人类基因组中功能性 RNA 编辑的生物信息学分析

蔡彦玮 [1]

指导教师：田卫东 [1]

(1. 生命科学学院)

摘要：RNA 编辑是一种非常重要的转录后修饰机制。目前通过 RNA 测序技术发现，RNA 编辑大量位于内含子、重复序列等低功能性的位点上，大部分不具有功能。然而有少数 RNA 编辑位点在不同细胞系中反复出现，或在不同物种中反复出现，这些位点是否具有功能，具有怎样的功能是本课题要探索的问题。本课题通过对 14 种细胞系中 RNA 编辑位点的分析，定义 RNA 编辑的细胞系保守性以衡量 RNA 编辑在细胞系中反复出现的程度。对高细胞系保守性位点的功能分析及进化保守性分析发现，细胞保守性强的 RNA 编辑倾向于在进化上保守，同时细胞保守性强的 RNA 编辑位点有明显的功能，大部分通过改变 microRNA 结合位点影响基因的功能。另外，细胞保守性强的 RNA 编辑位点对应的基因参与了重要的生物学通路，表明 RNA 编辑可能对细胞生长和发育是必要的。这些结论说明细胞系保守性强的 RNA 编辑位点具有特殊的性质，对 RNA 编辑领域的研究有启发作用。

关键词：RNA 编辑；细胞系；保守性；microRNA

Bioinformatics Analysis of Functional RNA Editing Sites in Human Genome

Cai Yanwei[1]

Advisor: Wei Dongtian[1]

(1. School of Life Sciences)

Abstract: RNA editing is a very important post-transcript regulation mechanism. Current RNA-seq data analysis has shown that RNA editing sites are massively located in introns, repeat sequences and that most RNA editing sites are non-functional. However, there do be a few RNA editing sites that occurred repeatedly in different cell lines, or appear over and over in different species. Whether these sites have functions and what is the function of these sites are the biological topic this project aim to explore. From RNA-seq data of 14 cell lines, RNA editing sites' cell-line conservation (CLC) is defined to describe the degree to which RNA editing sites repeatedly occur in different cell lines. Functional and evolutional analysis of high cell-line conservation sites show that cell-line conserved RNA editing sites tend to be evolutional

conserved, and that cell-line conserved sites have significant functions, mostly by the mechanism of altering microRNA target sequence. In addition, genes undergo high cell-line conservation editing involve in important biological pathways, suggesting that RNA editing could be essential to cell growth and development. These results illustrate that cell-line conserved RNA editing sites have special features, which may enlighten further research in the field of RNA editing.

Key words: RNA editing; Cell line; Conservation; microRNA

1 引言

1.1 研究背景

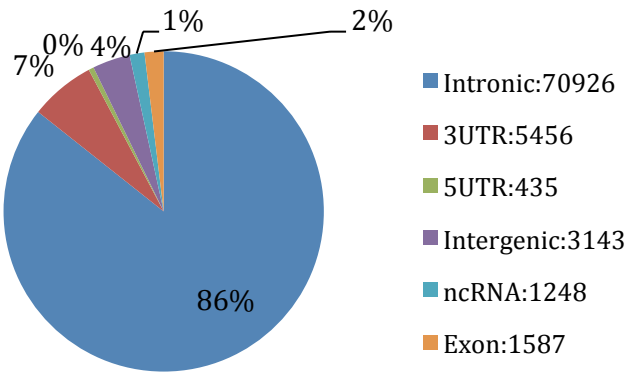
1.1.1 RNA 编辑现象与研究 RNA 编辑的意义

RNA 编辑指在转录后 RNA 水平上，通过对单个碱基的编辑作用改变前体 RNA 或 microRNA 序列，使成熟 mRNA 及 microRNA 上部分位点不同于基因组中编码的对应位点的现象。RNA 编辑是一种非常重要的转录后修饰机制，其可能造成蛋白质产物非同义突变，可变剪切，小 RNA 降解速率改变，microRNA 调控改变等多重作用[1]。RNA 编辑机理较复杂，其控制编辑的信号，编辑位点特异性选择的机制等都未研究清楚[2]。但 RNA 编辑在人类特别是人脑发育中非常重要，其发生受严密的时空特异性调控，与多种人类疾病相关[3]。在模式动物实验中，ADAR1$^{-/-}$的小鼠会在胚胎期致死[4]。这些结果说明了 RNA 编辑的复杂性与重要性。进一步研究 RNA 编辑有重要意义。

目前对 RNA 编辑的研究发现，哺乳动物中大部分（95%以上）RNA 编辑是由 ADAR(核糖腺苷脱氨酶)酶家族引导的 A 到 I 的编辑，即蛋白质的脱氨基结构域将腺嘌呤（A）转变为次黄嘌呤（I），因 I 可与 C 氢键配对，故在功能上类似于 G，又称 A 到 G 编辑[1-4]。ADAR 的作用机理研究相对较完备，已知其单体或二聚体结合于配对形成双链的 RNA 上，可选择性的对一些腺嘌呤进行编辑，其对位点的识别受 RNA 高级结构确定[5]。ADAR 酶的特异性作用区域是形成双链配对的 RNA，因此 RNA 编辑发生的区域 RNA 往往需要形成复杂的二级结构。人类基因组中很多短重复序列如 Alu 序列由于易形成互补配对的双链 RNA，往往参与到 ADAR 酶的编辑中[5,6]。

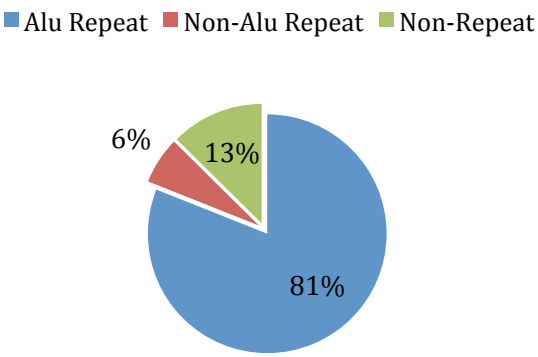
1.1.2 高通量测序数据注释的 RNA 编辑位点

值得引起注意的是，目前通过高通量数据分析发现，发生在外显子区域的 RNA 编辑事件仅占一小部分[6-8]。如图 1，编辑事件发生的区域从数量上最多的是内含子区域、之后排序依次是 3’非编码区、基因间区域、非编码 RNA、外显子区和 5’非编码区。另外，如图 2，大量编辑序列落于 Alu 短串联重复序列中，落在非重复序列上的 RNA 编辑事件相对较少。也就是说，大部分 RNA 编辑位点位于低功能性的区域，不直接改变蛋白质序列，这样的结论与已有的文献报道一致[7-9]。由此看，1.1.1 节中提到的 RNA 编辑重要性与本节展示的 RNA 编辑低功能性似有矛盾，需要具体关注已报到的 RNA 编辑的功能。



分布比例分别为内含子 89%，3’非编码区 5%，基因间区域 4%，外显子 1%，非编码 RNA1%，5’非编码区 0%

图 1. RNA 编辑在各基因组区域的分布



81%的位点分布于 Alu 短串联重复序列，13%的位点分布于非重复序列，6%的位点分布于非 Alu 的重复序列上

图 2. RNA 编辑在各基因组重复区域的分布。

1.1.3 RNA 编辑的功能性

1.1.3.1 文献报道的 RNA 编辑功能

目前一些对特定位点的 RNA 编辑研究发现，RNA 编辑主要有以下功能：改变蛋白质的氨基酸组成、影响 microRNA 引导的 RNA 沉默通路、影响 RNA 结合蛋白的结合、影响可变剪切、影响 RNA 降解速率等功能[1-4,11]。以下分段详叙。

单位点的编辑足够改变密码子从而改变蛋白质的氨基酸组成[1,2]。经典的例子包括编码谷氨酸受体 GluR2，5-羟色胺受体 HTR2C 等神经受体的 mRNA 在从线虫到人类的跨物种范围都需要正确编辑后才能行使相应的功能。研究聚焦的 GluR2 受体 mRNA 需要一个编辑产生 Val 到 Arg 的转变才能保证哺乳动物的正常脑发育和胚胎存活[1]。然而，在人类约 1600 个外显子上的编辑事件中，40%以上属于同义突变，剩下的也仅有一些案例证明编辑位点对蛋白质功能有明确意义[5,6]。

由 microRNA 介导的基因沉默通过是基因转录后调控的重要方式，成熟的 microRNA 通过结合 mRNA 上的靶点（通常在 3’非编码区）调控的基因的表达。对于成熟 microRNA 上的修饰可改变 microRNA 序列，增加或减少该 miRNA 的靶点，在不同条件下调控不同种转录本的表达[12]。同样的，对于基因 3’非编码区上的编辑也能对其成为某种 miRNA 靶点起开关的作用[12]。

RNA 编辑可改变 mRNA 组成影响 mRNA 与 RNA 结合蛋白的结合能力，如在前体小 RNA（pre-miRNA）上的修饰可显著降低其与 miRNA 成熟通路上蛋白 TRBP 的作用，影响其正确剪切，造成最终成熟 miRNA 量的显著下调[1,3]。同样的，RNA 编辑也可影响一些小 RNA 的降解速录，有文献报道某些小 RNA 在编辑前后核外表达量有数十倍的差异[13]。

RNA 编辑可通过在外显子内含子连接处的修饰，改变剪切位点的识别从而改变转录本构造并产生新的转录本。比如高度保守的 5’剪切二核苷酸识别序列 GU（AU=>IU=GU）或 3’剪切接受位点 AG（AA=>AI=AG）均可由单位点的编辑产生[4,11,13]。

1.1.3.2 RNA 编辑功能性相关的理论

与 RNA 编辑功能性相关的理论主要有编辑酶特异性理论及进化保守性理论。

由于 ADAR 蛋白本身识别双链 RNA 序列高级结构的特异性较低，RNA 编辑总体的特异性低，加上 Alu 序列上编辑位点的干扰，这些情况反应在数据上即各个样本间重复的 RNA 编辑数量少，大部分 RNA 编辑仅在某些样本中出现[7-9]。然而，通过详细分析数据发现，相应的有一些 RNA 编辑位点在各个样本中都受到编辑[8]。从 RNA 二级结构分析上，这些位点周围形成的高级结构可能对 ADAR 酶有高特异性，使得这些位点稳定的受到编辑，这一过程可能受到进化选择而固定，那么这些位点的编辑可能是比较关键的，即可能存在重要功能，这为信息学方法判定编辑位点特异性提供了理论依据[5,6]。

另一方面，从进化角度看 RNA 编辑，总体上，RNA 编辑是不保守的。Xu G *et al.*分析了人类基因外显子上的已知的编辑位点，认为大部分位点都处在受进化压力小的基因且该基因相对不重要的氨基酸位点上，这与 Pinto Y 等人的研究相一致，都说明大部分 RNA 编辑是不保守的[14,15]。然而，同样的，他们的数据中也发现有一些 RNA 编辑在不同物种的对应位点反复出现，从进化理论看这些编辑位点同样很可能存在重要功能[17]。

1.1.3.3 本课题涉及的 RNA 编辑的生物学问题

通过前文的描述，虽然对于 RNA 编辑的功能目前的研究有了一些眉目，但目前发现的大部分 RNA 编辑仍然都是没有功能的。在这样的背景下，存在少数的在不同的样本中反复出现，或在不同的物种中反复出现的 RNA 编辑位点，这些位点是否有功能，具有什么样的功能是一个值得研究的生物学问题。

1.2 研究思路

本课题将在细胞系 RNA 测序公共数据库与其他可借鉴公共数据的基础上，将在各细胞系中反复出现的 RNA 编辑定义为高细胞系保守性位点，将在不同物种中反复出现的 RNA

编辑定义为高进化保守性位点，首先判断细胞系保守性与进化保守性是否具有一致性，然后探究这些位点具有哪些功能。

课题研究中首先通过细胞系 RNA 测序数据确定细胞系保守性强的 RNA 编辑位点，然后分别进行进化保守性分析和功能性分析。在进化保守性分析中，通过比对其他物种的位点确定编辑的进化保守性，并检测细胞系保守性与进化保守性是否一致。在功能性分析中，对于五种文献中有报道的功能，对其每一种可能的功能机理，用相应的生物信息学手段进行分析。具体来说，通过检测蛋白质序列是否改变确定位点是否引起蛋白质密码子改变，通过结合能力预测确定位点是否与蛋白质结合和 miRNA 调控有关，通过表达分析确定位点是否与 RNA 降解调控和可变剪切调控有关，详细流程见图 3。

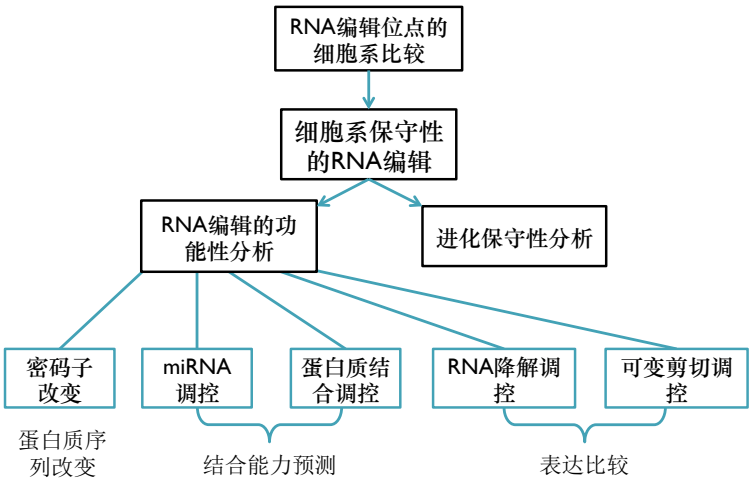


图 3. 本课题的研究思路

1.3 研究内容

根据课题的研究思路，本课题的研究内容将主要集中在三个方面：RNA 编辑的注释、RNA 编辑的细胞系保守性与进化保守性注释及比较、RNA 编辑的功能性注释与比较。

1.3.1 RNA 编辑的注释

RNA 编辑的注释参考 Ramaswami G *et al.*的方法，通过对 RNA 测序数据与基因组测序数据比较后得到[16]。这其中要注意测序数据的选择，比对、拼装过程中的算法参数选择，质量控制，以及对 SNP，同源基因等基因组层面易引起假阳性的因素的控制。

1.3.2 RNA 编辑的细胞系保守性与进化保守性的注释与比较

RNA 编辑的细胞系保守性通过比较编辑位点在各细胞系中出现的频数得到，频数越大说明编辑细胞系保守性越强。编辑的进化保守性分为两个部分，一是编辑位点在别的物种中是否也受编辑，其二是编辑位点在基因组上的对应位点是否在其他物种的基因组中也存在，对着这两者应分别比较阐述。

得到进化保守性位点后，通过观察随着细胞系保守性的增加，位点中进化保守位点的比例是否也有相应增加来分析细胞系保守性与进化保守性的关系。

1.3.3 RNA 编辑的功能注释与比较

RNA 编辑的功能性有关的数据分析中。判断是否改变蛋白质序列，即需比较对应的密码子是否改变；判断 miRNA 调控，通过 miRNA 靶基因预测软件比较编辑前与编辑后结合能力是否有变化来实现；判断蛋白质结合，通过运用蛋白质结合芯片数据量化结合能力；判断影响小 RNA 降解速率，通过比较小 RNA 在核内测序表达量与核外测序表达量是否有显著变化及该变化是否与编辑显著相关为依据；判断影响可变剪切，通过比较位点附近外显子在受编辑细胞系与不受编辑细胞系中表达量是否有显著变化以比较。

得到功能性位点注释后，类似于进化保守性分析，观察随着细胞系保守性的增加，位点中功能性位点的比例是否也有相应增加来分析细胞系保守性与进化保守性的关系。

2 数据与方法

2.1 实验数据

2.1.1 RNA-seq 数据

本课题的人类基因组 RNA-seq 数据来自于 ENCODE 计划 Caltech 测序数据[18]，测序采用 Illumina Genome Analyzer，75bp 配对（pair-end）测序。

表 1. 所用的细胞系数据情况表

细胞系名	GEO ID	Base 数	编辑位点数	AG%	细胞系注释
Gm12878	GSM958728	32.2G	20275	98.6	类淋巴母细胞
K562	GSM958729	33.4G	9952	97.6	永生粒细胞性白血病
Hepg2	GSM958732	26.5G	5090	95.4	原发性肝癌细胞
H1-hesc	GSM958733	30.7G	21252	98.1	胚胎干细胞
HUVEC	GSM958734	18.3G	10385	99.0	脐静脉内皮细胞
Hela-S3	GSM958735	20.3G	8280	98.4	增殖表皮癌细胞
NHEK	GSM958736	21G	19910	99.0	皮角质形成细胞
HSMM	GSM958744	23.4G	3466	95.4	骨骼肌细胞和成肌细胞
MCF-7	GSM958745	33.9G	7040	97.5	乳腺癌细胞
NHLF	GSM958746	20.4G	19910	99.0	肺成纤维细胞
GM12891	GSM958747	15G	5148	98.7	类淋巴母细胞
GM12892	GSM958748	20.3G	4711	98.1	类淋巴母细胞
HCT-116	GSM958749	37.3G	10385	99.0	结肠癌细胞
LHCN-M2	GSM958750	18.9G	3405	96.4	骨骼肌细胞

本课题使用的小 RNA 测序数据来源于 ENCODE 计划 CSHL small RNA-seq[18]，分别选取核内（Nucleus）测序数据与细胞质（Cytosol）测序数据，数据下载自 UCSC table browser[19]。

2.1.2 基因组注释数据

基因组注释数据下载于 UCSC table browser，基因定义与转录本定义来自于 GENCODE Genes V19[20]，重复序列来自于 RepeatMasker v4[21]。

2.1.3 保守性数据

本课题使用的基因组保守性分值数据来源于 UCSC table browser 上下载的 phastCons100

数据，包含由 100 种脊椎动物比较基因组学分析得出的保守性分值[22]。其他物种（黑猩猩，恒河猴，小鼠）的 RNA 编辑数据来源于 Ramaswami G *et al.*的工作[16]。

2.1.4 miRNA 数据

miRNA 数据来自 miRBase 定义的高可能性 miRNA 序列[23]，共 524 条序列。

2.1.5 RNA 结合蛋白结合数据

RNA 结合蛋白的模体数据来源于 Ray D *et al.*的数据[24]，该数据通过蛋白质结合芯片取得蛋白质对随机 RNA 序列的结合能力数据，分析总结出每个蛋白质的结合模体（motif），用 ES 值表征每种蛋白质对任意 7bp 序列的结合能力。

2.2 实验软件

2.2.1 系统环境: CentOS Linux 7.1.1503

2.2.2 编程语言

R 语言 版本 3.1.2

Perl 语言 版本 v5.16.3

2.2.3 课题中用到的软件及版本

samtools 0.1.18[25]：以 sam 格式进行测序片段比对、格式处理的工具

bwa 0.7.10-r789[26]：通过 Burrows-Wheelers 算法进行测序片段比对，拼装的工具

bcftools 0.1.17-dev (r973:277)[25]: 处理 VCF/BCF 格式文件的工具

cufflinks cuffdiff cuffnorm v2.0.2[27]: 转录本拼装，转录本差异表达计算工具

miRanda v3.3a[28]: microRNA 靶点预测工具

LEGO v1.0[29]: 基因功能富集的工具

fastq-dump 2.3.5:

处理 NCBI 下载格式的工具，来自 http://trace.ncbi.nlm.nih.gov/Traces/sra/

bedtools v2.17.0[30]: 处理 bed 格式文件的工具

bwtool v1.0[31]: 处理 bigwig 格式文件的工具

2.3 实验方法

2.3.1 RNA 编辑位点的注释

RNA-seq 数据由 NCBI GEO 平台下载.sra 文件后，通过 SRA 工具组中的 fastq-dump 工具解压为_1.fastq 和_2.fastq 文件，分别代表配对测序片段中的 1 号文件和 2 号文件。两个.fastq 分别用 bwa 的 aln 算法默认参数拼装到基因组上得到.sai 文件，.sai 文件通过 bwa sampe 对配对测序数据生成.sam 比对文件，该阶段的比对最多允许有四个错配（-n 4）。将.sam 文件由 samtools 工具经过整合（merge）、排序（sort）、去除测序 PCR 扩增片段（rmdup）后，用 mpileup 程序找出 CNV 位点（先生成.bcf 文件，再通过 bcftools 处理为.vcf 文件）。

对于包含所有 CNV 位点的.vcf 文件，通过质量控制筛除因序列比对问题造成的错配，具体阈值为：VDB（位点距离偏倚）大于 0.01，MQ（比对质量）大于 30，位点处的序列（reads）

数大于等于 4，支持位点为 CNV 的序列数大于等于 2。之后，通过比较位点在正链与负链上的出现的倾向性筛除基因组上 CNV 与在 RNA 水平上新生成的 CNV。

2.3.2 RNA 编辑位点所在基因组的注释

从 UCSC table browser 上下载得到基因组注释的.bed 文件，通过 bedtools 的 intersect 工具处理确定每个 RNA 编辑位点位于的基因组区域。

2.3.3 RNA 编辑的保守性分析

基因组保守性通过下载的.bigwig 文件用 bwtool extract 处理得到每个 RNA 编辑位点的保守性分值，以 0.5 作为阈值确定保守位点与非保守位点。对于编辑位点保守性数据，将其他物种的编辑位点信息通过 Ensembl Assembly Converter 网页工具进行比较基因组学比对。

2.3.4 RNA 编辑对 microRNA 靶点调控的影响

对于处在可能的 microRNA 靶点上的 RNA 编辑位点，向上下游分别扩展 30bp 得到编辑位点周围的序列，用 miRanda 预测工具分别预测编辑前与编辑前与编辑后位点周围序列与 microRNA 结合能力的变化（阈值设为配对分大于等于 160，自由能分小于等于-20，编辑后假设 I 等于 G）。

为了确定显著性水平阈值，选用随机位点进行测试，方法如下：对于每个基因选取相同数量的位于同一区域（同在 3'UTR 上等）的随机位点作为样本，计算发现配对分在 180 时约为 p 值 0.01 的阈值（实际 0.01201），故以此作为显著性阈值，将编辑前分值大于 180 编辑后小于 160 或编辑前小于 160 编辑后大于 180 的位点作为影响 microRNA 靶点调控的 RNA 编辑位点。

2.3.5 RNA 编辑对 microRNA 转录本调控的影响

对于处在 microRNA 上位点，分别用 miRanda 预测编辑前与编辑后调控谱的改变（同样采用 180 与 160 作为上下界阈值），将基因分为编辑前后都调控、编辑前后都不调控、编辑前调控编辑后不调控和编辑前不调控编辑后调控四类。然后通过四类基因在受编辑的细胞系和不受编辑的细胞系中的表达量变化作为位点确实影响 microRNA 调控的指标。具体的，将位点不受编辑的细胞系的.bam 文件作为对照组，将位点受编辑的细胞系的.bam 文件作为实验组，通过 cuffdiff 软件标准化后获得基因的表达量然后比较。

2.3.6 RNA 编辑对蛋白质结合的影响

截取编辑前与编辑后位点周围序列对蛋白质结合模体（motif）的分值的改变确定 RNA 编辑对蛋白质结合的影响。为确定显著性阈值，同样选取 mRNA 上的随机位点判断该位点受编辑后产生的结合能力（ES 值）变化，以 0.005 作为双侧阈值，获得新结合位点的变化阈值为+0.457，失去结合位点的变化阈值为-0.522，超过阈值的导致结合能力变化的位点定义为有功能的编辑位点。

2.3.7 RNA 编辑对可变剪切的影响

取在内含子-外显子连接处周围 5bp 内存在的 RNA 编辑位点研究，检测在受编辑与不受

编辑的细胞系中周围外显子片段的表达量是否有显著的变化。使用 cuffnorm 确定各细胞系标准化后特定外显子的表达量，对在受编辑的细胞系中的表达量与在不受编辑的细胞系中的表达量做 t 检验确定影响可变剪切的编辑位点。

2.3.8 RNA 编辑对小 RNA 降解速率的影响

对于小 RNA（sRNA）在细胞核内与细胞核外的表达量数据，使用 bedtools intersect 配对小 RNA 片段，若两片段互相重叠都超过 60%则认为是同一小 RNA（-f 0.6）。所有重叠的小 RNA 的细胞核内外表达量比取对数后具有稳定的分布，并且同种小 RNA 的该改变分值在不同细胞系中排序不同，若在受编辑的细胞系与不受编辑的细胞系中该值显著不同（秩和检验），则说明编辑影响了小 RNA 的降解速率。

2.3.9 受编辑基因的功能富集

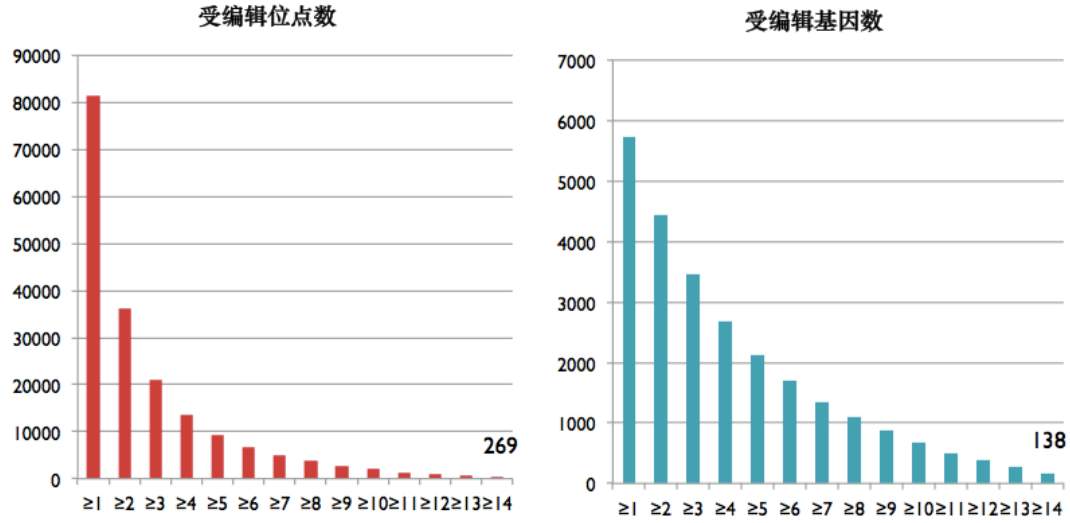
选用 LEGO server 工具进行受编辑基因的功能富集分析。富集数据集选择 KEGG 网络通路和基因本体的生物过程（GO-BP）注释。

3 研究结果

3.1 从 RNA-seq 数据注释 RNA 编辑位点

各细胞系中 RNA 编辑的情况见表 1。从 AG 百分比可以看出，定义的 RNA 编辑中大多数都为 A 到 G 的编辑，符合通常 RNA 编辑的分布。同时也可看出定义的 RNA 编辑数量与原始测序深度相关，测序深度越大，能获得的 RNA 编辑位点越多。

对于 14 种细胞系的数据，共出现 81353 个编辑位点，统计这些位点在各细胞系中的出现次数，如图 4，发现大部分编辑位点在 14 种细胞系中仅出现 1-2 次，在细胞系中反复出现的仅占不到 10%，而在每个细胞系中都出现的仅有 269 个位点。若对应到受编辑的基因数上，共有 5768 个基因受编辑，而那些在每个细胞系中都受编辑的位点对应到 138 个基因。

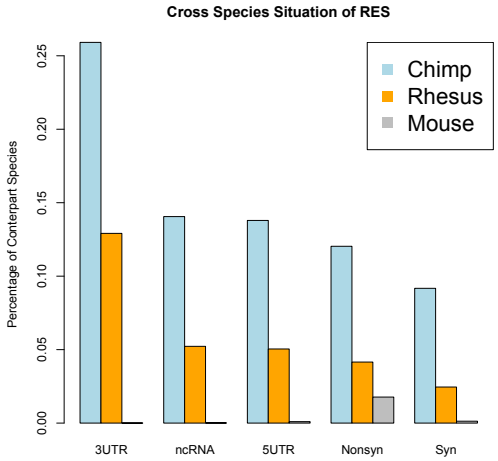


左图展示受编辑的位点数，右图展示对应的受编辑基因数

图 4. RNA 编辑在各细胞系中的出现频率。

3.2 RNA 编辑位点的细胞保守性与进化保守性关系

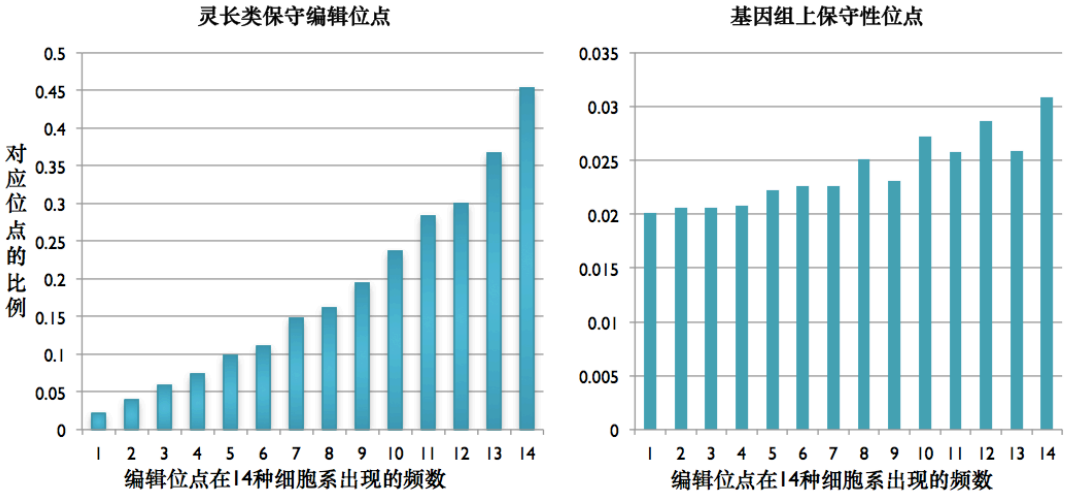
RNA 编辑的进化保守性分为编辑位点保守性和序列保守性。从 Ramaswami G *et al.*的不同物种 RNA 编辑分布数据中可以看出（图 5），对于在人类基因组中存在的位点，约 15%在黑猩猩中也存在，在恒河猴中存在的占不到 10%，而在小鼠中存在的位点不到 1%，这说明 RNA 编辑的编辑位点在灵长类中是保守的，在其他动物中是不保守，这与文献[9]中的结论相一致。这主要是由于 RNA 编辑大量发生在短串联重复序列 Alu 上，而只有灵长类基因组中才有 Alu 序列。对于序列保守性，选用来自 100 种脊椎动物基因组比对得出的 PhastCon 分值，发现仅有约 2%的 RNA 编辑位点位于 PhastCon 定义的保守序列中，低于人类基因组中的平均值 5.76%。



横轴表示对应的基因组区域，依次是 3’非编码区、非编码 RNA、5’非编码区、非同义突变位点、同义突变位点。蓝色、橙色、灰色分别代表黑猩猩、恒河猴与小鼠，纵轴表示人类基因组中横轴对应基因组区域编辑位点在其他物种中也存在的比例

图 5. RNA 编辑的编辑保守性在物种间的分布

为了确定 RNA 编辑位点的细胞保守性与进化保守性的关系，分别提取在 14 种细胞系中出现 1-14 次的位点，检测这些编辑位点位于进化保守位点的比例，得图 6。图 6 左图显示，随着 RNA 编辑的细胞系保守性上升（在细胞系见出现的次数增加），位点在灵长类基因组中也受编辑的比例显著增加，最多可达到 45%以上，说明 RNA 编辑的细胞保守性与编辑位点在 RNA 层面的保守性正相关。相应的，对应到图 6 右图基因组上的位点保守性，细胞系保守性增加对进化保守性的影响不大，说明 RNA 编辑的细胞保守性与脊椎动物基因组层面的保守性相关性不大，由于 Alu 的存在，进化保守性需要在存在 Alu 序列的灵长类范围中讨论。



横坐标越靠右表示编辑位点细胞系保守性越强，纵坐标表示在对应细胞系保守性的位点集中符合在灵长类中受编辑（左图）或位于基因组上的保守位点（右图）的比例

图 6. RNA 编辑细胞系保守性与进化保守性的关系

3.3 RNA 编辑功能性位点的注释

对于 RNA 编辑功能性位点的注释，共注释 3195 个位点。其中位于外显子影响蛋白质序列的有 920 个位点，位于 miRNA 上影响 miRNA 序列的有 3 个位点，位于 miRNA 靶点上影响 miRNA 调控的有 1364 个位点，位于蛋白质结合序列上影响蛋白质结合的有 887 个位点，位于可变剪切连接位点影响可变剪切的有 8 个位点，位于 sRNA 上影响 RNA 降解速率的位点有 15 个。

3.3.1 外显子上的 RNA 编辑位点

经注释，共有 1587 个处在外显子上的 RNA 编辑位点，其中有约 60%（920 个）为非编码突变。若考虑可变剪切，有 145 个位点始终位于外显子上。

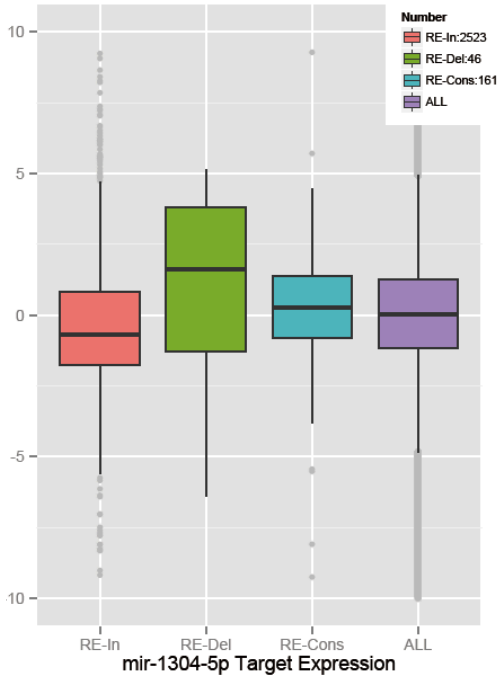
3.3.2 miRNA 上的 RNA 编辑位点

位于 miRNA 上的 RNA 编辑位点注释到三个位点，分别为 ch11:93466874 位于 miR-1304-5p 的第 16 位，chr15:59463452 位于 miR-2116-3p 的第 1 位，ch1:37966545 位于 miR-5581-5p 第 12 位。现对位点 ch11:93466874 作详细分析。

编辑位点位于 miR-1304-5p 的第 16 位，如图 7，虽然其不位于 miRNA 的核心区域（seed region），但通过 miRNA 结合预测软件预测其能显著改变 miR-1304-5p 的靶点（在 miRanda 阈值 160 的情况下，预测该位点的编辑会增加 715 个靶点，减少 148 个靶点，保留 641 个靶点）。

图 7. 位于 miR-1304-5p 上的编辑位点

为了证明该位点确实存在改变基因调控的功能，通过表达分析考察相关基因的表达变化。将该位点受编辑的 6 个细胞系作为实验组，不受编辑的 8 个细胞系作为对照组，观察对应基因的表达量变化。



横轴分别表示预测编辑前不受调控编辑后受调控的基因、预测编辑前受调控编辑后不受调控的基因、预测编辑前后都受调控的基因、所有基因。纵坐标为基因在实验组与对照组的表达比取以 10 为底的对数

图 8. miR-1304-5p 编辑前后基因的表达分布盒形图

由图 8，第四列（紫）展示集合中全部的基因表达量不变，说明表达量标准化正确。观察到第一列（红）受 RNA 编辑后新增的靶基因下调；第二列（绿）说明受 RNA 编辑后失去的靶点基因上调；第三列（蓝）显示 RNA 编辑前后都受调控的基因表达不变。这个结果下编辑导致受调控的基因下调，编辑导致不再受调控的基因上调，说明 mir-1304-5p 上的编辑位点确实影响了该 microRNA 的调控。

3.3.3 miRNA 靶点上的 RNA 编辑位点

通过 miRNA 结合位点预测软件的分析，共定义 1364 个位点可显著影响 miRNA 与靶点的结合能力，其中有 86.5%都位于 3’UTR 上，符合通常 miRNA 靶基因位点的分布。这些位点对应 330 个基因，89 个 miRNA。调控的基因未有显著的功能富集 ，各靶点被调控的 miRNA 数目相似，没有集中于某些特定的 miRNA 序列上。

3.3.4 影响蛋白质结合的 RNA 编辑位点

通过蛋白质结合数据共定义 887 个位点，其中 172 个为编辑后可被新蛋白质结合（模体获得），715 个为编辑后不能与原来的蛋白质结合（模体破坏）。在所有模体获得对应的蛋白

质中，出现频率前三的蛋白质分别为 RBFOX1（相对随机位点的奇异值比为 25.6，下同）、ANKHD1（7.13）、MSI1（6.00），这些基因的功能分别为可变剪切相关蛋白质、转录后脚手架蛋白及转录后调控蛋白。相应的，在模体破坏对应的蛋白质中，出现频率前三的蛋白分别是 A1CF（205.95）、PBFOX1（91.62）与 SNRPA（49.35），前者为 RNA 编辑相关蛋白，后两者为可变剪切相关蛋白。从以上结果可看出，RNA 编辑更多的作为模体破坏的手段，即相对于编辑后新增蛋白质结合位点，RNA 编辑倾向于编辑后破坏区域与原蛋白质的结合。从位点改变数排名高的蛋白质的功能分析可看出，RNA 编辑可通过改变蛋白质与 RNA 结合的方式行使调控可变剪切、影响其他转录后调控等具体功能。

3.3.5 影响可变剪切的 RNA 编辑位点

通过比对发现共有 274 个位点位于内含子-外显子连接处附近，其中可能影响可变剪切的位点仅 8 个。这可能是由于处在内孩子-外显子连接处的位点更容易发生错配，因此获得的 SNV 位点质量低被筛去，使得很多真阳性位点也被筛去了。

在得到的位点中，有一个值得讨论的案例。编辑位点 chr10:77930365 位于 POLR3A 基因上，其可影响第五外显子的剪切，p 值为 0.021，如图 9。同时该位点又可促进 mRNA 与 BRUNOL5 的结合（模体获得），BRUNOL5 的功能注释中提到其蛋白家族的其他基因有调控可变剪切的功能。由此，该编辑位点可能通过改变调控可变剪切的蛋白结合来参与可变剪切调控。

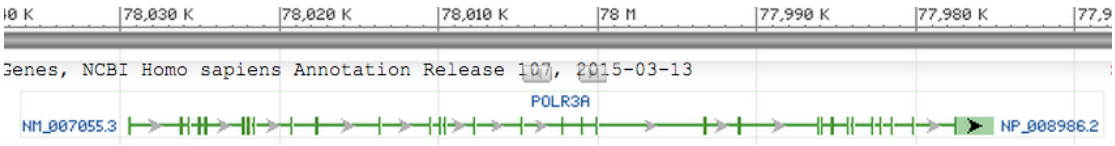


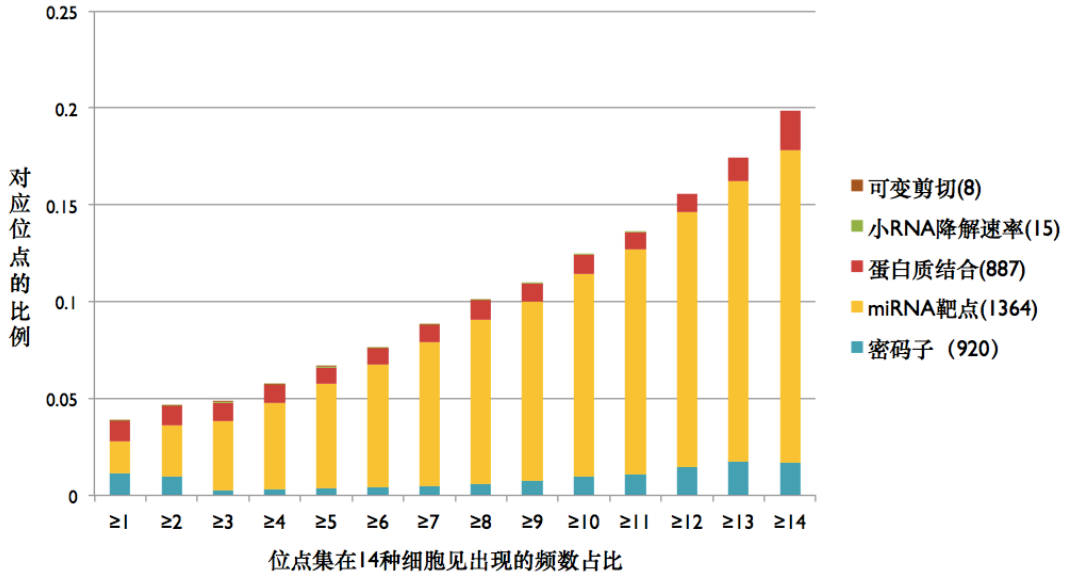
图 9. POLR3A 基因在基因组上的分布

3.3.6 影响小 RNA 降解速率的 RNA 编辑位点

通过对小 RNA 核内外的表达量比较，共筛选出编辑前与编辑后表达变化较大的位点 15 个。这个数量级说明影响小 RNA 降解速率可能不是 RNA 编辑的主要调控功能。

3.4 RNA 编辑细胞保守性与编辑位点功能性分析

将以上 3195 个高功能性 RNA 编辑位点与 RNA 编辑的细胞系保守性对照，得图 10。图中显示，随着细胞系保守性的增加，对应的功能位点比例增加。在所有位点中功能性位点仅占不到 4%，而那些高细胞系保守性的位点中有近 20%都是功能性位点，说明 RNA 编辑的细胞系保守性与功能性有强关联性，细胞保守性强的 RNA 编辑位点有明显的功能。在各功能性比较中，发现随着细胞系保守性的上升，密码子非同义突变、影响蛋白质结合、影响 miRNA 靶点三类功能都有不同程度的显著上升，其中以影响 miRNA 靶点调控功能为最显著，在高保守位点中占了约 16%，说明大部分高细胞系保守性的 RNA 编辑通过改变 miRNA 结合位点影响基因的功能。



横坐标表示编辑位点在细胞系中出现的频数，纵坐标表示这些位点中对应功能注释的比例

图 10. RNA 编辑的细胞系保守性与功能性的关系

3.5 高细胞系保守性 RNA 编辑位点的功能

通过对高细胞系保守性编辑位点对应的基因的功能富集分析，结果见图 11。图中排名前九位的功能集名称分别为核糖体、嘧啶代谢、错配修复、氰基氨基酸代谢、细胞质 DNA 感应通路、咖啡因代谢、卵母细胞减数分裂、酮体的合成与分解、黄体酮介导的卵母细胞突变。从这些功能可以看出，细胞保守性强的 RNA 编辑位点对应的基因参与了重要的生物学通路。

GO term	p-value
Cluster1:Ribosome - Homo sapiens (human)	1.9e-07
Cluster2:Pyrimidine metabolism - Homo sapiens (human)	4.8e-07
Cluster3:Mismatch repair - Homo sapiens (human)	5.3e-06
Cluster4:Cyanoamino acid metabolism - Homo sapiens (human)	8.1e-06
Cluster5:Cytosolic DNA-sensing pathway - Homo sapiens (human)	4.4e-05
Cluster6:Caffeine metabolism - Homo sapiens (human)	5.4e-05
Cluster7:Oocyte meiosis - Homo sapiens (human)	0.00021
Cluster8:Synthesis and degradation of ketone bodies - Homo sapiens (human)	0.00026
Cluster9:Progesterone-mediated oocyte maturation - Homo sapiens (human)	0.0014

图 11. 高细胞系保守性的 RNA 编辑位点对应基因的功能富集（KEGG 通路）

对于与核糖体相关的 52 个基因对应的 79 个 RNA 编辑位点，其中有 15 个高功能性位点（19.0%）。13 个位点处于 miRNA 结合位点上，2 个位于蛋白质结合位点上。比较发现，

对于高细胞系保守性的位点，位于 3’UTR 的编辑位点所占比例明显升高，达到了 16%，这也许是高细胞系保守性位点中有大量影响 miRNA 靶点的位点的原因。

4 讨论

RNA 编辑现象目前尚缺乏研究，目前已知其与多种人类疾病有关并与脑发育关系密切，具有很高的研究价值。在 RNA 编辑位点总体呈现低功能性的背景下，本探索性课题首先将研究对象缩小于那些具有细胞系保守性的编辑位点。从位点的分布图中可发现高细胞保守性的位点仅占所有编辑的一小部分，深入分析这部分编辑位点将对 RNA 编辑的总体研究有很大启发。

对于细胞保守性位点的进化分析发现，具有高细胞保守性的位点更倾向于在灵长类中也受编辑，说明细胞保守性的 RNA 编辑倾向于在进化上保守。由于大量 RNA 编辑位点分布于 Alu 序列上，而 Alu 序列仅在灵长类中存在，因此通过脊椎动物序列比对得出的基因组保守位点不适用于 RNA 编辑的分析，对 RNA 编辑的进化分析应聚焦于编辑在灵长类中是否受编辑的分析。

对细胞保守性位点的功能分析发现，细胞保守性强的 RNA 编辑位点有明显的功能。随着细胞系保守性的增强，功能占比都显著增加，其中比例最多的是改变 miRNA 结合位点影响基因的功能，说明调控 miRNA 靶点可能是高细胞系保守性位点的主要功能。

对于受编辑的基因的功能富集分析发现，细胞保守性强的 RNA 编辑位点对应的基因参与了重要的生物学通路，如核糖体相关基因，表明 RNA 编辑可能对细胞生长和发育是必要的。本课题这些结果对 RNA 编辑的后续研究分析有重要意义。

虽然取得了一些成果，本课题目前还存在一些问题，尚有改进之处。首先，RNA 编辑位点的确定过程还待优化。目前对于 RNA 编辑位点的确定还未有完美的解决方法，对于如何作跨外显子的测序片段的基因组比对，如何区分来自 RNA 编辑的 CNV 及来自 SNP、等位基因等基因组因素的 CNV 等问题还需进一步确定方案。其次，本课题的功能分析过多基于预测结果。功能性的分析中，数量最多的位点来自影响 miRNA 靶点的位点和影响蛋白质结合的位点，而这两部分都属于预测结果，缺乏表达数据或湿实验这样直接的证据，因此其中可能存在很多假阳性。相反，对于影响可变剪切及影响小 RNA 降解速率这两类功能，因分析流程不够细化，直接通过 p 值阈值筛选可能造成了很多假阴性。这样会造成功能分析存在偏倚。最后，本课题中对于 RNA 编辑中出现频率最高的 A 到 I 的编辑，在结合能力预测、序列比对等过程中统一作了 I 等于 G 的近似处理。事实上 I 与 G 虽然氢键结合相似，化学组成仍有很多不同点，对于细微的高级结构实际情况可能与近似处理的情况不同。

对于本课题的后续工作，应着重于以下几点。首先应细致优化流程，确认结论的可靠性，估计出前文所述的不足对结果的影响并加以修订。之后应验证细胞系保守性 RNA 编辑位点在其他独立细胞系中的受编辑一致性，若能验证出一致性则可进一步说明本课题找出的细胞

系保守位点是可靠的。然后应探索细胞系保守性 RNA 编辑位点对基因表达的影响，目前关于 RNA 编辑与基因表达的关系研究还较欠缺，这方面研究可能对本领域产生启示。最后，应从进化保守 RNA 编辑位点出发，重复上述研究。目前已知细胞系保守与进化保守是强相关的，但两者各自的侧重点未进行详细说明，相应的研究可填补这方面空白。

参考文献

[1] Nishikura K. Functions and regulation of RNA editing by ADAR deaminases[J]. *Annual review of biochemistry*, 2010, 79: 321.

[2] Nishikura K. Editor meets silencer: crosstalk between RNA editing and RNA interference[J]. *Nature Reviews Molecular Cell Biology*, 2006, 7(12): 919-931.

[3] Li J B, Church G M. Deciphering the functions and regulation of brain-enriched A-to-I RNA editing[J]. *Nature neuroscience*, 2013, 16(11): 1518-1522.

[4] Avesson L, Barry G. The emerging role of RNA and DNA editing in cancer[J]. *Biochimica et Biophysica Acta (BBA)-Reviews on Cancer*, 2014, 1845(2): 308-316.

[5] Nishikura K. Functions and regulation of RNA editing by ADAR deaminases[J]. *Annual review of biochemistry*, 2010, 79: 321.

[6] Tomaselli S, Locatelli F, Gallo A. The RNA editing enzymes ADARs: mechanism of action and human disease[J]. *Cell and tissue research*, 2014: 1-6.

[7] Kiran A, Baranov P V. DARNED: a DAtabase of RNa EDiting in humans[J]. *Bioinformatics*, 2010, 26(14): 1772-1776.

[8] Park E, Williams B, Wold B J, et al. RNA editing in the human ENCODE RNA-seq data[J]. *Genome research*, 2012, 22(9): 1626-1633.

[9] Ramaswami G, Li J B. RADAR: a rigorously annotated database of A-to-I RNA editing[J]. *Nucleic acids research*, 2014, 42(D1): D109-D113.

[10] Bazak L, Levanon E Y, Eisenberg E. Genome-wide analysis of Alu editability[J]. *Nucleic acids research*, 2014: gku414.

[11] Sakurai M, Ueda H, Yano T, et al. A biochemical landscape of A-to-I RNA editing in the human brain transcriptome[J]. *Genome research*, 2014.

[12] Chawla G, Sokol N S. ADAR mediates differential expression of polycistronic microRNAs[J]. *Nucleic acids research*, 2014: gku145.

[13] Rieder L E, Reenan R A. The intricate relationship between RNA structure, editing, and splicing[C]//Seminars in cell & developmental biology. *Academic Press*, 2012, 23(3): 281-288.

[14] Xu G, Zhang J. Human coding RNA editing is generally nonadaptive[J]. *Proceedings of the National Academy of Sciences*, 2014, 111(10): 3769-3774.

[15] Pinto Y, Cohen H Y, Levanon E Y. Mammalian conserved ADAR targets comprise only a small fragment of the human editosome[J]. *Genome biology*, 2014, 15(1): R5.

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

[17] Savva Y A, Reenan R A. Identification of evolutionarily meaningful information within the mammalian RNA editing landscape[J]. *Genome biology*, 2014, 15(1): 103.

[18] Djebali S, Davis C A, Merkel A, et al. Landscape of transcription in human cells[J]. *Nature*, 2012, 489(7414): 101-108.

[19] Karolchik D, Hinrichs A S, Furey T S, et al. The UCSC Table Browser data retrieval tool[J]. *Nucleic acids research*, 2004, 32(suppl 1): D493-D496.

[20] Harrow J, Frankish A, Gonzalez J M, et al. GENCODE: the reference human genome annotation for The ENCODE Project[J]. *Genome research*, 2012, 22(9): 1760-1774.

[21] Tarailo-Graovac M, Chen N. Using RepeatMasker to identify repetitive elements in genomic sequences[J]. *Current Protocols in Bioinformatics*, 2009: 4.10. 1-4.10. 14.

[22] Pollard K S, Hubisz M J, Rosenbloom K R, et al. Detection of nonneutral substitution rates on mammalian phylogenies[J]. *Genome research*, 2010, 20(1): 110-121.

[23] Kozomara A, Griffiths-Jones S. miRBase: annotating high confidence microRNAs using deep sequencing data[J]. *Nucleic acids research*, 2013: gkt1181.

[24] Ray D, Kazan H, Cook K B, et al. A compendium of RNA-binding motifs for decoding gene regulation[J]. *Nature*, 2013, 499(7457): 172-177.

[25] Li H, Handsaker B, Wysoker A, et al. The sequence alignment/map format and SAMtools[J]. *Bioinformatics*, 2009, 25(16): 2078-2079.

[26] Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform[J]. *Bioinformatics*, 2009, 25(14): 1754-1760.

[27] Trapnell C, Roberts A, Goff L, et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks[J]. *Nature protocols*, 2012, 7(3): 562-578.

[28] Betel D, Wilson M, Gabow A, et al. The microRNA. org resource: targets and expression[J]. *Nucleic acids research*, 2008, 36(suppl 1): D149-D153.

[29] Xinran Dong, Yun Hao, Billy Chang, and Weidong Tian, LEGO: a novel network-based weighting approach for gene set enrichment analysis. *Unpublished*

[30] Quinlan A R, Hall I M. BEDTools: a flexible suite of utilities for comparing genomic features[J]. *Bioinformatics*, 2010, 26(6): 841-842.

[31] Pohl A, Beato M. bwtool: a tool for bigWig files[J]. *Bioinformatics*, 2014, 30(11): 1618-1619.

后记（致谢）：

参加望道计划让我收获良多。感谢我的指导教授田卫东教授。田老师在学术上的建树与拼劲给我们全实验室树立了榜样。本论文的完成离不开他对生物信息领域的深刻见解与丰富经验。作为我的指导老师，田老师在科研方面总是可以给与我正确并及时的意见。感谢陈靖祺学姐教会我编程技巧和科研思维，为本课题的进行打下了坚实的基础。感谢卢宇蓝学长、董欣然学姐不厌其烦的帮我解决问题。感谢张峰同学对本课题的帮助，你设计的方法对本课题有非常重要的意义。感谢 FDUROP 对本课题的资助，感谢卢大儒教授在中期报告时对本课题的宝贵意见。

指导教师评语
RNA 编辑是重要的生物学现象。蔡彦玮同学经过一年多的探索，总结了现有研究结果，独创性得提出了以细胞系保守性作为切入点的研究思路，并设计了合理的验证流程予以支持，得到了令人信服的结果，对 RNA 编辑的后续研究有启示作用，同意望道项目结题。 **指导教师：田卫东**