# RNA editing

New sources of genomic variation, software and application

Scott Funkhouser

Ernst Lab

Genetics and Quantitative Biology PhD Candidate

# Outline

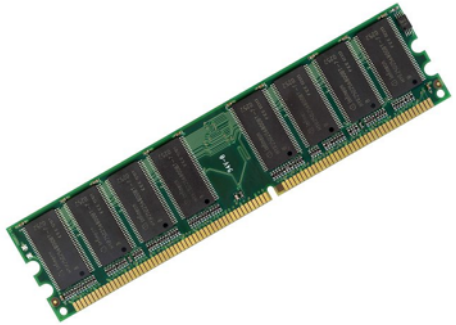
- What is RNA editing?
- *editTools:* an R package for RNA editing analysis
- Application of *editTools* to the swine model

# The central dogma

DNA

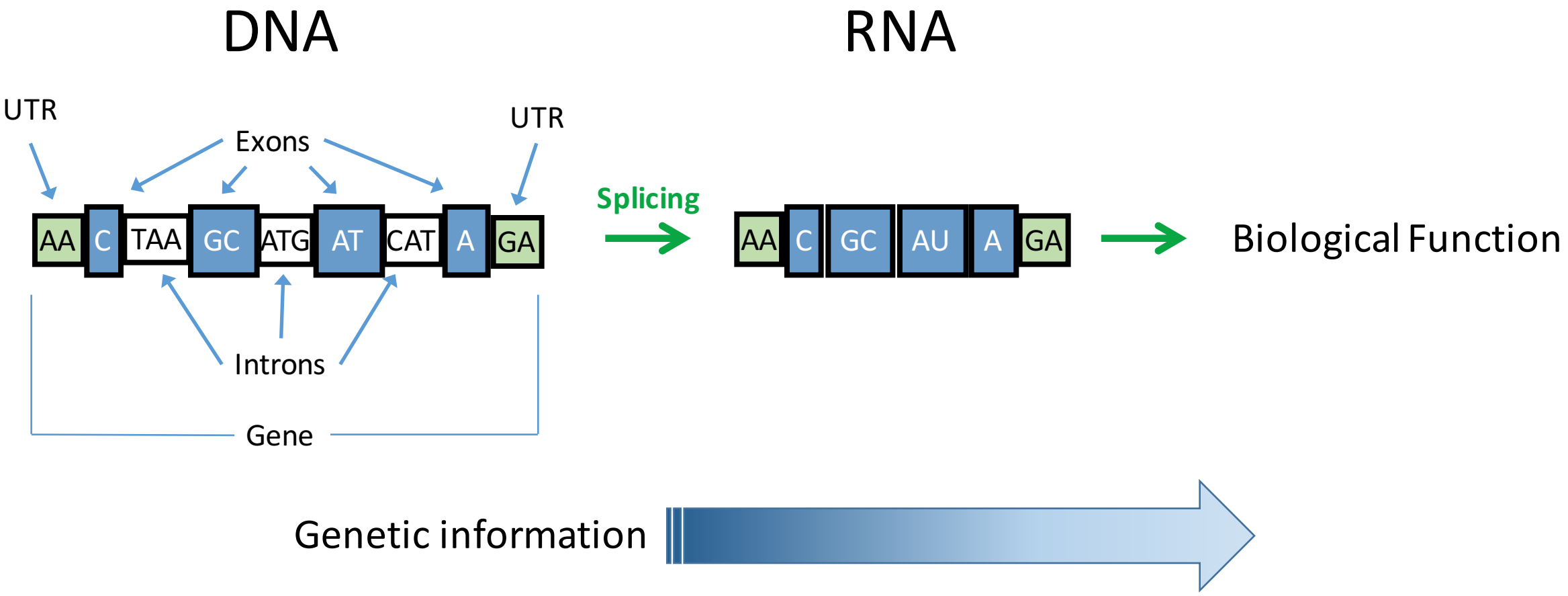
RNA

Biological Function

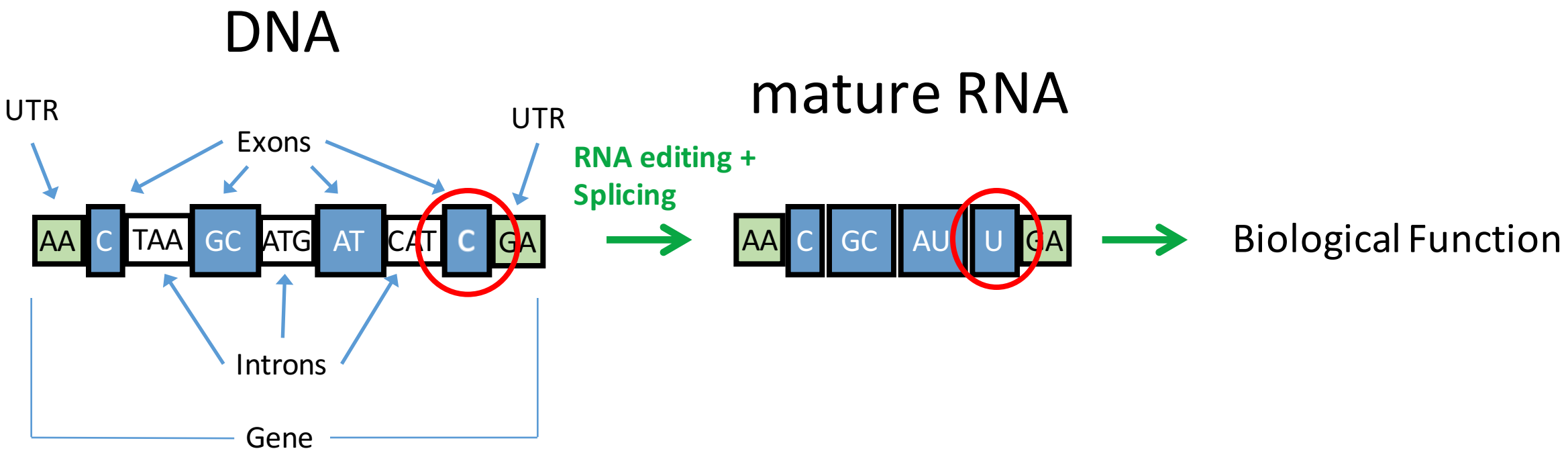


Software Program

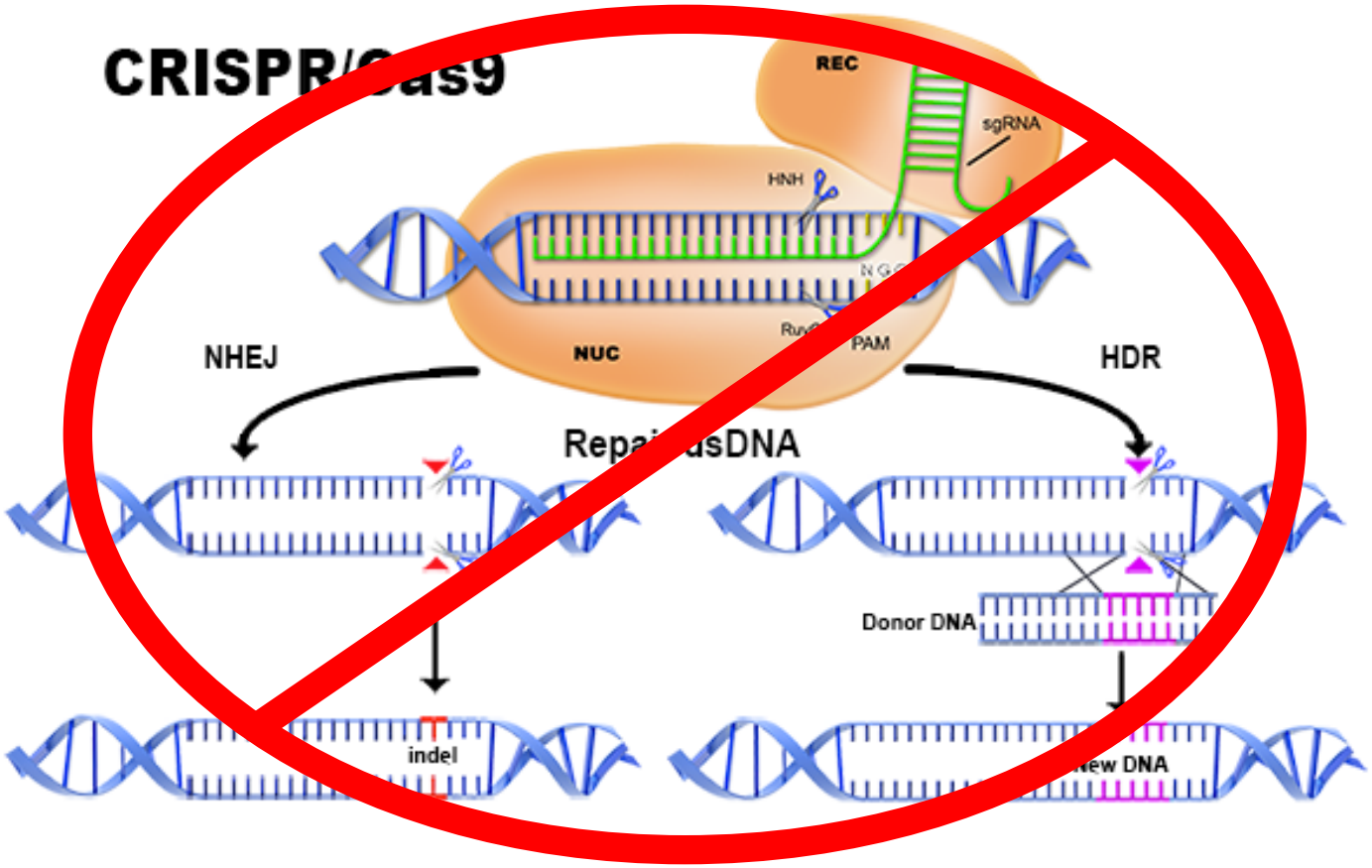
# The central dogma



# RNA editing


DNA
UTR
Exons
UTR
AA
C
TAA
GC
ATG
AT
CAT
C
GA
Introns
Gene
RNA editing +
Splicing
mature RNA
AA
C
GC
AU
U
GA
Biological Function

# Not to be confused with “genome editing”



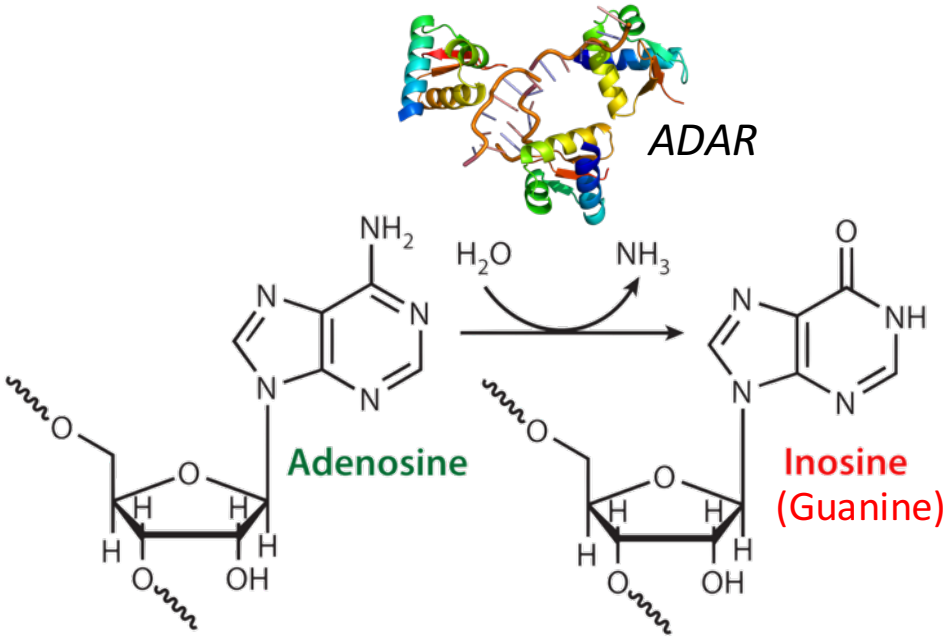

Genome editing by CRISPR cas9

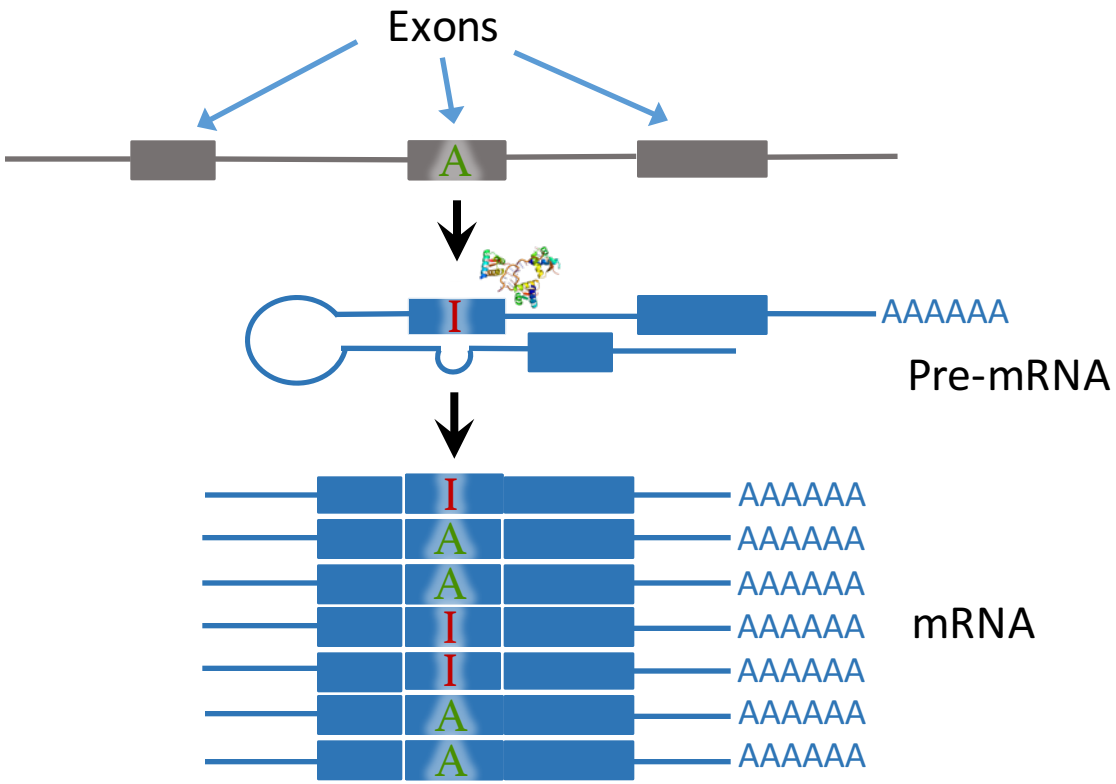
- Editing performed on the genome
- Artificial technology

RNA editing

- Editing performed on the transcriptome
- Natural. Done intentionally by the cell.

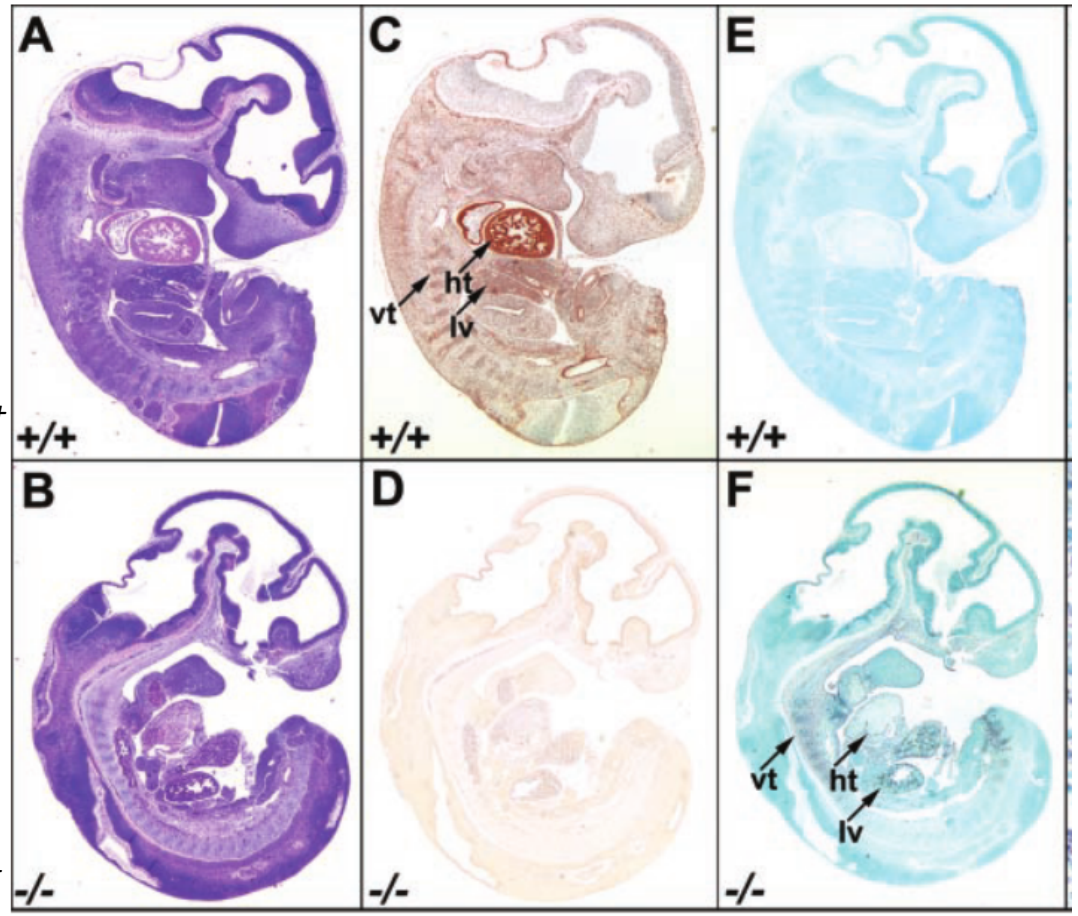
# RNA editing in mammalian models



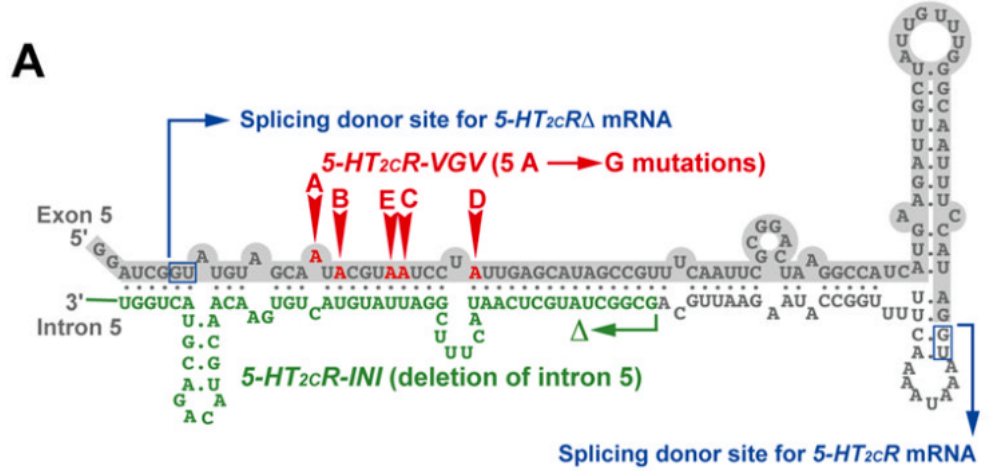

Editing level:
Fraction of edited transcripts

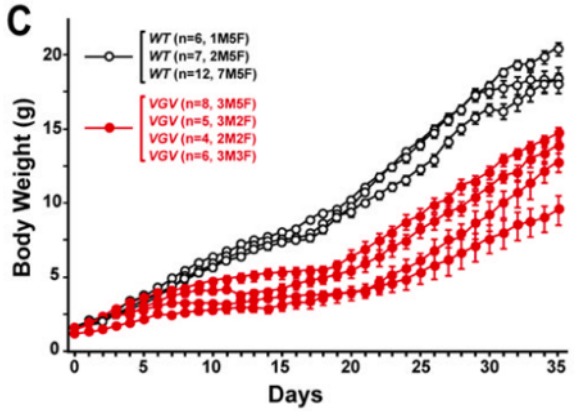
# RNA editing in mammalian models

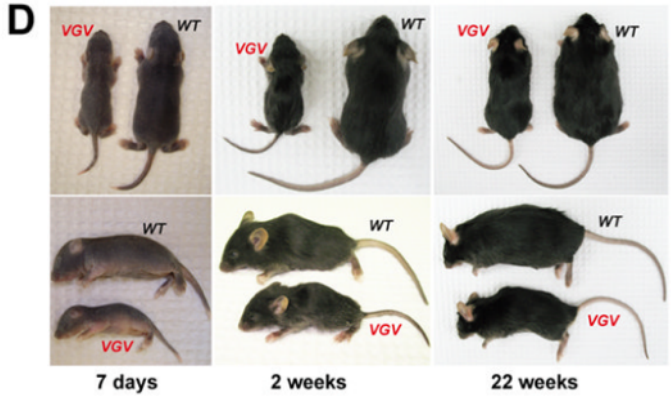



*ADAR*$^{+/+}$

*ADAR*$^{-/-}$

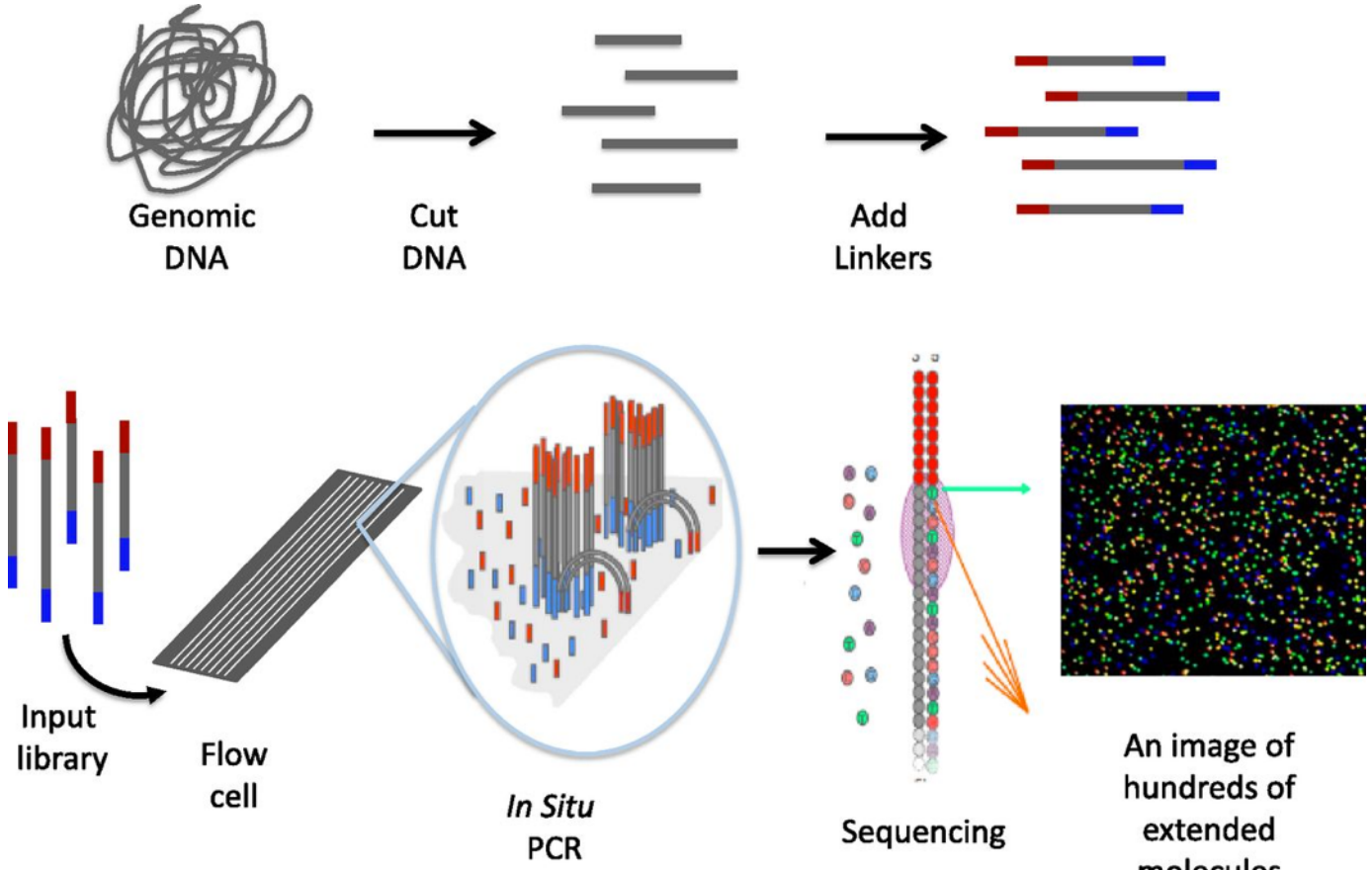


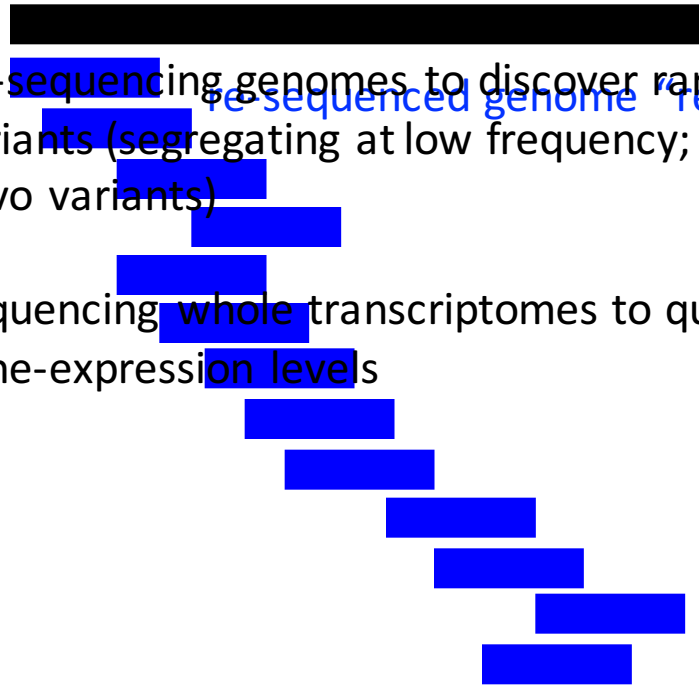

VGV mice = 100% Editing level at sites A-D

# Second-generation sequencing




Common uses of second-generation sequencing:

- Re-sequencing genomes to discover rare variants (segregating at low frequency; de-novo variants)
- Sequencing whole transcriptomes to quantify gene-expression levels



# A-to-I RNA editing occurs at over a hundred million genomic sites, located in a majority of human genes

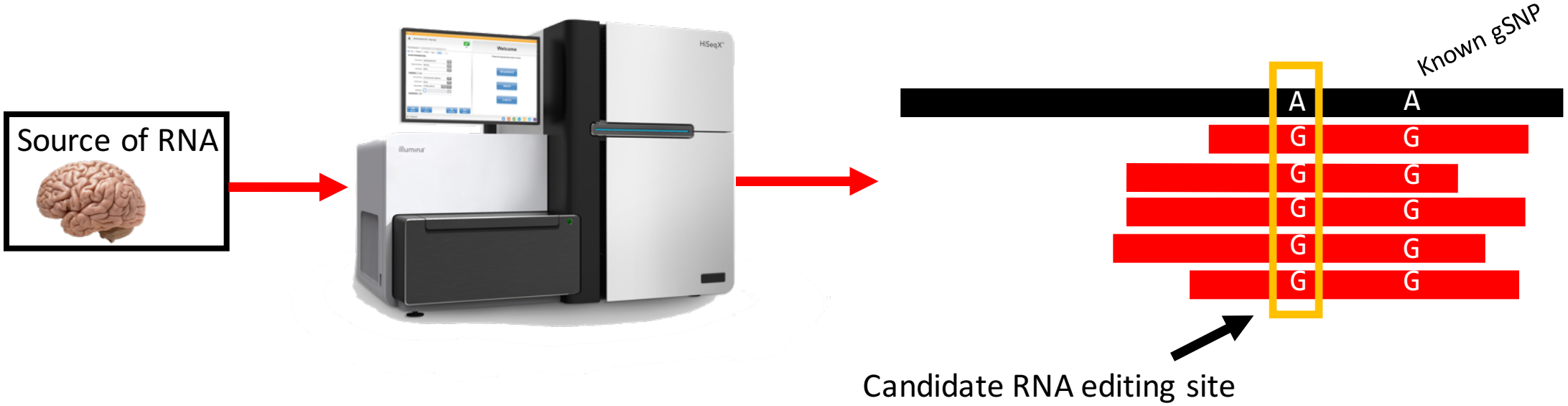
Lily Bazak,[1] Ami Haviv,[1] Michal Barak,[1] Jasmine Jacob-Hirsch,[1,2] Patricia Deng,[3] Rui Zhang,[3] Farren J. Isaacs,[4] Gideon Rechavi,[2,5] Jin Billy Li,[3] Eli Eisenberg,[6,7,8] and Erez Y. Levanon[1,7,8]

*[1]Mina and Everard Goodman Faculty of Life Sciences, Bar-Ilan University, Ramat Gan 52900, Israel; [2]Cancer Research Center, Chaim Sheba Medical Center, Tel Hashomer 52621, Israel; [3]Department of Genetics, Stanford University School of Medicine, Stanford, California 94305, USA; [4]Department of Molecular, Cellular and Developmental Biology and Systems Biology Institute, Yale University, New Haven, Connecticut 06520, USA; [5]Sackler School of Medicine, Tel Aviv University, Tel Aviv 69978, Israel; [6]Raymond and Beverly Sackler School of Physics and Astronomy and Sagol School of Neuroscience, Tel Aviv University, Tel Aviv 69978, Israel*

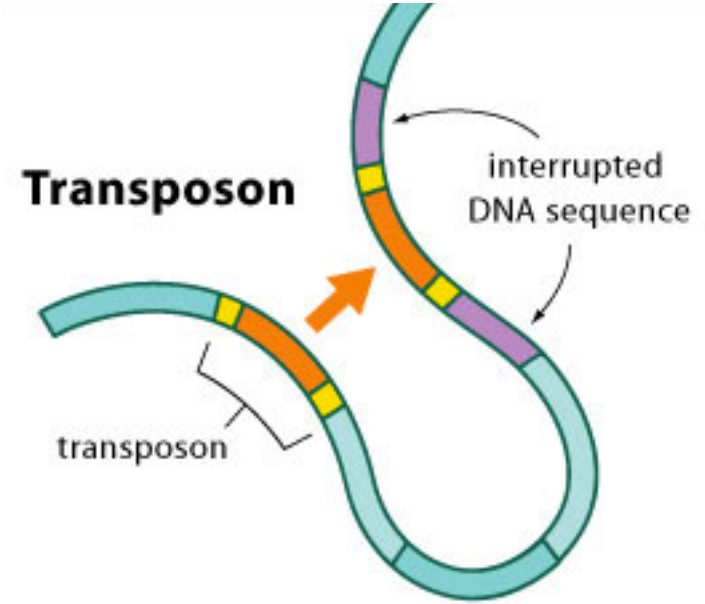
1.6 Million "edited sites"

>99% of all edited sites in Alu elements

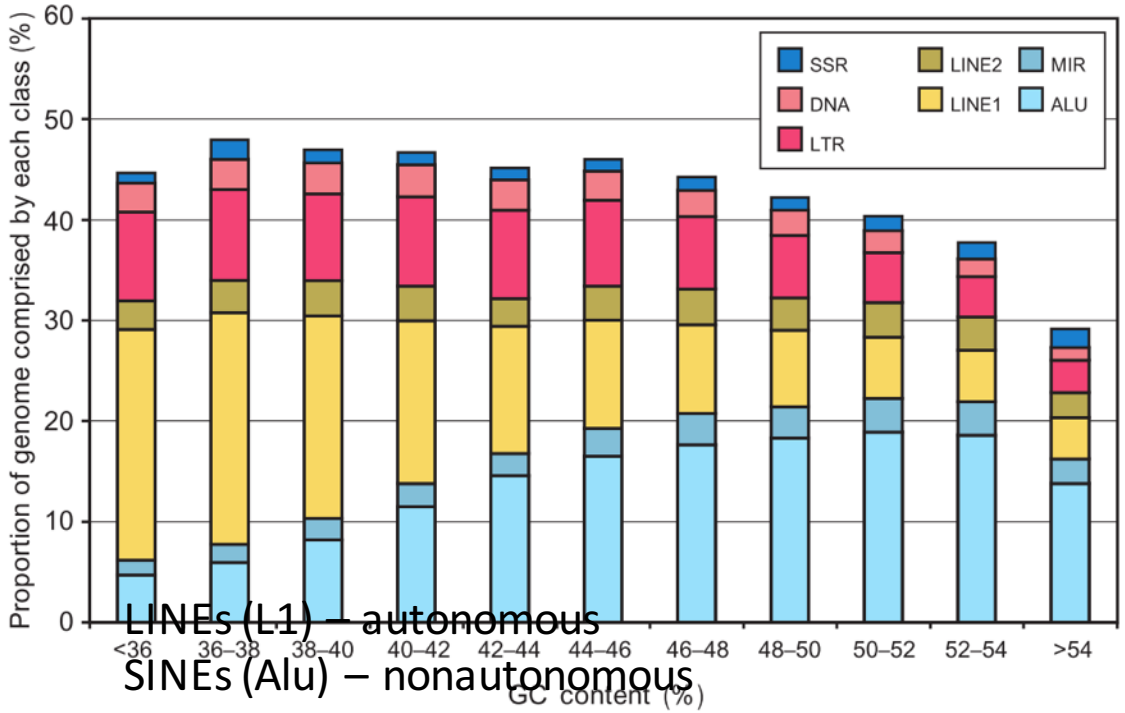
Nearly all adenosines within Alu elements appear editable. Estimated 100M edited sites.



# Transposable elements ("Junk DNA")




Human Genome




LINEs (L1) – autonomous

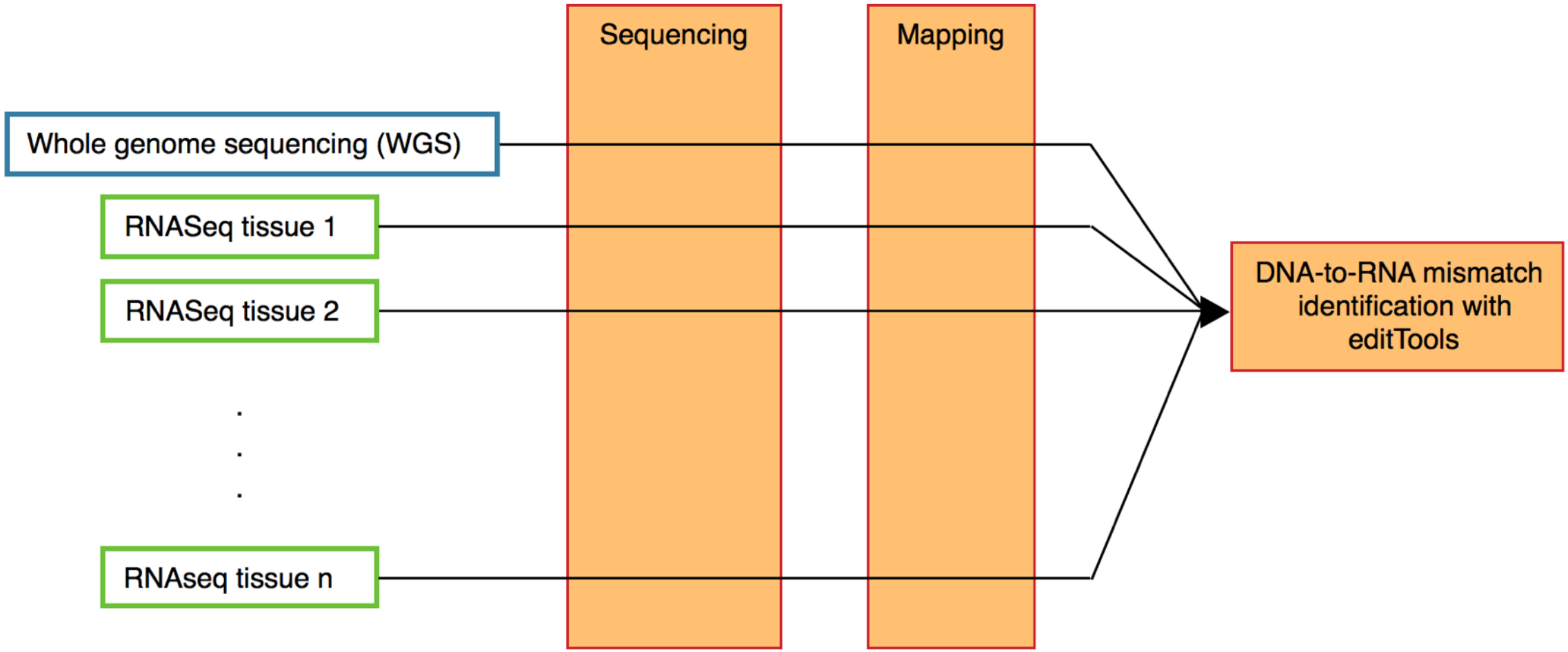
SINEs (Alu) – nonautonomous

# Currently…

- Most high-throughput RNA editing studies are done using human, other primates and mouse.

- Software designed to analyze RNA editing is nearly nonexistent

www.github.com/funkhou9/editTools




- Custom C++ libraries for processing of variant call format (VCF)
- *Rcpp: Seamless R and C++ Integration*

# Variant call format

```
##fileformat=VCFv4.0
##fileDate=20110705
##reference=1000GenomesPilot-NCBI37
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=.,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
```

| #CHROM | POS | ID | REF | ALT | QUAL | FILTER | INFO | FORMAT | Sample1 (DNA) | Sample2 (RNA_1) | Sample3 (RNA_2) |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 2 | 4370 | rs6057 | G | A | 29 | . | NS=2;DP=13;AF=0.5;DB;H2 | GT:GQ:DP:HQ | 0\|0:48:1:52,51 | 1\|0:48:8:51,51 | 1/1:43:5:.,. |
| 2 | 7330 | . | T | A | 3 | q10 | NS=5;DP=12;AF=0.017 | GT:GQ:DP:HQ | 0\|0:46:3:58,50 | 0\|1:3:5:65,3 | 0/0:41:3 |
| 2 | 110696 | rs6055 | A | G,T | 67 | PASS | NS=2;DP=10;AF=0.333,0.667;AA=T;DB | GT:GQ:DP:HQ | 1\|2:21:6:23,27 | 2\|1:2:0:18,2 | 2/2:35:4 |
| 2 | 130237 | . | T | . | 47 | . | NS=2;DP=16;AA=T | GT:GQ:DP:HQ | 0\|0:54:7:56,60 | 0\|0:48:4:56,51 | 0/0:61:2 |
| 2 | 134567 | microsat1 | GTCT | G,GTACT | 50 | PASS | NS=2;DP=9;AA=G | GT:GQ:DP | 0/1:35:4 | 0/2:17:2 | 1/1:40:3 |

```
class Rna {
    char genotype;
    int depth;
    int var_depth
    …
}
```

```
class Variant {
    char genotype;
    …
    std::list<Rna> r_list;
}
```

Single Animal
MSUPRP

RNAseq

150bp x 2 (~ 100M reads, SS)

Data cleaning (trimming)

Mapping (SS 10.2 – TopHat)

Extract "unique" alignments

+ strand alignments

- strand alignments

WGS

150bp x 2 (30X Coverage)

Data cleaning (trimming)

Mapping (SS 10.2 – Bowtie)

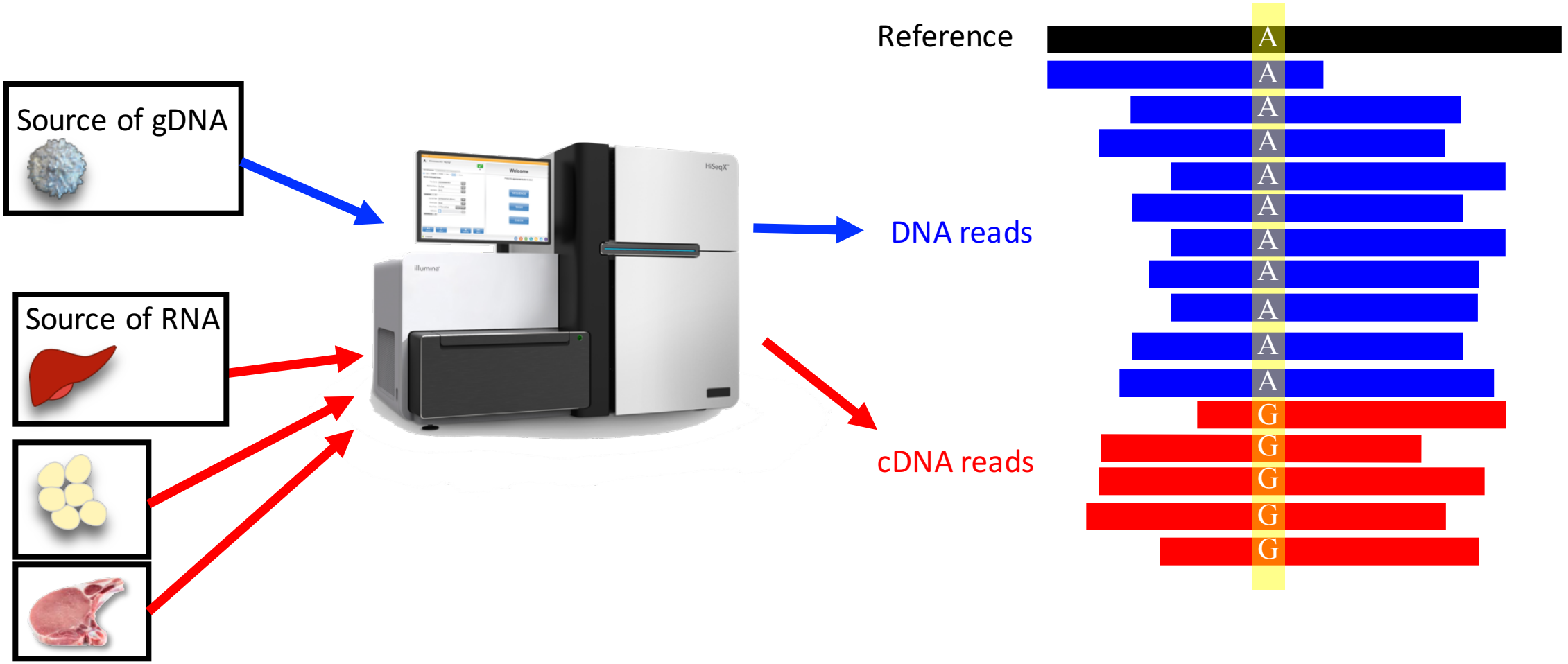
Extract "unique" alignments

Simultaneous
Variant Calling
(Samtools)

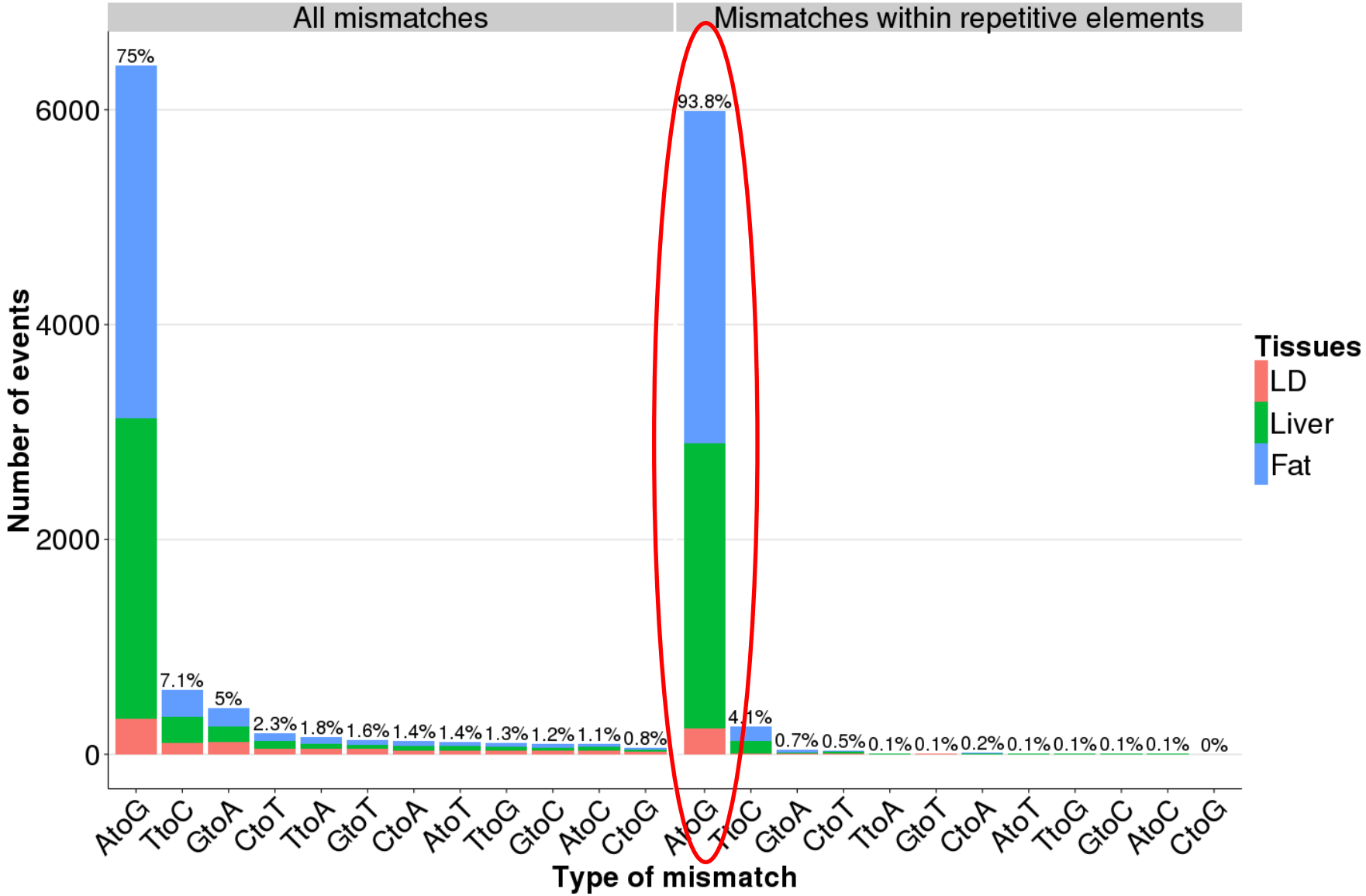
- DNA RNA mismatch detection
- RNA editing calling
- Visualization of edited loci counts, editing levels
- Merge with annotation files:
  - Transcript
  - Repeatmasker

("editTools")

# Calling RNA edited sites from WGS + RNASeq

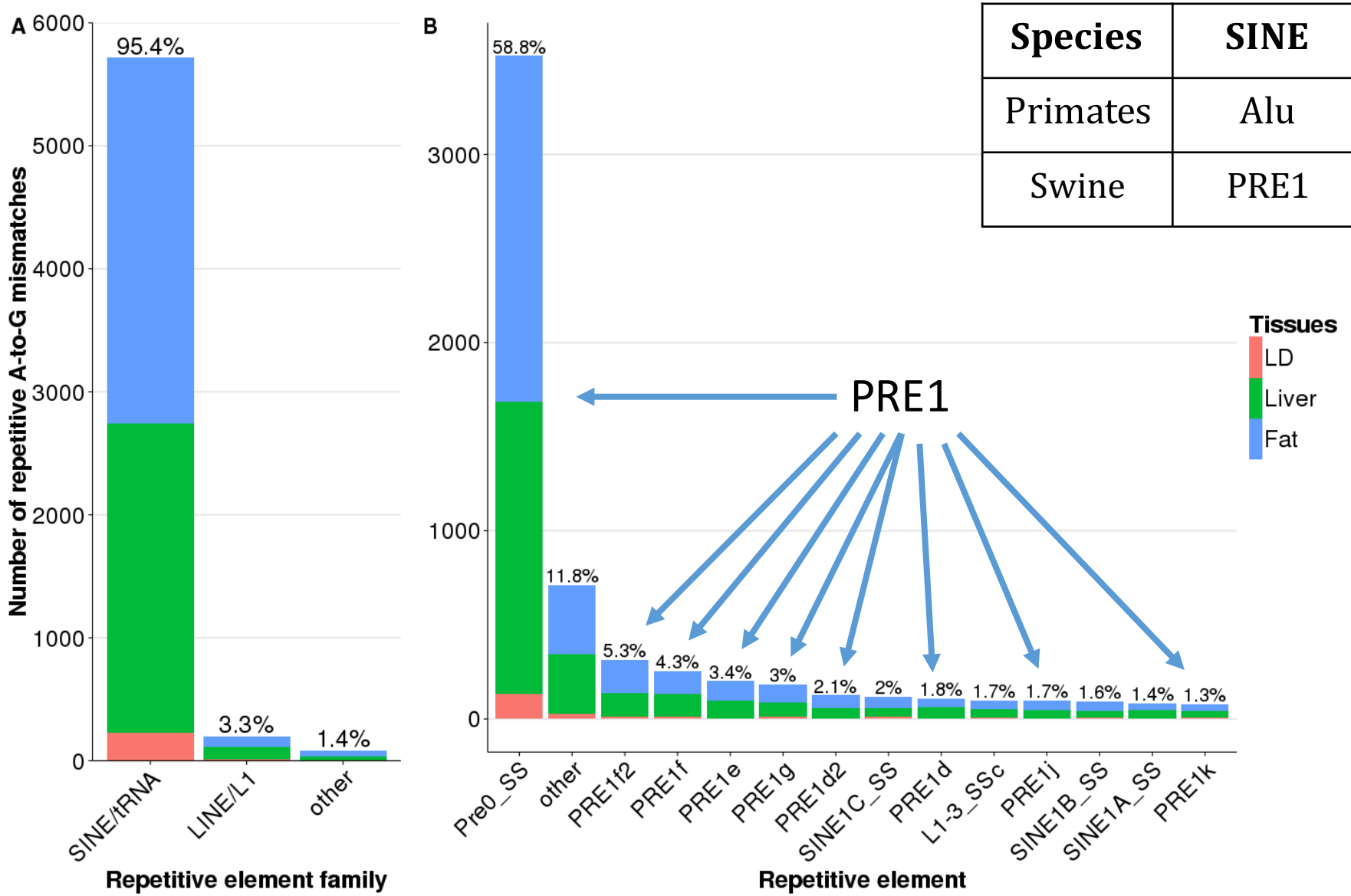


# Total DNA-to-RNA mismatch counts



# Repetitive RNA editing distribution

| Species | SINE | length | Copy number |
|---|---|---|---|
| Primates | Alu | 282bp | $1.1x10^6$ |
| Swine | PRE1 | 246bp | $1x10^6$ |



# Coding RNA editing distribution

- *GRIA2* – Normally edited at 100%. Lower editing levels associated with glioblastomas.
- *BLCAP* – Overediting associated with hepatocarcinoma.
- *NEIL1* – Unobservable in mouse

| Position | Gene symbol/ID | AA | SIFT | Tissues |
|---|---|---|---|---|
| 1:63408856 | ENSSSCG00000029003 | L/P | tolerated(1) | Fat LD Liver |
| 1:125424444 | ENSSSCG00000024660 | Q/R | tolerated(1) | Fat LD Liver |
| **11:22178068** | ***COG3*** | **I/V** | **tolerated(1)** | **Fat LD Liver** |
| 12:20231860 | *AOC3* | Q/R | tolerated(1) | Liver |
| 13:131377159 | *EIF2B5* | Q/R | tolerated(1) | Fat |
| 13:156760971 | *UBE2B* | D/G | tolerated(0.48) | Fat LD Liver |
| 13:206979572 | *SON* | R/G | - | Fat |
| 14:40832826 | *PLBD2* | R/G | tolerated low confidence(0.12) | Fat |
| 14:52398588 | *IGLV-3* | E/G | tolerated(0.05) | Fat |
| 14:59613334 | *LYST* | S/G | - | LD |
| 14:81796679 | *OIT3* | S/G | tolerated(1) | Liver |
| 15:59811585 | *HNRNPA2B1* | L/P | tolerated(0.35) | Fat LD Liver |
| 15:98217885 | ENSSSCG00000028949 | R/G | tolerated low confidence(1) | Fat LD Liver |
| 16:29335640 | ENSSSCG00000016869 | N/D | tolerated(1) | Fat LD |
| 16:42512978 | *ELOVL7* | S/G | tolerated(1) | Fat |
| **17:46041505** | ***BLCAP*** | **Y/C** | **deleterious(0)** | **Fat Liver** |
| 2:12622576 | *LDHB* | I/M | tolerated(1) | Fat LD Liver |
| 2:49316285 | *ARNTL* | K/E | tolerated low confidence(1) | Liver |
| **4:98044799** | ***COPA*** | **I/V** | **deleterious(0.02)** | **Fat** |
| 5:42375023 | *KRR1* | I/T | deleterious(0.01) | Liver |
| 6:92516721 | *PTPRM* | K/R | tolerated(1) | Fat |
| 6:146168578 | *NDC1* | E/G | deleterious(0.01) | Liver |
| **7:62951442** | ***NEIL1*** | **K/R** | **deleterious(0.02)** | **Fat LD** |
| 7:81602273 | ENSSSCG00000002045 | C/R | tolerated(1) | Fat LD Liver |
| 7:102789222 | *ACOT4* | T/A | tolerated(0.61) | Fat |
| 7:129322238 | *RPS21* | C/R | - | Fat LD Liver |
| 8:28015971 | ENSSSCG00000008767 | H/R | tolerated(1) | Fat LD Liver |
| 8:31629014 | *TLR1* | I/V | tolerated(1) | Liver |
| 8:32309809 | *RPL9* | I/V | tolerated(0.4) | Fat |
| 8:32309814 | *RPL9* | E/G | deleterious(0.01) | Fat |
| **8:48244993** | ***GRIA2*** | **Q/R** | **tolerated(0.07)** | **Fat** |
| 9:41146365 | ENSSSCG00000023913 | Q/R | deleterious(0.04) | Fat |
| 9:74510703 | ENSSSCG00000015294 | K/R | tolerated(0.13) | Liver |
| 9:83273454 | *SLC25A13* | E/G | deleterious(0.02) | LD |

# Future goals

- *Evidence for substantial RNA editing transcriptome-wide among Sus scrofa PRE-1 SINE elements.* Intended for a "Report" in the journal *RNA.*
- Enhance the functionality and usability of *editTools*. Targeted for Bioconductor.
- FAANG – 6 Pigs. Each with WGS and RNASeq for *liver, fat, muscle, spleen, cortex, cerebellum, hypothalamus, and lung.*
- USDA-MARC – 3rd Generation Sequencing data. Compare to 2nd generation sequencing data from *hypothalamus, spleen, thymus and small intestine*.

# Future of RNA editing

- Using RNA editing levels as predictors of quantitative traits
- $p \sim$ 1K or 1M or 100M
- At the current cost of 2nd Gen sequencing, $n$ would typically be very small.