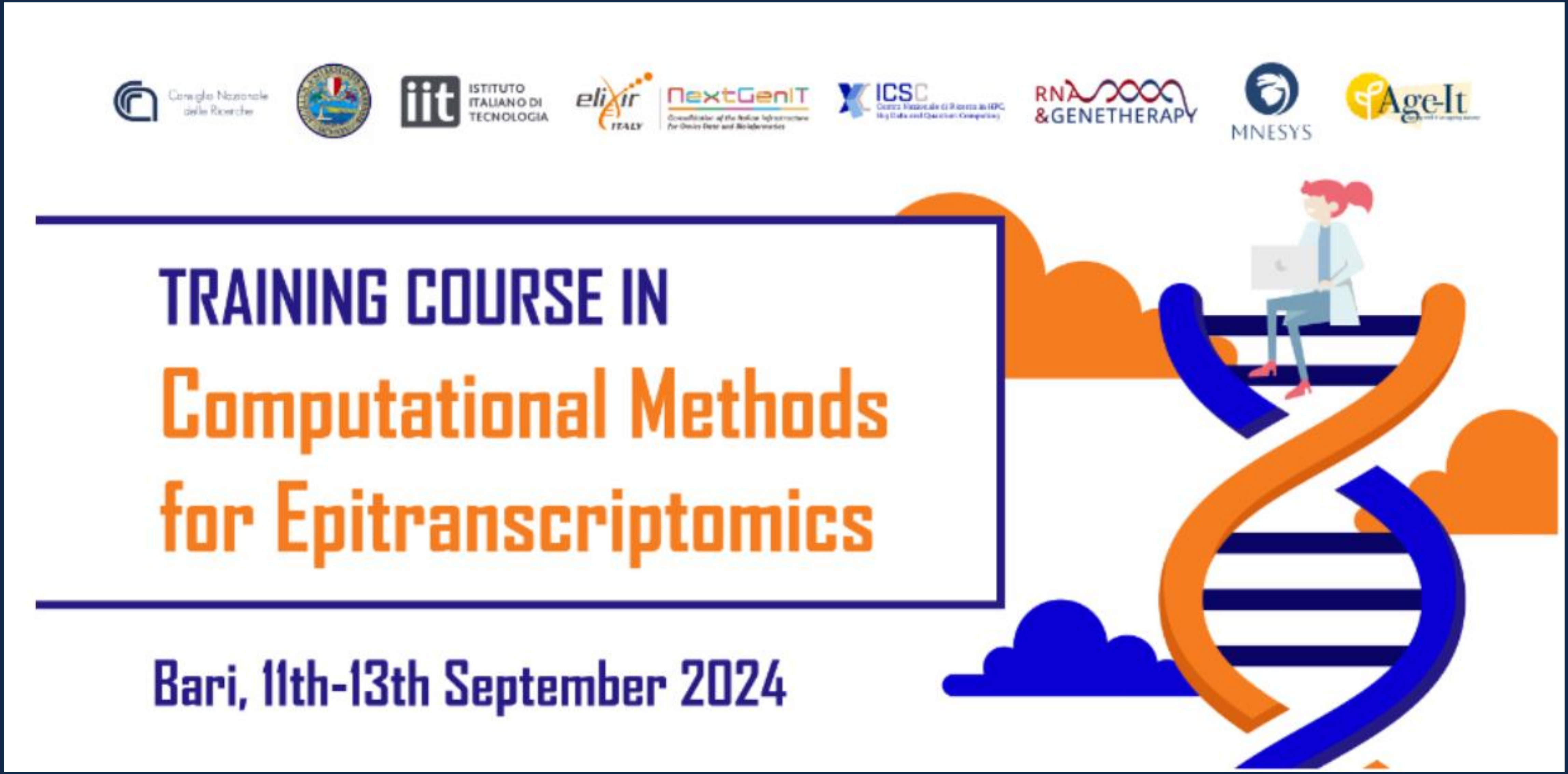
Computational approaches for detection and quantification of A-to-I RNA-editing


Consiglio Nazionale delle Ricerche
IIT ISTITUTO ITALIANO DI TECNOLOGIA
elixir ITALY
NextGenIT
ICSC
RNA &GENETHERAPY
MNESYS
Age-It

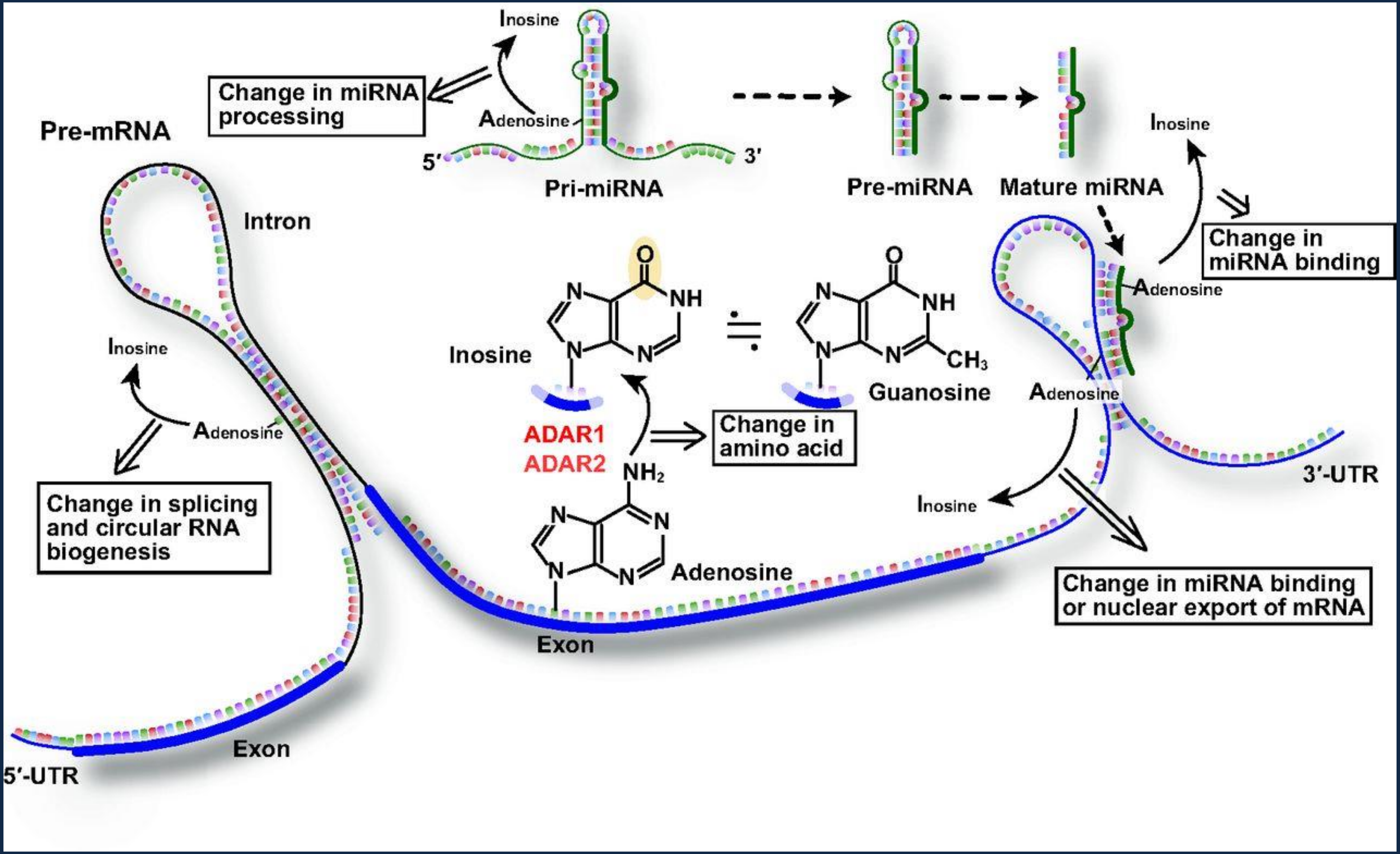
TRAINING COURSE IN

Computational Methods for Epitranscriptomics

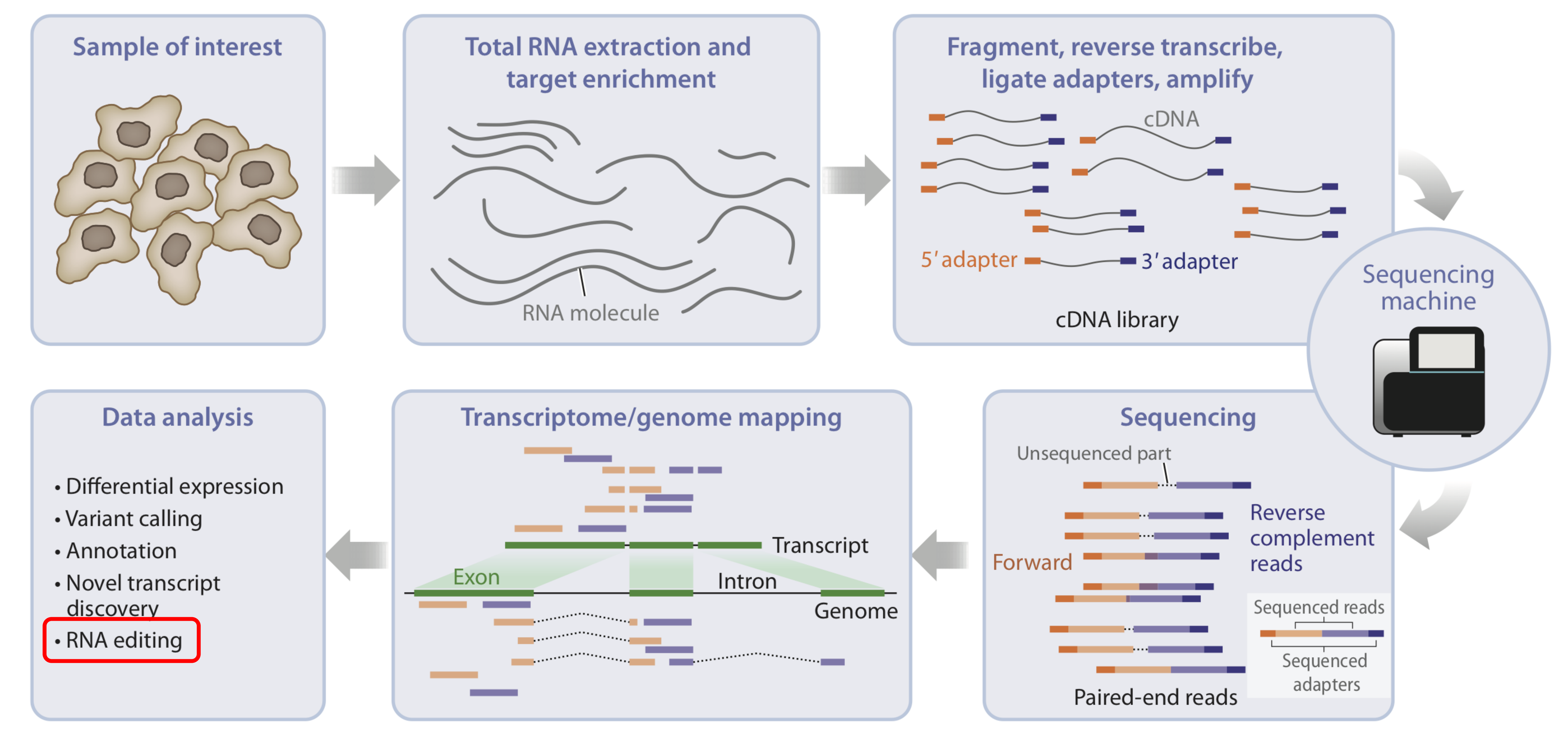
Bari, 11th-13th September 2024

Dr. Domenico Alessandro Silvestris PhD RTD-A
Dipartimento di Bioscienze, Biotecnologie e Ambiente (DBBA), UNIBA   email: domenico.silvestris@uniba.it

Functional significance of A-to-I RNA editing



From sample to data analysis



Sample of interest

Total RNA extraction and target enrichment

Fragment, reverse transcribe, ligate adapters, amplify

Sequencing machine

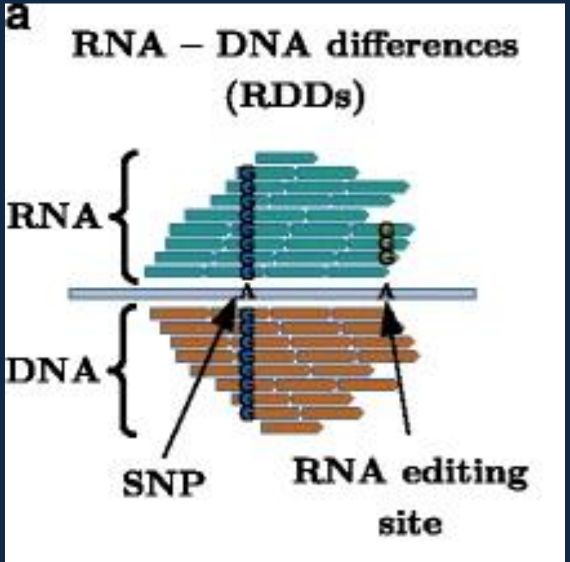
Sequencing

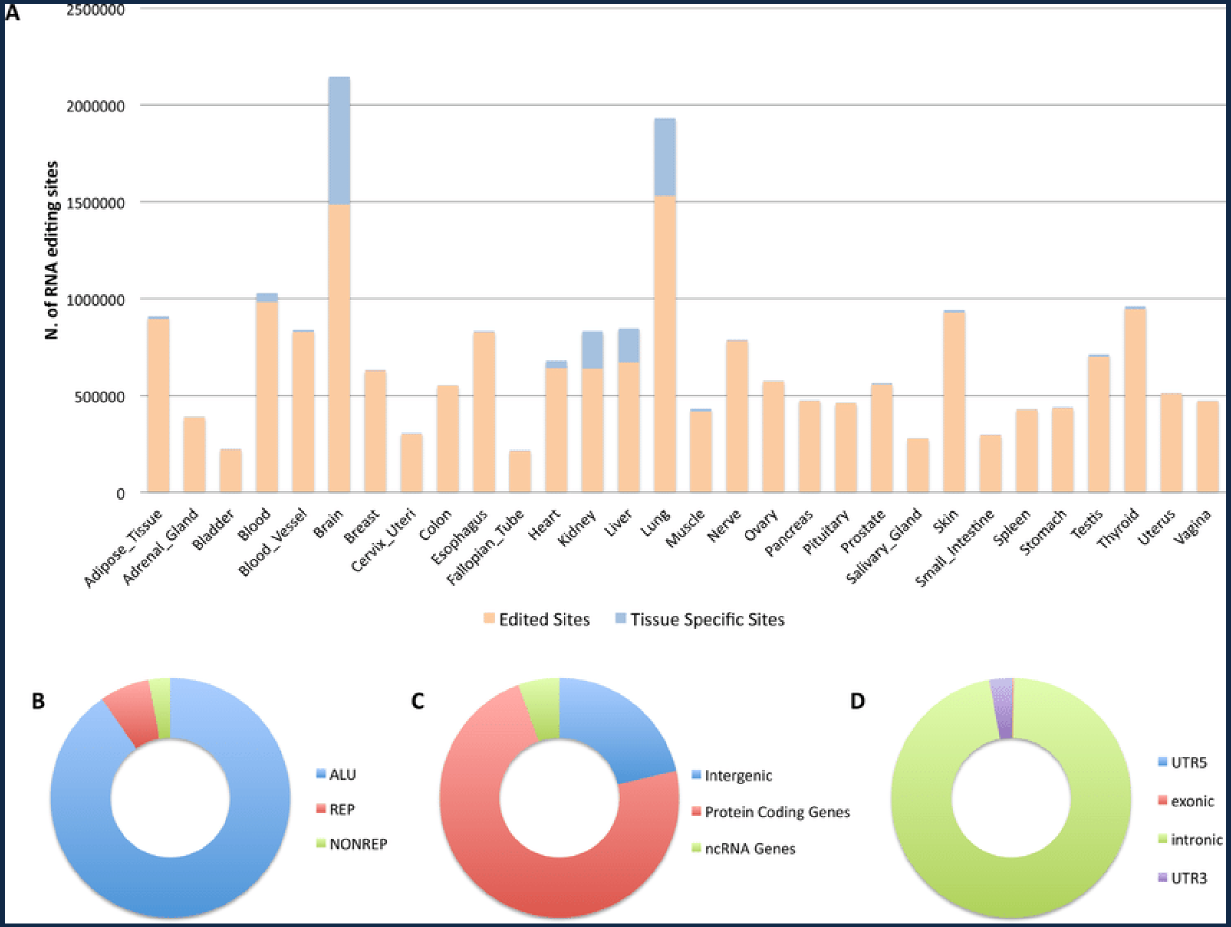
Transcriptome/genome mapping

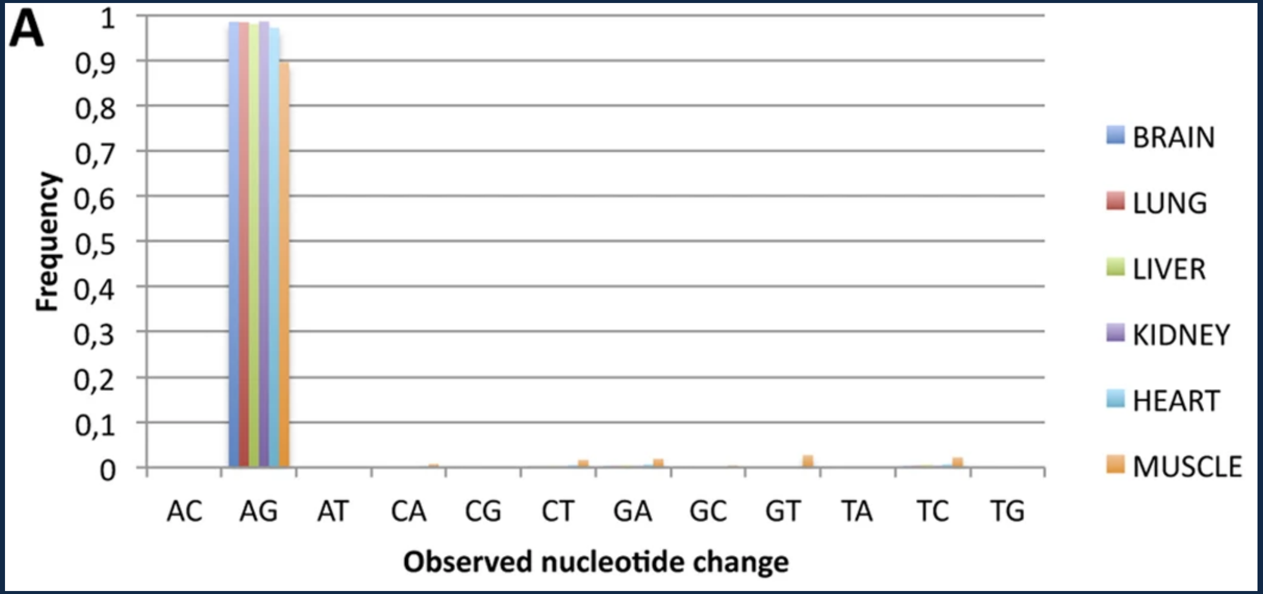
Data analysis

- Differential expression
- Variant calling
- Annotation
- Novel transcript discovery
- RNA editing

Design of RNA sequencing experiments for identifying RNA editing



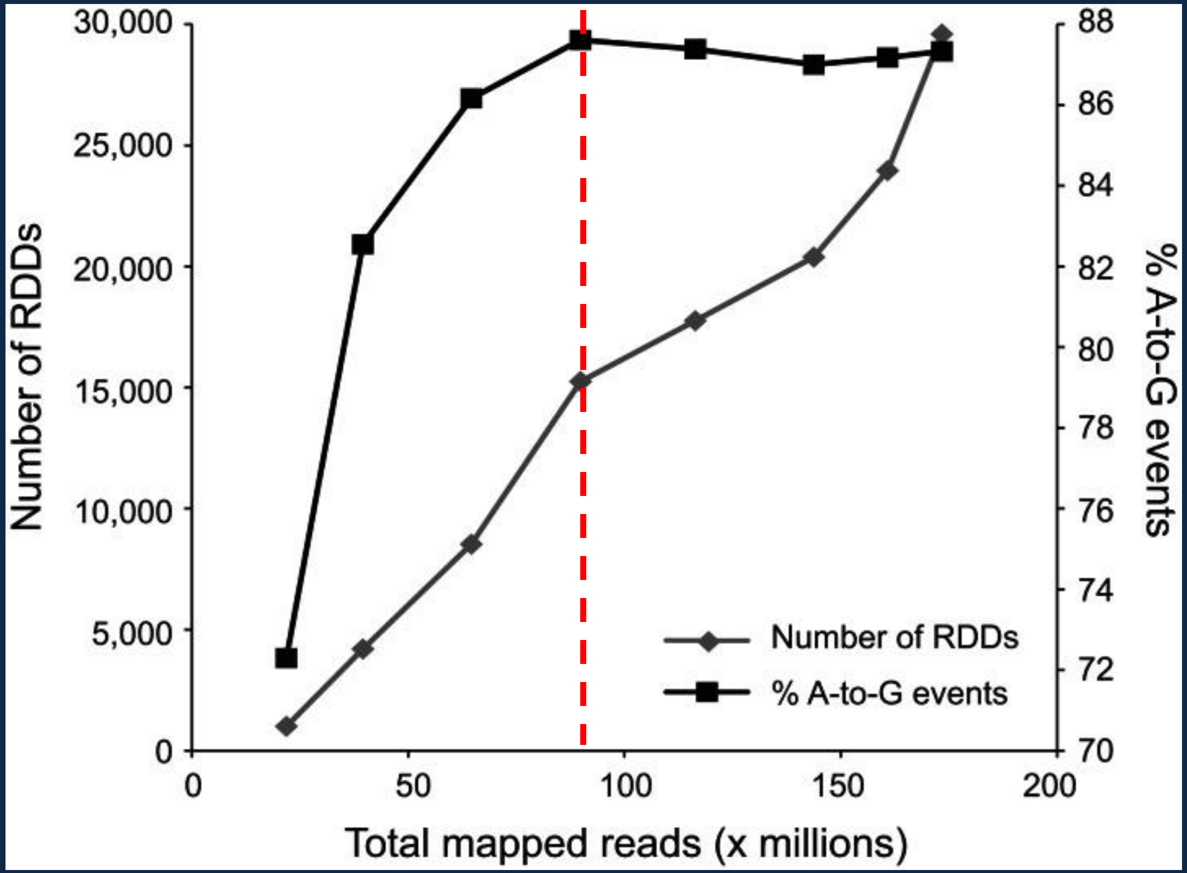




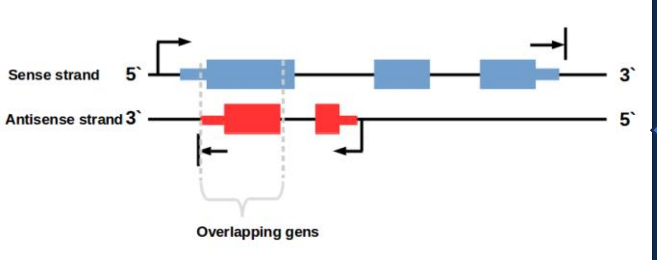
RNA editing distribution along human tissues and a graphical overview of sites stored in REDIportal. A-to-I events collected in REDIportal derive from RNAseq data encompassing 55 human body sites grouped in 30 different tissues.

12 types of RNA–DNA differences (RDDs) in human cells

Design of RNA sequencing experiments for identifying RNA editing



Number of RDDs and % of A-to-G events identified in RNA-Seq closely depend on the sequencing depth.



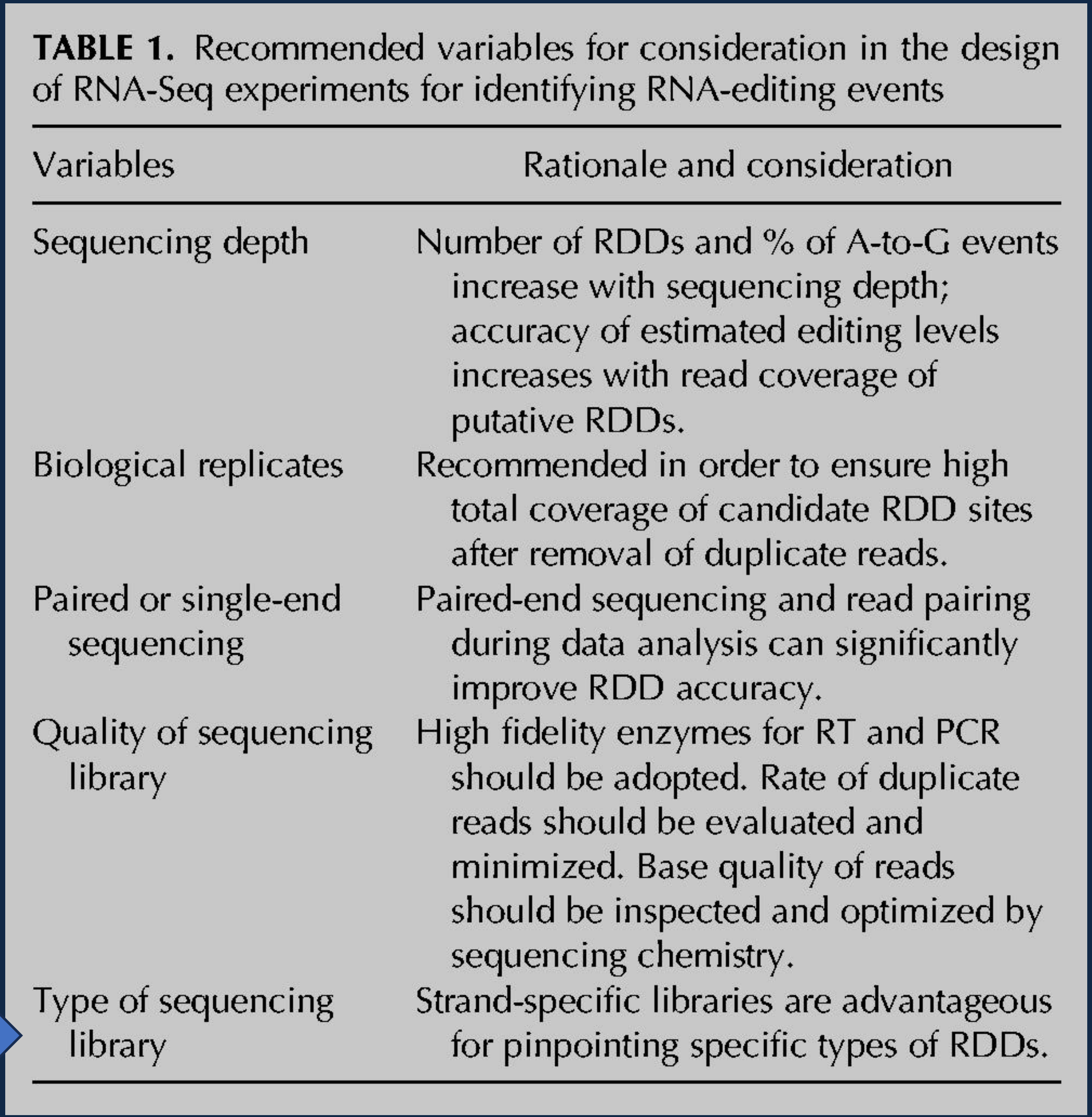


TABLE 1. Recommended variables for consideration in the design of RNA-Seq experiments for identifying RNA-editing events

Variables	Rationale and consideration
Sequencing depth	Number of RDDs and % of A-to-G events increase with sequencing depth; accuracy of estimated editing levels increases with read coverage of putative RDDs.
Biological replicates	Recommended in order to ensure high total coverage of candidate RDD sites after removal of duplicate reads.
Paired or single-end sequencing	Paired-end sequencing and read pairing during data analysis can significantly improve RDD accuracy.
Quality of sequencing library	High fidelity enzymes for RT and PCR should be adopted. Rate of duplicate reads should be evaluated and minimized. Base quality of reads should be inspected and optimized by sequencing chemistry.
Type of sequencing library	Strand-specific libraries are advantageous for pinpointing specific types of RDDs.

Lee JH, Ang JK, Xiao X. Analysis and design of RNA sequencing experiments for identifying RNA editing and other single-nucleotide variants. RNA. 2013 Jun;19(6):725-32.

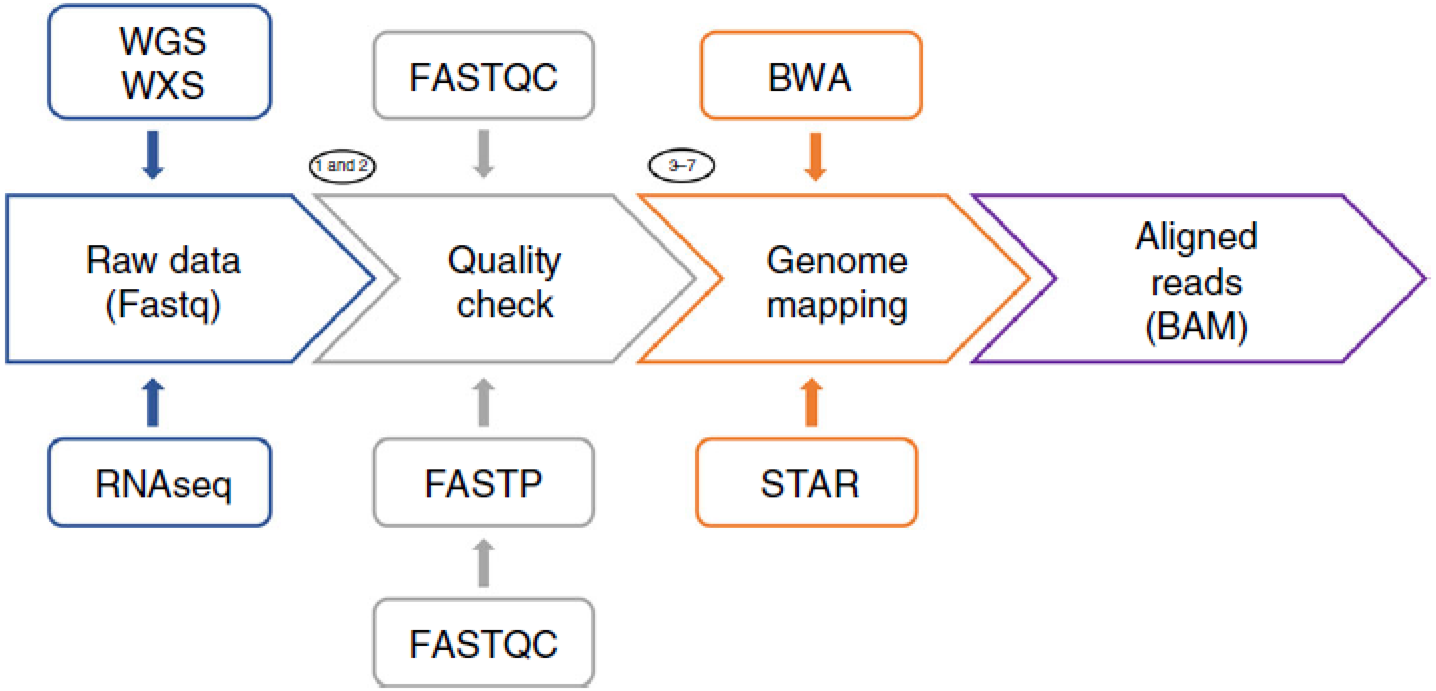
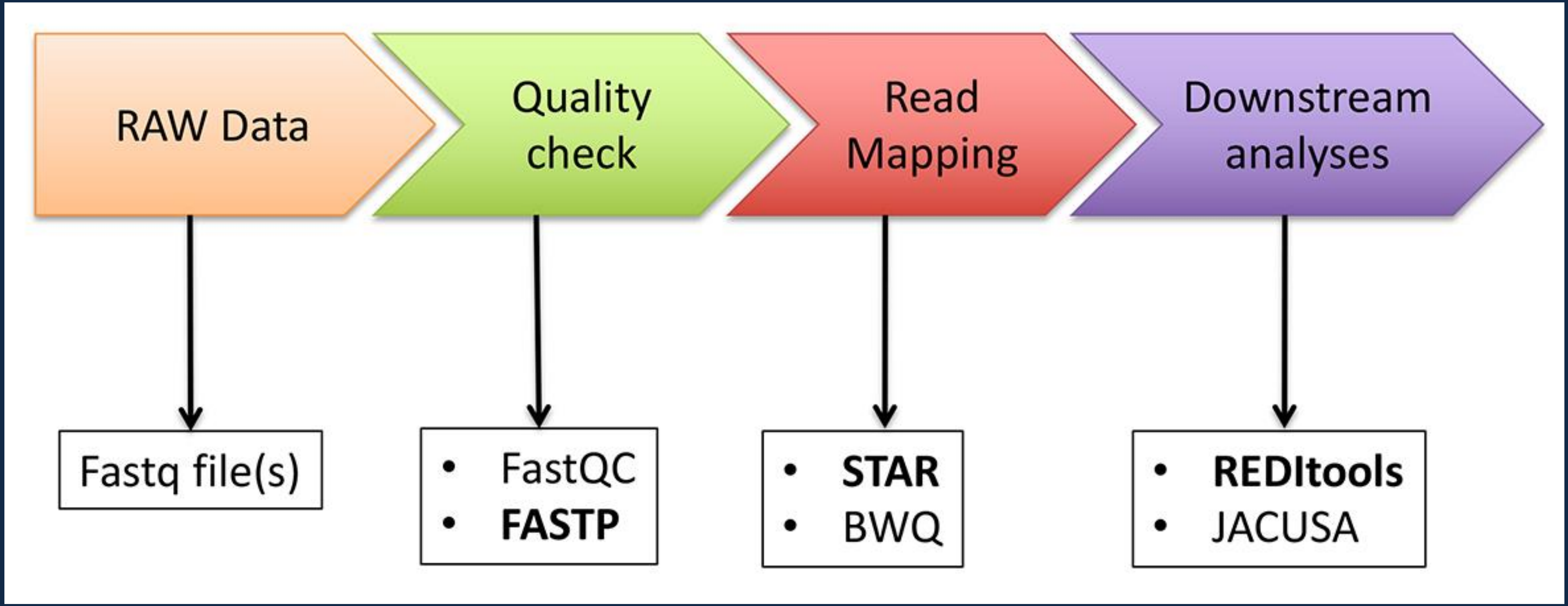


Fig. 1 | Overview of the bioinformatics workflow to preprocess data. Reliable RNA editing calls require good quality WGS and RNAseq reads. Once obtained from public databases, raw reads in fastq format are quality checked using FASTQC and cleaned using FASTP (Steps 1 and 2, Procedures 1 and 2). Then, RNAseq reads are aligned to the reference genome using a splice-aware software like STAR, while WGS reads are mapped using BWA (Steps 3–7, Procedures 1 and 2). Finally, aligned reads are converted into the standard BAM format for the downstream detection of RNA editing.

Lo Giudice, C., Tangaro, M.A., Pesole, G. *et al.* Investigating RNA editing in deep transcriptome datasets with REDItools and REDIportal. *Nat Protoc* **15**, 1098–1131 (2020)



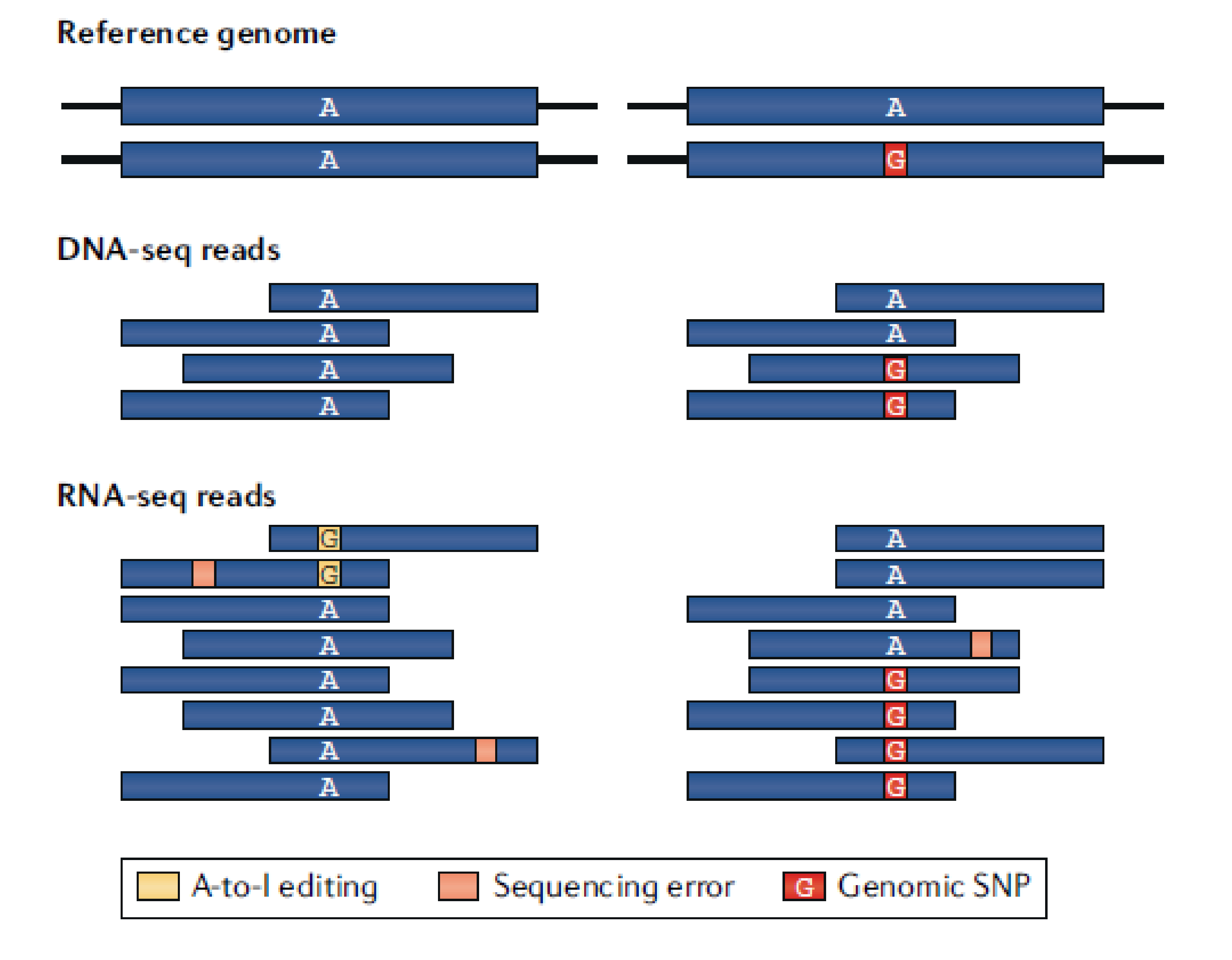
RNA editing detection and quantification

***De novo* approach**

RNA editing candidates can be detected using REDItools. There are two current versions: 1) REDItools 1.3 or 2) REDItools 2.0. REDItools2 is a faster re-implementation of REDItools1 for HPC clusters. Its serial version is about ten times faster than REDItools1.

“Known” approach

While the *de novo* approach provides a list of most likely editing candidates, the "known" approach focuses on a limited pool of known events in order to better investigate RNA editing dynamics in different experimental contexts. The "known" approach can be carried out using the REDItools package and a list of events from own data or from public databases such as DARNED, RADAR and REDIportal.

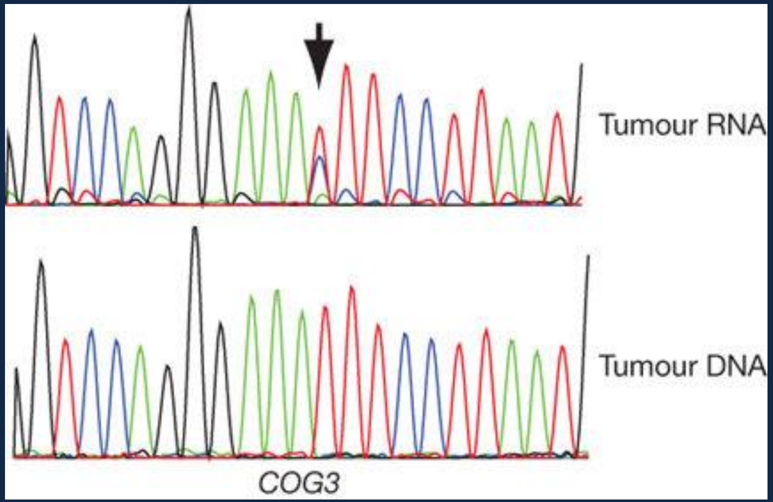


Reverse transcription replaces inosines in mRNA with guanosines in the cDNA. Thus, the hallmark of RNA editing is a consistent A → G mismatch between RNA sequencing (RNA-seq) data and the reference genomic sequence to which it is aligned. However, most of these mismatches arise from sequencing errors and genomic polymorphisms, including somatic mutations and incorrect alignment. Matched DNA sequencing (DNA-seq) data may be utilized to distinguish between editing events and genomic polymorphisms. At an editing site, the DNA reads agree with the genome reference (left), while a genomically polymorphic site exhibits mismatches in both DNA-seq and RNA-seq data.

Eisenberg, E., Levanon, E.Y. A-to-I RNA editing — immune protector and transcriptome diversifier. Nat Rev Genet 19, 473–490 (2018).

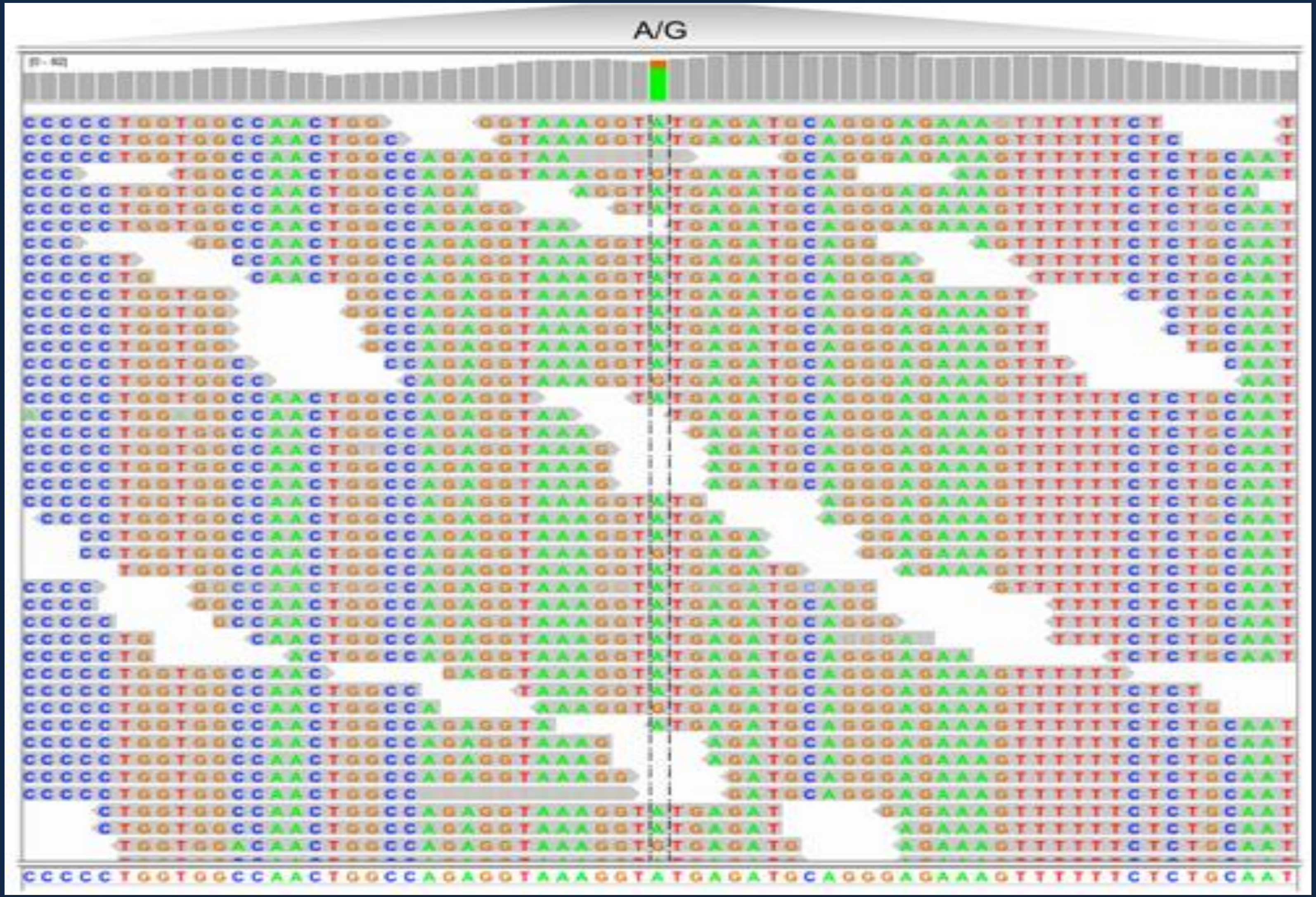
RNA EDITING

RNA editing quantification by RNA-Seq



Sanger validation of editing candidates

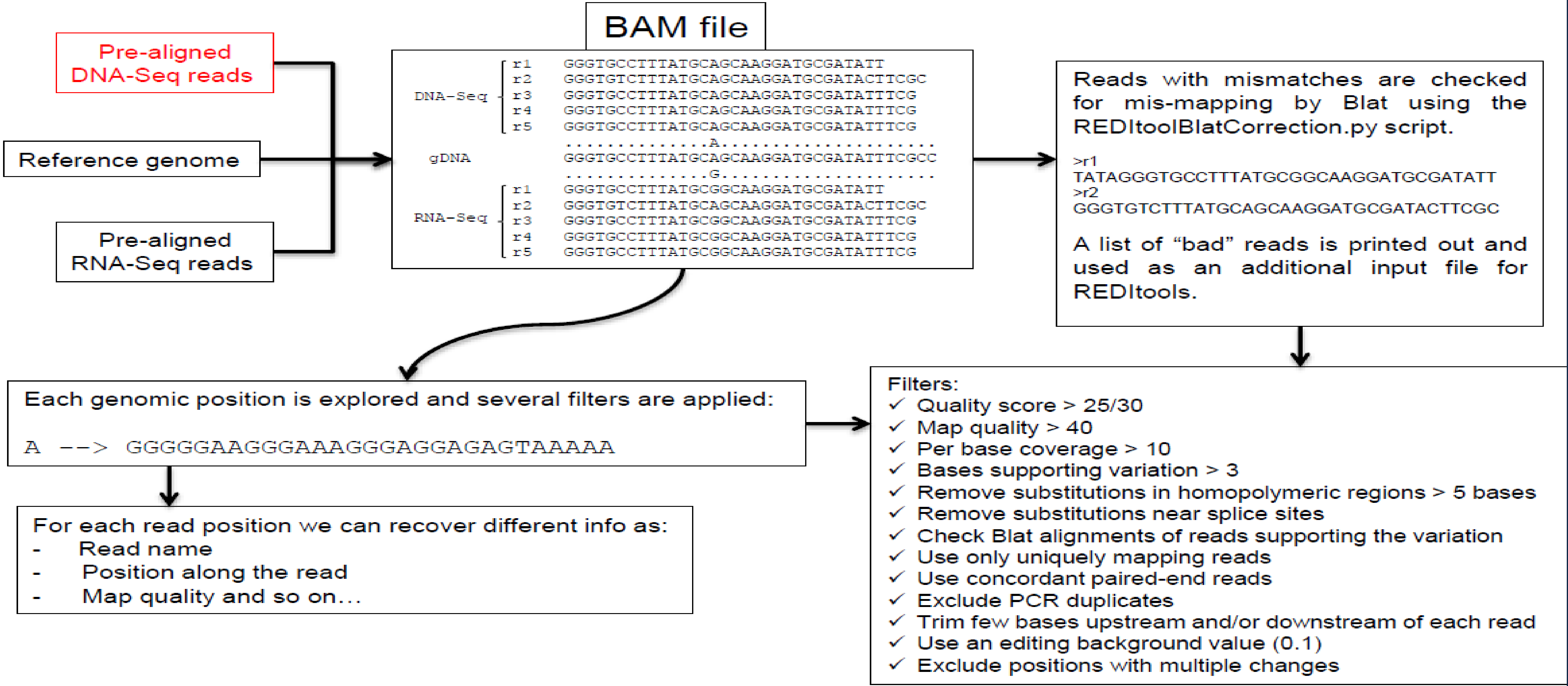
r1
r2
r3
r4
r5
r6
r7
r8
r9
r10
r11
r12
r13
r14
r15
r16
r17
r18
r19
r20
r21
r22
r23
r24
....


A/G
CCCCCTGGTGGCCAACTGGCCAGAGGTAAAGGTATGAGATGCAGGGAGAAAGTTTTTTTCTCTGCAAT

coverage 32: n.A 27 n.G 5 -> editing frequency 15,6%

RNA editing and NGS

Workflow to call RNA editing by REDItools.



Pre-aligned DNA-Seq reads

Reference genome

Pre-aligned RNA-Seq reads

BAM file

```
          [ r1   GGGTGCCTTTATGCAGCAAGGATGCGATATT
          [ r2   GGGTGTCTTTATGCAGCAAGGATGCGATACTTCGC
DNA-Seq  -[ r3   GGGTGCCTTTATGCAGCAAGGATGCGATATTTCG
          [ r4   GGGTGCCTTTATGCAGCAAGGATGCGATATTTCG
          [ r5   GGGTGCCTTTATGCAGCAAGGATGCGATATTTCG
                 ..............A......................
 gDNA            GGGTGCCTTTATGCAGCAAGGATGCGATATTTCGCC
                 ..............G......................
          [ r1   GGGTGCCTTTATGCGGCAAGGATGCGATATT
          [ r2   GGGTGTCTTTATGCAGCAAGGATGCGATACTTCGC
RNA-Seq  -[ r3   GGGTGCCTTTATGCGGCAAGGATGCGATATTTCG
          [ r4   GGGTGCCTTTATGCGGCAAGGATGCGATATTTCG
          [ r5   GGGTGCCTTTATGCGGCAAGGATGCGATATTTCG
```

Reads with mismatches are checked for mis-mapping by Blat using the REDItoolBlatCorrection.py script.

```
>r1
TATAGGGTGCCTTTATGCGGCAAGGATGCGATATT
>r2
GGGTGTCTTTATGCAGCAAGGATGCGATACTTCGC
```

A list of “bad” reads is printed out and used as an additional input file for REDItools.

Each genomic position is explored and several filters are applied:

```
A  -->  GGGGGAAGGGAAAGGGAGGAGAGTAAAAA
```

For each read position we can recover different info as:

- Read name
- Position along the read
- Map quality and so on…

Filters:

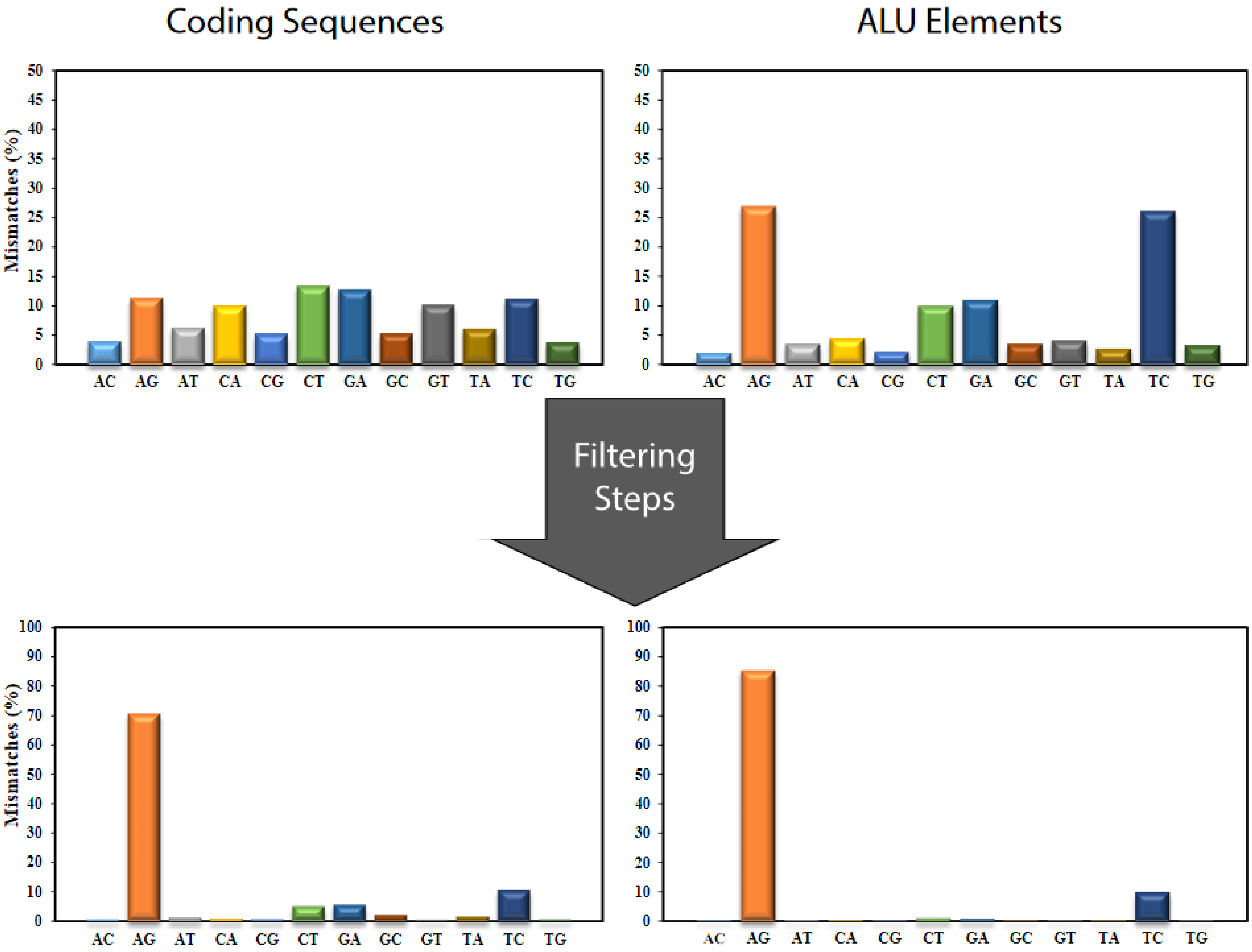
- ✓ Quality score > 25/30
- ✓ Map quality > 40
- ✓ Per base coverage > 10
- ✓ Bases supporting variation > 3
- ✓ Remove substitutions in homopolymeric regions > 5 bases
- ✓ Remove substitutions near splice sites
- ✓ Check Blat alignments of reads supporting the variation
- ✓ Use only uniquely mapping reads
- ✓ Use concordant paired-end reads
- ✓ Exclude PCR duplicates
- ✓ Trim few bases upstream and/or downstream of each read
- ✓ Use an editing background value (0.1)
- ✓ Exclude positions with multiple changes

Suggested filteri

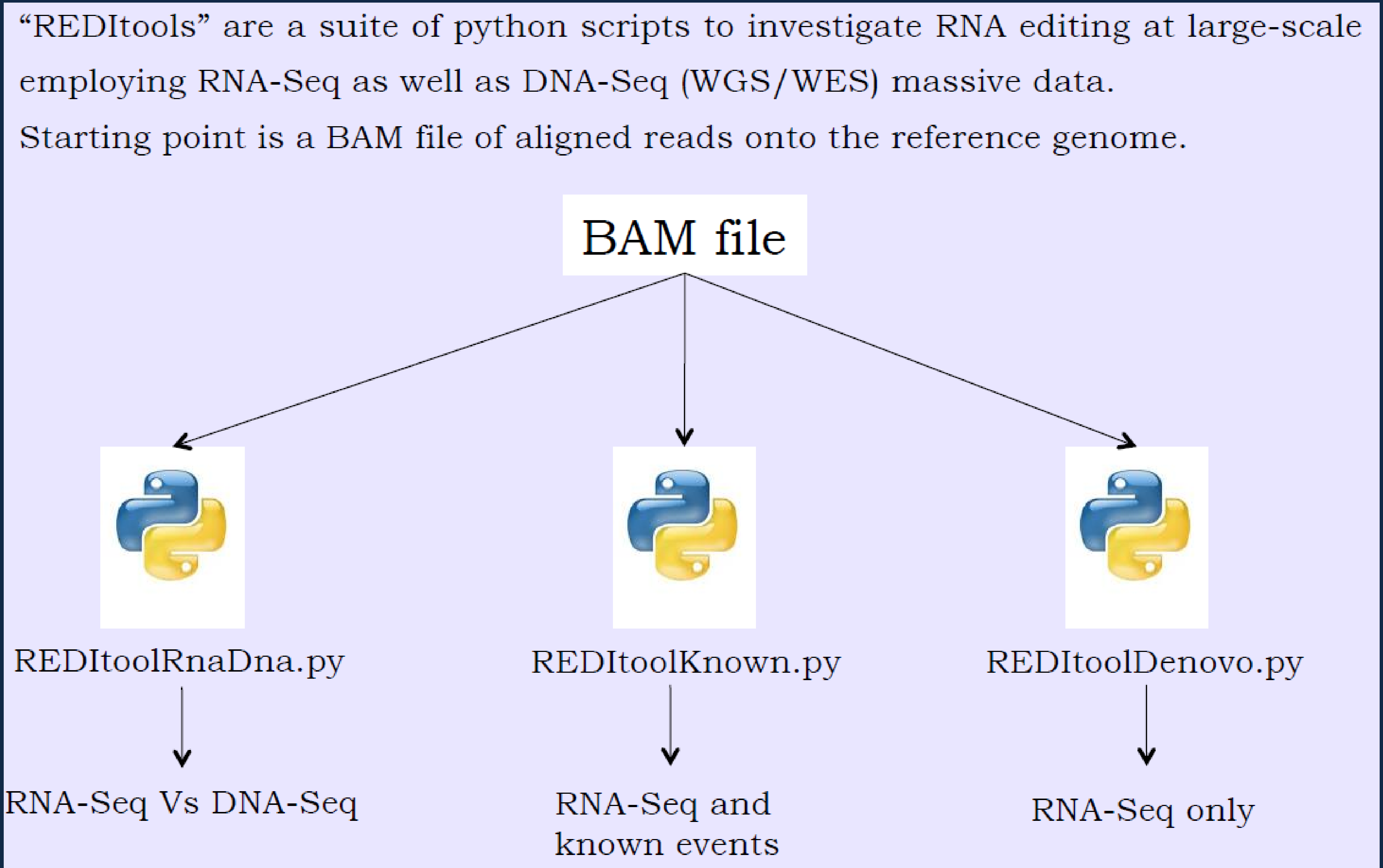
Filter

Min. read covera
Min. quality sco
Min. mapping sc
Min. number of
Min. editing freq
Exclude multi-m
Exclude duplicat
Exclude genomi
Trim bases in re
Sites near splice d.
Exclude position errors
Clip overlapping
Remove sites in
Remove sites in
Remove sites wi alogous regions
Filter DNA varia

* Filtering ste
** Require ma



Using REDItools to Detect RNA Editing Events in NGS Datasets



"REDItools" are a suite of python scripts to investigate RNA editing at large-scale employing RNA-Seq as well as DNA-Seq (WGS/WES) massive data.

Starting point is a BAM file of aligned reads onto the reference genome.

BAM file

REDItoolRnaDna.py

RNA-Seq Vs DNA-Seq

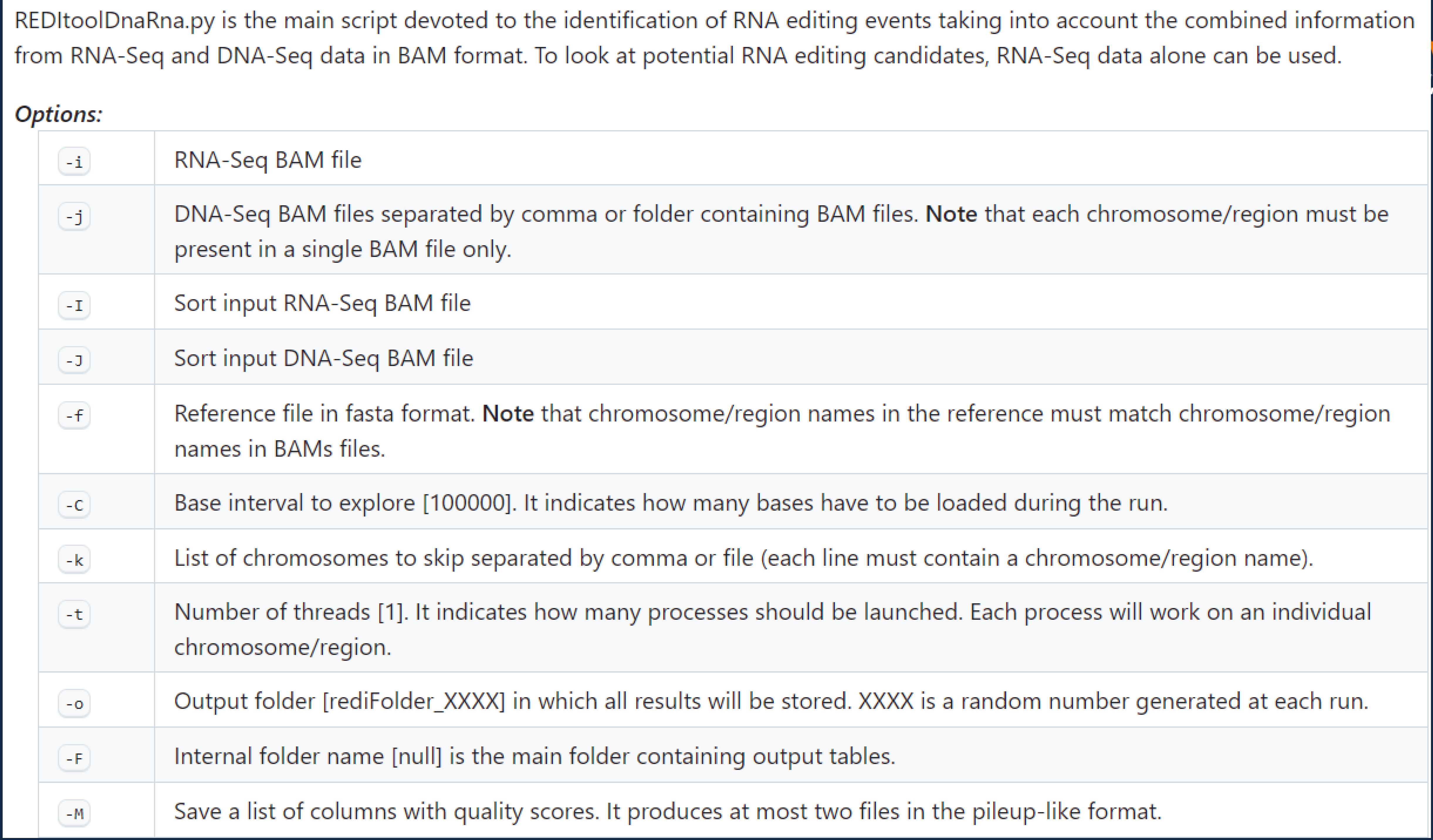
REDItoolKnown.py

RNA-Seq and known events

REDItoolDenovo.py

RNA-Seq only

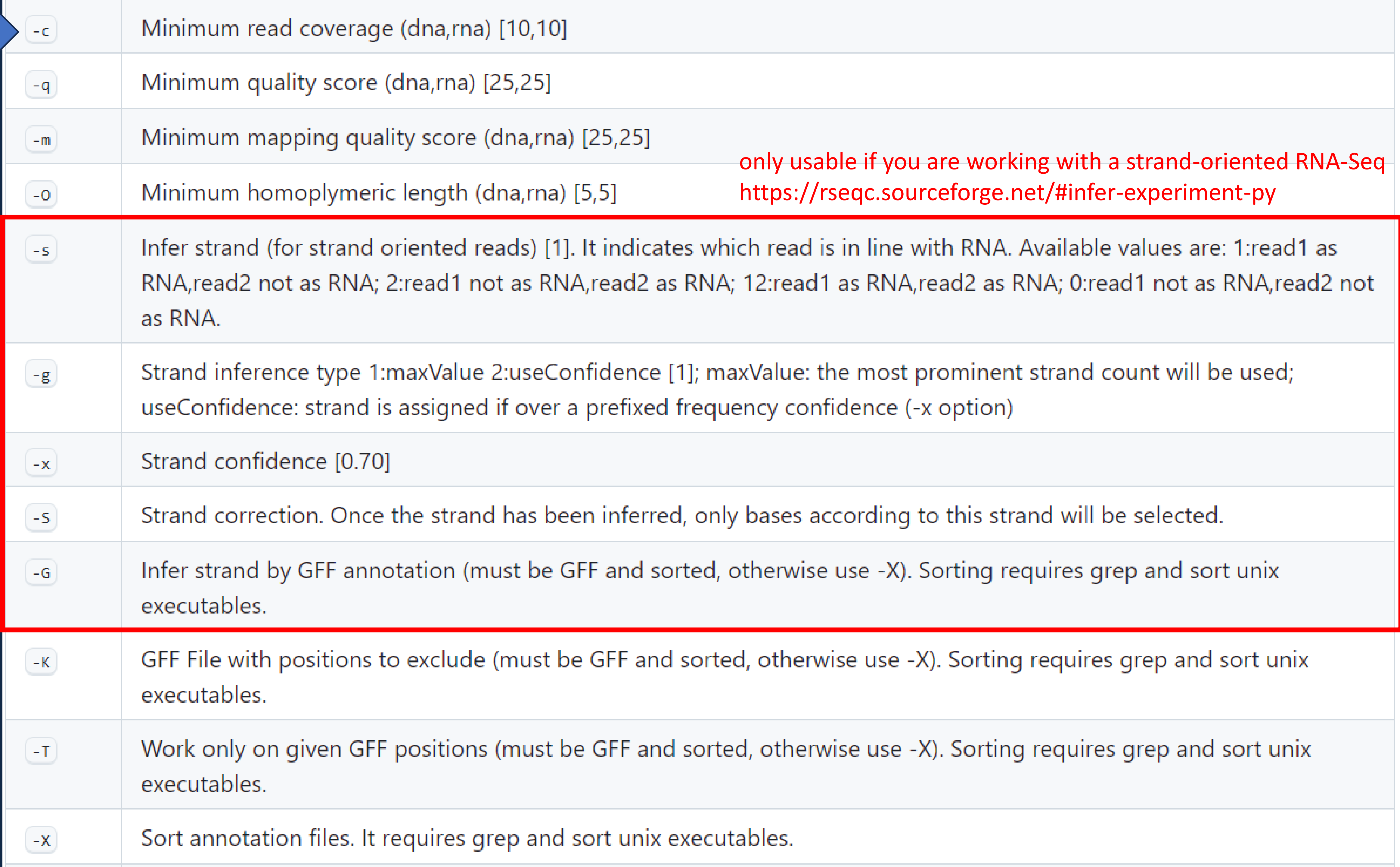
https://github.com/BioinfoUNIBA/REDItools



REDItoolDnaRna.py is the main script devoted to the identification of RNA editing events taking into account the combined information from RNA-Seq and DNA-Seq data in BAM format. To look at potential RNA editing candidates, RNA-Seq data alone can be used.

Options:

`-i`	RNA-Seq BAM file
`-j`	DNA-Seq BAM files separated by comma or folder containing BAM files. **Note** that each chromosome/region must be present in a single BAM file only.
`-I`	Sort input RNA-Seq BAM file
`-J`	Sort input DNA-Seq BAM file
`-f`	Reference file in fasta format. **Note** that chromosome/region names in the reference must match chromosome/region names in BAMs files.
`-C`	Base interval to explore [100000]. It indicates how many bases have to be loaded during the run.
`-k`	List of chromosomes to skip separated by comma or file (each line must contain a chromosome/region name).
`-t`	Number of threads [1]. It indicates how many processes should be launched. Each process will work on an individual chromosome/region.
`-o`	Output folder [rediFolder_XXXX] in which all results will be stored. XXXX is a random number generated at each run.
`-F`	Internal folder name [null] is the main folder containing output tables.
`-M`	Save a list of columns with quality scores. It produces at most two files in the pileup-like format.

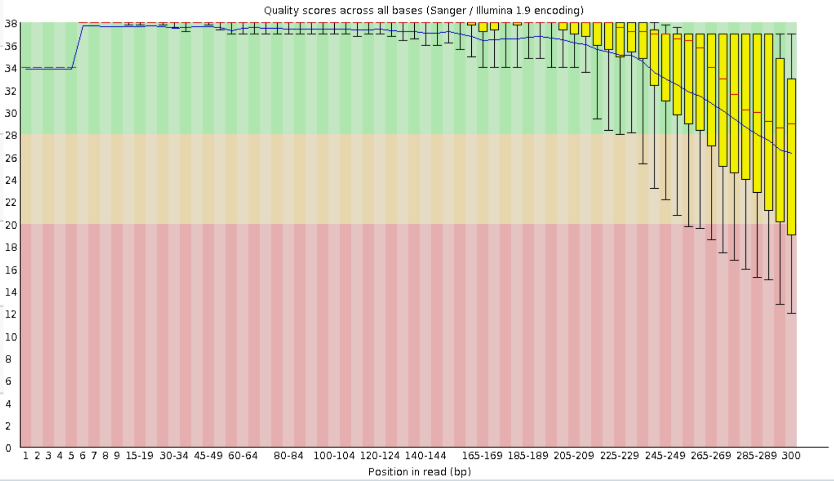


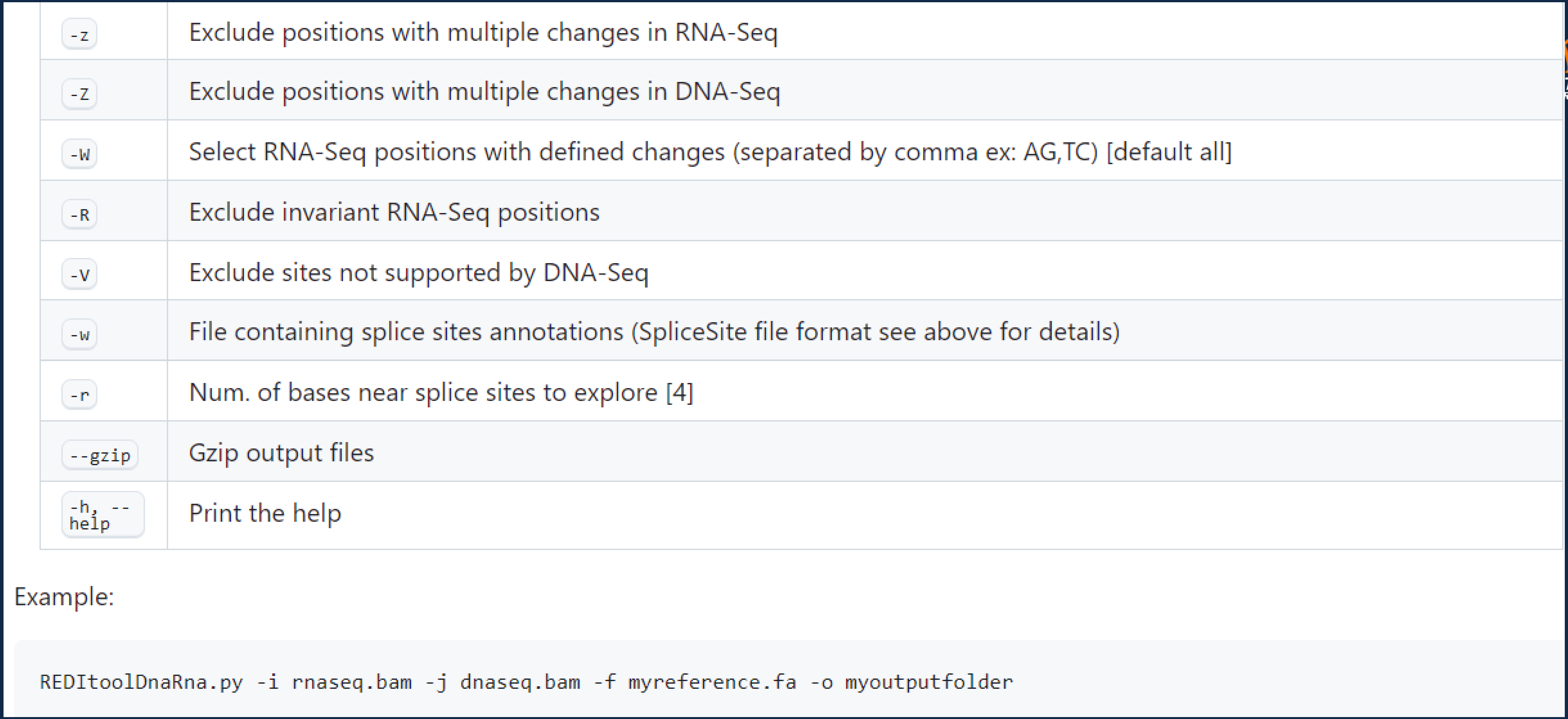
`-c`	Minimum read coverage (dna,rna) [10,10]
`-q`	Minimum quality score (dna,rna) [25,25]
`-m`	Minimum mapping quality score (dna,rna) [25,25]
`-O`	Minimum homoplymeric length (dna,rna) [5,5]
`-s`	Infer strand (for strand oriented reads) [1]. It indicates which read is in line with RNA. Available values are: 1:read1 as RNA,read2 not as RNA; 2:read1 not as RNA,read2 as RNA; 12:read1 as RNA,read2 as RNA; 0:read1 not as RNA,read2 not as RNA.
`-g`	Strand inference type 1:maxValue 2:useConfidence [1]; maxValue: the most prominent strand count will be used; useConfidence: strand is assigned if over a prefixed frequency confidence (-x option)
`-x`	Strand confidence [0.70]
`-S`	Strand correction. Once the strand has been inferred, only bases according to this strand will be selected.
`-G`	Infer strand by GFF annotation (must be GFF and sorted, otherwise use -X). Sorting requires grep and sort unix executables.
`-K`	GFF File with positions to exclude (must be GFF and sorted, otherwise use -X). Sorting requires grep and sort unix executables.
`-T`	Work only on given GFF positions (must be GFF and sorted, otherwise use -X). Sorting requires grep and sort unix executables.
`-X`	Sort annotation files. It requires grep and sort unix executables.

only usable if you are working with a strand-oriented RNA-Seq
https://rseqc.sourceforge.net/#infer-experiment-py

-e	Exclude multi hits in RNA-Seq
-E	Exclude multi hits in DNA-Seq
-d	Exclude duplicates in RNA-Seq
-D	Exclude duplicates in DNA-Seq
-p	Use paired concardant reads only in RNA-Seq
-P	Use paired concardant reads only in DNA-Seq
-u	Consider mapping quality in RNA-Seq
-U	Consider mapping quality in DNA-Seq
-a	Trim x bases up and y bases down per read [0-0] in RNA-Seq
-A	Trim x bases up and y bases down per read [0-0] in DNA-Seq
-b	Blat folder for correction in RNA-Seq
-B	Blat folder for correction in DNA-Seq
-l	Remove substitutions in homopolymeric regions in RNA-Seq
-L	Remove substitutions in homopolymeric regions in DNA-Seq
-v	Minimum number of reads supporting the variation [3] for RNA-Seq
-n	Minimum editing frequency [0.1] for RNA-Seq
-N	Minimum variation frequency [0.1] for DNA-Seq

PCR duplicates must first be marked, for example, with Picard MarkDuplicates or Samtools





`-z`	Exclude positions with multiple changes in RNA-Seq
`-Z`	Exclude positions with multiple changes in DNA-Seq
`-W`	Select RNA-Seq positions with defined changes (separated by comma ex: AG,TC) [default all]
`-R`	Exclude invariant RNA-Seq positions
`-V`	Exclude sites not supported by DNA-Seq
`-w`	File containing splice sites annotations (SpliceSite file format see above for details)
`-r`	Num. of bases near splice sites to explore [4]
`--gzip`	Gzip output files
`-h, --help`	Print the help

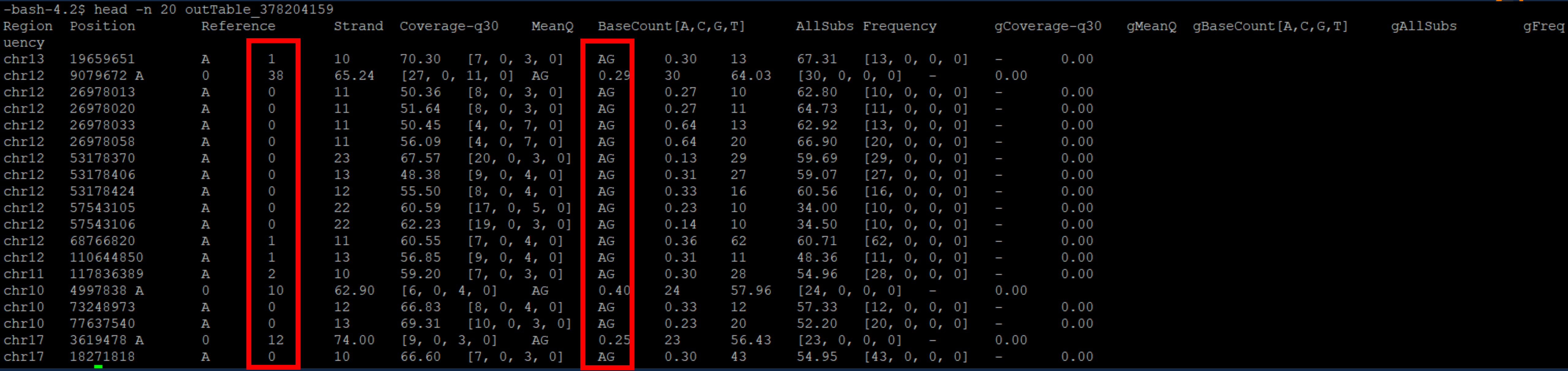
Example:

```
REDItoolDnaRna.py -i rnaseq.bam -j dnaseq.bam -f myreference.fa -o myoutputfolder
```

Mapping quality value may change according to the aligner:

- For Bowtie use 255
- For Bowtie2 use 40
- For BWA use 30
- For RNA-STAR use 255
- For HiSAT2 use 60
- For Tophat1 use 255
- For Tophat2 use 50
- For GSNAP use 30

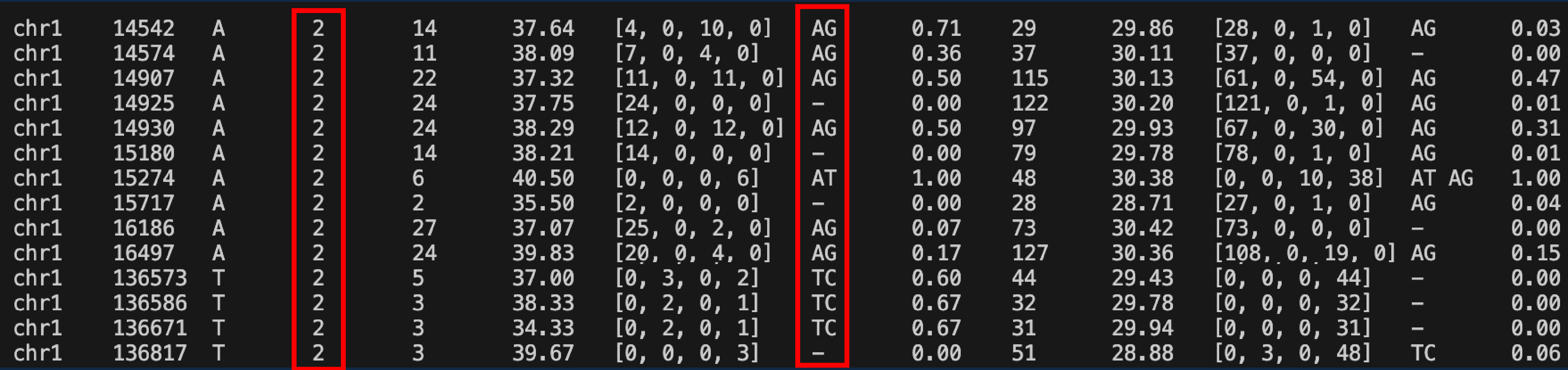
Output strand oriented RNA-Seq



```
-bash-4.2$ head -n 20 outTable_378204159
Region  Position        Reference       Strand  Coverage-q30    MeanQ   BaseCount[A,C,G,T]      AllSubs Frequency       gCoverage-q30   gMeanQ  gBaseCount[A,C,G,T]     gAllSubs        gFreq
uency
chr13   19659651        A       1       10      70.30   [7, 0, 3, 0]    AG      0.30    13      67.31   [13, 0, 0, 0]   -       0.00
chr12   9079672 A       0       38      65.24   [27, 0, 11, 0]  AG      0.29    30      64.03   [30, 0, 0, 0]   -       0.00
chr12   26978013        A       0       11      50.36   [8, 0, 3, 0]    AG      0.27    10      62.80   [10, 0, 0, 0]   -       0.00
chr12   26978020        A       0       11      51.64   [8, 0, 3, 0]    AG      0.27    11      64.73   [11, 0, 0, 0]   -       0.00
chr12   26978033        A       0       11      50.45   [4, 0, 7, 0]    AG      0.64    13      62.92   [13, 0, 0, 0]   -       0.00
chr12   26978058        A       0       11      56.09   [4, 0, 7, 0]    AG      0.64    20      66.90   [20, 0, 0, 0]   -       0.00
chr12   53178370        A       0       23      67.57   [20, 0, 3, 0]   AG      0.13    29      59.69   [29, 0, 0, 0]   -       0.00
chr12   53178406        A       0       13      48.38   [9, 0, 4, 0]    AG      0.31    27      59.07   [27, 0, 0, 0]   -       0.00
chr12   53178424        A       0       12      55.50   [8, 0, 4, 0]    AG      0.33    16      60.56   [16, 0, 0, 0]   -       0.00
chr12   57543105        A       0       22      60.59   [17, 0, 5, 0]   AG      0.23    10      34.00   [10, 0, 0, 0]   -       0.00
chr12   57543106        A       0       22      62.23   [19, 0, 3, 0]   AG      0.14    10      34.50   [10, 0, 0, 0]   -       0.00
chr12   68766820        A       1       11      60.55   [7, 0, 4, 0]    AG      0.36    62      60.71   [62, 0, 0, 0]   -       0.00
chr12   110644850       A       1       13      56.85   [9, 0, 4, 0]    AG      0.31    11      48.36   [11, 0, 0, 0]   -       0.00
chr11   117836389       A       2       10      59.20   [7, 0, 3, 0]    AG      0.30    28      54.96   [28, 0, 0, 0]   -       0.00
chr10   4997838 A       0       10      62.90   [6, 0, 4, 0]    AG      0.40    24      57.96   [24, 0, 0, 0]   -       0.00
chr10   73248973        A       0       12      66.83   [8, 0, 4, 0]    AG      0.33    12      57.33   [12, 0, 0, 0]   -       0.00
chr10   77637540        A       0       13      69.31   [10, 0, 3, 0]   AG      0.23    20      52.20   [20, 0, 0, 0]   -       0.00
chr17   3619478 A       0       12      74.00   [9, 0, 3, 0]    AG      0.25    23      56.43   [23, 0, 0, 0]   -       0.00
chr17   18271818        A       0       10      66.60   [7, 0, 3, 0]    AG      0.30    43      54.95   [43, 0, 0, 0]   -       0.00
```

Output unstranded RNA-Seq

strand (+ or -). You can also indicate strand by 0 (strand -), 1 (strand +) or 2 (+ and - or unknown)

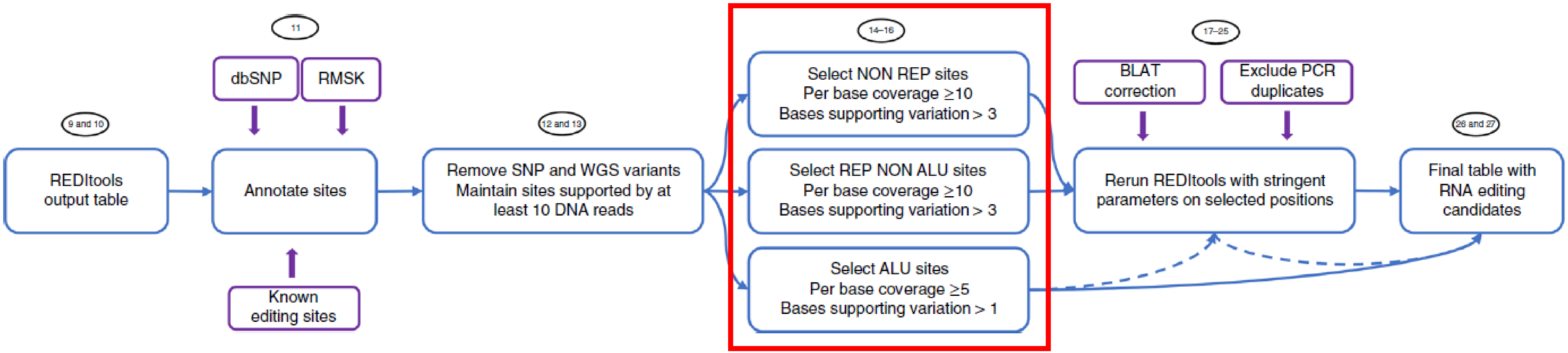


```
chr1    14542   A   2   14  37.64   [4, 0, 10, 0]   AG  0.71    29  29.86   [28, 0, 1, 0]   AG      0.03
chr1    14574   A   2   11  38.09   [7, 0, 4, 0]    AG  0.36    37  30.11   [37, 0, 0, 0]   -       0.00
chr1    14907   A   2   22  37.32   [11, 0, 11, 0]  AG  0.50    115 30.13   [61, 0, 54, 0]  AG      0.47
chr1    14925   A   2   24  37.75   [24, 0, 0, 0]   -   0.00    122 30.20   [121, 0, 1, 0]  AG      0.01
chr1    14930   A   2   24  38.29   [12, 0, 12, 0]  AG  0.50    97  29.93   [67, 0, 30, 0]  AG      0.31
chr1    15180   A   2   14  38.21   [14, 0, 0, 0]   -   0.00    79  29.78   [78, 0, 1, 0]   AG      0.01
chr1    15274   A   2   6   40.50   [0, 0, 0, 6]    AT  1.00    48  30.38   [0, 0, 10, 38]  AT AG   1.00
chr1    15717   A   2   2   35.50   [2, 0, 0, 0]    -   0.00    28  28.71   [27, 0, 1, 0]   AG      0.04
chr1    16186   A   2   27  37.07   [25, 0, 2, 0]   AG  0.07    73  30.42   [73, 0, 0, 0]   -       0.00
chr1    16497   A   2   24  39.83   [20, 0, 4, 0]   AG  0.17    127 30.36   [108, 0, 19, 0] AG      0.15
chr1    136573  T   2   5   37.00   [0, 3, 0, 2]    TC  0.60    44  29.43   [0, 0, 0, 44]   -       0.00
chr1    136586  T   2   3   38.33   [0, 2, 0, 1]    TC  0.67    32  29.78   [0, 0, 0, 32]   -       0.00
chr1    136671  T   2   3   34.33   [0, 2, 0, 1]    TC  0.67    31  29.94   [0, 0, 0, 31]   -       0.00
chr1    136817  T   2   3   39.67   [0, 0, 0, 3]    -   0.00    51  28.88   [0, 3, 0, 48]   TC      0.06
```

Accurate identification of human *Alu* and non-*Alu* RNA editing sites

Even though the editing level of *Alu* sequences is typically low (0.6% on average) and varies considerably, almost any one of the adenosines within these sequences will potentially be targeted by ADARs.

STRATEGY: apply different filtering criteria according to the type of region (repeated or not) in which the putative edited site falls



9 and 10

REDItools output table

11

dbSNP

RMSK

Annotate sites

Known editing sites

12 and 13

Remove SNP and WGS variants Maintain sites supported by at least 10 DNA reads

14–16

Select NON REP sites
Per base coverage ≥10
Bases supporting variation > 3

Select REP NON ALU sites
Per base coverage ≥10
Bases supporting variation > 3

Select ALU sites
Per base coverage ≥5
Bases supporting variation > 1

17–25

BLAT correction

Exclude PCR duplicates

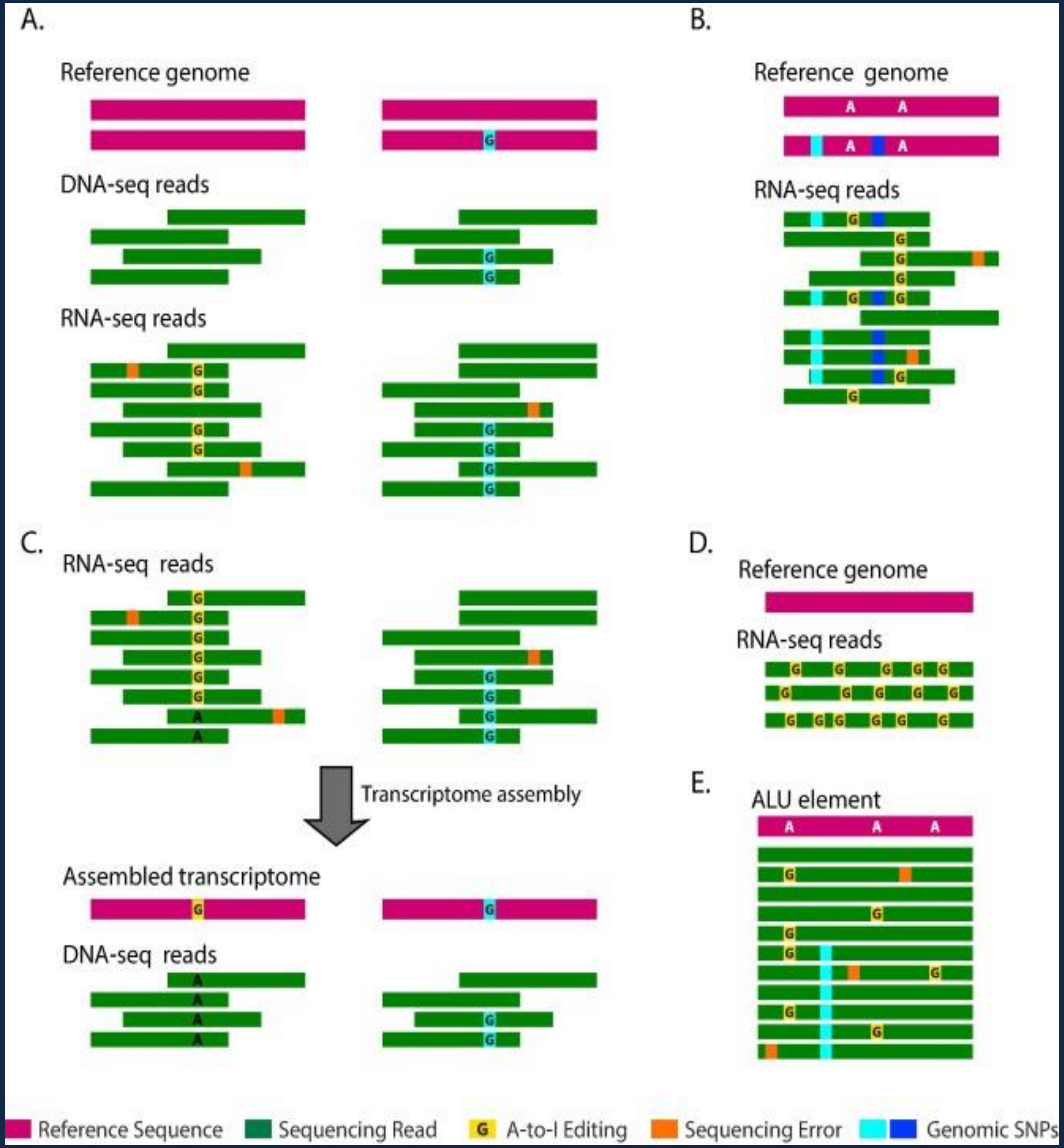
Rerun REDItools with stringent parameters on selected positions

26 and 27

Final table with RNA editing candidates

Fig. 3 | Filtering of REDItools tables to call RNA editing events. REDItools tables (Steps 9 and 10, Procedure 1) are generally subjected to further filters to remove potential artifacts and errors. The procedure begins with the annotation of all individual positions using known SNP sites, repeated elements in RepeatMasker and known editing events stored in the REDIportal database (Step 11, Procedure 1). Then, SNPs and sites not supported by ≥10 WGS reads are removed (Steps 12 and 13, Procedure 1) and divided into three groups: ALU, REP NON ALU and NON REP (Steps 14–16, Procedure 1). NON REP and REP NON ALU sites undergo more stringent call criteria than ALU sites that take into account mis-mapping reads and PCR duplicates (Steps 17–25, Procedure 1). Optionally, stringent filters can also be applied to ALU sites (Steps 20 and 21, Procedure 1). Finally, filtered positions are collected in the final list of RNA editing candidates (Steps 26 and 27, Procedure 1).

Lo Giudice, C., Tangaro, M.A., Pesole, G. *et al.* Investigating RNA editing in deep transcriptome datasets with REDItools and REDIportal. *Nat Protoc* **15**, 1098–1131 (2020)



Quantifying RNA editing in deep transcriptome datasets

The quantification of RNA editing is important to compare values across samples and study its potential role in different experimental conditions or in human disorders.

Determine the fraction of edited transcripts of a site (editing levels) by dividing the number of the ‘G’-containing transcripts that map to the site, by the total number of transcripts mapped to the position.

For example, the editing levels of the leftmost editing site in Fig. E is 4/11 as we found evidence for editing in 4 reads out of 11 reads.

The accuracy of measuring the editing levels of a site depends on the site coverage in the RNA-seq dataset, which in turn is determined mainly by the sequencing depth and the expression levels of the transcript of interest. Unfortunately, sufficient coverage for each editing site is often not available in a typical RNA-seq. In order to overcome this limitation have been developed **metrics for unbiased RNA editing quantification in a sample**:

Overall editing level

The overall editing is defined as the total number of reads with G at all known editing positions over the number of all reads covering the positions without imposing specific sequencing coverage criteria. It can be calculated using REDItools tables obtained imposing loosing parameters.

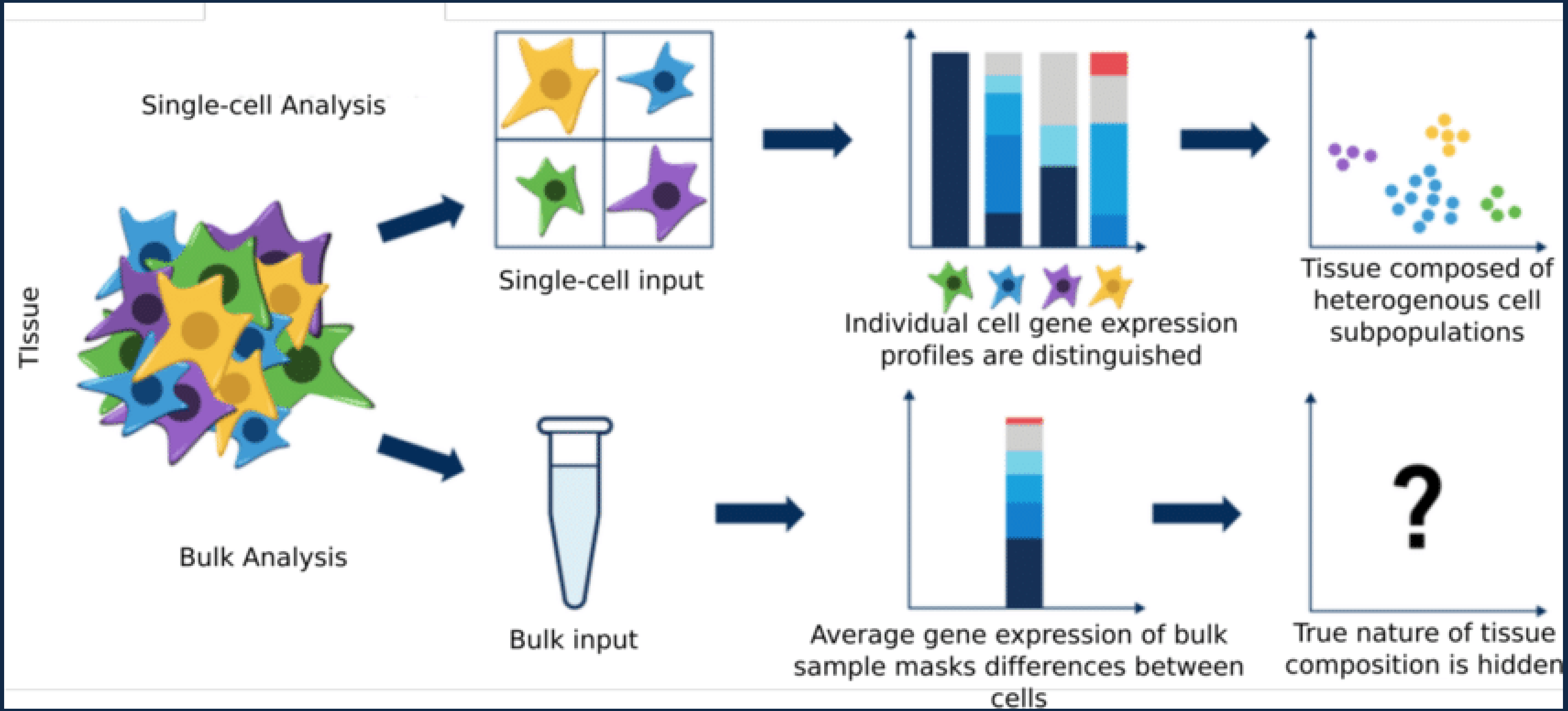
ALU editing index

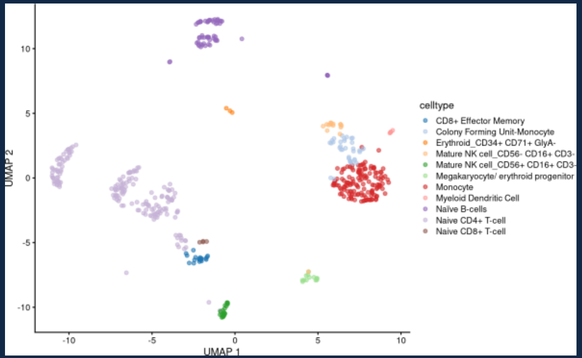
The Alu editing index (AEI) is a metric to quantify the global RNA editing activity of sample and is defined as the weighted average of editing events occurring in all Alu elements. The pipeline to calculate AEI is described in Roth et al. (2019) and available here.

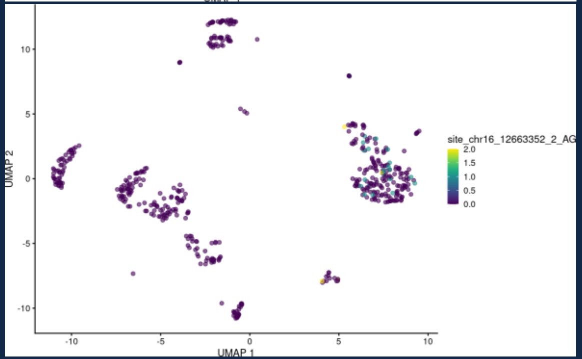
Recoding index

The overall editing calculated at recoding positions residing in coding protein genes is named recoding index (REI). It has been initially described in Silvestris et al. (2019). This metric, used to investigate the activity of ADAR2, can be calculated using REDItools tables obtained imposing loosing parameters and a list of recoding sites from REDIportal.

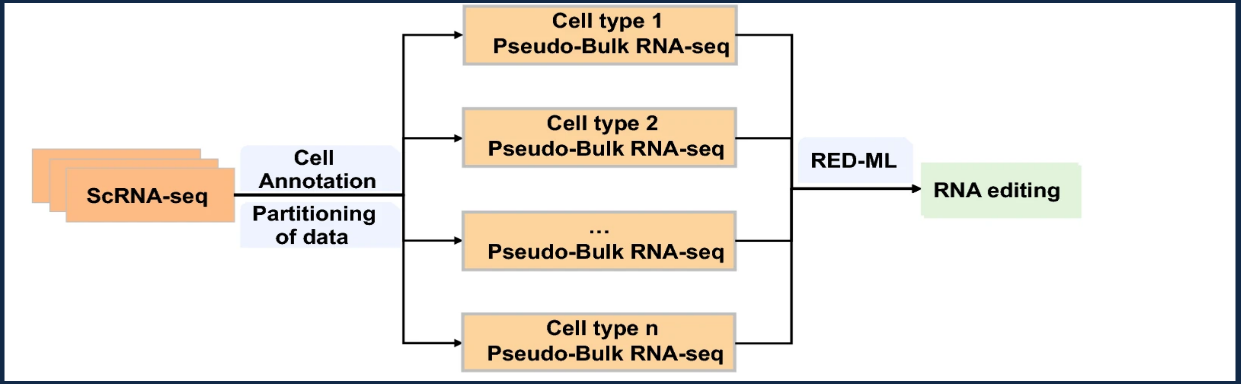
RNA editing at single-cell resolution (with 10x data)


Single-cell Analysis
Tissue
Single-cell input
Individual cell gene expression profiles are distinguished
Tissue composed of heterogenous cell subpopulations
Bulk Analysis
Bulk input
Average gene expression of bulk sample masks differences between cells
?
True nature of tissue composition is hidden





The cell type annotation information was used to combine the mapped reads of the same cell type in scRNA-seq to obtain pseudo-Bulk RNA-seq for each cell type.


ScRNA-seq

Cell Annotation

Partitioning of data

Cell type 1 Pseudo-Bulk RNA-seq

Cell type 2 Pseudo-Bulk RNA-seq

... Pseudo-Bulk RNA-seq

Cell type n Pseudo-Bulk RNA-seq

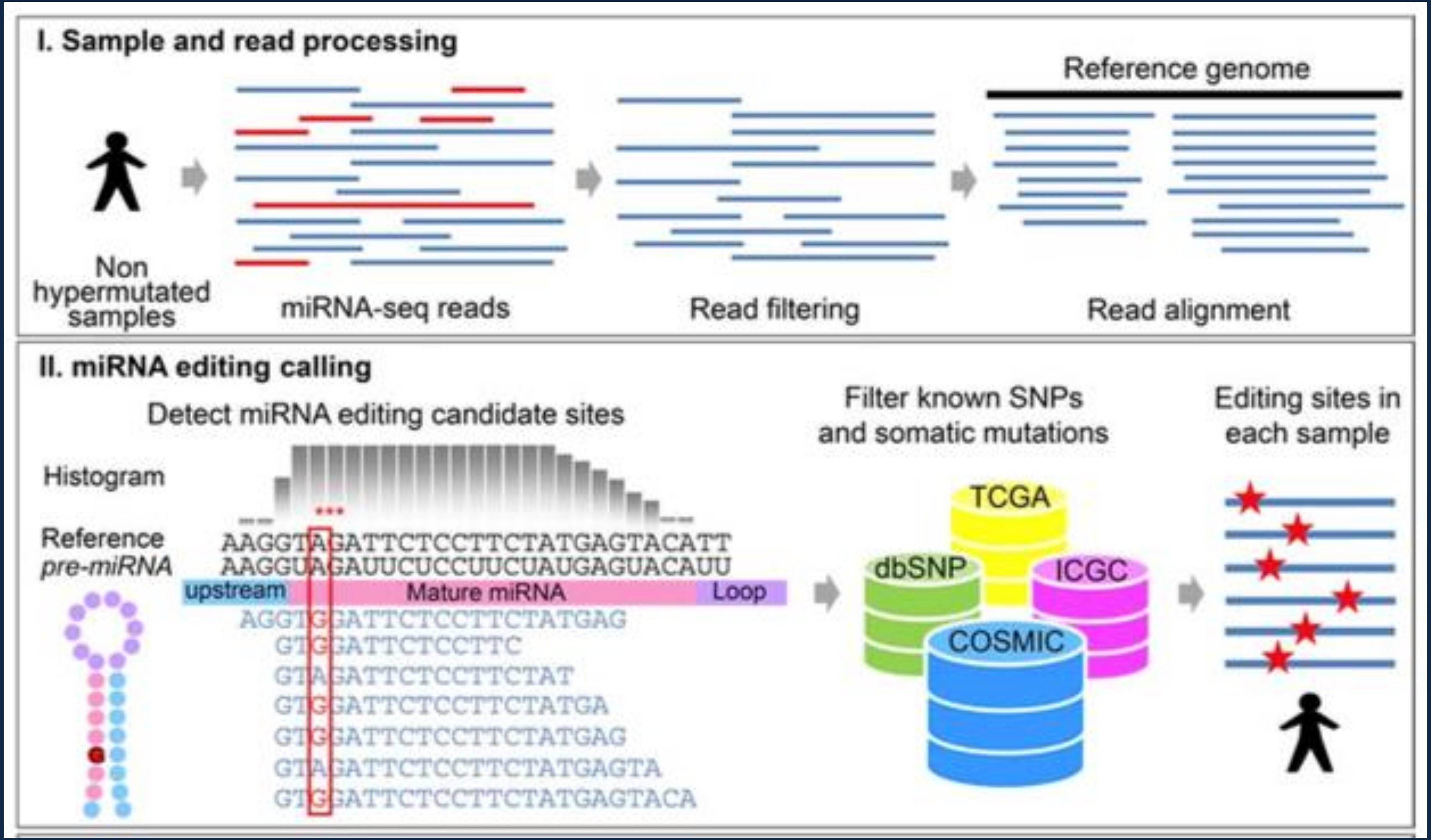
RED-ML

RNA editing

https://bioconductor.org/packages/release/bioc/vignettes/raer/inst/doc/raer.html

small RNA-seq

Identifying RNA Editing Sites in miRNAs by Deep Sequencing



https://www.tau.ac.il/~elieis/miR_editing/

Pipeline

step1. Filtering Low- Quality Reads and Trimming Sequence Adapters

```
perl Process_reads.pl Input_fastq_file The_filtered_fastq_file
```

step2. Aligning the reads against the genome

```
bowtie -n 1 -e 50 -a -m 1 --best --strata --trim3 2 The_bowtie_folder/The_genome_indexes
The_filtered_fastq_file > The_output_file
```

step3. Mapping the mismatches against the pre-miRNA sequences

```
perl Analyze_mutation.pl The_output_file main_output.txt
```

step4. Using binomial statistics to remove sequencing errors

```
perl Binomial_analysis.pl main_output.txt >binomial_output.txt
```

THANK YOU FOR YOUR ATTENTION!