A survey of genomic traces reveals a common sequencing error, RNA editing, and DNA editing

Alexander Wait Zaranek, E. Y. Levanon, T. Zecharia, T. Clegg, and G. M. Church

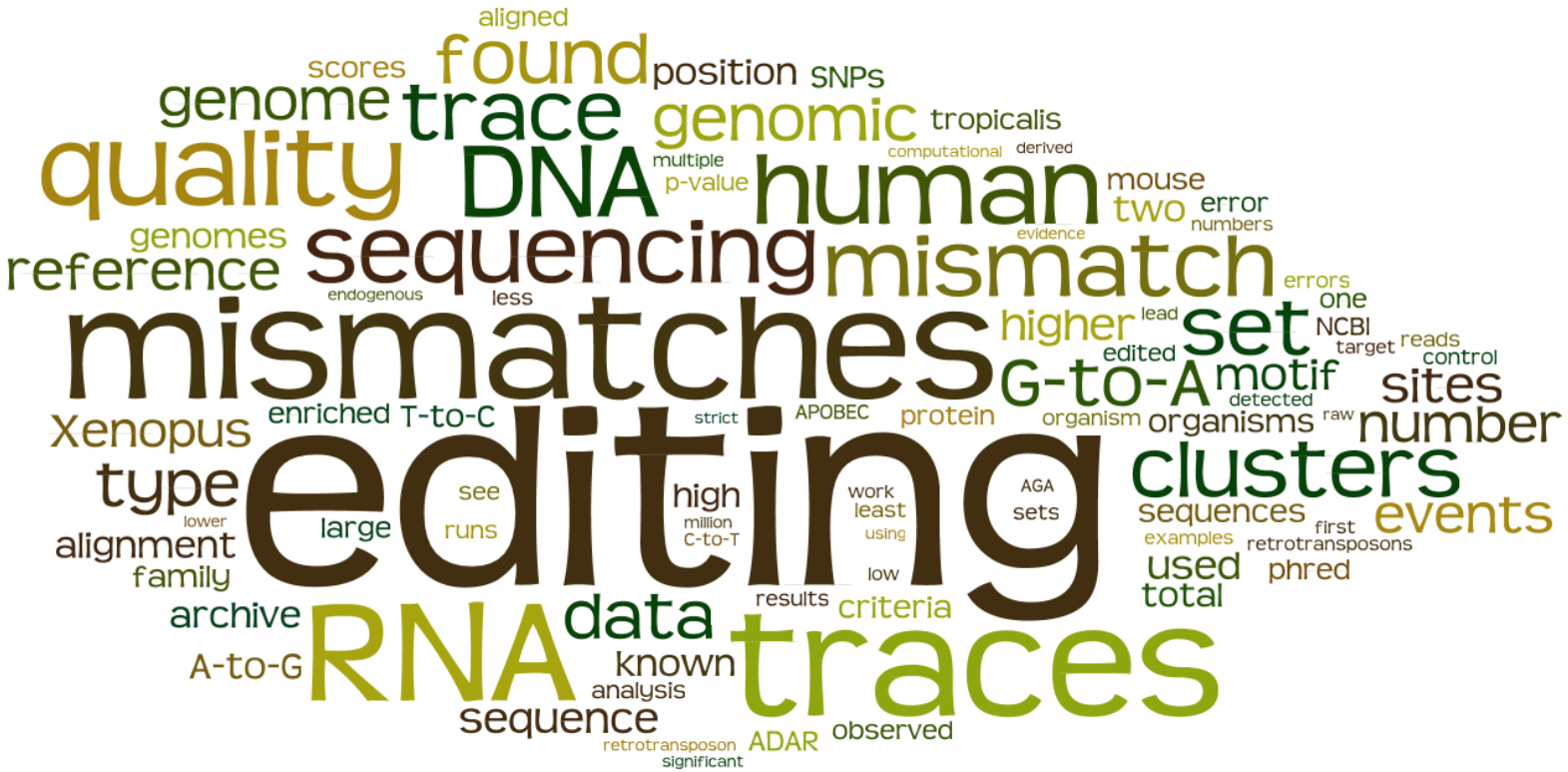
PLoS Genetics **6** (2010) 1–11.

Speaker: Joseph Chuang-Chieh Lin

The Comparative & Evolutionary Genomics/Transcriptomics Lab.
Genomics Research Center, Academia Sinica
Taiwan

26 September 2012

The word cloud



Outline

- It is widely believed that an organism's genomic content should be fixed throughout its lifetime with the exception of infrequent somatic mutations.
- However, proteins that can modify genomic content have been identified in human and many other organisms.

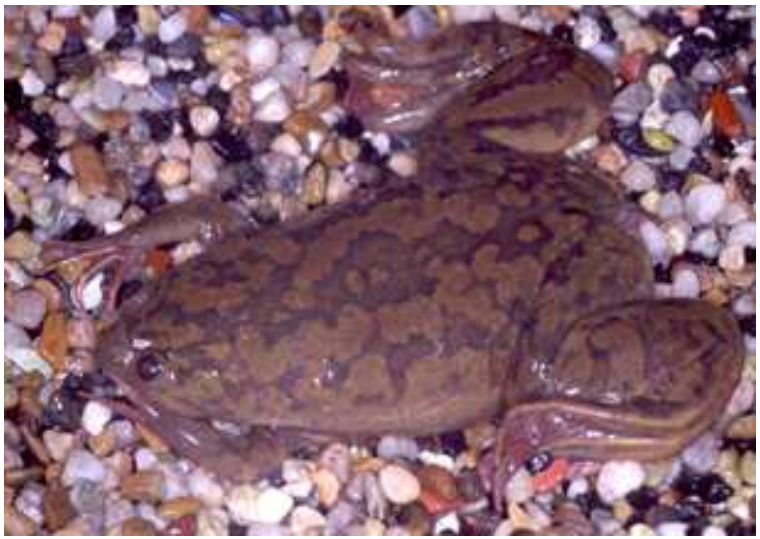
Proteins which can modify DNA/RNA

- The family of **a**denosine **d**eaminase **a**cting on **R**NA (ADAR).
  - Adenosine (A) $\rightarrow$ Inosine (I) (read as Guanosine (G) in turn).
  - On RNA nucleotides.
- The families of **a**ctivation-**i**nduced **d**eaminase (AID) & **apo**lipoprotein **B** **e**dting **c**omplex (APOBEC) deaminase.
  - Cytosine (C) $\rightarrow$ Uracil (U).
  - On both DNA & RNA nucleotides.

Summary of the paper

- Analyze the raw data used to assemble the reference genomes (in NCBI Trace Archive) of ten organisms to discover:
  - Sequencing error;
  - DNA editing;
  - RNA editing.
- The ten organisms:
  - Mosquito (anoGam1), Marmoset (calJac1), Dog (canFam2), Drosophila (dm3), Chicken (galGal3), Human (hg18), Mouse (mm9), Chimp (panTro2), Fugu (fr2), and *Xenopus tropicalis* (xenTro2).
- The criteria of clusters of consecutive mismatches of the same type.
- The first investigation of extensive RNA editing in *Xenopus tropicalis*.

African/Western (Tropical) Clawed Frogs



Xenopus laevis

Xenopus tropicalis

- ADAR activity was first observed in *Xenopus laevis* oocytes [Bass & Weintraub *Cell* 1987].

More about AID/APOBEC family of deaminases

APOBEC1:

The first family member to be found and studied.

- Edit the apolipoprotein B (ApoB) RNA, which is involved in **lipid transport**.
  - Navaratnam *et al.* *J. Biol. Chem.* 1993.
  - Teng *et al.* *Science* 1993.
- Deaminate cytidine in DNA.
  - Harris *et al.* *Mol. Cell* 2002.

AID:

Discovered to be vital for antigen-driven diversification of immunoglobulin genes in the vertebrate adaptive immune system.

- Muramatsu *et al.* Jm Biol. Chem. 1999 & *Cell* 2000.
- Revy *et al.* Cell 2000.

More about AID/APOBEC family of deaminases (contd.)

APOBEC3s:

Involved in the **restriction** of retrovirus proliferation in primates.

- Jarmuz *et al.* *Genomics* 2002.
- Sheehy *et al.* *Nature* 2002.

APOBEC3G:

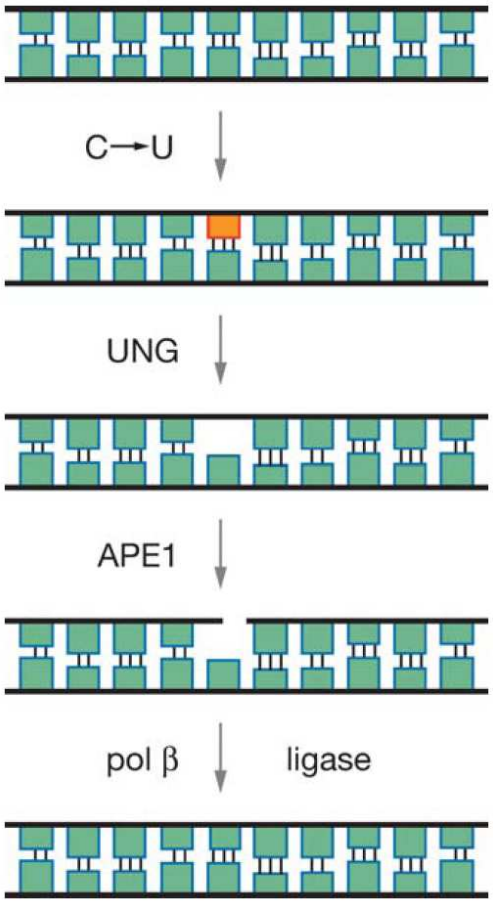
Serve as a potent **inhibitor** of a wide range of retroviruses, including endogenous retrotransposons.

- Harris *et al.* *Cell* 2003.
- Mangeat *et al.* *Nature* 2003.
- ...

Capable of editing the mouse IAP retrotransposon.

- Esnault *et al.* *Nature* 2005.
- IAP (intracisternal A-particle): endogenous sequences: retrovirus-like mobile elements; $\approx$ 1,000 copies in the mouse genome.

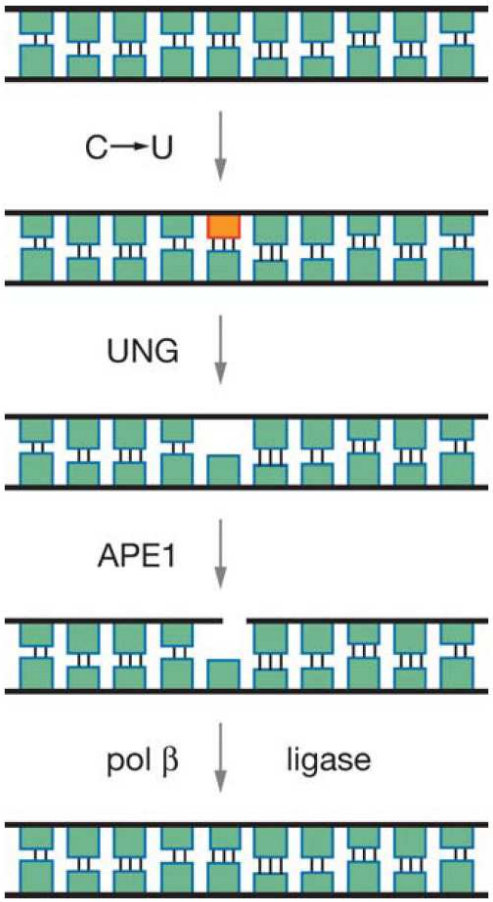
Faithful repair of uracil in DNA



⋆ Uracil is repaired by a conserved and ubiquitous pathway: uracil nucleoside glycosylase (UNG) removes the uracil base (orange), AP endonuclease 1 (APE1) cleaves the phosphodiester backbone at the abasic site, and DNA polymerase and ligases repair the gap.

[Refer to N. Maizels: Immunoglobulin Gene Diversification *Annu. Rev. Genet.* 2005]

AID overrides uracil DNA repair in *E. coli*



⋆ AID expression can overwhelm the normally efficient uracil DNA repair pathway to cause mutagenesis in *E. coli*.

[Refer to N. Maizels: Immunoglobulin Gene Diversification *Annu. Rev. Genet.* 2005]

"C-to-U" vs. "G-to-A" (DNA editing)

- C-to-U DNA editing by various APOBEC protein families is characterized by clusters of "G-to-A" mismatches between the reference genome and the edited sequence.
- These mismatches are the end product of deamination of "C" into "U" in the other DNA strand (newly formed) after reverse transcription.

Materials & methods

Materials & methods (data preparation)

- Obtain all traces for 10 organisms (603,249,815 traces in total) in NCBI Trace Archive (May 2008) and align them with their reference genomes.
  - $\approx$ 300 million that aligned uniquely.
- Download **SCF** raw binary data from the trace archive and analyze them using Phred version 0.071220.b.
  - SCF data: chromatogram files used to store data from DNA sequencing.
  - Phred: generate an alternative base call for every position in the trace.
- Align the two sequences from the same trace separately and look for a large alignment with a single bp off-set.

Materials & methods (mapping tool)

- The applied sequence alignment tool: MegaBlast (v.2.2.13).
  - Optimized for aligning sequences that *differ slightly*.
  - More efficient to handle *much longer* DNA sequences than the *blastn* of traditional BLAST algorithm.
- Parameters:
  - alignment length $\geq$ 400bp
  - identity $\geq$ 97%
  - no regions to be masked
  - gap penalty: 25
  - gap extension penalty: 10

$\star$ Only unique alignments matching the above criteria were retained.

Materials & methods (computation facilities)

- Two computational clusters were used:
  - 96 nodes w/ (predominantly) 4 $\times$ 1.8GHZ Opteron cores, 4–16GB RAM/node, 0–3750GB disk/node.

⋆ The human analysis consumed 347 node days and 530GB of space (reduced to 22GB by further processing).

⋆ The mouse analysis consumed greater than **4.2TB**.
  - Many mouse traces may not place uniquely.

Materials & methods (contd.)

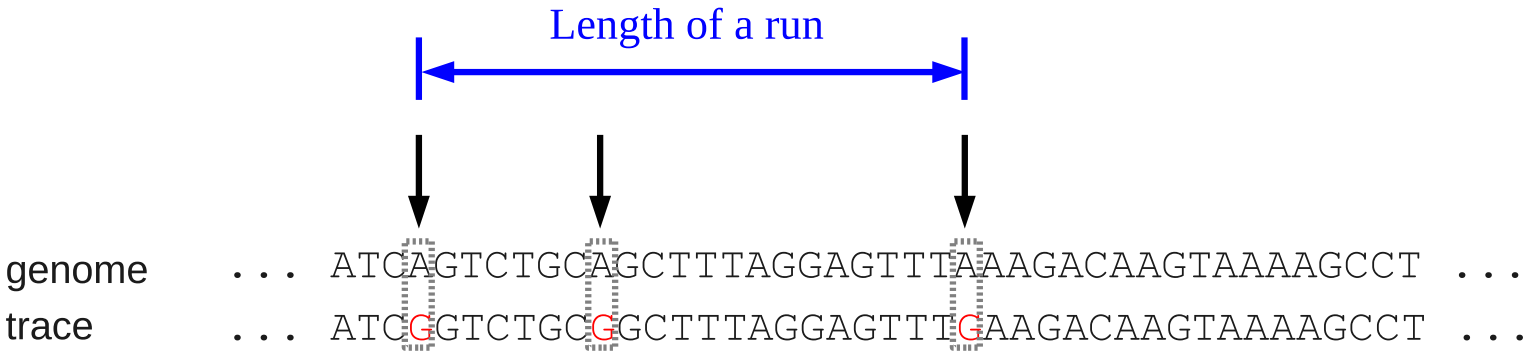
Table 1. Summary of computation.

Organism name	#reference bp (millions)	#unique traces (millions)	Mean coverage	Space (Gb)	Time (millions of node seconds)
Anopheles gambiae	260	4.3	9.9	13	0.56
Callithrix jacchus	2,900	22	4.6	160	1.5
Canis familiaris	2,400	33	8.3	370	3.4
Drosophila melanogaster	160	0.67	2.5	2.5	0.06
Gallus gallus	1,000	12	7.2	30	1.3
Homo sapiens	2,900	85	18	530	30
Mus musculus	2,600	93	21	4,200	114
Pan troglodytes	2,900	32	6.6	150	7.0
Takifugu rubripes	350	2.5	4.2	6.4	1.2
Xenopus tropicalis	1400	14	6.0	360	4.8
Total		298.47		5821.90	163.82

Total data generated from analysis of 603,249,815 traces, 30% of the total number of traces at NCBI (outside the short-read archive). Approximately half were placed uniquely while applying our cutoffs, with total data consuming six terabytes of disk and more than five "node years"of CPU time. The computation on mouse traces produced the bulk of the data.
doi:10.1371/journal.pgen.1000954.t001

Materials & methods (editing enrichment criteria)

- Editing enrichment criteria:
  - Runs of ≥ 3 consecutive mismatches of the same type.
  - ⋆ Clusters of consecutive mismatches of the same type care common in APOBEC/ADAR targets.



- $\approx$ 20.7 million traces of human were potentially enriched for editing.
- Augment the data by downloading auxiliary information and quality scores for the $\approx$ 20.7 million traces.

Materials & methods (filtering runs by three constraints)

Consider the 20.7 million human traces potentially enriched for editing.

- $\star_1$ ≥ 5 consecutive mismatches
  - $\star$ 657,826 traces left;
  - $\star$ 218,595 (33%): G-to-A.
- $\star_2$ Discard runs of length < 100bp & traces where the mismatch site (ref. or trace) were 'N'.
  - $\because$ Sequencing errors tend to form **short** mismatch clusters.
- $\star_3$ Restrict to traces with identical 3-bp motif centered at each mismatch site.
  - $\because$ Editing enzymes have a preferred sequence content.
  - $\star$ "AGA-to-AAA" (26,694; 49.8%) & "AGG-to-AAG" (21,274; 39.7%).

♠ 53,639 traces left.

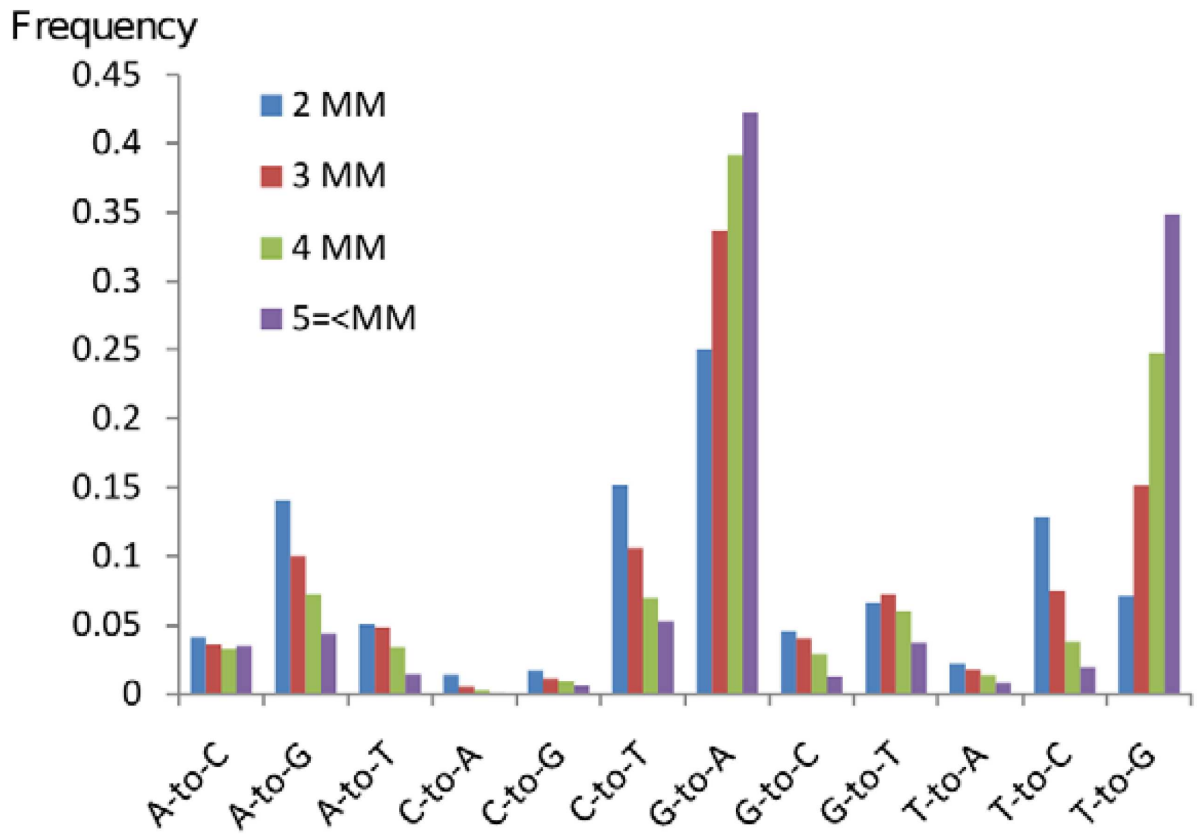
- $\star$ 46,483 (82%): G-to-A.

Materials & methods (filtering runs by three constraints)

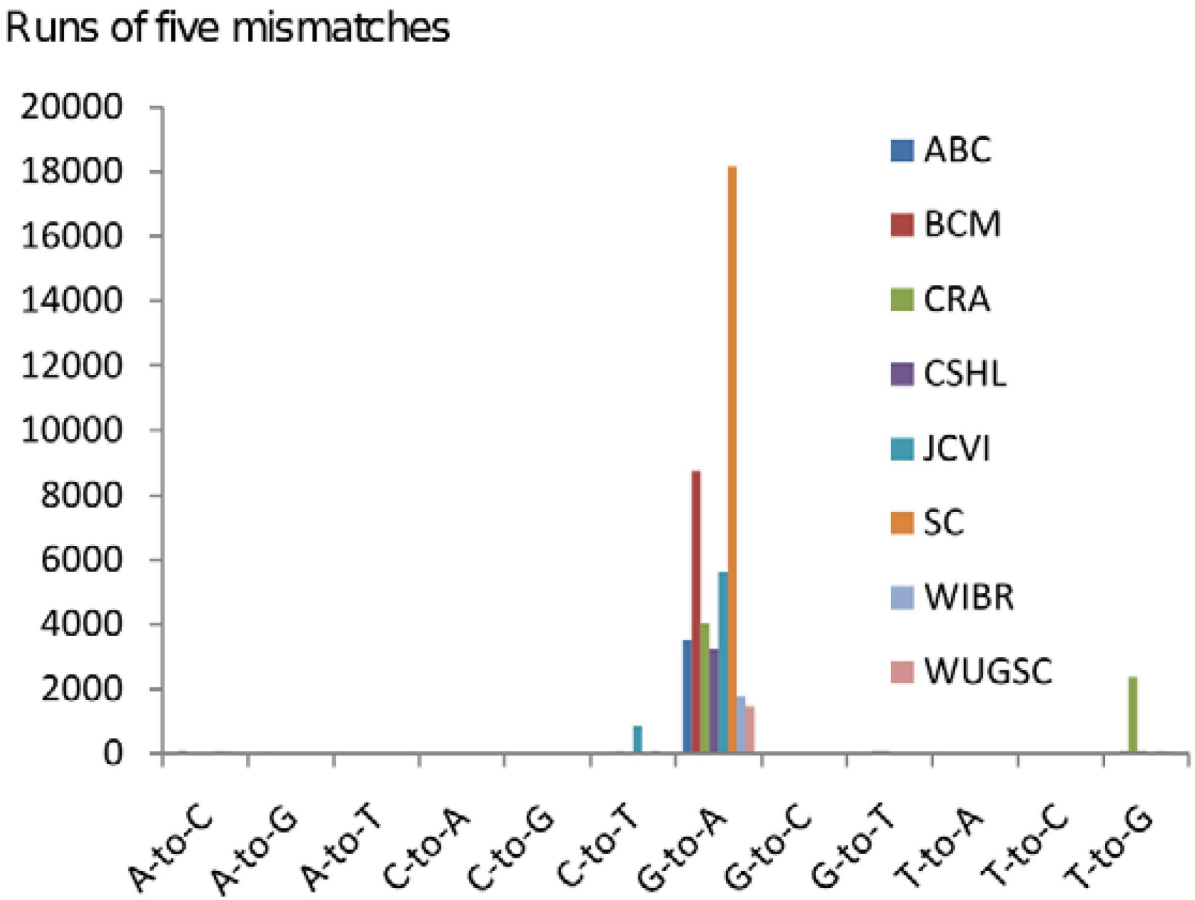
Consider the 20.7 million human traces potentially enriched for editing.

- $\star_1$ ≥ 5 consecutive mismatches
  - $\star$ 657,826 traces left;
  - $\star$ 218,595 (33%): G-to-A.
- $\star_2$ Discard runs of length < 100bp & traces where the mismatch site (ref. or trace) were 'N'.
  - $\because$ Sequencing errors tend to form **short** mismatch clusters.
- $\star_3$ Restrict to traces with identical 3-bp motif centered at each mismatch site.
  - $\because$ Editing enzymes have a preferred sequence content.
  - $\star$ "AGA-to-AAA" (26,694; 49.8%) & "AGG-to-AAG" (21,274; 39.7%).
- ♠ 53,639 traces left.
  - $\star$ 46,483 (82%): G-to-A.

Materials & methods (contd.)

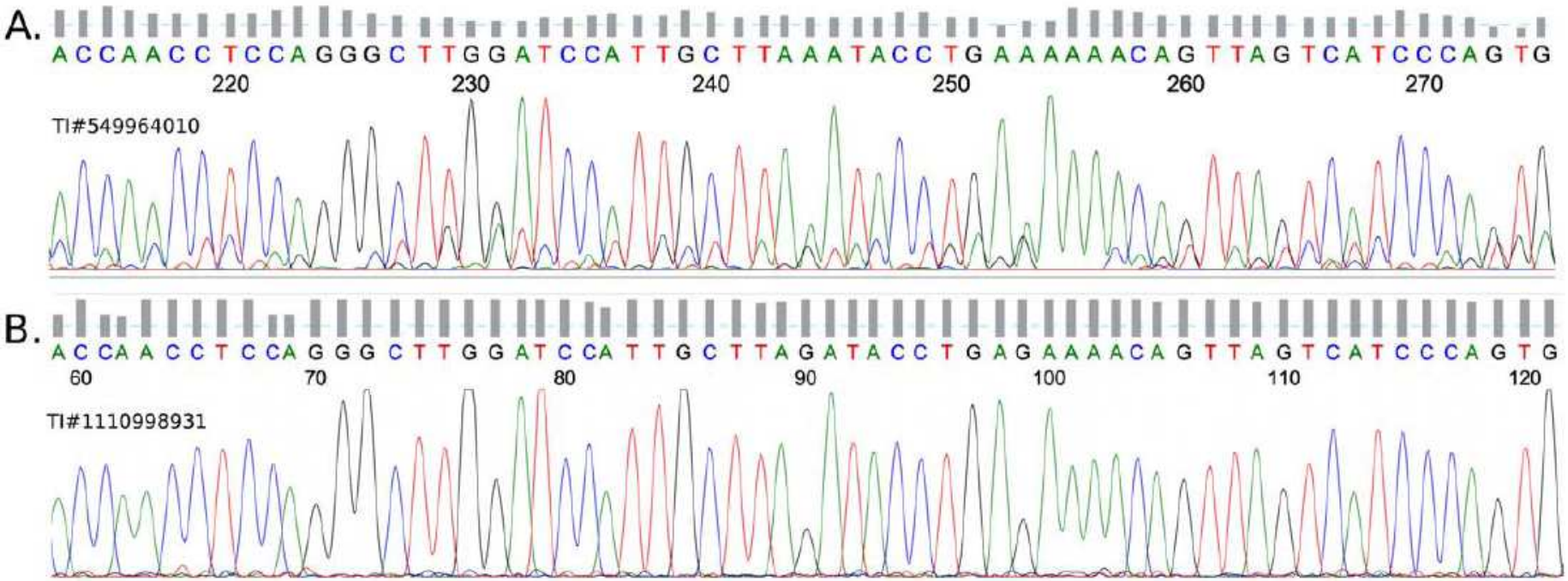


Materials & methods (contd.)



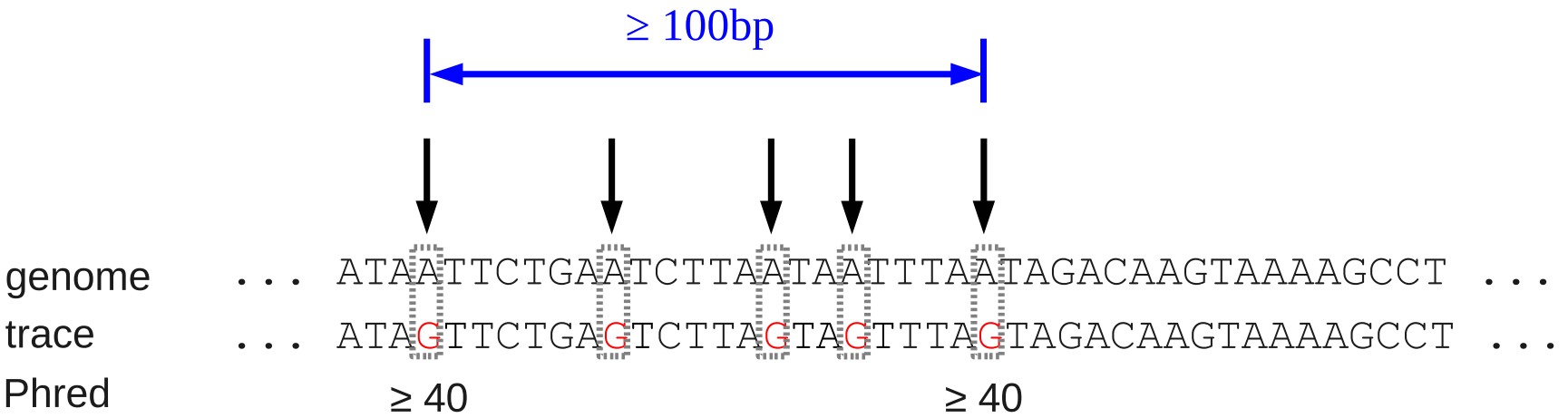
- Yet, traces are derived from both DNA strands (G-to-A ↔ C-to-T symmetric?)

An example of G-to-A sequencing artifact

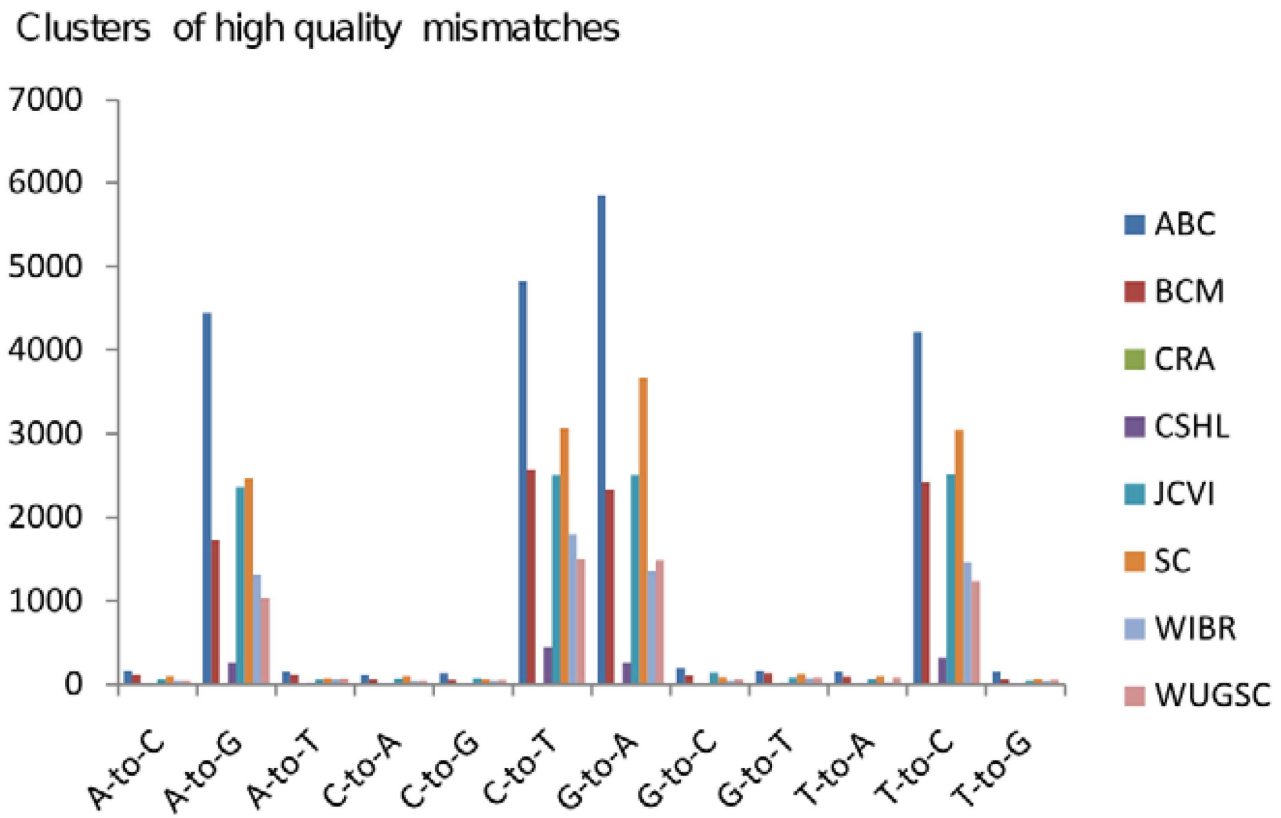


Materials & methods (incorporating Phred quality scores)

- Phred $a \cdot 10$:
  - **Pr**[a base call is incorrect] $= 10^{-a}$.



Materials & methods (contd.)



Results

Editing enriched traces of high quality

Table 2. Editing enriched traces—higher quality.

Reference genome version	G-to-A	C-to-T	A-to-G	T-to-C	Other
anoGam1	2836	2830	2907	3098	440
calJac1	3012	3362	2735	3133	145
canFam2	3170	3777	3270	3027	212
dm3	1	1	0	1	0
galGal3	1290	878	1026	1760	48
hg18	17719(82)	16778(72)	13701(188)	15301(419)	700(8)
mm9	1801(219)	1644(272)	1346(276)	1411(346)	76(11)
panTro2	3485	3120	2918	4046	240
fr2	467	449	390	482	45
xenTro2	1483(202)	1574(262)	1461(1289)	1631(1066)	269(28)

Number of traces by mismatch type with two or more mismatches at or above a quality threshold of phred 40, spanning 100bp or more. All mismatches belong to runs of three consecutive mismatches of the same type of any quality. The number of traces from the next largest substitution type, or the largest substitution type if it is not one of A-to-G, T-to-C, G-to-A, or C-to-T, is shown in the "other" column for comparison. The numbers in parentheses indicate traces of RNA origin. See Materials and Methods for more details.

Sequencing artifact may disrupt the accuracy of genomic assemblies

- Each position in the reference genome: determined by majority voting of the supporting traces.
- In genomic projects with **low coverage**: the error could not be detected.
- There are genomes with lower coverage tended to be free of G-to-A mismatches (most striking in *drosophila*).

The effect in the assignment of SNPs

- A sequencing error in one genomic trace will not usually lead to the determination of a SNP at this position.
- However, many of the "AGA" mismatches have a quality score of phred ≥ 20, which is considered an acceptable quality.
  - Some of them might be classified as SNPs.

⋆ Evidence:

- In 26,694 traces with identical 3bp G-to-A motif in runs of ≥ 5: $\approx$ 260,000 G-to-A mismatches with the 3-bp motif AGA-AAA.
  - 28,722 appear in dbSNP (11,145 in HapMap; genotyped in 4 populations) $\Rightarrow$ not real SNPs.
  - ⋆ 10,532 (94%) in HapMap are homozygous for the reference allele (G) with no representation of other SNP allele in any of the 90 individuals genotyped in the Yoruba population.

The effect in the assignment of SNPs

- A sequencing error in one genomic trace will not usually lead to the determination of a SNP at this position.
- However, many of the "AGA" mismatches have a quality score of phred ≥ 20, which is considered an acceptable quality.
  - Some of them might be classified as SNPs.

⋆ Evidence:

- In 26,694 traces with identical 3bp G-to-A motif in runs of ≥ 5: $\approx$ 260,000 G-to-A mismatches with the 3-bp motif AGA-AAA.
  - 28,722 appear in dbSNP (11,145 in HapMap; genotyped in 4 populations) $\Rightarrow$ not real SNPs.
  - ⋆ 10,532 (94%) in HapMap are homozygous for the reference allele (G) with no representation of other SNP allele in any of the 90 individuals genotyped in the Yoruba population.

The effect in the assignment of SNPs

- A sequencing error in one genomic trace will not usually lead to the determination of a SNP at this position.
- However, many of the "AGA" mismatches have a quality score of phred ≥ 20, which is considered an acceptable quality.
  - Some of them might be classified as SNPs.

⋆ Evidence:

- In 26,694 traces with identical 3bp G-to-A motif in runs of ≥ 5: $\approx$ 260,000 G-to-A mismatches with the 3-bp motif AGA-AAA.
  - 28,722 appear in dbSNP (11,145 in HapMap; genotyped in 4 populations) $\Rightarrow$ not real SNPs.
  - ⋆ 10,532 (94%) in HapMap are homozygous for the reference allele (G) with no representation of other SNP allele in any of the 90 individuals genotyped in the Yoruba population.

The effect in the assignment of SNPs

- A sequencing error in one genomic trace will not usually lead to the determination of a SNP at this position.
- However, many of the "AGA" mismatches have a quality score of phred ≥ 20, which is considered an acceptable quality.
  - Some of them might be classified as SNPs.

⋆ Evidence:

- In 26,694 traces with identical 3bp G-to-A motif in runs of ≥ 5: $\approx$ 260,000 G-to-A mismatches with the 3-bp motif AGA-AAA.
  - 28,722 appear in dbSNP (11,145 in HapMap; genotyped in 4 populations) $\Rightarrow$ not real SNPs.
  - ⋆ 10,532 (94%) in HapMap are homozygous for the reference allele (G) with no representation of other SNP allele in any of the 90 individuals genotyped in the Yoruba population.

DNA editing

- In the mouse genome:
  - (A-to-G/T-to-C m.m., C-to-T/G-to-A m.m.) = (7,860, 9,799).
  - In IAP regions: (49, 114). [p-value: 0.00018]
  - ⋆ The origin of the mismatches: a result of editing by APOBEC after reverse transcription of the retrotranposons.
- In human genome:
  - (A-to-G/T-to-C m.m., C-to-T/G-to-A m.m.) = (79,401, 91,120).
  - In HERVK retrotransposon elements: (129, 247). [p-value: 1.7×10^{-6}]
  - ⋆ Two examples of the editing events in HERVL-A1 and in *AluY* (the most active SINE family) are present.
    - ◇ HERVs: Human Endogenous RetroVirus-Like sequences;
    - ◇ SINE: Short INterspersed Elements.

DNA editing in human HERVL-A1

```
Query  1    TGACAGTGGATTATCATAAGCTTAATCAAGTGGTGACTCCAATTTCAGCTGCTGTACCAG  60
Sbjct  1    ............................................................  60

Query  61   ATGTGGTTTCATTGCTTGAGCAAATTAACACATCTGGTACCTGGTATGCAGCCACTGACT  120
Sbjct  61   .....................................G......................  120

Query  121  TGGCCTTCGGAGCCTTTGGCAGGCTCCCATAAGTGAATCACAGTGGAGGCCTGTAGGATT  180
Sbjct  121  ....G.........................................................  180

Query  181  TTGGAGCAAGGCCCTACCATCTTCTGAAAATAACTACTCTCCTTTTGACAGACAGCTCTT  240
Sbjct  181  ............................................................  240

Query  241  GGCCTGTTACTGGGCTTTGGTGGAAACTGAATGTTTGACTATGGGTCATCAAGTCACCAT  300
Sbjct  241  ............................................................  300

Query  301  GCGACCTGAACTGTCTATCATGCACTGGATGTTTTCTGACCCATCTGGTCATAAAGTGGG  360
Sbjct  301  ............................................................  360

Query  361  TCATGCACAGCAGCATTCCATCATCAAATGGAAGTGGTATATATGTGATCGGGCTCGAGC  420
Sbjct  361  ............................................................  420

Query  421  CGGTCCTGAAGGCACAAGTAAGTTACATGAGGAAGTGGCTCAAGTGCCCATGGTCTCTAC  480
Sbjct  421  ............................................................  480

Query  481  TCCTGCCACCTGCCTTCTCTCCCCTAGCCTGCACCGATGGCCTCATGGGGAGTTCCCTGT  540
Sbjct  481  ............................................................  540

Query  541  GATCAGTTGACAGAGGAAGGGAAGACTAGGCCCTGGTTCAGAGATGGTTCTACATGATAT  600
Sbjct  541  ............................................................  600

Query  601  GCAGGCACCACCCGGAAGTGGACAGCTGCAGGACTACAGCCCTTTCTAGGACATCCCTGA  660
Sbjct  601  ............A.............A.............................  660

Query  661  AGGACAGCGGTGGAGGGAACTTCCCAGTGGGCAGAACTTCGAGCAGTGCACCTGGTTATG  720
Sbjct  661  ............AA............................................  720

Query  721  CACTTTGCATGGAAGGAGAAATGGCCAGATGTCTGATTATATACTGATTCATGGGCTGCA  780
Sbjct  721  .............................................A..............  780

Query  781  GCCAATGGTTTGGCTGGATGGTCAGGGACTTGGAAGAAGCATGATTGGAAAATGTGTGAC  840
Sbjct  781  ......A...........A....A......A.............AA.........A..  840

Query  841  AAAGAAATCTAGGGAAGAAGTATGTGGATGGACCTCTCTGAGAGGTCAAAAACTGTGAAG  900
Sbjct  841  ..........A.................................................  900

Query  901  ATATTTGTATCCCATGTGAGTGCTCACCAATGGGTGACCTCAGCAGAGGGGGATTTTAAC  960
Sbjct  901  .............................................A.A.............  960

Query  961  AATCAAGTGGATAGGAT  977
Sbjct  961  .................  977
```

DNA editing in human *AluY*

```
Genome   1   ATGTAATTTGGACACAAGCATATTCTCTGGTCTGTTGTTCATCTAAGAGTTTTCATTTCA  60
Trace    1   ............................................................  60

Genome  61   GGAAAATTCAAGAGAATAACAGGATCATTTAGGAAAGAATATTGTGTAGTGATAACCATA 120
Trace   61   ............................................................ 120

Genome 121   ATGCTGTTAGATTATTATTATTATCGACAAGCTAAAATAGATGTCACAAATCAAGATTTG 180
Trace  121   ............................................................ 180

Genome 181   CTTAGACAATGTGCCACAGTATAAGAAAACAGGATTTGAGATTGAGAAATATAATTTTTG 240
Trace  181   ............................................................ 240

Genome 241   ACTCAGAATAACTTGCTAGCTACTCAAGAAGTCAGTGTGAACTCAAGGTATTGTCGAGCA 300
Trace  241   ............................................................ 300

Genome 301   GAGATAAAGGTGGGTTGGACCAGGCTTATGGTGCTGTGTCCTTTGCTCTGGGCTGGAGCA 360
Trace  301   ............................................................ 360

Genome 361   GGTGGAGGAAGGTCTCCTTAAGCAGATGGGTTGCTTGGCCTCCAGAAATCCCTCAGGCGG 420
Trace  361   ............................................................ 420

Genome 421   AGCTACCATGGCTGTCAGCCCTTTGTGGCTGTCCTTCTGAGCAGATGGGGCAGGATGGAG 480
Trace  421   ............................................................ 480

Genome 481   TAAACCATCCAGCAGCCCAGACTTCCTCTCTCCTCAGCAACCGCGTCACCTGTGGAATCC 540
Trace  481   ............................................................ 540

Genome 541   TCAGTCTAGGAGCCACCCCGCTTCCCTTCTCCACCGTCAGCCTGTGGAGCGTTCCTGCAG 600
Trace  541   ..........................................AA.A.......... 600

Genome 601   ACTGGGCACATTGAGCATTTATGCCAGTCCGGTTTCTTTTTTCTTTTTTCTTTTTTTTTG 660
Trace  601   ...A........A.............................................A 660

Genome 661   ATGGAGTCTCACTCTGTCACCCAGGCTGGAGTGCAGTGACGCGATCTCGGCTCATTGCAA 720
Trace  661   ..AA.................A...AA.A.A....A....A................. 720

Genome 721   CCTCTTCTCCACCGAGTTCAAGTGATTCTCCTGCCTCAGCCTCCGGAGTAGCTGGGACTA 780
Trace  721   ............A.........A.................CA.......AAA.... 780

Genome 781   CAGGCACCTGCCACCACACCTGGCTATTTTTTTTTTTTTAGTAGAGACAGGGTTACACCA 840
Trace  781   ........................-..................................... 839

Genome 841   TGTTAGCCAGGATGGTCTCGATCTCCTGACCTCGTGATCTGCCCGCCTCGGCCTCCCAAA 900
Trace  840   ............................................................ 899

Genome 901   GTGCTGGGATTATAGGTGTGAGCCACCGTGCGTGG  935
Trace  900   ..................................  934
```

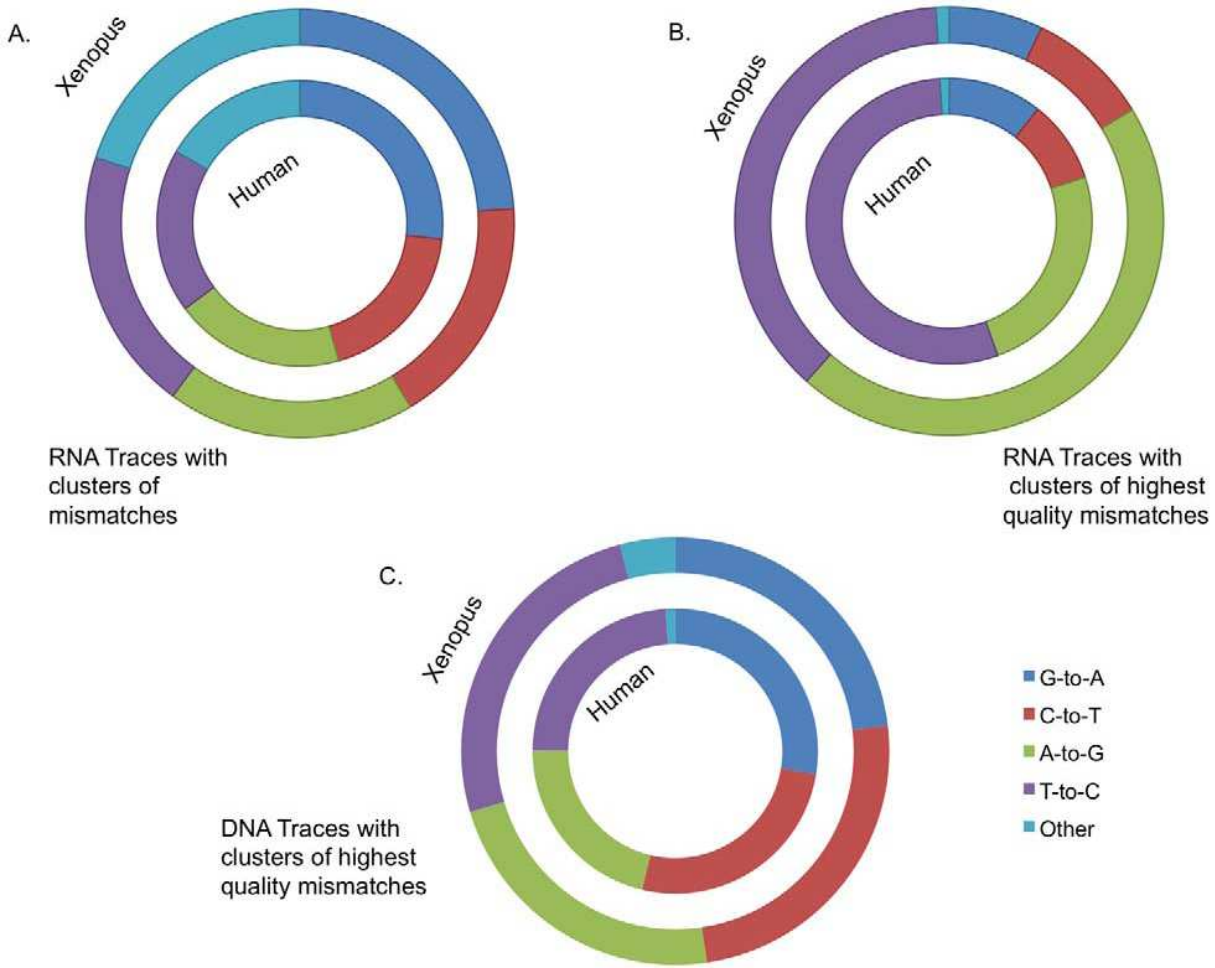
RNA editing

- A fraction of the human, mouse, and *Xenopus tropicalis* traces are derived from RNA.

organism	**human**	**mouse**	**Xenopus tropicalis**
passed traces	250K	513K	454K

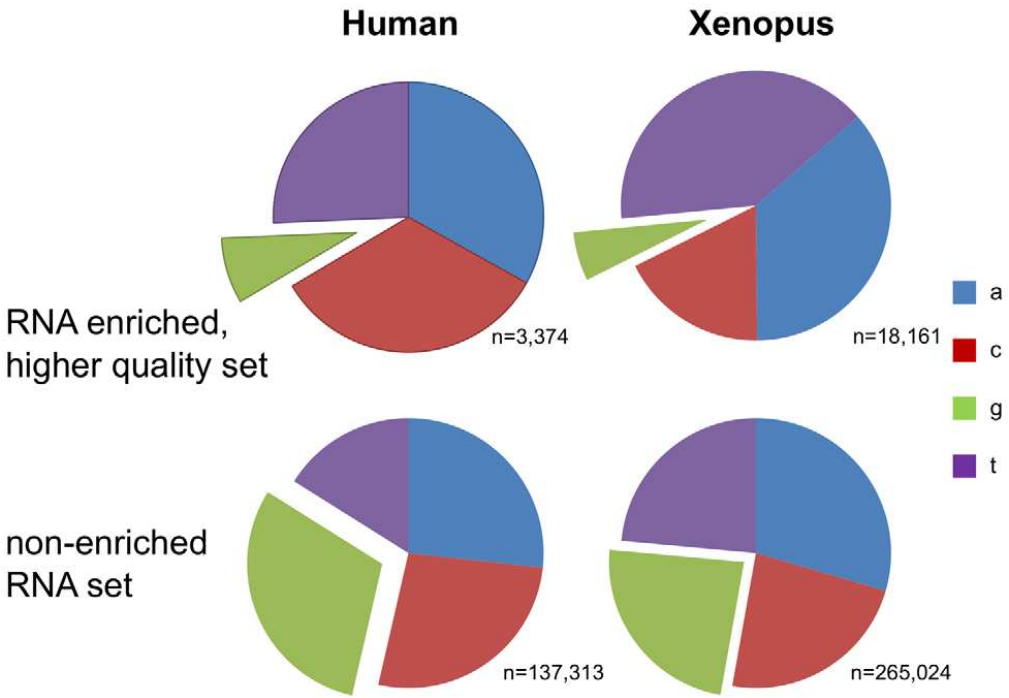
  - ⋆ *passed traces*: number of traces passing the stringent alignment criteria.

Evidence for RNA editing in the cDNA traces



Further evidences

- ADAR signature in the cDNA edited traces.



- 72% of the mismatches in the higher quality set are located in *Alu* repeats;
  - $\Leftrightarrow$ *Alu*/human Genome $\approx 10\%$; p-value: 1.7×10^{-110}.

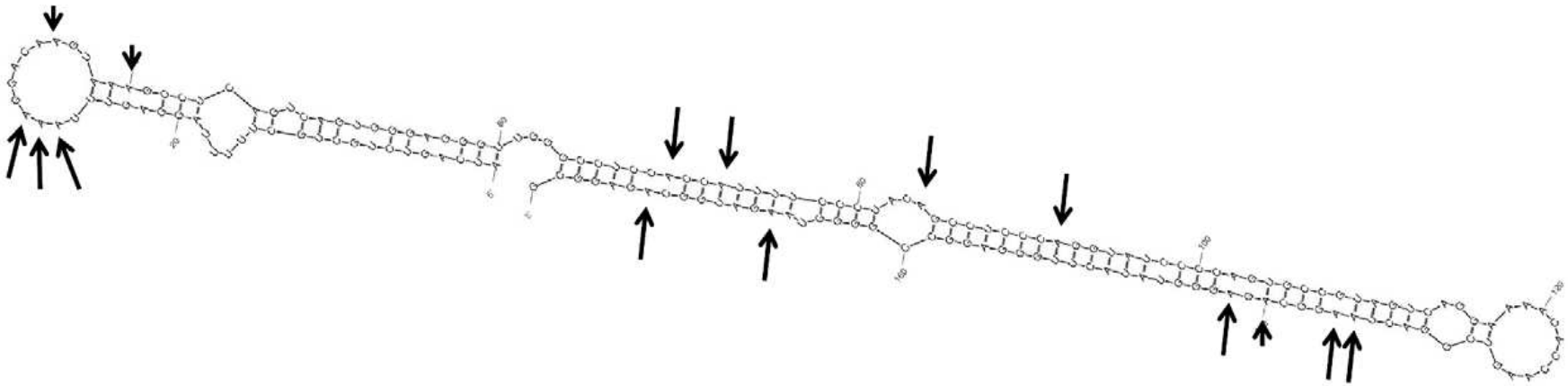
RNA editing in Xenopus tropicalis

```
X.T Genome      ATCAGTCTGCTGCTTTTTAGGAGTTTAAAGGACAAGTAAAGCCTCAGTCAGTGGGAGGGT
1810455972      ........................ggg.....g....g..................... 234
1810477902      .........................g.........g..................... 334
1065483469      ........................ggg..................................... 283
1065490247      ........................ggg..................................... 328
1065466398      ........................ggg..................................... 524
1065471353      .........................g..................................... 535
1669879253      .........................g.....g..................... 464

X.T Genome      TGGGCCTCCACCATTTTCCCTACAGCCTCCCAGGTATCCCCAGTGCCGTAGTCAGGAAAA
1810455972      .........g..g.........g..................................... 294
1810477902      .........g..................................... 394
1065483469      .........g..g.........g.......g..................... 343
1065490247      .........g.............g..................................... 388
1065466398      .........g..g.........g.......g..................... 584
1065471353      .........g.............g..................................... 595
1669879253      .........g..................................... 524

X.T Genome      CACCAAGTCGGACTAAGGCAGAGGGTATACTTGGGAGGCCGGGGTAAGATGGCAGAGGCG
1810455972      ..............gg.....g.....................g......g...... 354
1810477902      ..............gg.....g..... 454
1065483469      ..............gg.....g.....................g......g...... 403
1065490247      ..............gg.....g.....................g......g...... 448
1065466398      ..............gg.....g.....................g......g...... 644
1065471353      ..............gg...g.g.....................g......g...... 655
1669879253      ..............g......g.....................g..... 584
```

RNA editing in Xenopus tropicalis



- Total 18,161 mismatches in the editing enriched, higher quality set;
  - ⋆ 10,001 of them in clusters of ≥ 10 sites.

Discussion & conclusion

Discussion & conclusion

- The NCBI Trace Archive can be used in the search for DNA & RNA editing.
- ⋆ The NCBI Short-Read Archive (SRA) might be considered in the future.
  - The analysis will be much more challenging.

Discussion & conclusion (contd.)

- The availability of computational resources for carrying out the analysis was essential to this paper.
  - 6TB disk space and $>$ 5 node years of CPU time.

⋆ Do with further computational effort to combine:
  - the data in the trace archive
  - the NGS data

  in order to:
  - improve genomic databases
  - eliminate the sequencing errors.

Discussion & conclusion (contd.)

- Using well-calibrated quality scores to investigate editing events.
- ⋆ Using quality scores, many additional genomes can be surveyed for editing.

Discussion & conclusion (contd.)

- *Xenopus tropicalis*:
  - The non-human organism with the largest number of known editing sites so far.

Discussion & conclusion (contd.)

- The actual number of editing sites could be significantly underestimated.
- ⋆ Refine the criteria and perform comprehensive detection of RNA editing.
  - The comparison of editing levels (ratios) in different tissues, disease conditions, etc.

Discussion & conclusion (contd.)

- In this work, evidence for the events of DNA editing was found.
- ⋆ To survey how leakage of DNA editing events, outside retroelements or immunoglobulins, could cause many simultaneous mutations in the genome ($\longrightarrow$ eventually lead to cancer).

Thank you.