RNA editing 位点检测原理与方法

北京基因组研究所：连明

Content

- Introduction of RNA editing

- 2 strategies to identify RNA editing sites
  - Genome sequence-based method
  - Genome sequence–independent method

- Pipeline or tools
  - Genome sequence-based method
    - GATK4
    - REDItool
  - Genome sequence–independent method
    - GIREMI

Introduction of RNA editing

- 广义：an important post-transcriptional mechanism that alters primary RNAs through the **insertion/deletion** or **modification of specific nucleotides**
- 狭义：主要为A->G

2 strategies: 1^{st}

1^{st}: Genome sequence-based method

- 将转录本与其对应的基因组序列进行比较
- 挑战：在存在测序错误与mapping不准确的干扰下，怎样从基因组范围内的SNPs中鉴定出真正的RNA editing位点？
- 解决方法：
  - use DNA-Seq data from single individuals
  - annotations in dbSNPs and several stringent filters

2 strategies: 1st

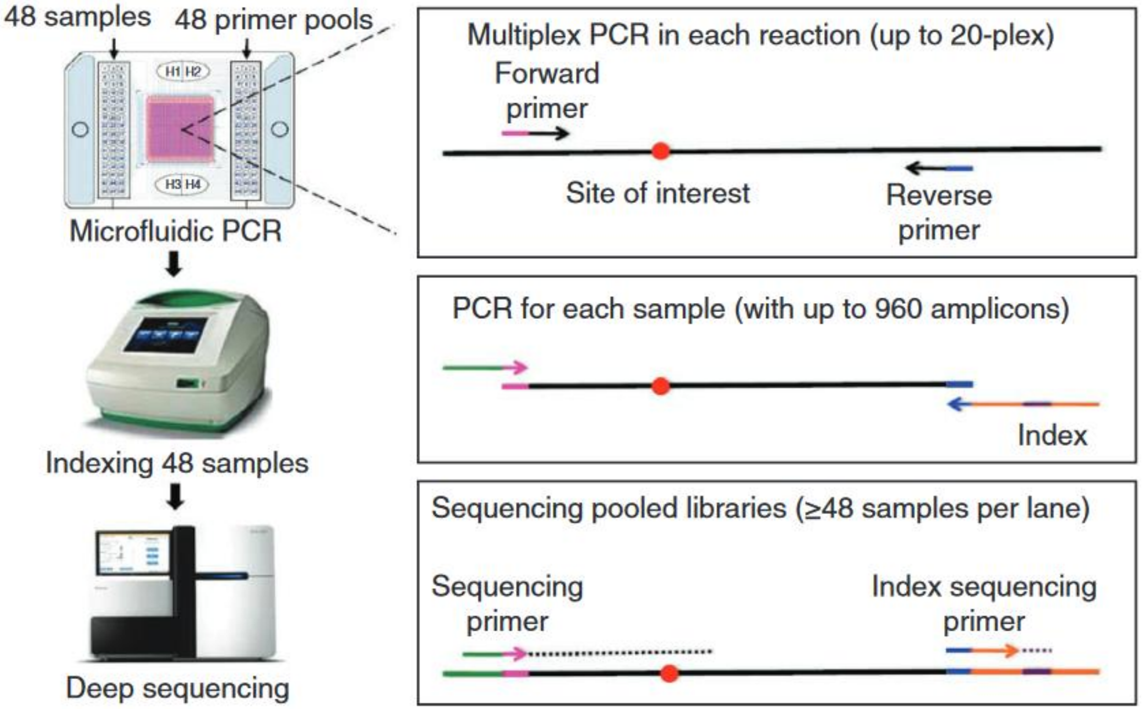
- 1st: Genome sequence-based method

优点：依据比较重测序结果与RNA-seq结果，从而获得RNA editing位点，是目前最为准确的鉴定方法

缺点：① 额外的重测序，大大增加了检测成本；②即使提供了genome sequence data，但是由于测序覆盖度（sequencing coverage）不一致等原因，使得仍然无法完全去除SNPs的干扰

2 strategies: 1st

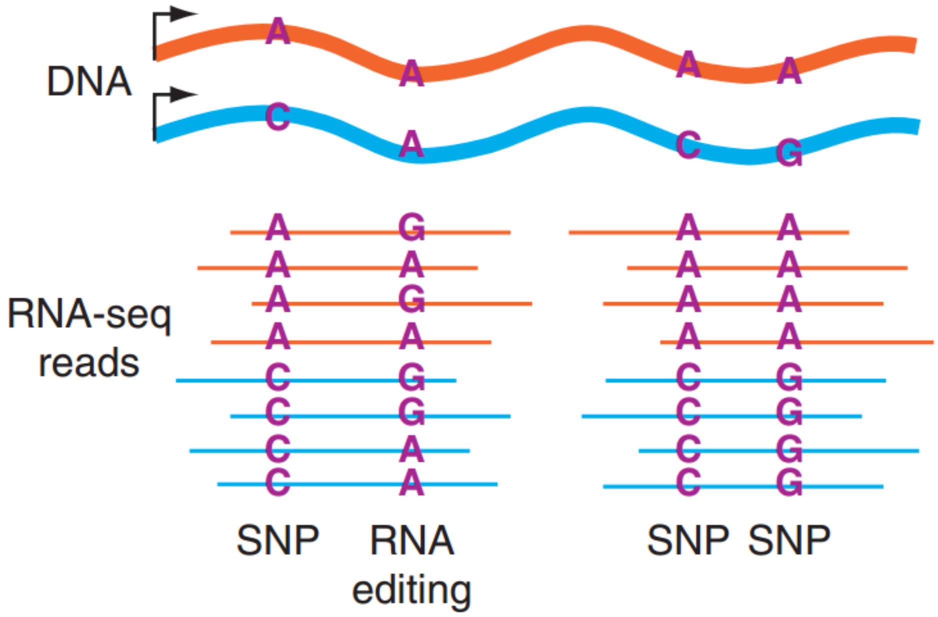
针对测序覆盖度（sequencing coverage）不一致的解决方法——mmPCR-seq



这个测序技术的关键在于进行类似454测序中用到的乳化PCR，即让每个RNA片段处于一个独立的PCR反应环境中，从而实现成比例扩增RNA片段，而不影响基因表达水平的相对定量，同时能提高对低丰度RNA的灵敏度

2 strategies: 2nd

2nd: Genome sequence–independent method

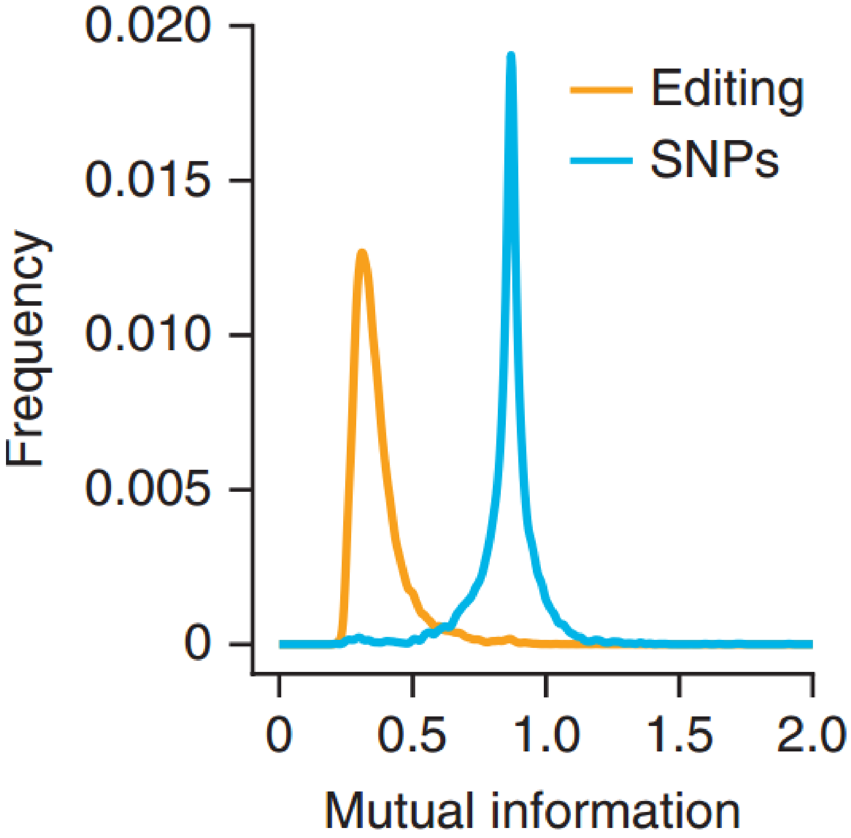


当一个SNVs pair是SNPs pair时，即不存在RNA editing时，保持严格的等位基因连锁；

而当SNVs pair中存在RNA editing时，连锁关系被破坏，连锁关系呈现出更大的随机性

2 strategies: 2nd

2nd: Genome sequence–independent method



SNPs与Editing位点的互信息分布存在显著性差异

可以利用它们互信息的差异来区分鉴别SNPs和RNA editing位点

Pipeline or tools—GATK4

- **1. Mapping of RNA-seq**
  - ➢ **Mapping**：用BWA将RNA-seq reads mapping到reference genome和已知的剪接区域附近的exonic sequences
  - ➢ **过滤**：过滤mapping结果，保留高质量的mapping结果（uniquely mapped且q > 10），并用samtools rmdup过滤PCR重复
  - ➢ **重比对与重校正**：用GATK中的 IndelRealigner 和 TableRecalibration 对保留下来的高质量的Unique reads进行局部重比对（local realignment）和碱基值重校正（base score recalibration）
- **2. Identification of editing sites from RNA-seq data**
  - ➢ **variant calling**：用GATK中的UnifiedGenotyper来call variants，与普通的variant calling不同，这里采用了比较宽松的选项：stand_call_conf 0, stand_emit_conf 0, and output mode EMIT_VARIANTS_ONLY
  - ➢ **remove all known SNPs**：利用dbSNP数据，过滤已知的SNPs

Pipeline or tools—GATK4

- **remove false positive variant calls**：过滤因技术操作原因导致的variant calling中的假阳性结果
  - required a variant call quality q > 20
  - 若variants落在read的头6个碱基里，过滤掉
  - 除去落在重复区域的variants
  - 过滤intron中离剪接位点4bp范围内的variants

Pipeline or tools—REDItools

- 三个主要脚本
  - **REDItoolDnaRNA.py**：检测候选的RNA editing位点，通过比较pre-aligned RNA-Seq 和 DNA-Seq reads（BAM format）获得
  - **REDItoolKnown.py**：explore the RNA editing potential of RNA-Seq experiments by looking at known events only
  - **REDItoolDenovo.py**：不需要重测序数据，只利用RNA-seq数据进行RNA editiong的denovo检测，检测原理类似于后面提到的基于GATK4的方法

Pipeline or tools—GIREMI

How GIREMI works ?

- **calculates the mutual information (MI)** of the mismatch pairs identified in the RNA-seq reads to distinguish RNA editing sites and SNPs.
- **trains a generalized linear model (GLM)** to achieve enhanced predictive power, which makes use of
  - sequence bias information
  - difference between the mismatch ratio of the unknown single nucleotide variants (SNVs) and the estimated allelic ratio of the gene.

Pipeline or tools—GIREMI

GIREMI底层依赖的工具

- HTSlib：这是用于对SAM/BAM文件进行读写操作的库

  注意：请将库文件所在的路径添加进 $LD_LIBRARY_PATH 环境变量

- samtools：用于构建参考基因组的faidx索引

  在运行GIREMI前，请提前用 samtools faidx命令构建好参考基因组的faidx索引，而且要保证参考基因组的fasta文件与faidx文件要位于同一文件夹下

- R：用于GLM（广义线性模型）的训练与预测

Pipeline or tools—GIREMI

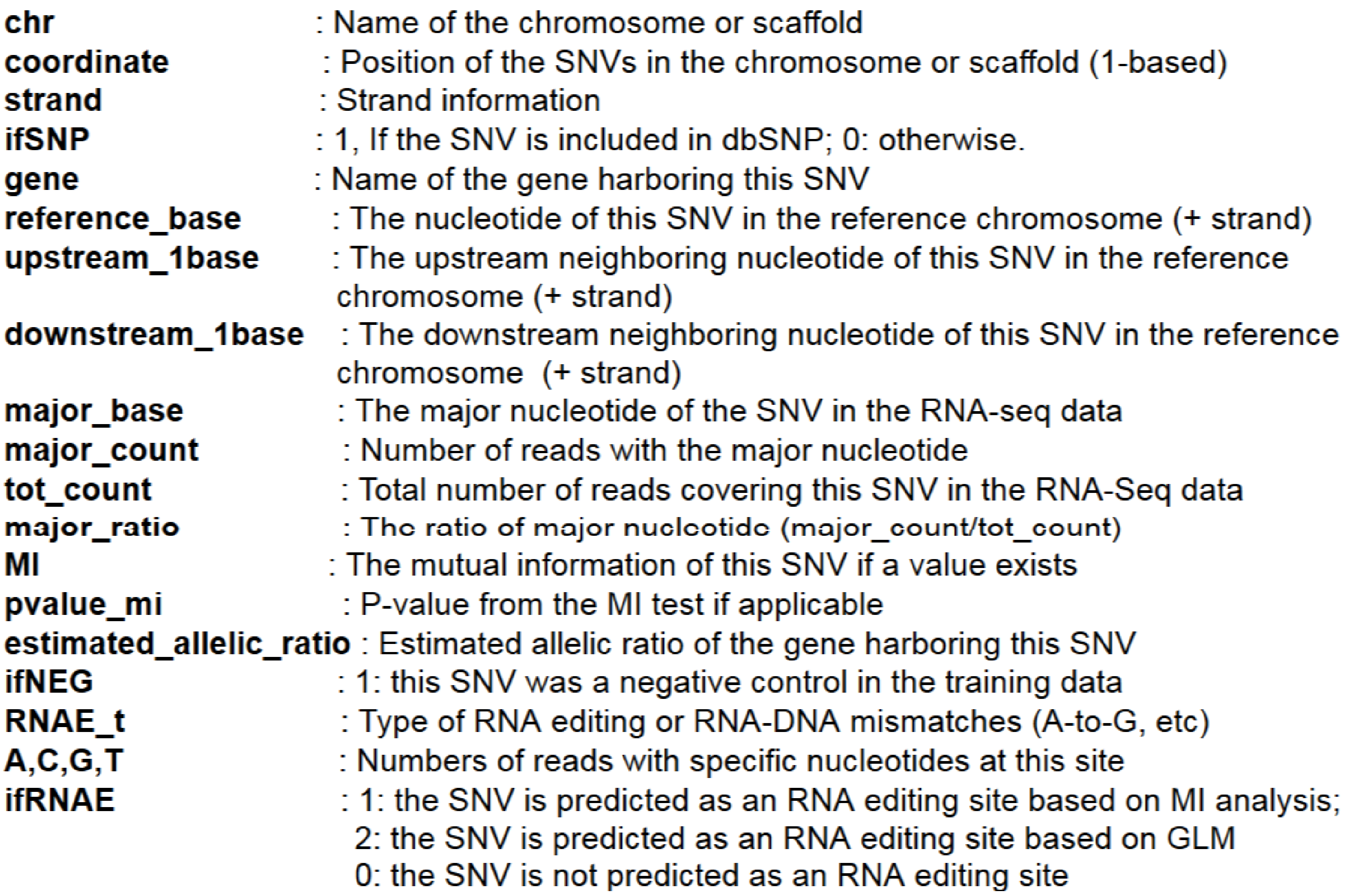
Usage：

```
giremi [options] in1.bam [in2.bam [...]]
```

重要参数：

- -f, --fasta-ref FILE reference genome sequence file in fasta format
- -m, --min INT minimal number of total reads covering candidate editing sites [default: 5]
- -p, --paired-end INT 1:paired-end RNA-Seq reads; 0:single-end [default: 1]

Pipeline or tools—GIREMI



chr	: Name of the chromosome or scaffold
coordinate	: Position of the SNVs in the chromosome or scaffold (1-based)
strand	: Strand information
ifSNP	: 1, If the SNV is included in dbSNP; 0: otherwise.
gene	: Name of the gene harboring this SNV
reference_base	: The nucleotide of this SNV in the reference chromosome (+ strand)
upstream_1base	: The upstream neighboring nucleotide of this SNV in the reference chromosome (+ strand)
downstream_1base	: The downstream neighboring nucleotide of this SNV in the reference chromosome (+ strand)
major_base	: The major nucleotide of the SNV in the RNA-seq data
major_count	: Number of reads with the major nucleotide
tot_count	: Total number of reads covering this SNV in the RNA-Seq data
major_ratio	: The ratio of major nucleotide (major_count/tot_count)
MI	: The mutual information of this SNV if a value exists
pvalue_mi	: P-value from the MI test if applicable
estimated_allelic_ratio	: Estimated allelic ratio of the gene harboring this SNV
ifNEG	: 1: this SNV was a negative control in the training data
RNAE_t	: Type of RNA editing or RNA-DNA mismatches (A-to-G, etc)
A,C,G,T	: Numbers of reads with specific nucleotides at this site
ifRNAE	: 1: the SNV is predicted as an RNA editing site based on MI analysis; 2: the SNV is predicted as an RNA editing site based on GLM 0: the SNV is not predicted as an RNA editing site

- 该文献调研笔记保存在github中：

https://github.com/Ming-Lian/NGS-analysis/blob/master/%E6%96%87%E7%8C%AE%E8%B0%83%E7%A0%94%EF%BC%9ARNA%20editing.md

持续更新中