

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

G. Ramaswami, W. Lin, R. Piskol, M. H. Tan, C. Davis, and J. B. Li

Nature Methods **9** (2012) 579–581.

Speaker: Joseph Chuang-Chieh Lin

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

4 July 2012



Outline

- 1 Introduction
 - The contribution of the paper
- 2 Materials & methods
- 3 Results
- 4 Discussions



Introduction

- RNA editing: a post-transcriptional modification of RNA molecules.
- Two types of RNA editing (catalyzed by **deaminases**) are known:
 - Cytosine-to-uracil (C-to-U) editing.
 - Catalyzed by **APOBEC1**.
 - Rare and specific to small intestine enterocytes.
 - Adenosine-to-inosine (A-to-I; A-to-G) editing.
 - Catalyzed by the adenosine deaminases acting on RNA (**ADARs**).
 - Pervasive in *Alu* repeats [Levanon *et al.*, *Nature Biotechnology* 2004].



Introduction

- RNA editing: a post-transcriptional modification of RNA molecules.
- Two types of RNA editing (catalyzed by **deaminases**) are known:
 - Cytosine-to-uracil (C-to-U) editing.
 - Catalyzed by **APOBEC1**.
 - Rare and specific to small intestine enterocytes.
 - Adenosine-to-inosine (A-to-I; A-to-G) editing.
 - Catalyzed by the adenosine deaminases acting on RNA (**ADARs**).
 - Pervasive in *Alu* repeats [Levanon *et al.*, *Nature Biotechnology* 2004].



Introduction

- RNA editing: a post-transcriptional modification of RNA molecules.
- Two types of RNA editing (catalyzed by **deaminases**) are known:
 - Cytosine-to-uracil (C-to-U) editing.
 - Catalyzed by **APOBEC1**.
 - Rare and specific to small intestine enterocytes.
 - Adenosine-to-inosine (A-to-I; A-to-G) editing.
 - Catalyzed by the adenosine deaminases acting on RNA (**ADARs**).
 - Pervasive in *Alu* repeats [Levanon *et al.*, *Nature Biotechnology* 2004].



Introduction (contd.)

- Identification of human RNA editing events *outside* of the *Alu* repeats has been challenging [Silberberg & Ohman, *Current Opinion in Genetics & Development* 2011].
 - Hundreds sites in non-*Alu* regions were found using high-throughput sequencing (next-generation sequencing; NGS) data [Li *et al.*, *Science* 2009].



Introduction (contd.)

- Computational approaches to identify human RNA editing sites of all $4 \times 3 = 12$ possible types by comparing gDNA and RNA sequencing data from the same individuals:
 - ▷ Li *et al.*, *Science* 2011.
 - ▷ Ju *et al.*, *Nature Genetics* 2011.
 - ▷ Bahn *et al.*, *Genome Research* 2012.
 - ▷ Peng *et al.*, *Nature Biotechnology* 2012.
 - ▷ ...



Introduction (contd.)

- Subsequent analyses suggest that many of the identified are likely **false positives** (particularly in non-*Alu* regions).
 - ▷ Schrider, Gout & Hahn, *PLoS One* 2011.
 - ▷ Lin *et al.*, *Science* 2012.
 - ▷ Kleinman & Majewski, *Science* 2012.
 - ▷ Pickrell, Gilad & Pritchard, *Science* 2012.
- Major challenges:
 - Sequencing error;
 - Mapping error.



Introduction (contd.)

- Subsequent analyses suggest that many of the identified are likely **false positives** (particularly in non-*Alu* regions).
 - ▷ Schrider, Gout & Hahn, *PLoS One* 2011.
 - ▷ Lin *et al.*, *Science* 2012.
 - ▷ Kleinman & Majewski, *Science* 2012.
 - ▷ Pickrell, Gilad & Pritchard, *Science* 2012.
- Major challenges:
 - Sequencing error;
 - Mapping error.



Contribution of this paper

- A computational framework to robustly identify RNA editing sites.
 - Using transcriptome and genome deep-sequencing data from the same individual.
- Editing of non-*Alu* sites appears to be **dependent** on *nearby* edited *Alu* sites.
 - Possibly because of the locally formed dsRNA structure.
- The **first** systematic examination of **repetitive non-*Alu*** editing sites in humans.
- The advantage over others:
 - Even more striking in **nonrepetitive** regions where identification of editing sites is more challenging.



Contribution of this paper

- A computational framework to robustly identify RNA editing sites.
 - Using transcriptome and genome deep-sequencing data from the same individual.
- Editing of non-*Alu* sites appears to be **dependent** on *nearby* edited *Alu* sites.
 - Possibly because of the locally formed dsRNA structure.
- The **first** systematic examination of **repetitive non-*Alu*** editing sites in humans.
- The advantage over others:
 - Even more striking in **nonrepetitive** regions where identification of editing sites is more challenging.



Contribution of this paper

- A computational framework to robustly identify RNA editing sites.
 - Using transcriptome and genome deep-sequencing data from the same individual.
- Editing of non-*Alu* sites appears to be **dependent** on *nearby* edited *Alu* sites.
 - Possibly because of the locally formed dsRNA structure.
- The **first** systematic examination of **repetitive non-*Alu*** editing sites in humans.
- The advantage over others:
 - Even more striking in **nonrepetitive** regions where identification of editing sites is more challenging.



Contribution of this paper

- A computational framework to robustly identify RNA editing sites.
 - Using transcriptome and genome deep-sequencing data from the same individual.
- Editing of non-*Alu* sites appears to be **dependent** on *nearby* edited *Alu* sites.
 - Possibly because of the locally formed dsRNA structure.
- The **first** systematic examination of **repetitive non-*Alu*** editing sites in humans.
- The advantage over others:
 - Even more striking in **nonrepetitive** regions where identification of editing sites is more challenging.



Characteristics

- The characteristic features distinguishing the authors' approaches:
 - The choice of the short-read mapper: **Burrow-Wheeler Aligner (BWA)**.
 - High specificity & speed for RNA-seq reads
 - Mapping across splicing junctions.
 - Careful parameter-tuning.
 - Incorporation of several filtering steps tailored for *Alu* & **non-Alu** regions.



Materials

- GM12878
 - A lymphoblastoid cell line.
 - Deeply sequenced genome and RNA.
 - poly(A)⁺ RNA-seq data from whole-cell GM12878 (ENCODE).
 - Strand-specific RNA-seq libraries were made.
 - The transcriptome was deeply sequenced (paired-end) with Illumina HiSeq in two biological replicates.
- Han Chinese individual (YH)
 - A lymphoblastoid cell line.
 - Used by Peng *et al.*, *Nature Biotechnology* 2012.
 - An unstranded poly(A)⁺ library & a strand-specific poly(A)⁻ library.



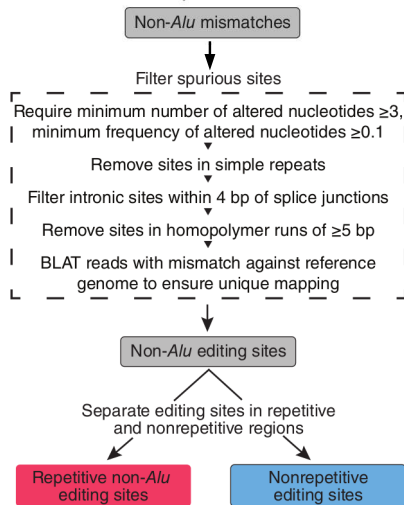
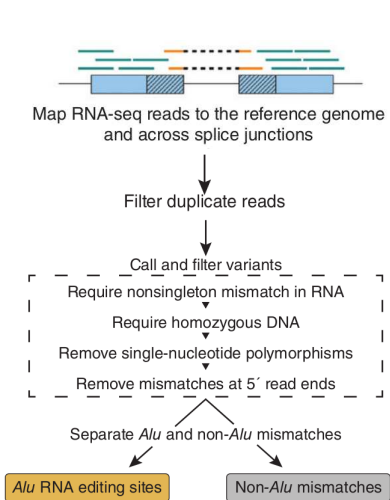
Materials (contd.)

	GM12878 Replicate 1	GM12878 Replicate 2	GM12878 Combined
# reads	235.7 million	263.7 million	499.4 million
Read length	76 bp	76 bp	76 bp

	YH poly(A) ⁺	YH poly(A) ⁻	YH Combined
# reads	323.6 million	843.6 million	1167.2 million
Read length	75, 100 bp	90 bp	75, 90, 100 bp



The pipeline



Validation of the sites

- PCR & Sanger sequencing to validate only a small subset of candidate sites.
 - 7 non-A-to-G sites were validated (all fail).
 - 12 A-to-G sites (with $> 10\%$) were validated (11 successfully validated).



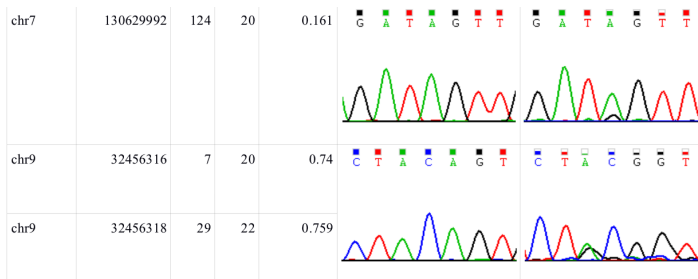
Validation of the sites (contd.)

Validation of the A-to-G sites.

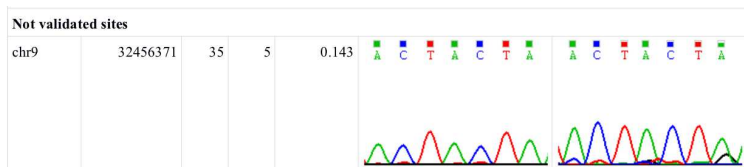
Chr	Position	#A	#G	Editing frequency	gDNA trace	cDNA trace
Validated sites						
chr1	214529740	63	41	0.651		
chr1	214529774	61	12	0.197		
chr7	130629624	122	89	0.730		



Validation of the sites (contd.)

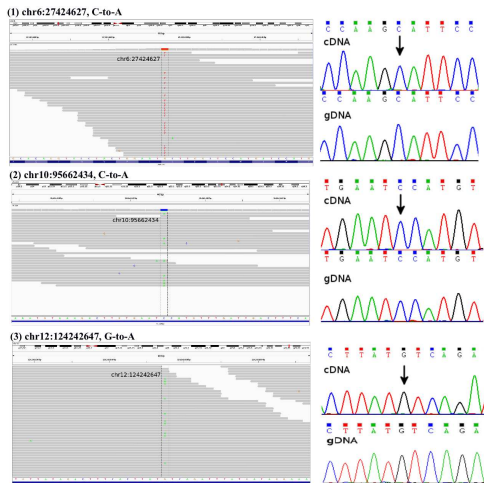


Validation of the sites (contd.)



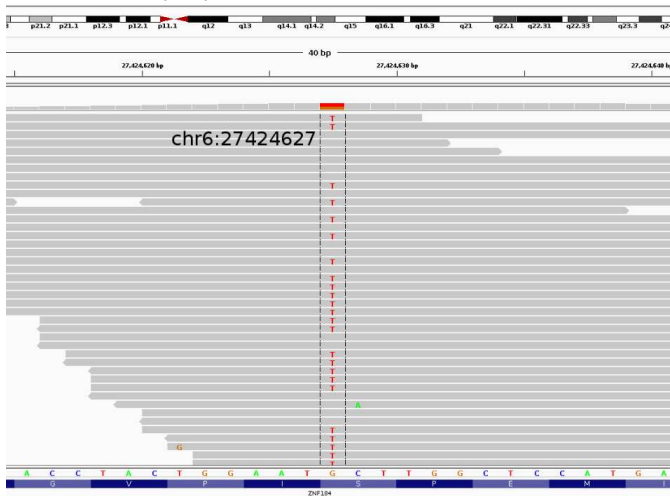
Validation of the sites (contd.)

Validation of the non-A-to-G sites



Validation of the sites (contd.)

Integrative Genomics Viewer (IGV).



Application to the Han Chinese (YH) data

- **Goal:** Evaluate the performance of the authors' method.
- Apply the pipeline to the YH genome & RNA-seq data obtained from Peng *et al.* [*Nature Biotechnology* 2012.]
- The RNA-seq data consists of:
 - an unstranded poly(A)⁺ library, and
 - a strand-specific poly(A)⁻ library.
- Consider three subsets of the data:
 - i. poly(A)⁺ reads only;
 - ii. poly(A)⁻ reads only;
 - iii. poly(A)⁺ reads combined with poly(A)⁻ reads.
- The DNA-RNA mismatch type:
 - Case ii: determined by the strandness of the edited reads;
 - Case i & iii: determined based on RefSeq, UCSC Genes, & Gencode annotations.



Statistical analysis

- **Goal:** Evaluate the significance of the overlap between *Alu* and non-*Alu* A-to-G site containing genes.
- Using the cumulative probability of the **hypergeometric distribution**:

$$\sum_{i=k}^m \frac{\binom{m}{i} \binom{N-m}{n-i}}{\binom{N}{n}}.$$

- ★ *N*: the total number of loci;
 - ★ *n*: the number of genes with *Alu* A-to-G sites;
 - ★ *m*: the number of genes with non-*Alu* A-to-G sites;
 - ★ *k*: the number of genes with both *Alu* and non-*Alu* A-to-G sites.
- Apply the *one-tailed Mann-Whitney test*.



Statistical analysis

- **Goal:** Evaluate the significance of the overlap between *Alu* and non-*Alu* A-to-G site containing genes.
- Using the cumulative probability of the **hypergeometric distribution**:

$$\sum_{i=k}^m \frac{\binom{m}{i} \binom{N-m}{n-i}}{\binom{N}{n}}. \quad ??$$

- ★ *N*: the total number of loci;
 - ★ *n*: the number of genes with *Alu* A-to-G sites;
 - ★ *m*: the number of genes with non-*Alu* A-to-G sites;
 - ★ *k*: the number of genes with both *Alu* and non-*Alu* A-to-G sites.
- Apply the *one-tailed Mann-Whitney test*.



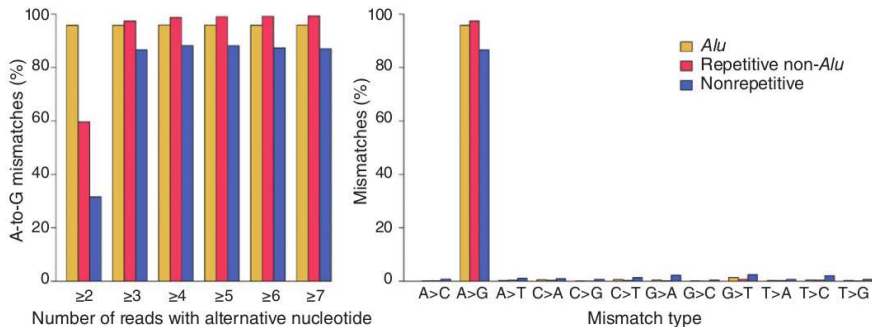
Statistical analysis (contd.)

- **Goal:** Evaluate the significance of the overlap between *Alu* and non-*Alu* A-to-G site containing genes.
- My opinion:

$$\sum_{j=0}^{m-k} \frac{\binom{m-k}{j} \binom{N-k-j}{n-k-j}}{\binom{N}{n}}.$$

- ★ N : the total number of loci;
 - ★ n : the number of genes with *Alu* A-to-G sites;
 - ★ m : the number of genes with non-*Alu* A-to-G sites;
 - ★ k : the number of genes with both *Alu* and non-*Alu* A-to-G sites.
- Apply the *one-tailed* **Mann-Whitney test**.





	GM12878 Replicate 1	GM12878 Replicate 2	GM12878 Combined	YH poly(A)⁺	YH poly(A)⁻	YH Combined³
Number of reads	235.7 million	263.7 million	499.4 million	323.6 million	843.6 million	1167.2 million
Number of filtered reads ¹	202.5 million	239.9 million	457 million	295.3 million	735.6 million	1030.9 million
Number of reads after duplicate removal ²	57.5 million	64.8 million	122.3 million	94.1 million	174.6 million	268.7 million
Read length	76 bp	76 bp	76 bp	75, 100 bp	90 bp	75, 90, 100 bp
Number of RNA editing sites in <i>Alu</i> regions	52,347 (96.2%) ⁴	88,276 (97.3%) ⁴	147,029 (95.8%) ⁴	82,186 (90.9%) ⁴	348,224 (94.2%) ⁴	400,349 (91.0%) ⁴
Number of RNA editing sites in repetitive non- <i>Alu</i> regions	1,086 (96.6%) ⁴	1,190 (98.3%) ⁴	2,385 (97.4%) ⁴	1,486 (88.9%) ⁴	4,037 (85.2%)	4,854 (89.0%) ⁴
Number of RNA editing sites in non-repetitive regions	629 (85.7%) ⁴	827 (86.5%) ⁴	1,451 (86.6%) ⁴	1,078 (74.2%) ⁴	3,102 (77.4%) ⁴	3,774 (78.6%) ⁴

¹ after removal of reads that cannot be uniquely mapped by BWA

² after removal of identical reads that map to exactly the same location

³ reads pooled from poly(A)⁺ and poly(A)⁻ libraries

⁴ numbers in parenthesis: A-to-G percentage



	<i>Alu</i> sites		
	Total	A>G	Percentage A>G
Li <i>et al.</i> ⁶		Not investigated	
Ju <i>et al.</i> ⁷	1,012	806	79.6
Bahn <i>et al.</i> ⁸	3,979	3,589	90.2
Peng <i>et al.</i> (YH) ⁹	19,408	18,919	97.5
This study (GM12878)	147,029	140,825	95.8
This study (YH) ^a	446,670	414,533	92.8 ^b

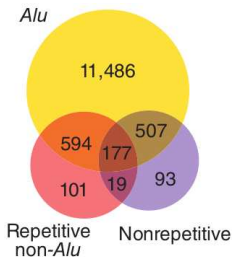


	Repetitive non-Alu sites		
	Total	A>G	Percentage A>G
Li <i>et al.</i> ⁶		Not investigated	
Ju <i>et al.</i> ⁷	163	78	47.9
Bahn <i>et al.</i> ⁸	477	284	59.5
Peng <i>et al.</i> (YH) ⁹	1,544	1,390	90.0
This study (GM12878)	2,385	2,324	97.4
This study (YH) ^a	5,975	5,406	90.5 ^b

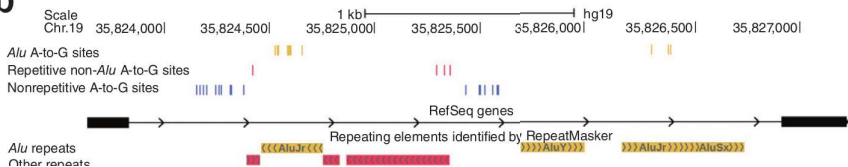


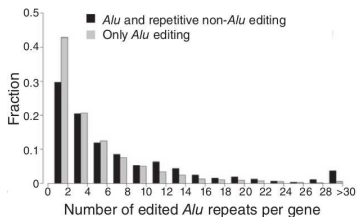
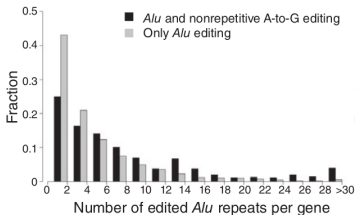
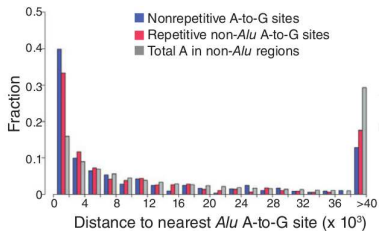
	Nonrepetitive sites		
	Total	A>G	Percentage A>G
Li <i>et al.</i> ⁶	10,210	2,328	22.8
Ju <i>et al.</i> ⁷	646	102	15.8
Bahn <i>et al.</i> ⁸	1,049	268	25.5
Peng <i>et al.</i> (YH) ⁹	1,734	802	46.3
This study (GM12878)	1,451	1,257	86.6
This study (YH) ^a	4,433	3,438	77.6 ^b



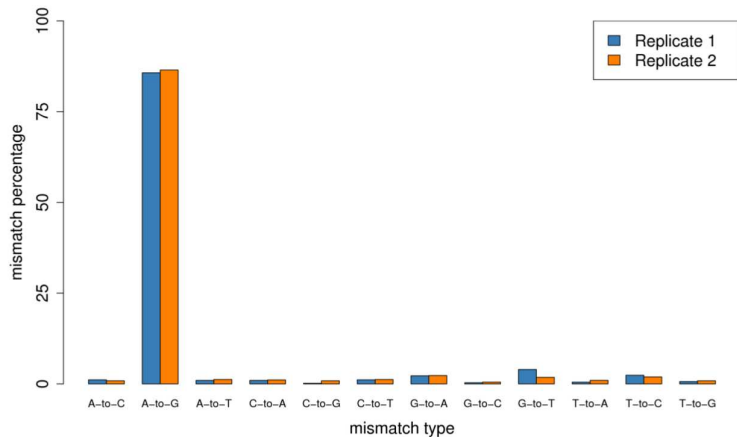


b

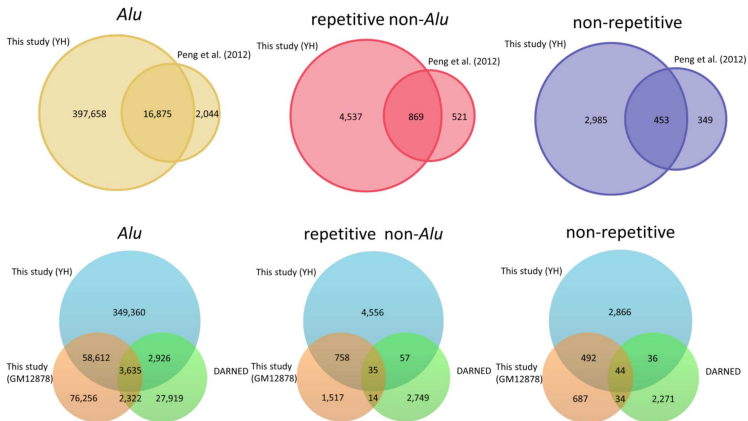




Supplementary Figure 3. Percentage of all 12 mismatch types in non-repetitive sites in each biological replicate of GM12878.



Comparison with the other work



Analysis of the dsRNA structure

- To determine whether sites are located within a dsRNA:
 - Use BLAST to search a region of 100 base pairs up- and downstream of each site against an extended region of 2000 base pairs up- and downstream of the same position [refer to Li *et al.*, *Science* 2009].
 - We considered a site to be within a dsRNA structure if:
 - the second best BLAST hit has an E-value < 0.1 and is located on the reverse strand of the editing site.

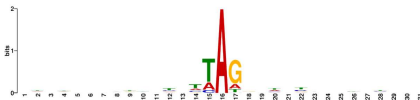


Multiple Em for Motif Elicitation (MEME)

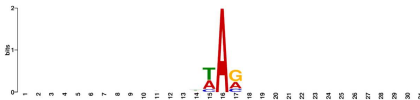
The motif flanking A-to-I RNA editing sites based on 140,825 *Alu* sites. (most likely wrong)



The motif flanking A-to-I RNA editing sites based on 2,324 repetitive non-*Alu* sites.

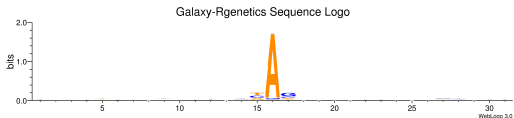


The motif flanking A-to-I RNA editing sites based on 1,257 non-repetitive sites (GM12878)

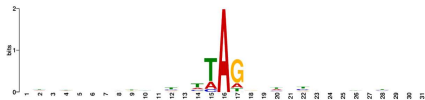


Multiple Em for Motif Elicitation (MEME)

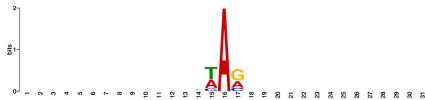
The motif flanking A-to-I RNA editing sites based on 140,825 *Alu* sites.



The motif flanking A-to-I RNA editing sites based on 2,324 repetitive non-*Alu* sites.



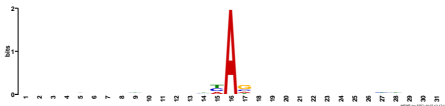
The motif flanking A-to-I RNA editing sites based on 1,257 non-repetitive sites (GM12878)



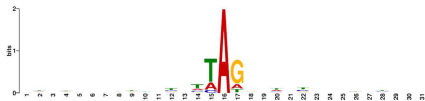
Multiple Em for Motif Elicitation (MEME)

The motif flanking A-to-I RNA editing sites based on 140,825 *Alu* sites.

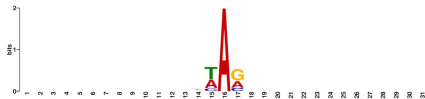
Flanking regions of the **first 3,000**, mid 3,000, and & last 3,000 *Alu* editing sites



The motif flanking A-to-I RNA editing sites based on 2,324 repetitive non-*Alu* sites.

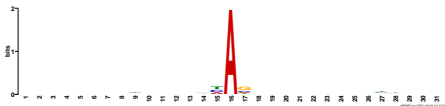


The motif flanking A-to-I RNA editing sites based on 1,257 non-repetitive sites (GM12878)

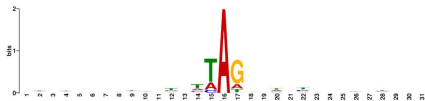


Multiple Em for Motif Elicitation (MEME)

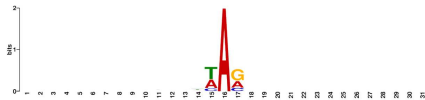
The motif flanking A-to-I RNA editing sites based on 140,825 *Alu* sites.
Flanking regions of the first 3,000, mid 3,000, and & last 3,000 *Alu* editing sites



The motif flanking A-to-I RNA editing sites based on 2,324 repetitive non-*Alu* sites.

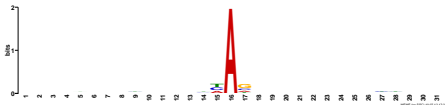


The motif flanking A-to-I RNA editing sites based on 1,257 non-repetitive sites (GM12878)

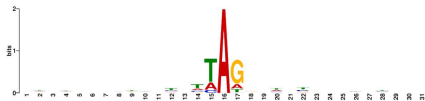


Multiple Em for Motif Elicitation (MEME)

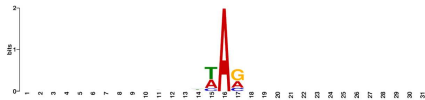
The motif flanking A-to-I RNA editing sites based on 140,825 *Alu* sites.
Flanking regions of the first 3,000, mid 3,000, and & last 3,000 *Alu* editing sites



The motif flanking A-to-I RNA editing sites based on 2,324 repetitive non-*Alu* sites.



The motif flanking A-to-I RNA editing sites based on 1,257 non-repetitive sites (GM12878)



STEME: Accurate efficient motif finding in large data sets

STEME

[home](#) | [search](#) |

[next](#) | [index](#)

Accurate efficient motif finding in large data sets

STEME started life as an approximation to the Expectation-Maximisation algorithm for the type of model used in motif finders such as [MEME](#). STEME's EM approximation runs an order of magnitude more quickly than the MEME implementation for typical parameter settings. STEME has now developed into a fully-fledged motif finder in its own right.

STEME's source code can be found at its [PyPI page](#). The latest version of STEME's documentation is at its [Python package page](#). An installation of STEME is [available](#) to run over the web.

Contents

- [Installation](#)
 - [Prerequisites](#)
 - [Download STEME](#)
 - [Configure, build, install](#)
 - [Darwin/MacOS specific installation instructions](#)
- [Using STEME](#)
 - [Quick start](#)
 - [Multiple motifs](#)
 - [Controlling the running time](#)
 - [Characteristics of the motifs](#)
 - [STEME as an implementation of the EM algorithm](#)
- [STEME Options](#)
 - [Multiple motifs](#)
 - [Output](#)
 - [Background model](#)
 - [Start finding](#)
 - [EM](#)

Table Of Contents

[Accurate efficient motif finding in large data sets](#)
[Contents](#)

Why use STEME?

- Proven motif finding techniques
 - Designed for large data sets
 - Fast
 - Flexible motif models
 - Easy to use
 - Accurate significance calculations
 - Available as a web service
- Publication

Next topic

[Installation](#)


Quick search

Go

Enter search terms or a module, class or function name.



STEME: Accurate efficient motif finding in large data sets



Home [New job](#) [List jobs](#)

Name to identify this analysis:
(optional)

Input sequences:
(FASTA format):

Maximum number of motifs to find:

Order of background Markov model:

Background sequences:
(optional but recommended; FASTA format)

Minimum number of sites:
(use 0 for default settings)

Maximum number of sites:
(use 0 for default settings)

Minimum motif width:

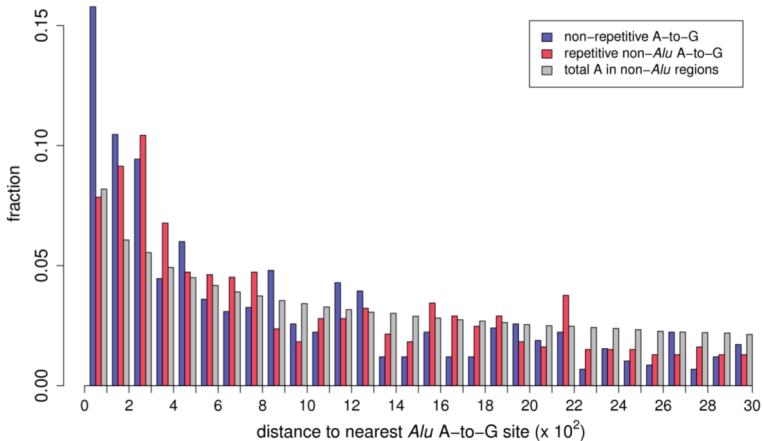
Maximum motif width:



STEME: Accurate efficient motif finding in large data sets

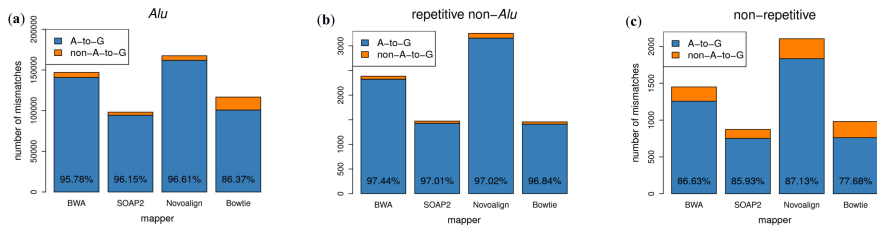
Flanking regions of first 1,000 Alu editing sites





Discussions

BWA, SOAP2, Novoalign, Bowtie?



Thank you.



Appendix



Mapping of RNA-seq reads

- Mapping RNA-seq reads to:
 - the reference genome,
 - the transcriptome,
 - the reference genome + exonic sequences surrounding all currently known splicing junctions from available gene models.
 - Annotations from Gencode, RefSeq, Ensembl, and UCSC Genes.
- The length of the splicing junction regions are set slightly shorter than the RNA-seq reads.
 - For 76 bp reads, a region of 75 bp up- and downstream was chosen.
 - Avoid simultaneous hits to the reference genome & splicing junctions.
 - Note that regions might be extended across multiple exons.



Mapping of RNA-seq reads (contd.)

- Only uniquely mapped reads are considered.
- Use *samtools'* **rmdup** to remove identical reads (due to PCR duplicates) mapping to the same location.
 - The read with the highest mapping quality was retained.



Identification of the candidates

Several major causes leading to erroneous editing sites:

- Undetected genomic variation;
- Misalignment of spliced reads;
- Mismatches introduced by the RNA sequencing protocol
- Local misalignment of reads (especially at read ends);
- Global misalignment of reads (mismapping to duplicated or repetitive regions).



Identification of the candidates (contd.)

- Only the sites at homozygous positions in gDNA of the same individual were inspected.
- How to determine homozygous positions in the gDNA?
 - Use the read mapping data provided by the 1000 Genomes Project.
 - 44x coverage of GM12878.
 - A homozygous site:
 - ★ ≥ 10 reads contained the same base that represented $> 95\%$ of the complete coverage.
 - ★ ≤ 2 alternative bases were present at the same position.



Identification of the candidates (contd.)

- Each variant must be supported by ≥ 2 reads.
 - base quality score: ≥ 25 .
 - mapping quality score: ≥ 20 .
- Remove all known SNPs present in dbSNP (v.135).
 - In theory, this is not necessary because we got gDNA. However, not filtering them results in a drastically lowered A-to-G fraction in the final non-repetitive sites.
- Truncate the first 6 bases of each read.
 - To avoid the effect of random-hexamer priming at the 5' read ends.
- Remove sites with conflicting annotation of editing types.



Identification of the candidates (non-*Alu* regions)

- ≥ 3 variant reads & mismatch frequency ≥ 0.1 .
- Remove sites in simple repeats (by *RepeatMasker* annotation).
- Discard intronic candidates located within 4 bp of all known splicing junctions (according to RefGene, UCSC Genes, Gencode).
- Remove sites in *homopolymers* (≥ 5 bp).
- Remove sites in regions of high similarity to other parts of the genome.
 - “BLAT” all reads that overlap the candidate site and, at the same time, show a mismatch from the reference.
 - For each read, require that:
 - the best hit overlaps the candidate site, and
 - the second-best hit having a score $< 95\%$ of the best hit.



The End

