. Cell Mol Life Sci, 2011, **68**(4): 567−586

[4] Kleinberger Y, Eisenberg E. Large-scale analysis of structural

sequence and thermodynamic characteristics of A-to-I RNA editing sites in human Alu repeats. BMC Genomics, 2010, **11**: 453

[5] Gott J M.Expanding genome capacity *via* RNA editing. C R Biol, 2003, **326**(10-11): 901-908

[6] Eisenberg E, Nemzer S, Kinar Y, *et al*. Is abundant A-to-I RNA editing primate-specific?. Trends Genet, 2005, **21**(2): 77-81

[7] Guryev V, Koudijs M J, Berezikov E, *et al*. Genetic variation in the zebrafish. Genome Res, 2006, **16**(4): 491-497

[8] Enstero M, Akerborg O, Lundin D, *et al*. A computational screen for site selective A-to-I editing detects novel sites in neuron specific Hu proteins. BMC Bioinformatics, 2010, **11**: 6

[9] Yang W, Chendrimada T P, Wang Q, *et al*. Modulation of microRNA processing and expression through RNA editing by ADAR deaminases. Nat Struct Mol Biol, 2006, **13**(1): 13-21

[10] Maas S, Kawahara Y, Tamburro K M, *et al*. A-to-I RNA editing and human disease. RNA Biol, 2006, **3**(1): 1-9

[11] Paz N, Levanon E Y, Amariglio N, *et al*. Altered adenosine-to-inosine RNA editing in human cancer. Genome Res, 2007, **17**(11): 1586-1595

[12] Gallo A, Galardi S. A-to-I RNA editing and cancer: form pathology to basic science. RNA Biol, 2008, **5**(3): 135-139

[13] Hoopengardner B, Bhalla T, Staber C, *et al*. Nervous system targets of RNA editing identified by comparative genomics. Science, 2003, **8**: **301**(5634): 832-836

[14] Kim D D, Kim T T, Walsh T, *et al*. Widespread RNA editing of embedded alu elements in the human transcriptome. Genome Res, 2004, **14**(9): 1719-1725

[15] Levanon E Y, Eisenberg E, Yelin R, *et al*. Systematic identification of abundant A-to-I editing sites in the human transcriptome. Nat Biotechnol, 2004, **22**(8): 1001-1005

[16] Athanasiadis A, Rich A, Maas S. Widespread A-to-I RNA editing of Alu-containing mRNAs in the human transcriptome. PLoS Biol, 2004, **2**(12): e391

[17] Blow M, Futreal P A, Wooster R, *et al*. A survey of RNA editing in human brain. Genome Res, 2004, **14**(12): 2379-2387

[18] Shendure J, Ji H. Next-generation DNA sequencing. Nat Biotechnol, 2008, **26**(10): 1135-1145

[19] Mardis E R. Next-generation DNA sequencing methods. Annu Rev Genomics Hum Genet, 2008, **9**: 387-402

[20] Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet, 2009, **10**(1): 57-63

[21] Li J B, Levanon E Y, Yoon J K, *et al*. Genome-wide identifcation of human RNA editing sites by parallel DNA capturing and sequencing. Science, 2009, **324**(5931): 1210-1203

[22] Wahlstedt H, Daniel C, Enstero M, *et al*. Large-scale mRNA sequencing determines global regulation of RNA editing during brain development. Genome Res, 2009, **19**(6): 978-986

[23] 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

[24] Rosenberg B R, Hamilton C E, Mwanqi M M, *et al*. Transcriptome-wide sequencing reveals numerous APOBEC1 mRNA-editing targets in transcript 3′ UTRs. Nat Struct Mol Biol, 2011, **18**(2): 230-236

[25] Wetterbom A, Ameur A, Feuk L, *et al*. Identification of novel exons and transcribed regions by chimpanzee transcriptome sequencing. Genome Biol, 2010, **11**(7): R78

[26] Nagalakshmi U, Wang Z, Waern K, *et al*. The transcriptional landscape of the yeast genome defned by RNA sequencing. Science, 2008, **320**(5881): 1344-1349

[27] Mortazavi A, Williams B A, McCue K, *et al*. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methords, 2008, **5**(7): 621-628

[28] Chepelev I, Wei G, Tang Q, *et al*. Detection of single nucleotide variations in expressed exons of the human genome using RNA-Seq. Nucleic Acids Res, 2009, **37**(16): e106

[29] Pan Q, Shai O, Lee L J, *et al*. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nat Genet, **40**(12): 1413-1415

[30] Ramani A K, Calarco J A, Pan Q, *et al*. Genome-wide analysis of alternative splicing in *Caenorhabditis elegans*. Genome Res, 2011, **21**(2): 342-348

[31] Twine N A, Janitz K, Wilkins M R, *et al*. Whole transcriptome sequencing reveals gene expression and splicing differences in brain regions affected by Alzheimer's disease. PLoS One, 2011, **6**(1): e16266

[32] Chimpanzee sequencing and analysis consortium. Initial sequence of the chimpanzee genome and comparison with the human genome. Nature, 2005, **437**(7055): 69-87

[33] Paz-Yaacov N, Levanon EY, Nevo E, *et al*. Adenosine- to-inosine RNA editing shapes transcriptome diversity in primates. Proc Natl Acad Sci USA. 2010, **107**(27): 12174-12179

[34] Trapnell C, Pachter L, Salzberg S L. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics, 2009, **25**(9): 1105-1111

[35] Tay S K, Blythe J, Lipovich L. Global discovery of primate-specific genes in the human genome. Proc Natl Acad Sci USA, 2009, **106**(29): 12019-12024

[36] Schwartz S, Kent W J, Smit A, *et al*. Human-mouse alignments with BLASTZ. Genome Res, 2003, **13**(1): 103-107

[37] Griffiths-Jones S, Grocock R J, van Dongen S, *et al*. miRBase: microRNA sequences, targets and gene nomenclature. Nucleic Acids Res, 2006, **34**(Database issue): D140-144

[38] Vacic V, laloucheva L M, Radivojac P. Two Sample Logo: a graphical representation of the differences between two sets of sequence alignments. Bioinformatics, 2006, **22**(12): 1536-1537

[39] Cribier B, Peltre B, Grosshans E, *et al*. On the regulation of hair keratin expression: Lessons from studies in pilomatricomas. J Invest Dermatol, 2004, **122**(5): 1078-1083

[40] Griffiths-Jones S. The microRNA registry. Nucleic Acids Res, 2004, **32**(Database issue): D109-111

[41] Ambros V. The functions of animal microRNAs. Nature, 2004, **431**(7006): 350-355

[42] Chen D, Dou Q P. The ubiquitin-proteasome system as a prospective molecular target for cancer treatment and prevention. Curr Protein

Pept Sci, 2010, **11**(6): 459−470

[43] Arik D, Kulaçoglu S. P53, bcl-2, and nm23 expressions in serous ovarian tumors: correlation with the clinical and histopathological parameters. Turk Patoloji Derq, 2011, **27**(1): 38−45

[44] Burk R F, Hill K E. Selenoprotein P-expression, functions, and roles in mammals. Biochim Biophys Acta, 2009, **1790**(11): 1441−1447

[45] Sherry S T, Ward M H, Kholodov M, *et al*. dbSNP: The NCBI database of genetic variation. Nucleic Acids Res, 2001, **29** (1): 308−311

[46] Pfeifer S. PanMap: Mapping genomic variation in chimpanzee. Annual Meeting of the Society for Molecular Biology and Evolution, Lyon, France, 2010

Identification of RNA Editing Sites in Chimpanzee by Transcriptome-wide Sequencing Data*

WANG Duan-Qing[1,2], HE Tao[2], WANG Li[2]**, WANG Yu-Min[2], SHAO Wei-Dong[1]**

([1] *School of Electronics and Information, Soochow University, Suzhou* 215006, China;

[2] *Institute of Biotechnology, Academy of Military Medical Sciences, Beijing* 100071, China)

Abstract RNA editing is a widespread post-transcriptional modification mechanism that alters genetic information at the RNA level by nucleotide insertions, deletions or substitutions, which can contribute to the diversification of the transcriptome and proteome. Although tens of thousands of A-to-I RNA editing events have been found in humans, there is limited knowledge of RNA editing in other nonhuman primates. For exploring the mechanism as well as potential functions of the RNA editing events in chimpanzee, we identified RNA editing sites based on chimpanzee RNA-Seq data here. By aligning between RNA-Seq data and chimpanzee genome sequences with TopHat software, all RNA-DNA mismatch sites were regarded as a candidate set. Low quality sites were filtered out by using both genome and transcriptome sequencing quality scores. The other filters containing uncertainty of sequencing at 3'-terminial positions, read coverage, SNP sites and estimated editing level were also applied on the candidate set. Statistical tests based on the Binomial distribution and Bonferroni multiple testing correction were performed on each candidate site to remove random errors between genome and transcriptome. Then, we detected tissue- and sex-specific RNA editing sites using bioinformatics approaches based on the Fisher's exact test and the Bonferroni multiple testing correction. The Two Sample Logo software was used to analyze the feature of the sequences surrounding the RNA editing site. A total of 8 334 RNA editing sites were identified in chimpanzee transcriptome and all 12 possible categories of discordances were observed. The top four distributions were A-to-G, U-to-C, G-to-A and C-to-U editing sites, which contained 1 995, 1 452, 1 293 and 1 101 sites, respectively. Forty-one editing sites alter amino acid residues, one of them creates a new stop codon which may shorten the KRT31 protein and affect its activity. Three editing sites damage the binding of microRNA potentially. Six hundred and forty and eight hundred and seventy-two RNA editing sites were identified to be tissue-specific and sex-specific respectively. The analysis of base frequencies indicated that all substitution editings have preferences for certain neighbouring nucleotides. RNA editing is widespread in chimpanzee and has important biology function. Our findings paved the way for further exploration of the mechanism of RNA editing in primates.

Key words chimpanzee, RNA editing, RNA-Seq, computational identification

DOI: 10.3724/SP.J.1206.2011.00328

*This work was supported by grants from The National Natural Science Foundation of China (30800644), Hi-Tech Research and Development Program of China (2007AA022204) and Chinese Key Program for Drug Invention (2008ZXJ09007-001).

**Corresponding author.

SHAO Wei-Dong. Tel: 86-18913146453, E-mail: shaowd@suda.edu.cn

WANG Li. Tel: 86-10-66948793, E-mail: liwang@tsinghua.edu.cn

Received: July 14, 2011 Accepted: November 4, 2011