Pre-meeting Workshop

Bioinformatic analysis of RNA editing

Zhoufeng Gao & Xin Li Rui Zhang lab
Sun-Yat-Sen University

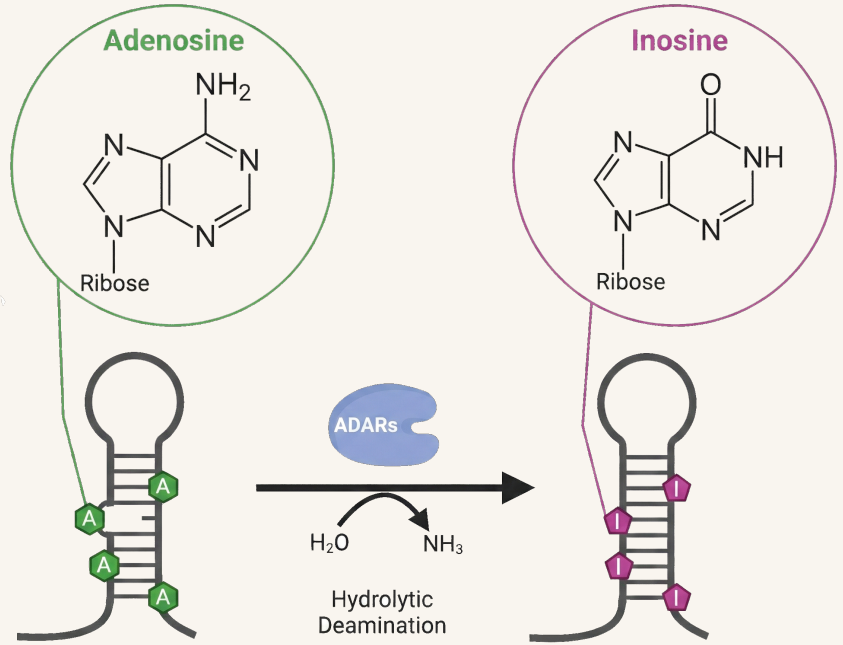
Main contents

Part1 Theoretical Basis

1. Background & Challenges
2. Genome-based Identification
3. RNA-seq Alone Approaches
4. Advanced Topics & Resources

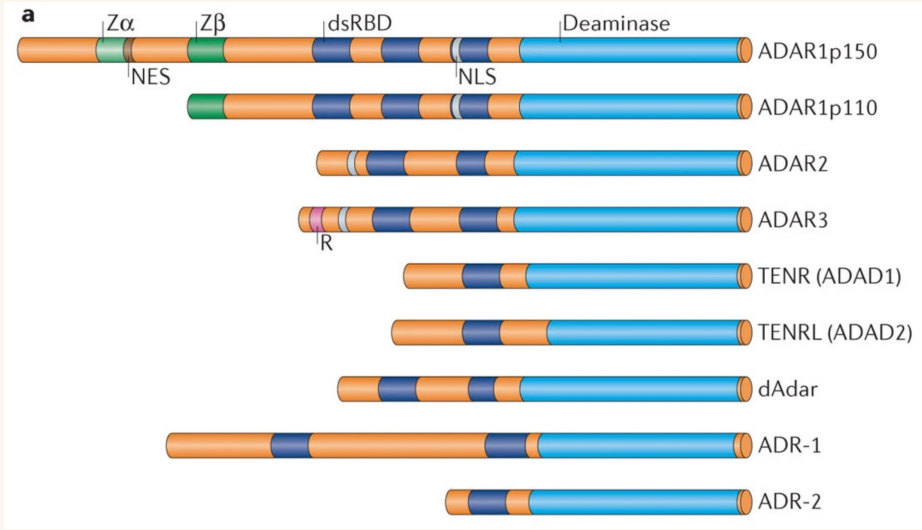
Part2 Hands-on Practice

What is RNA A-to-I editing?



Mechanism: ADAR enzymes convert Adenosine (A) to Inosine (I) in dsRNA.

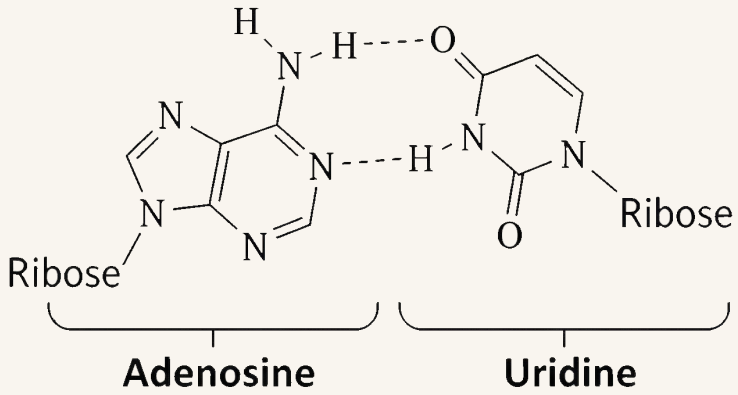
Vesely et al. 2021, genes

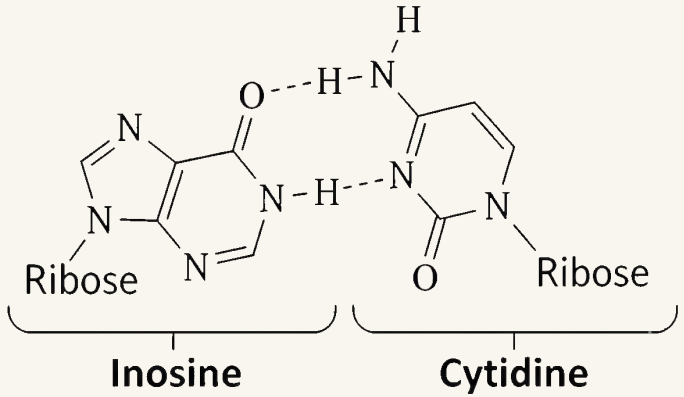


ADAR family contains multiple ADARs

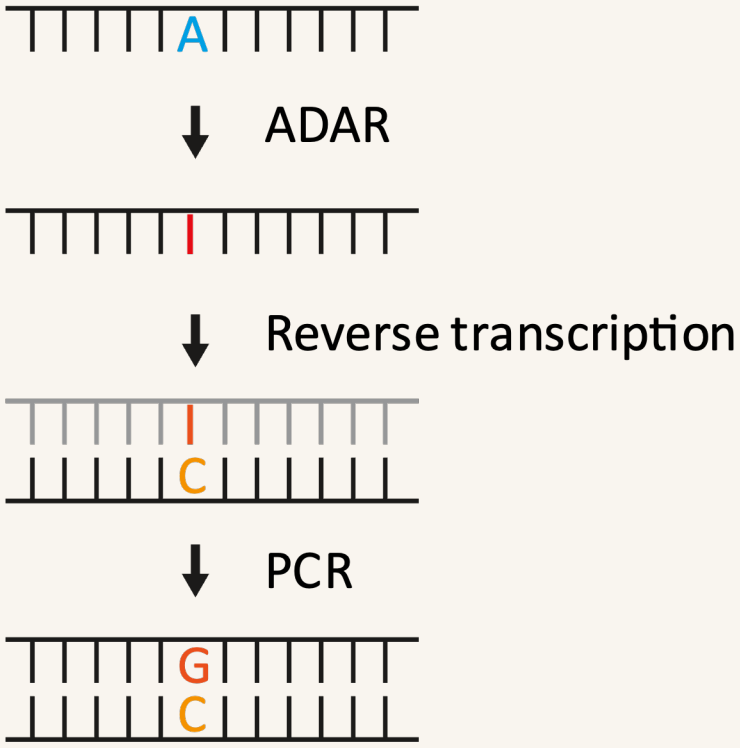
Nishikura , 2016, Nat Rev Mol Cell Biol

How A-to-I editing are detected?



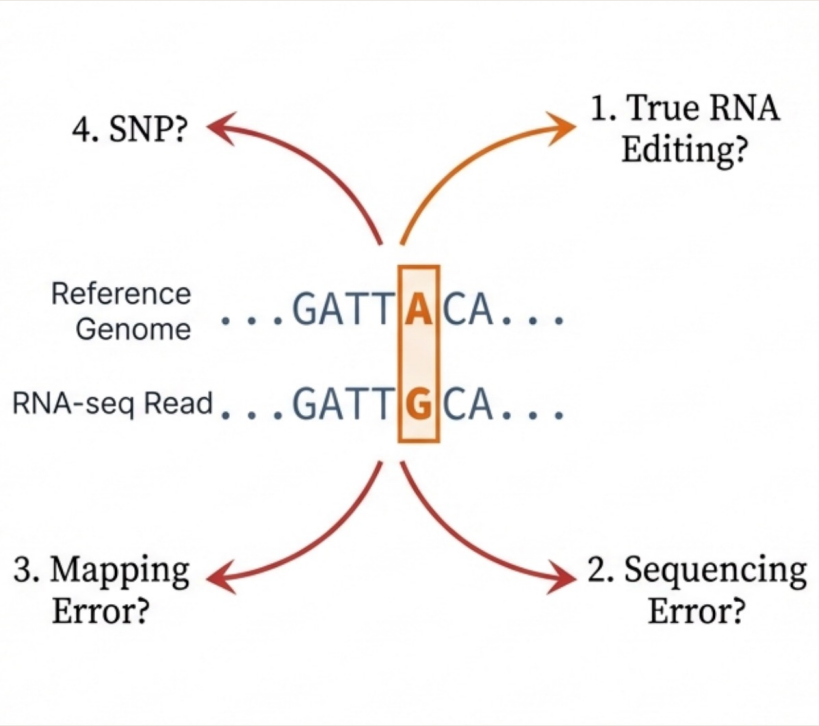


Nishikura et al. 2016, Nat Rev Mol Cell Biol

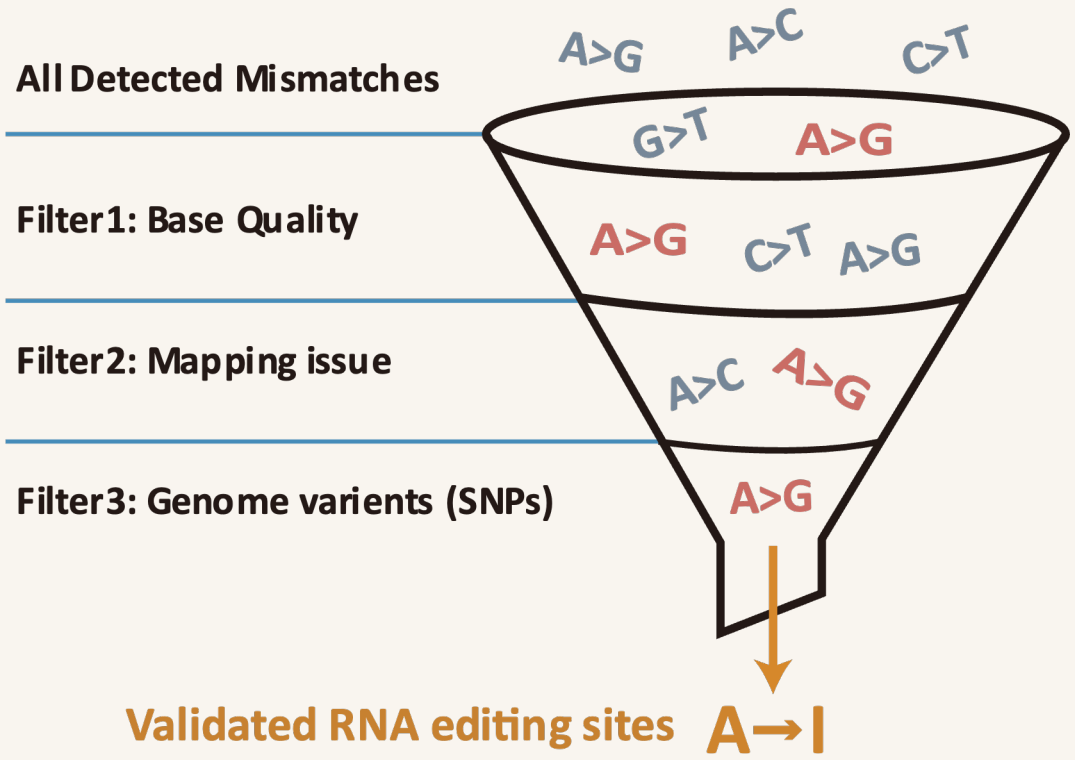


A → I editing appears as an A → G mismatch

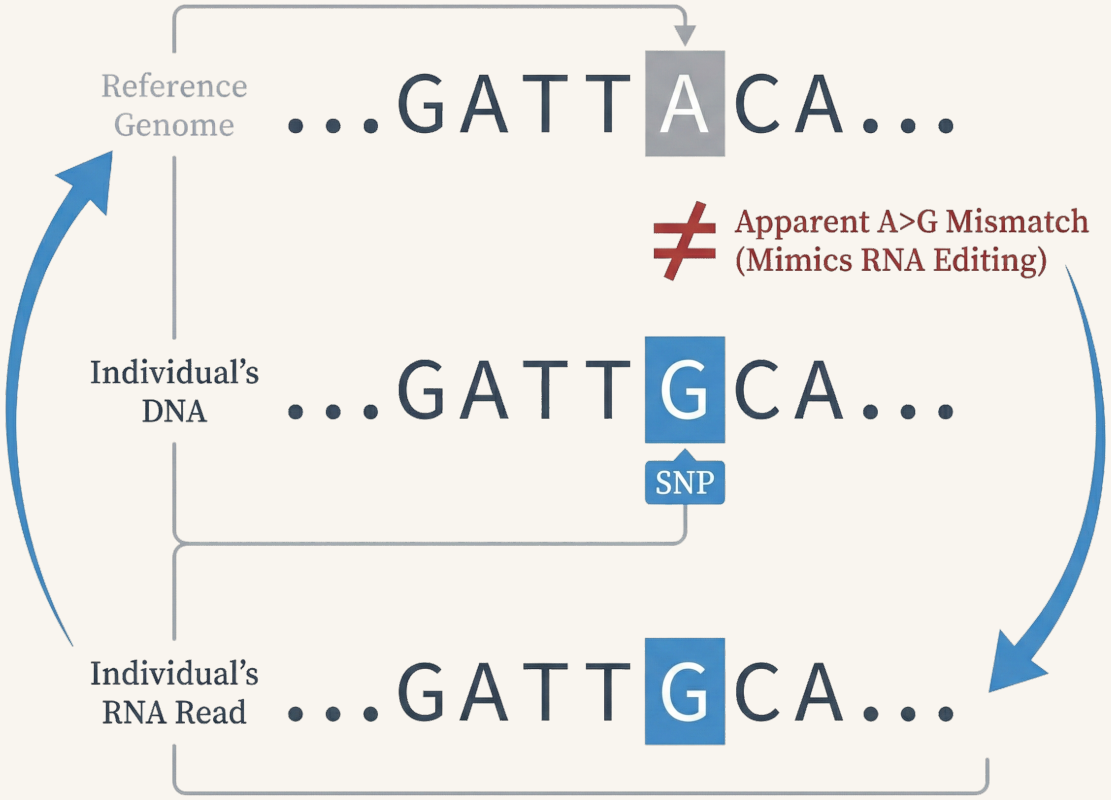
Signal vs noise



An A-to-G mismatch can be a false positive arising from multiple sources.



Noise1: Mismatch on Genome: SNPs



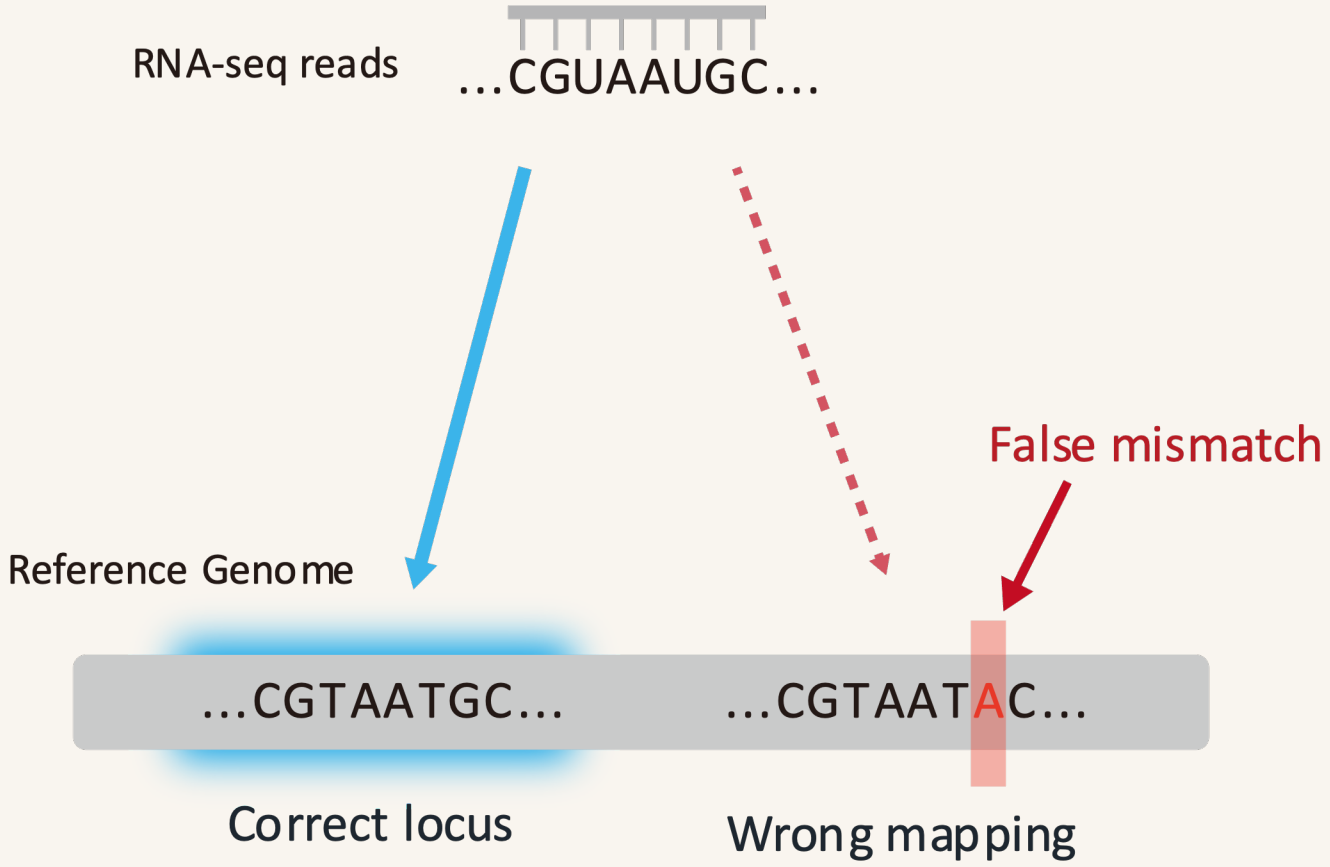
What it is

SNP is a inherited DNA variation where a single nucleotide differs from the reference genome.

Why it's a problem

Inherited SNPs are indistinguishable from dynamic RNA editing events, because both appear as A-to-G mismatches in sequencing data.

Noise2: Mapping Error



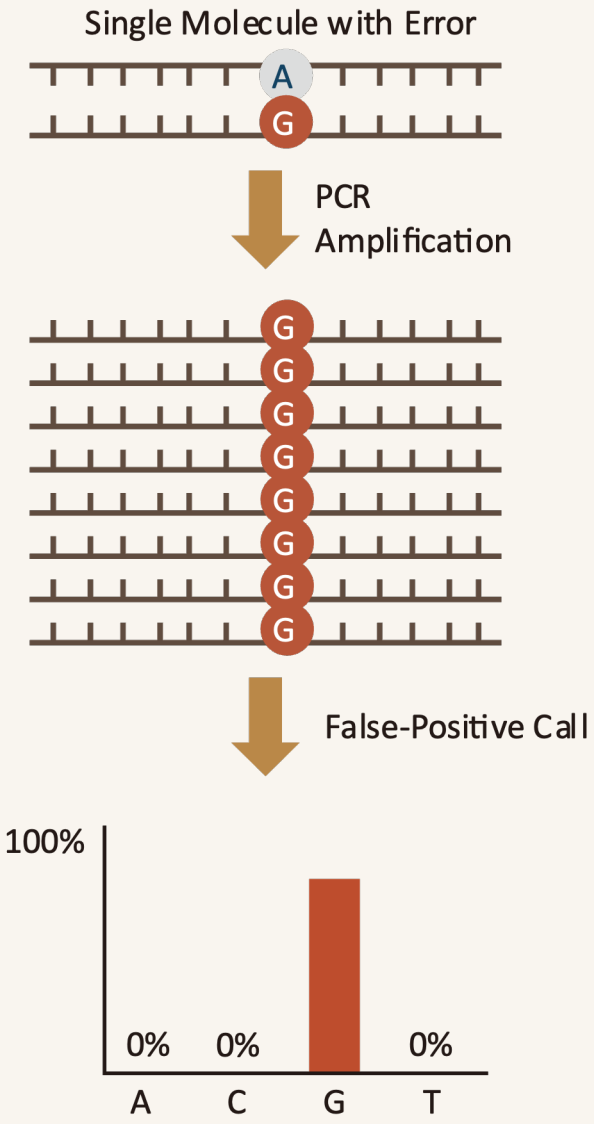
What it is

Mapping error is Incorrect alignment of sequencing reads to non-originating genomic loci due to high sequence similarity.

Why it’s a problem

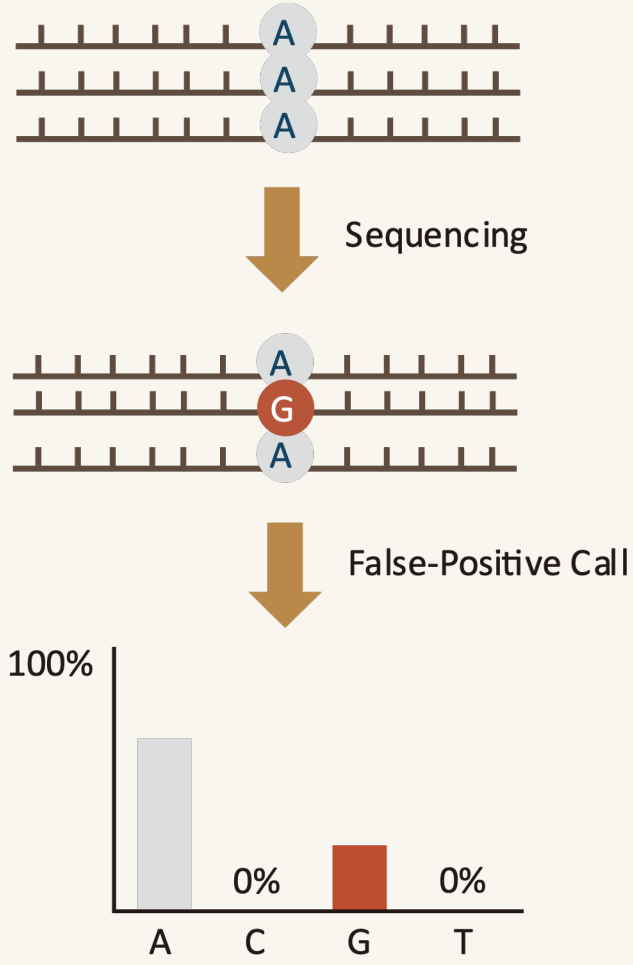
Ref(A)-vs-Read(G) artifacts are indistinguishable from biological editing events.

Noise3: Random Error



PCR errors

Polymerase mistakes during amplification introduce spurious A-to-G variants.

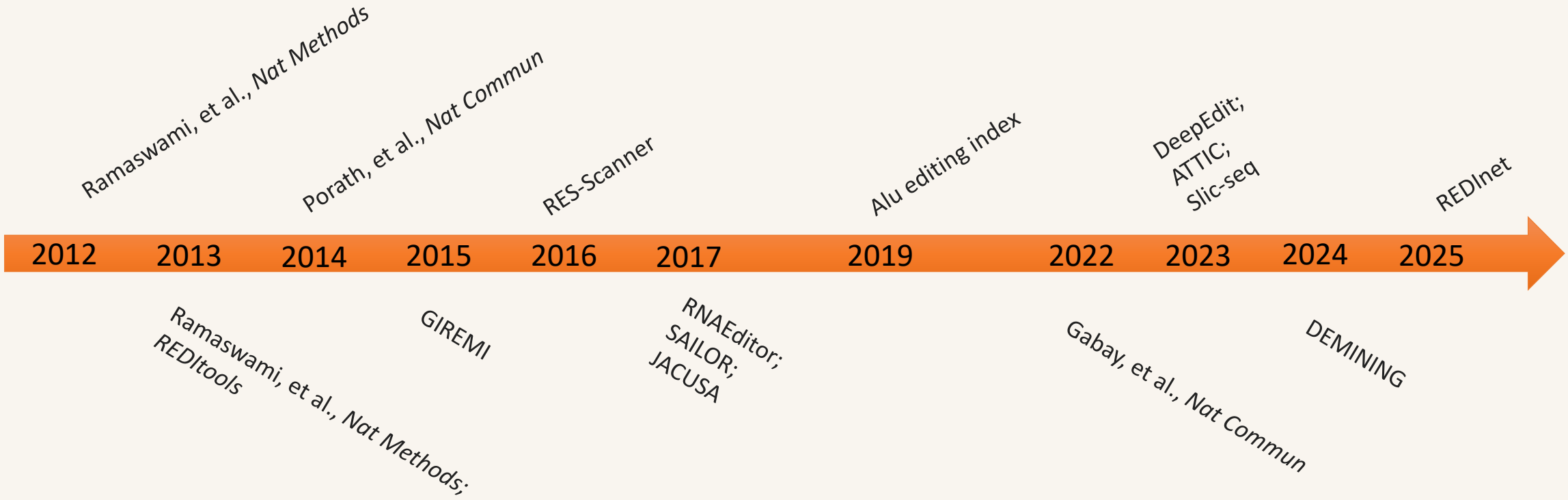


Sequencing error

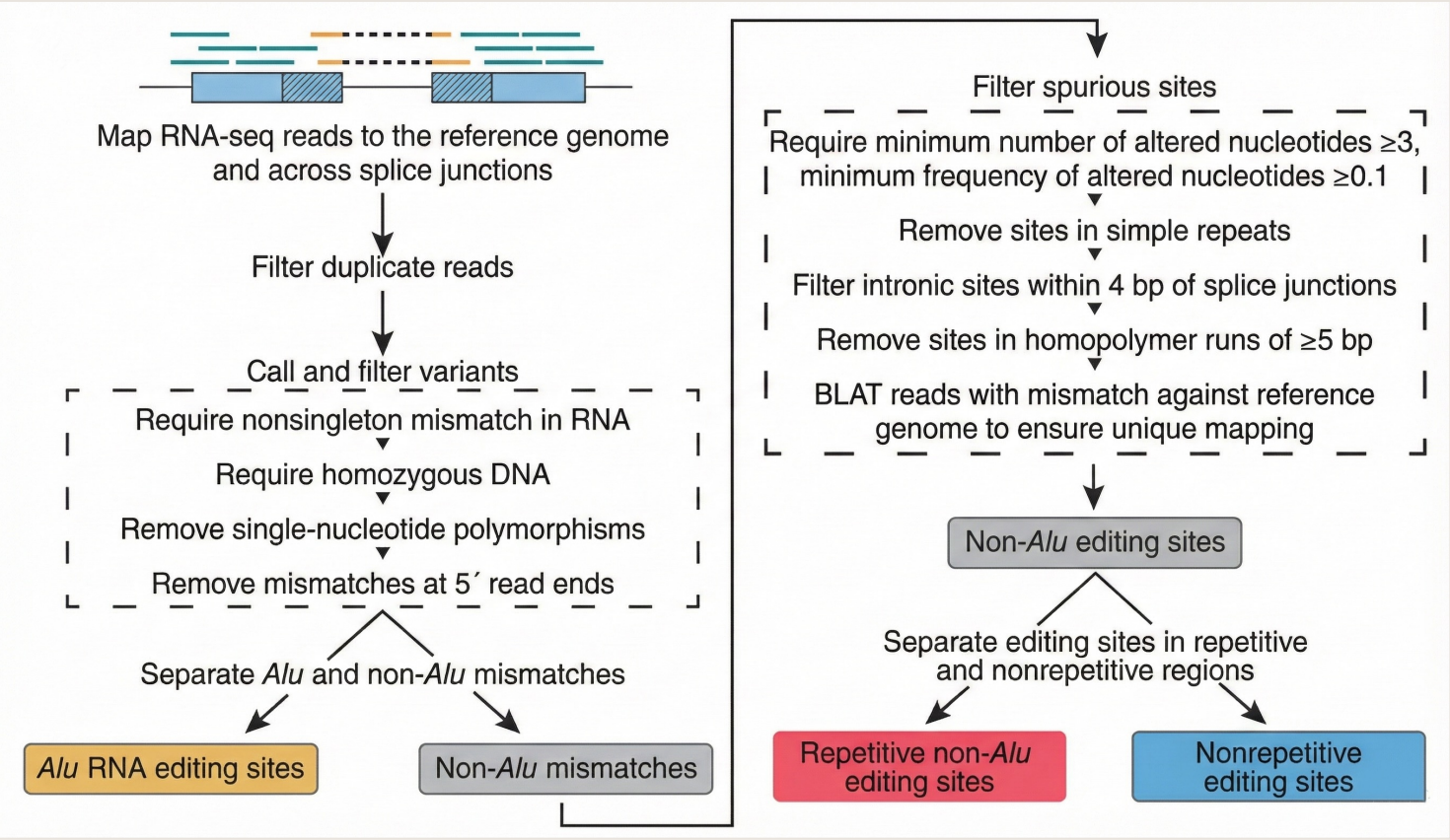
Instrumental base-calling errors randomly misread 'A' sites as 'G'.

Theoretical Basis: Contents

Evolution of Identification Strategies



Comparative Genomic and Transcriptomic Analysis



Ramaswami et al., 2012, *Nat Methods*

Editing sites separate into Alu and non-Alu

Alu Region

High Signal-to-noise ratio → fewer filter

Non-Alu Region

low Signal-to-noise ratio → more filter

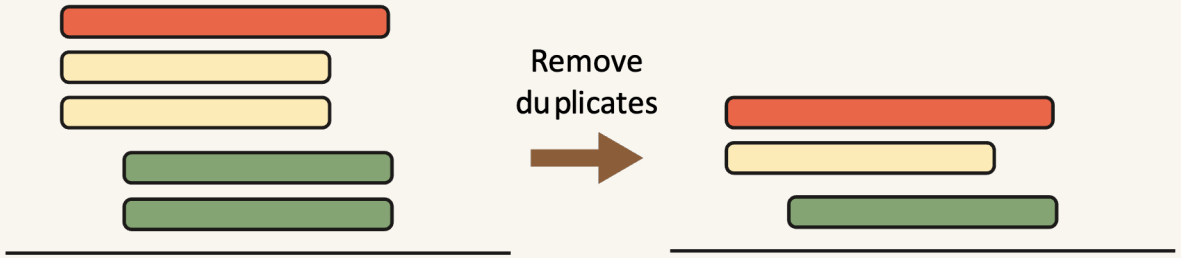
Remove PCR duplicates

Computational Removal

Removes duplicates based solely on identical mapping coordinates in the alignment.

e.g.

samtools dedup
picard markduplicate

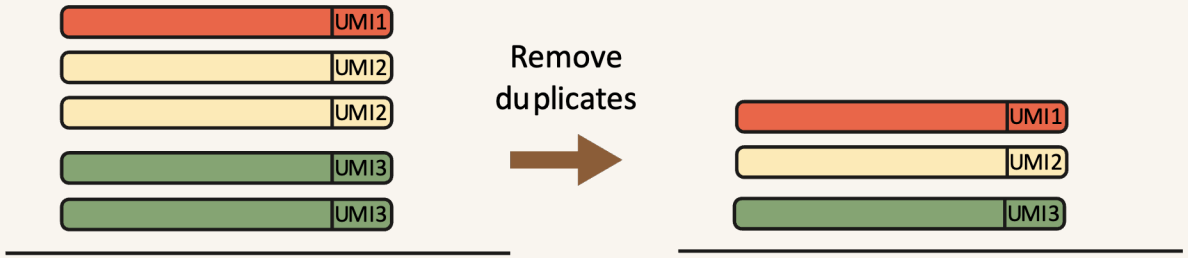


Experimental Removal

Uses unique molecular tags to distinguish PCR artifacts from true biological duplicates.

e.g.

umitools



And with a coverage/ratio cutoff

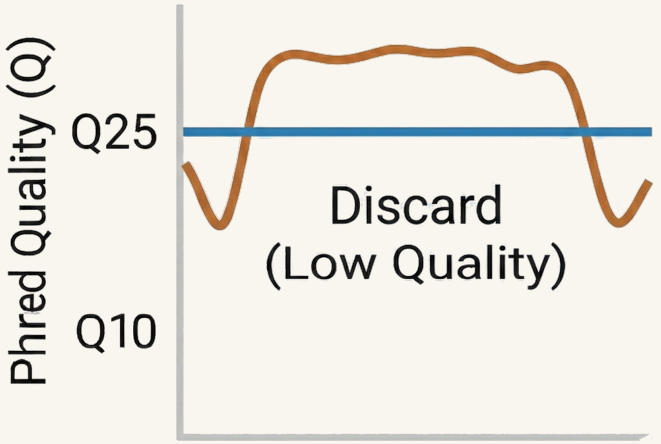
Filter to drop Sequencing error

Base Quality

$$Q = -10 \log_{10} P_{error}$$

BaseQ	Error probability
Q40	0.0001 （1 in 10,000）
Q30	0.001 (1 in 1,000)
Q20	0.01 (1 in 100)
Q10	0.1 (1 in 10)

A base quality score (Q-score) is a prediction of the probability of an error in base calling.



Trimming seq ends

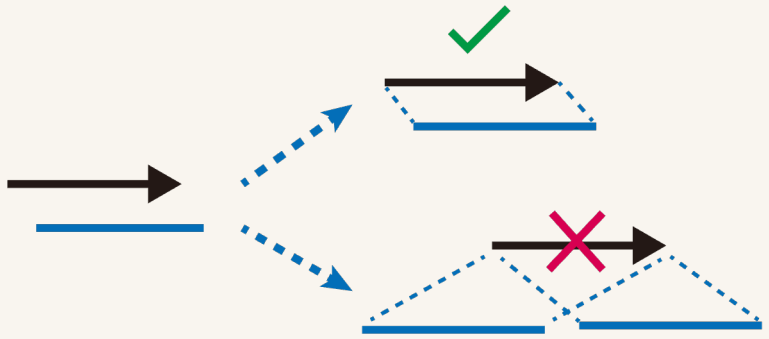
GGGGAGGGG

Remove homopolymers

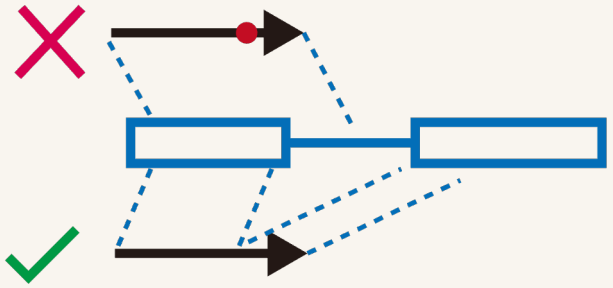
Filtering homopolymers eliminates artifacts caused by polymerase slippage in low-complexity regions.

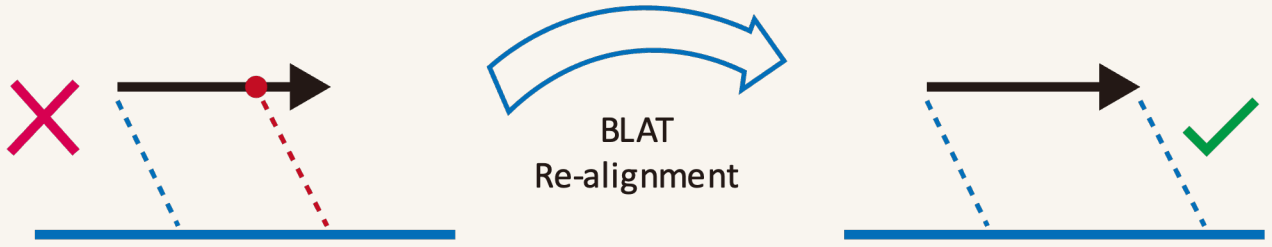
Filter to drop Mapping error

Only reserve unique mapping reads



Remove intron mismatches within 4bp of splice junctions





Use BLAT to re-align reads

SNP calling

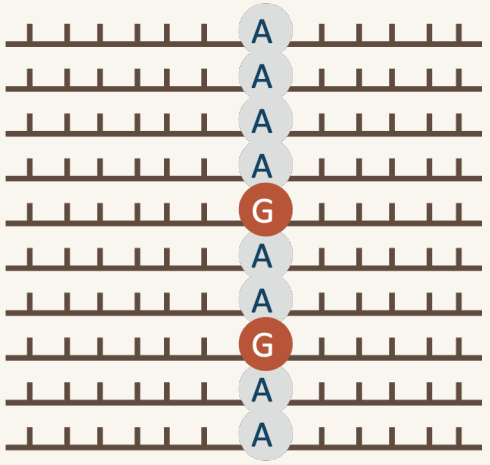
Get common SNPs in database

dbSNP database

1000 Genomes Project

Remove the mismatch if is in common SNPs list

Call SNPs from DNA-seq



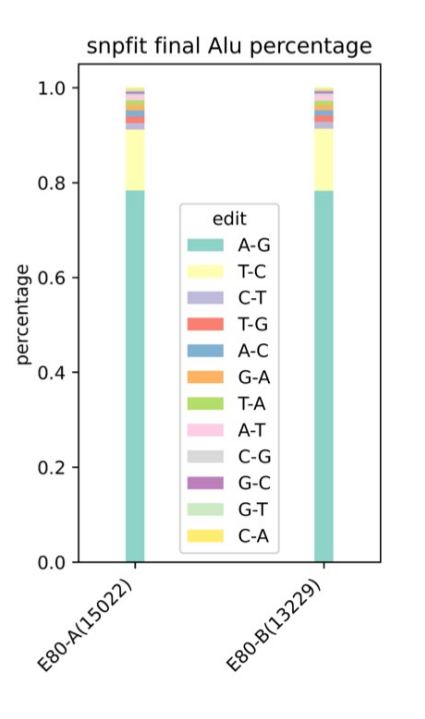
Call mismatch using DNA-seq

Using low ratio cutoff

e.g.
Coverage >= 10
Mismatch >= 2

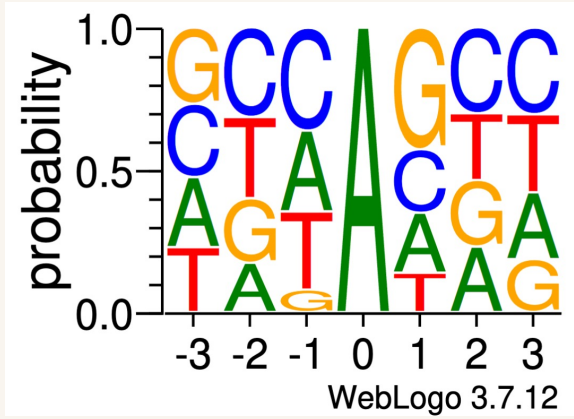
- Samtools / BCFtools
- GATK

Methods to assess quality of editing calling



A-to-G fraction

Real A-I editing will be conseved
Noises will be filtered

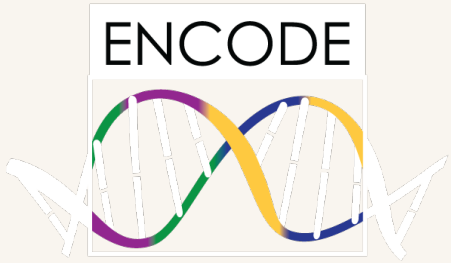


motif

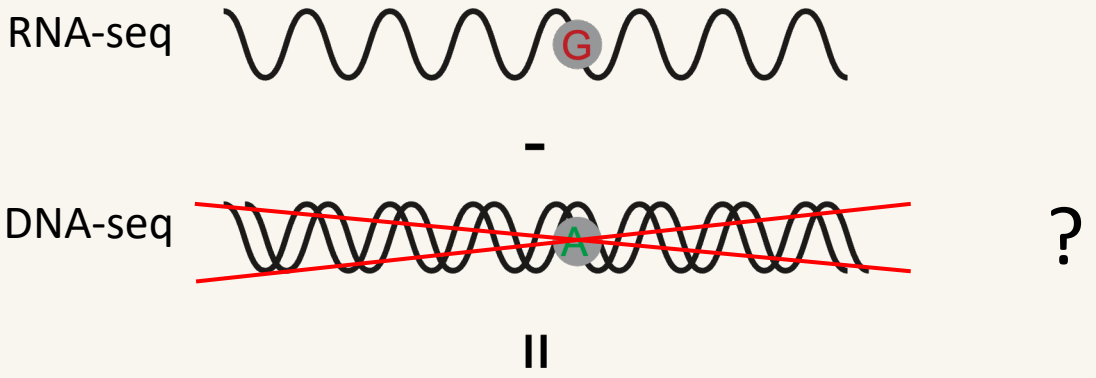
ADAR binding & deaminase preference

Theoretical Basis: Contents

De Novo Identification from RNA-Seq Data Alone



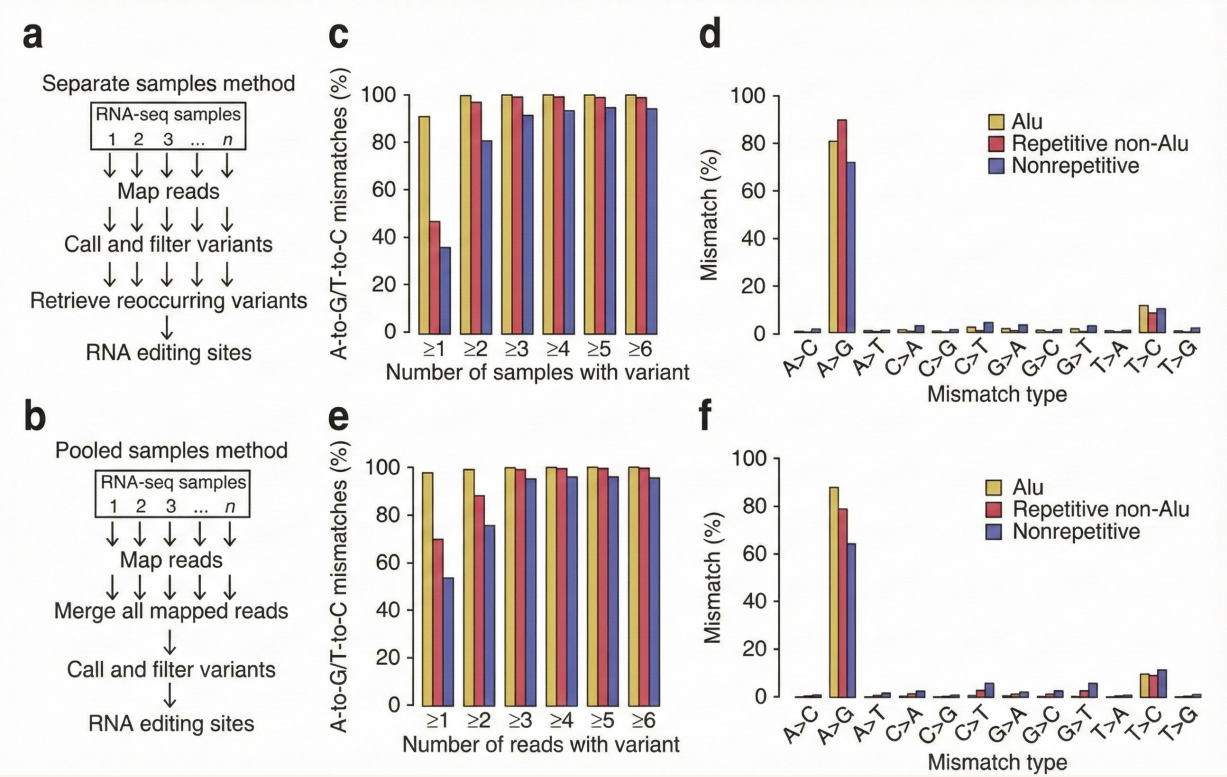
Large database are commonly used in data mining



True editing sites

In these database paired DNA sequencing and RNA sequencing data are often unavailable.

Merge samples

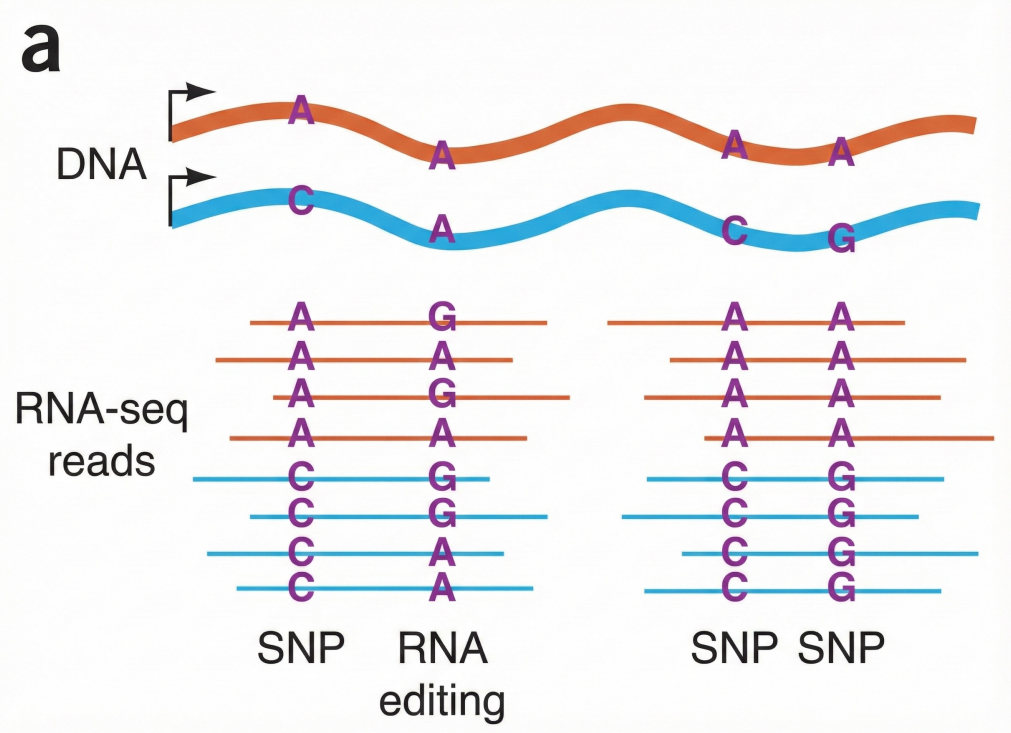


Ramaswami, et al., 2013, *Nat Methods*

Separate samples strategy:
RNA editing sites typically appear repeatedly in multiple individuals. Rare SNPs typically exist in only a single individual.

Pooled samples strategy:
Pool sequencing data from multiple individuals to improve coverage, we can remove rare SNPs because it appear at very low frequencies

Filter SNPs by Linkage Disequilibrium

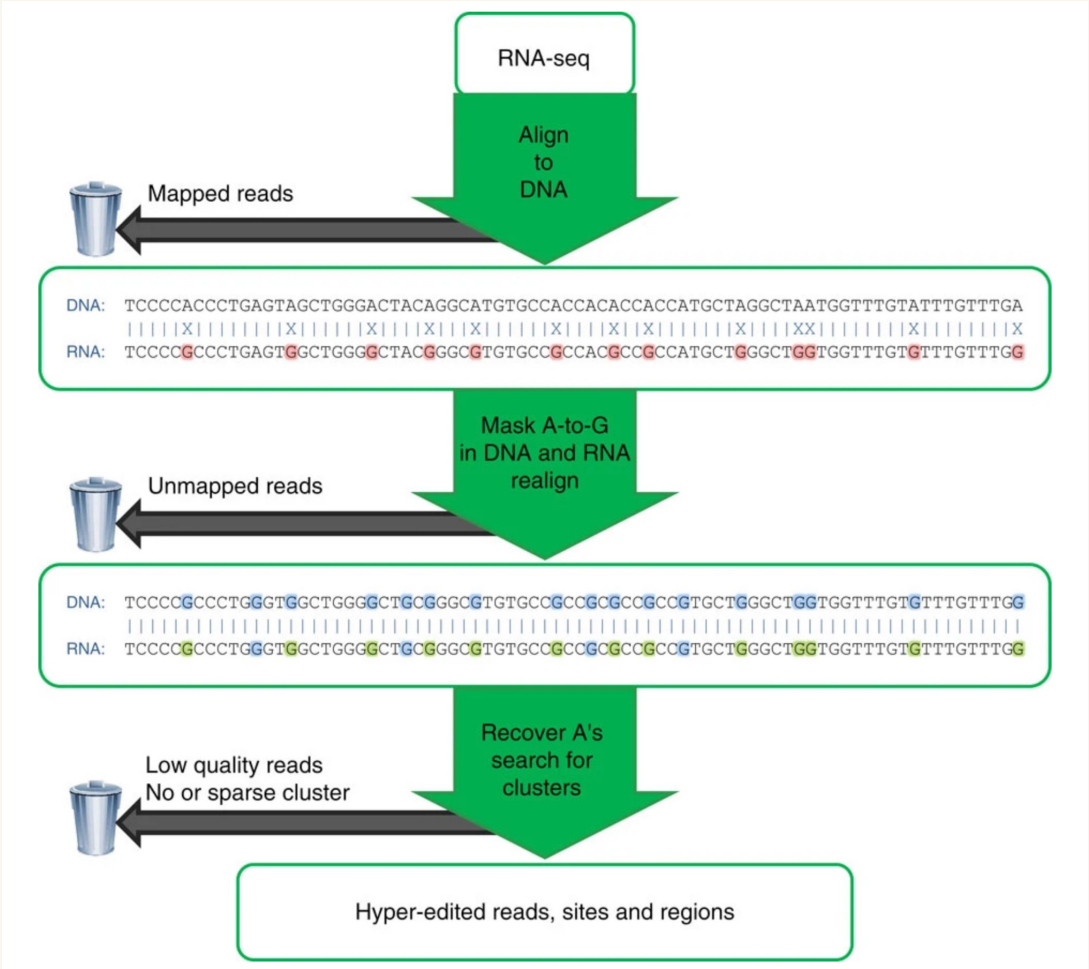


GIREMI:
SNPs typically exhibit strong linkage disequilibrium while RNA editing sites tend to be randomly distributed

Zhang, et al., 2015, *Nat Methods*

Theoretical Basis: Contents

Rescue hyper-editing reads

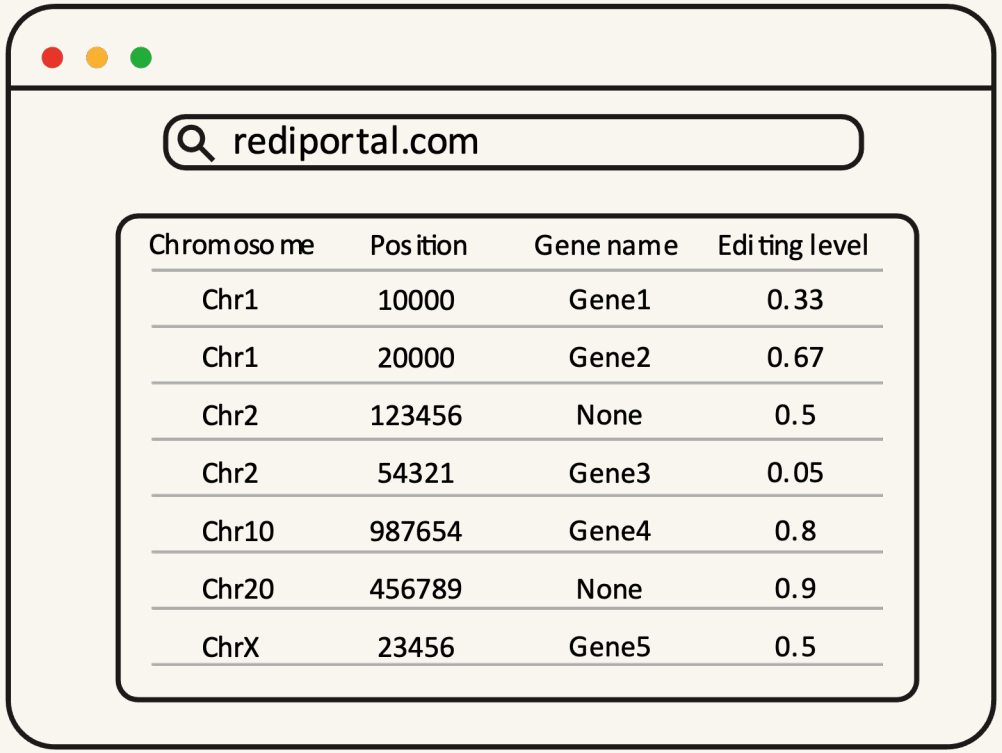


Porath, et al., 2014, *Nat Commun*

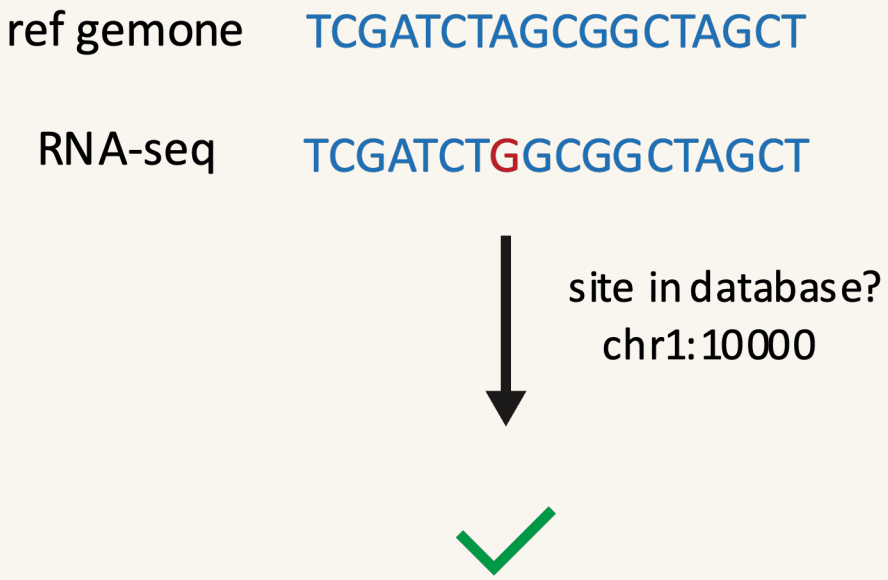
Hyper-editing:
ADARs could extensively edit dense adenosine residues, creating numerous A-to-G mismatches that can cause reads to fail standard alignment.

By computationally converting all adenosine (A) residues to guanine (G), this strategy rescue the reads that were previously unalignable due to hyper-editing.

Call from known sites

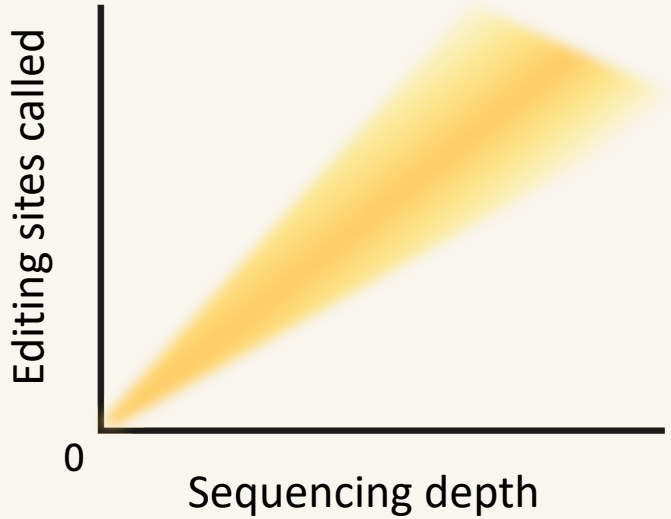


REDIPortal



Direct quantification of known sites serves as a straightforward, first-pass assessment of editing activity.

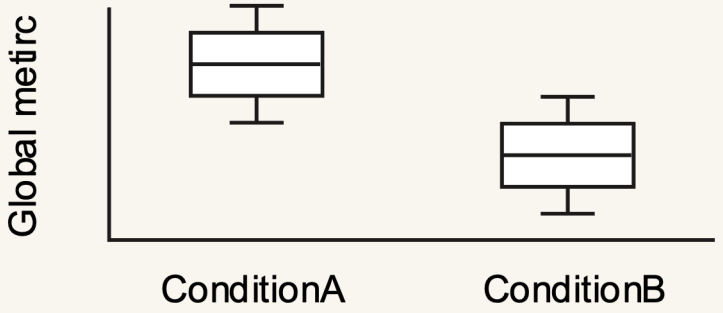
Beyond counting: Global metric



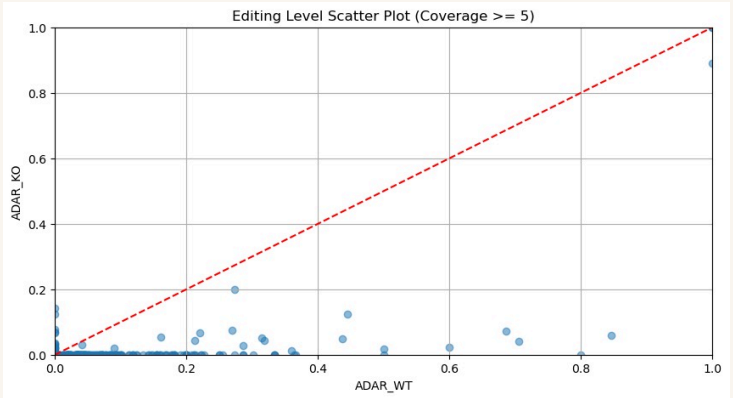
The number of editing Sites called increase with sequencing depth

More Editing Sites ≠ Higher Editing Efficiency

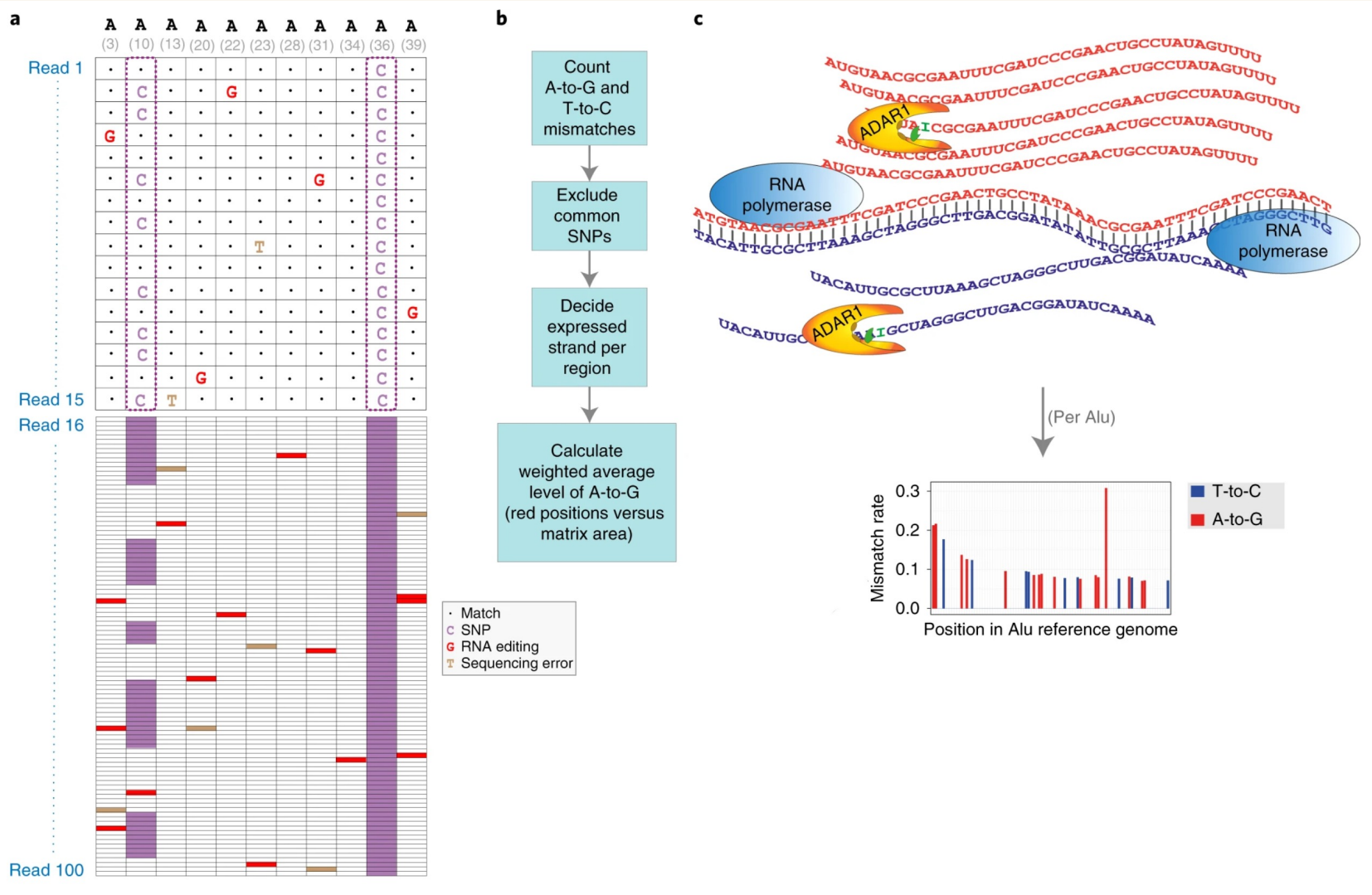
Compute a global metric



Globally editing shift



Alu editing index



The AEI measures the aggregate editing level of all Alu sites (total mismatches / total coverage).

This provides a robust metric of global ADAR activity that is insensitive to sequencing depth.

Roth, et al., 2019, *Nat Methods*

Acknowledgment

Next: Part2: Hands-on Practice