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]*[#], Erez Y. Levanon[1]*[#]

[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, CT 06520, USA
[5]Sackler School of Medicine, Tel Aviv University, Tel Aviv 69978, Israel
[6]Raymond and Beverly Sackler School of Physics and Astronomy, Tel-Aviv University, Tel Aviv 69978, Israel

* These authors contributed equally to this work
[#] To whom correspondence should be addressed. E-mail:
elieis@post.tau.ac.il (E.E.); Erez.levanon@biu.ac.il (E.Y.L.).

Supplemental Material

Contents

Low overlap between editing sites in previous studies

We analyzed the overlap between seven recently published editing datasets (Ramaswami et al. 2012; Ju et al. 2011; Bahn et al. 2012; Peng et al. 2012; Park et al. 2012; Carmi et al. 2011) and the DARNED database (Kiran and Baranov 2010)) (downloaded from UCSC genome browser, representing sites identifies in older published studies). The six recent sets of editing sites were obtained based on analysis of different RNA-seq datasets and applying various algorithmic approaches for editing detection. Overall, we found very low overlaps between the sets, as shown in Table 1. This low overlap can be explained by the different methods for identification of editing, usage of different tissue samples, or varying coverage levels. For example, we re-analyzed the YH RNA-seq data (Peng et al. 2012) using our own editing-identification scheme in *Alu* elements (see below) and found 818,078 editing sites. The same dataset was previously analyzed using two other methods (Ramaswami et al. 2012; Peng et al. 2012), leading to sets of 414,533 and 21,111 sites, respectively. Comparing sets of editing sites in *Alu*'s obtained using the three approaches on the same sample results in higher overlaps, but the overlap with other datasets is still low (Table 2). These observations suggest that the full scope of the editosome is much larger than that identified in all datasets currently available.

Table 1: Overlaps between different published datasets of editing sites. Each entry presents the number of sites found in the two compared sets, and the percentage of the smaller of the two compared sets that is reproduced in the larger one.

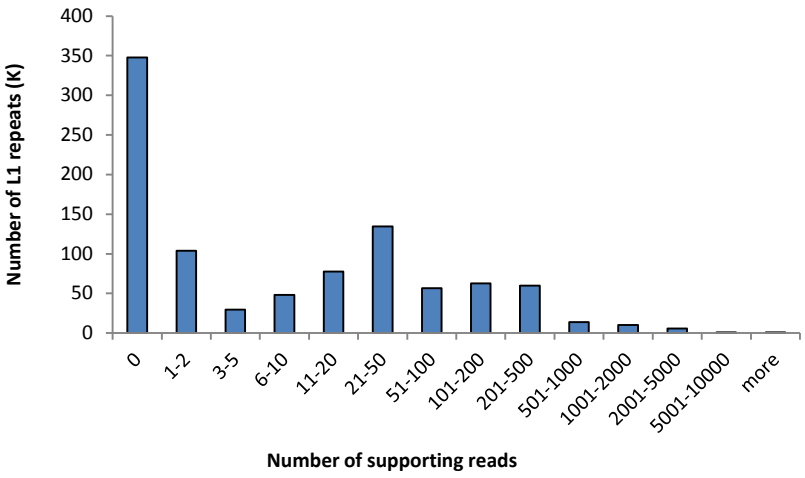
	Ramaswami *et al,* 2012 (Ramaswami et al. 2012) **(YH) 424,755**	**Ju *et al,* 2011**(Ju et al. 2011) **(Supp T12) 1,002**	**Bahn *et al,*** (Bahn et al. 2012) **(Supp T6) 4,141**	**Peng *et al,* 2012**(Peng et al. 2012) **(YH) 21,111**	**Park *et al,* 2012**(Park et al. 2012) **(GM12878) 10,083**	**Carmi *et al,,* 2011**(Carmi et al. 2011) **(Supp 002) 14538**	**DARNED 42,039**
Ramaswami *et al,* 2012 (Ramaswami et al. 2012) **(GM12878) 144,406**	**64244** **44.48%**	**615** **61.38%**	**2200** **53.13%**	**10016** **47.44%**	**2312** **22.93%**	**1042** **7.18%**	**6078** **14.46%**
Ramaswami *et al,* 2012 (Ramaswami et al. 2012) **(YH) 424,755**		**582** **58.08%**	**2211** **53.39%**	**18302** **86.69%**	**2343** **23.24%**	**1326** **9.12%**	**6759** **16.08%**
Ju *et al,* 2011(Ju et al. 2011) **(Supp T12) 1,002**			**110** **10.98%**	**311** **31.04%**	**145** **14.47%**	**7** **0.70%**	**300** **29.94%**
Bahn *et al,* (Bahn et al. 2012) **(Supp T6) 4,141**				**809** **19.54%**	**434** **10.48%**	**87** **2.10%**	**617** **14.90%**
Peng *et al,* 2012(Peng et al. 2012) **(YH) 21,111**					**1202** **11.92%**	**160** **1.10%**	**1399** **6.63%**
Park *et al,* 2012(Park et al. 2012) **(GM12878) 10,083**						**69** **0.68%**	**744** **7.38%**
Carmi *et al,,* 2011(Carmi et al. 2011) **(Supp 002) 14,538**							**870** **5.98%**

Table 2: Overlaps between three lists of editing sites obtained using the same dataset of expression data (YH) by three analysis methods. Each entry presents the number of sites in *Alu*'s found in the two compared sets, and the percentage of the smaller of the two compared sets that is reproduced in the larger one. In addition, a comparison to the results obtained in the present study is shown, as applied to the HBM data.

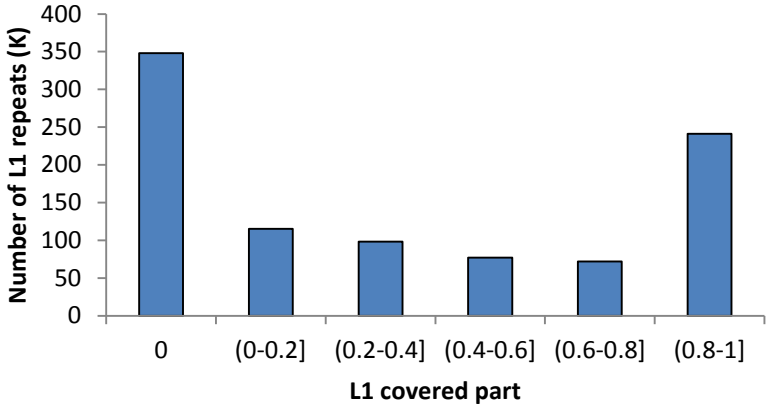
	Peng et al, 2012(Peng et al. 2012) **(YH) 18,919**	**This study YH 818,078**	**This study HBM 993,052**
Ramaswami *et al,* 2012 (Ramaswami et al. 2012) **(YH) 415,890**	**16976 89.73%**	**267959 64.43%**	**109107 26.23%**
Peng et al, 2012(Peng et al. 2012) **(YH) 18,919**		**15286 80.80%**	**10058 53.16%**
This study YH 818,078			**161845 19.78%**

Supplemental Figure 1: (a) Distribution of *L1* repeats according to their read coverage per repeat. (b) Distribution of the *L1* repeats according to the fraction of the length covered by at least a single read.

(a)

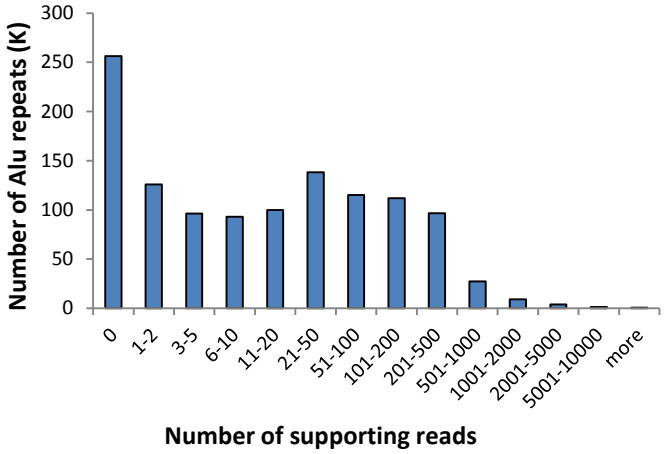


(b)

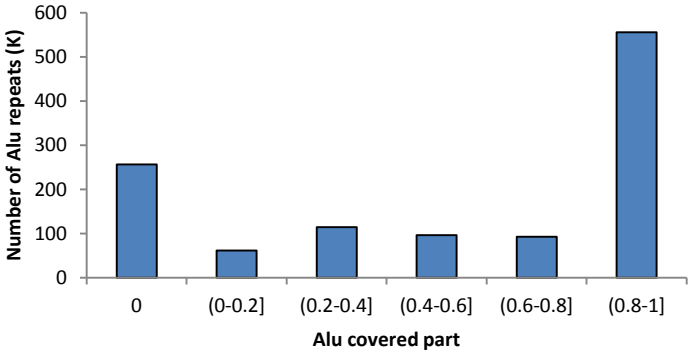


Supplemental Figure 2: (a) Distribution of *Alu* repeats according to their read coverage (number of reads covering, even partially, the *Alu* repeat). Almost 22% of the genomic *Alu*s, ~250,000 of them, are not covered at all, and more than 25% are covered by one to 10 reads. Only a small fraction of the *Alu*s are highly covered (>1000 reads). (b) Distribution of the *Alu* repeats according to the fraction of the *Alu* length that is covered by at least a single read. About 46% of the *Alus* in the genome are fully covered; the rest (32%) are only partially covered.

(a)

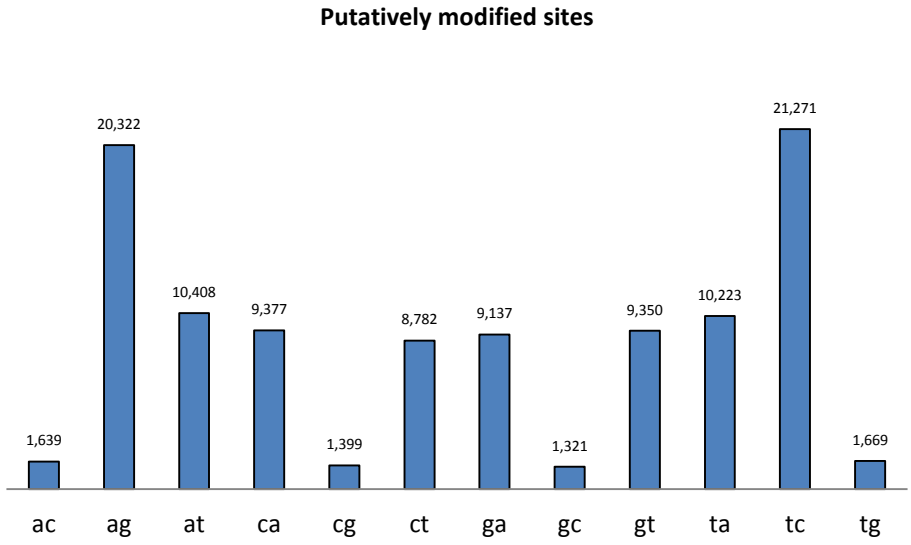


(b)

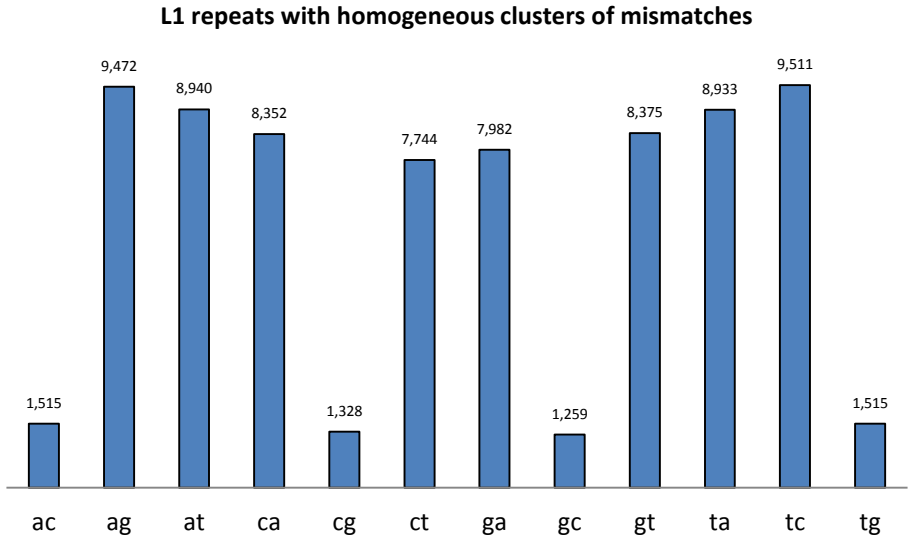


Supplemental Figure 3: (a) Mismatch distribution for the putatively modified *L1* sites. AG and TC mismatches demonstrate minor enrichment, an order of magnitude less than in *Alu*. (b) Distribution of *L1* repeats exhibiting homogeneous clusters of mismatches, by mismatch type (one dominant mismatch type in the *L1* repeat). No dominant type is seen. (c) Distribution of the number of sites in clusters of mismatches for putative editing sites (AG and TC) and control GA and CT sites, showing only tiny fractions of *L1* have clusters of editing.

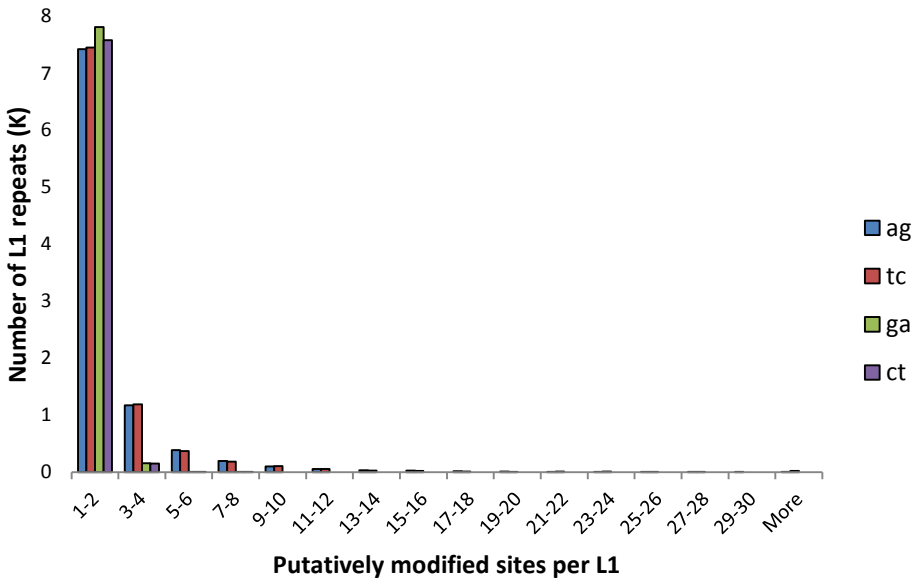
(a)



(b)

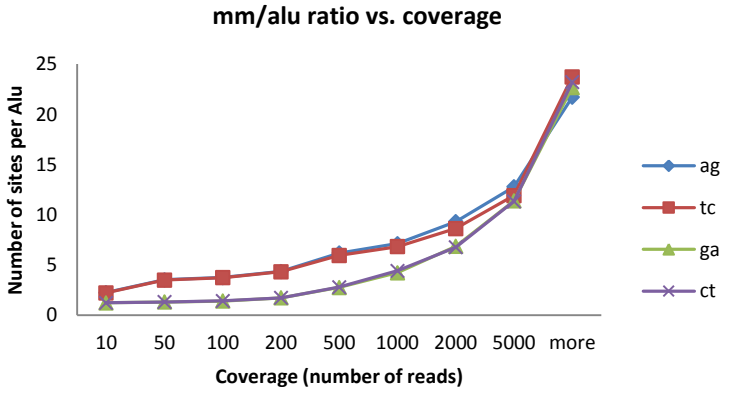


(c)

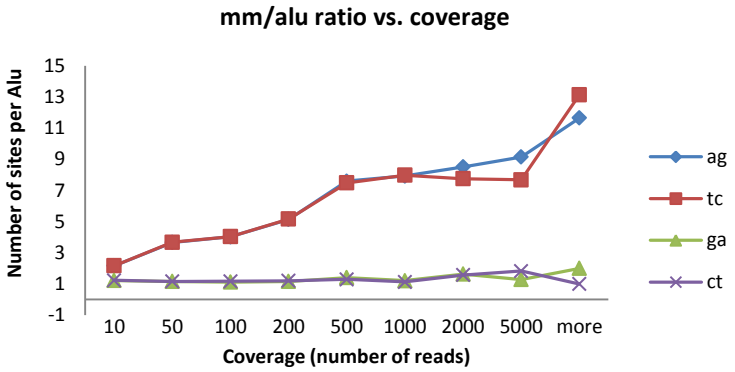


Supplemental Figure 4: Ratio of mismatch sites per *Alu* vs. coverage (#reads per site) in the HBM data. (a) Data before filtering: counts of all mismatch types increase with coverage. (b) Data after filtering and clustering: the higher the coverage, more editing sites (AG and TC) are detected per *Alu*, while no change is observed for GA and CT mismatches (Also presented in the main text, Figure 6a).

(a)



(b)



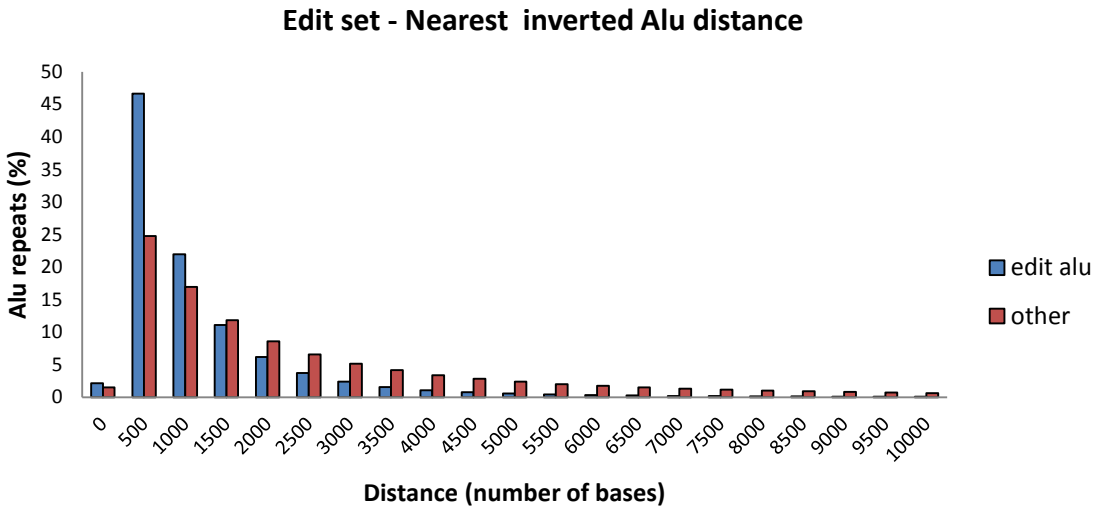
Editable Alu parameters

In order to characterize the features of editable *Alu*s, we used the HBM data to determine which features have the greatest impact on the probability of an *Alu* element being edited. We compared the *Alu*s detected as edited and those not edited, in terms of the following properties: (i) *Alu* length, (ii) *Alu* family type, (iii) similarity to the consensus, (iv) distance to the nearest inverted *Alu* repeat, and the (v) number of such inverted repeats located up to 10kb apart of the *Alu* repeat (data not shown). We found that two criteria best distinguish between *Alu*s detected as edited and others: similarity with the consensus sequence, and distance to the nearest inverted repeat (Supplemental Figure 5). In agreement with previous studies (Levanon 2004; Kim 2004; Athanasiadis 2004; Blow et al. 2004), we conclude that editing is much more prevalent when the distance to such a neighbor is less than 3500bp. In addition, we require the *Alu* to have high similarity with the consensus (SW score >= 1300). Out of the total of 1,175,329 *Alu* repeats in the human genome, 761,244 (64.8%) meet these criteria (editable *Alu*s) and 414,085 do not. Of the editable *Alu*s, 25.5% are detected as edited using the HBM data, compared with only 9.9% of the other *Alu*s.

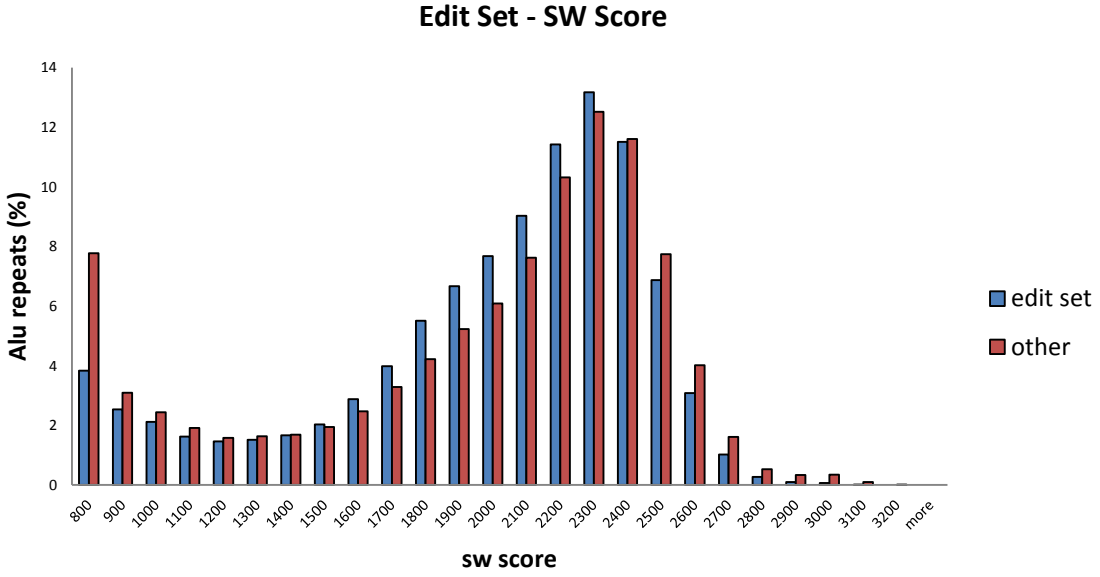
In order to further validate these criteria, we looked at the results obtained for the YH data. Indeed, 21.2% of the editable *Alu*s were detected as edited in the YH data, compared to only 7.8% of other *Alu*s.

Supplemental Figure 5: (a) Distribution of the distance to the nearest inverted *Alu* repeat for *Alu* repeats detected as edited (in HBM), and other *Alu* repeats. (b) Distribution of the divergence from the *Alu* consensus (quantified by the Smith Waterman (SW) score) for *Alu* repeats detected as edited (in HBM), and other *Alu* repeats

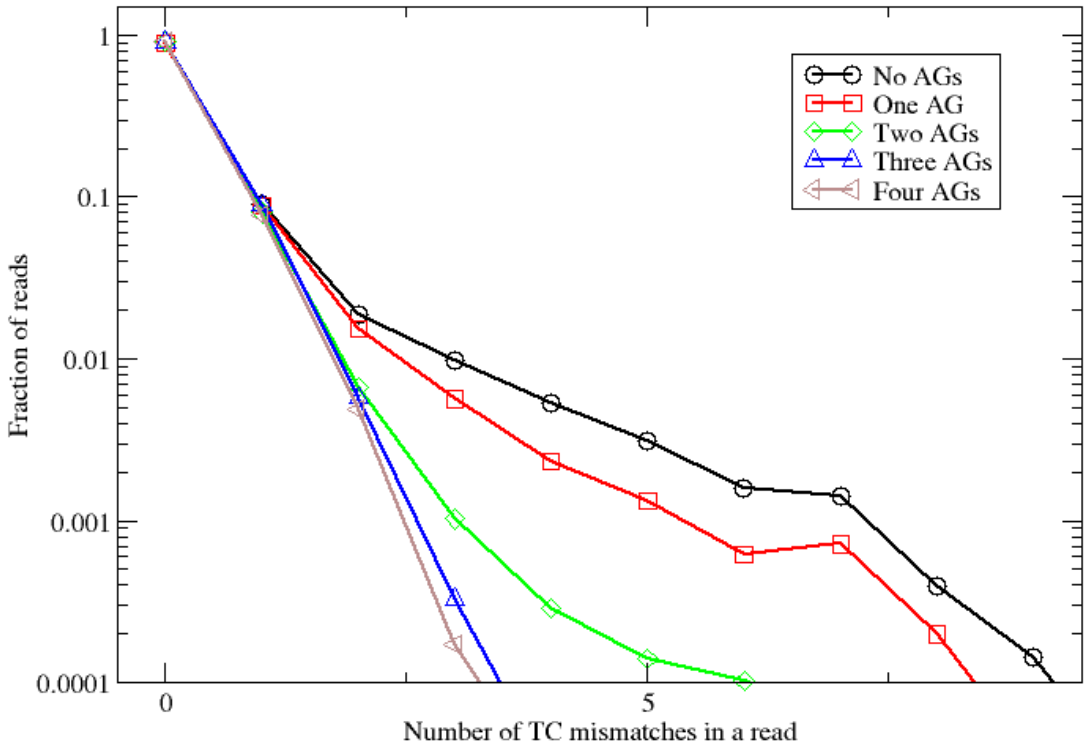
(a)



(b)



Supplemental Figure 6: Distribution of number of TC mismatches per read, stratified by the number of AG mismatches in the same read. Reads with many AGs are likely to have been edited, and to be transcribed from the sense strand. In these reads, the TC mismatches are mostly due to sequencing errors and their number per read follows an exponential distribution (a straight line in the semi-log graph presented). Reads with one (or zero) AGs may have come from both strands, and indeed show clustering of TC mismatches, where the probability of having a large number of TC mismatches decrease slower than exponentially. In other words, as sequencing error rate is approximately 0.1% (for Phred score >30), about 8% of all 75bp contain a sequencing error, including the edited reads. A single AG mismatch is therefore not a very good indication for AG editing. Therefore, one observes only a small difference in the number of TC's between reads with zero-AG and one-AG. However, reads with 2 AGs or (even better) 3 or 4 AGs are almost surely edited, and thus the TCs in these reads result only from sequencing errors. Accordingly, the number of TC's in these reads follows an exponential curve, as expected.



Set of *Alu* predicted to be edited

A group of 52 *Alu*s was chosen. These *Alus* were not found to be edited according to the HBM data, but were expected to be edited based on their features. We looked at *Alu*s in this group that had median As and Ts coverage > 1000 (30 *Alu*s). We used a random read selection to reach each of the presented coverage levels and looked for the edited *Alu*s according to the selected reads. We repeated the random selection 100 times, and found the average predicted percentage among the edited *Alu*s at each coverage level (Supplemental Figure 7).

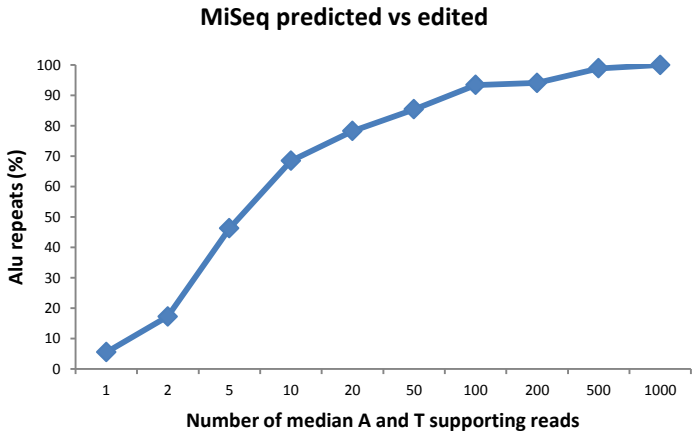
Alu elements within lncRNAs

Seven *Alu* elements within lncRNAs (large non coding RNAs) were selected for deep sequencing. All were found to be edited; five of them were edited on both strands.

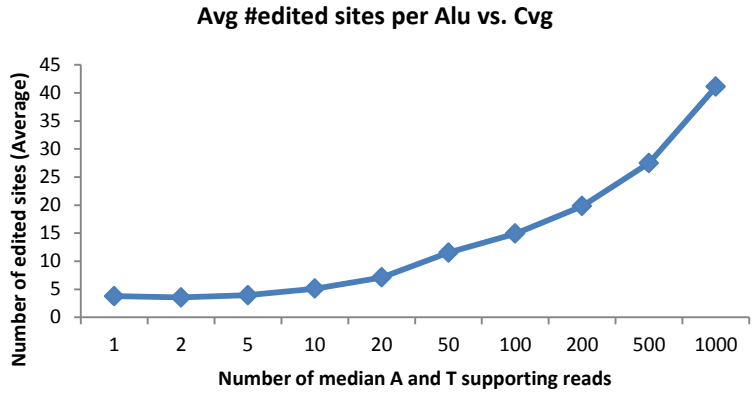
It was shown that lncRNAs can regulate mRNA degradation via creation of dsRNA structure between the *Alu* element in the lncRNA and in the targeted mRNA. Edited *Alu*s within the lncRNA might affect this regulation.(Gong and Maquat 2011)

Supplemental Figure 7: Editing as a function of coverage. (a) Fraction of *Alu* repeats exhibiting putative editing. (b) Number of sites per *Alu* exhibiting putative editing. (c) Number of sites per *Alu* exhibiting putative editing (before clustering). Data are based on random partial sampling of the full MiSeq data.

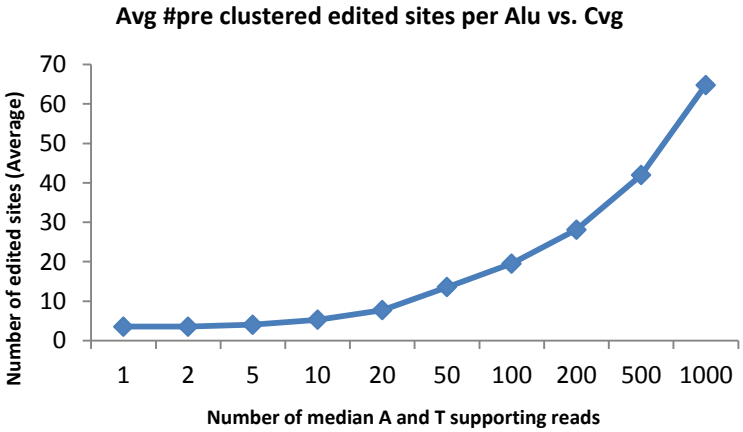
(a)



(b)

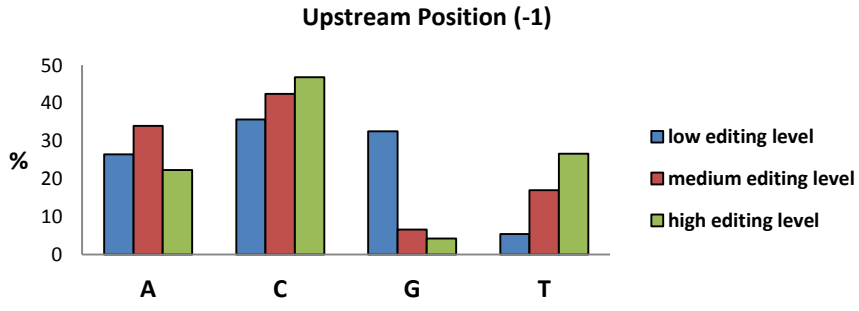


(c)

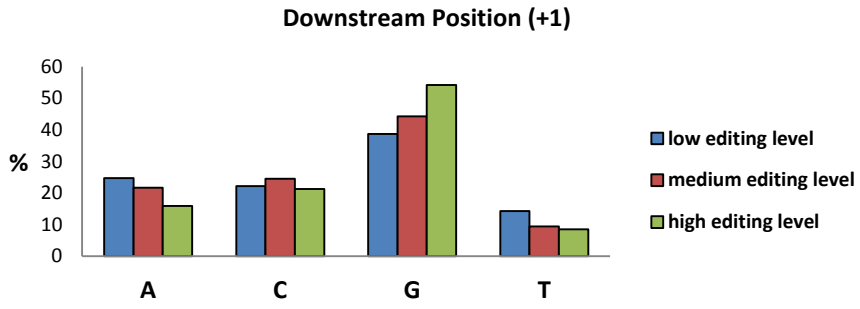


Supplemental Figure 8: Editing site motifs (in MiSeq experiment results). Three editing levels were defined: low (<=10%), medium (>10% and <40%) and high (>=40%). (a) Edit site upstream base distribution; (b) edit site downstream base distribution.

(a)



(b)



Number of unique transcripts and the information entropy per *Alu*

[We describe the process for the case of AG editing; the TC case is handled similarly.]

For a given *Alu*, we have an initial set of editing vectors corresponding to reads covering that *Alu*. The vector corresponding to a single read has N characters, one for each potential editing site within the *Alu*. Characters are either A (no editing), G (editing), or X (no coverage by that read or low-quality data). Vectors corresponding to two pair mates are united into a single vector. Several reads (or read pairs) may result in the same editing vector P, and their count is denoted by mult(P).

Our purpose is to obtain a "minimal" set of editing vectors that are consistent with all of the observed vectors and to estimate the probabilities of transcripts in this set.

As a first step, we detect the editing vectors that are maximal with respect to containment (vector P contains a vector Q if all of the non-X sites of Q are non-X in P as well, and the A/G values in those sites coincide).

Next, starting with the set of maximal editing vectors, several "merging rounds" are performed: In each merging round, a pair of editing vectors, P and Q are merged if S, the set of sites that are non-X in both P and Q, is at least 5 sites in size, P and Q coincide over S, and sites in P or Q outside S have only A's (This is a "conservative" merging criterion). Merging rounds continue until no two editing vectors can be merged, yielding the final set of editing vectors.

Technical note: Merging can be done directly on the initial set of editing patterns with the same results, but passage to maximal editing vectors greatly reduces the number of vectors and improves processing time.

Each of the final editing vectors is attached a multiplicity, as follows:

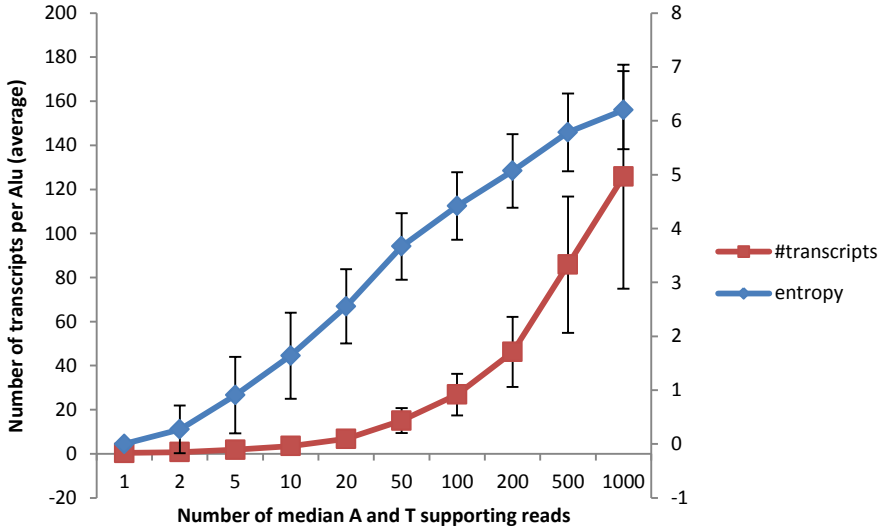
1. For a final editing vector Q, denote by C the set of initial editing vectors that are contained in Q. Then mult(Q) = sum_{P in C} mult(P).
2. For a final editing vector Q, denote by C the set of initial editing vectors that are contained in Q. For each initial vector P, compute n = #{final editing vectors containing P} and set mult'(P) = mult(P) / n. Then mult(Q) = sum_{P in C} mult'(P).
3. For a final editing vector Q, denote by U the set of initial editing vectors that are uniquely contained in Q. Then mult(Q) = sum_{P in U} mult(P).
4. Denoting by F the set of all final editing vectors, probabilities are naturally assigned by: Prob(Q) = mult(Q) / sum_{R in F} mult(R).

The information entropy is then calculated in the standard way:

Entropy = sum_{sites Q} - Prob(Q) ln(Prob(Q)).

Note that this procedure ignores the possibility of some of the unedited reads being transcribed from the antisense strand. Thus, it underestimates the editing probabilities and the information entropy.

Supplemental Figure 9: Entropy and number of different transcript variants per *Alu*, as a function of the reads' coverage. No saturation is observed even for ultra-high coverage.



Supplemental Figure 10: Mismatches in strand specific sequencing (three runs of the YH data). Virtually all mismatches are consistent with AG editing in the expressed strand. Data are presented for the '+' strand; similar results were found for the '-' strand.

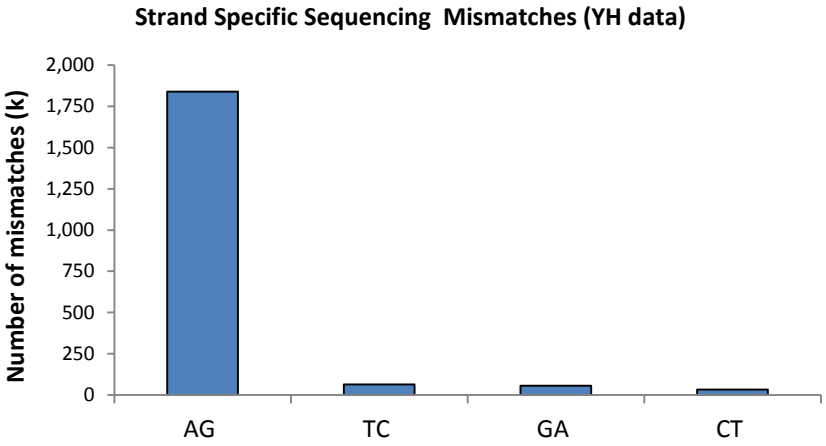


Table 3: (a) HBM alignment data; (b) YH alignment data

method	**tissue**	**# read pairs processed**	**# aligned pair reads**	**% aligned pair reads**	**# reported pair alignments**	**# unique aligned pair reads**	**% unique aligned pair reads**	**# reads overlap *Alu***	**% reads overlap *Alu***	**# Covered *Alu***	**% Covered *Alu***
50_PE_FCA_run_s_1	thyroid	81,912,887	49,285,018	60.17	190,402,044	36,752,476	74.57	3,007,919	4.09	225,069	19.15
50_PE_FCA_run_s_2	testes	81,836,199	51,553,298	63	198,467,008	37,333,936	72.42	2,681,632	3.59	264,025	22.46
50_PE_FCA_run_s_3	ovary	80,946,260	53,269,165	65.81	242,745,428	38,279,942	71.86	3,323,283	4.34	292,071	24.85
50_PE_FCA_run_s_4	WBC	81,217,148	55,631,426	68.5	265,775,458	39,167,872	70.41	2,844,506	3.63	65,733	5.59
50_PE_FCA_run_s_5	skeletal_muscle	82,111,139	57,069,375	69.5	211,758,354	36,655,648	64.23	1,876,867	2.56	42,407	3.61
50_PE_FCA_run_s_6	prostate	82,334,076	58,942,802	71.59	265,756,986	42,599,576	72.27	3,053,531	3.58	130,167	11.07
50_PE_FCA_run_s_7	lymph_node	82,078,157	57,278,417	69.79	258,707,812	39,360,053	68.72	3,580,809	4.55	130,448	11.10
50_PE_FCA_run_s_8	lung	79,296,905	55,936,694	70.54	238,170,510	40,462,422	72.34	2,941,206	3.63	123,484	10.51
50_PE_FCB_run_s_1	adipose	77,300,072	49,222,986	63.68	199,104,664	32,928,443	66.9	2,442,534	3.71	182,284	15.51
50_PE_FCB_run_s_2	adrenal	74,472,871	50,608,465	67.96	209,476,596	37,204,274	73.51	4,851,764	6.52	362,579	30.85
50_PE_FCB_run_s_3	brain	73,513,047	52,343,398	71.2	175,444,776	39,059,434	74.62	2,783,455	3.56	304,049	25.87
50_PE_FCB_run_s_4	breast	75,862,215	50,159,508	66.12	188,925,252	35,276,236	70.33	3,004,876	4.26	241,152	20.52
50_PE_FCB_run_s_5	colon	82,437,443	57,639,169	69.92	253,167,094	36,051,373	62.55	2,180,315	3.02	131,986	11.23
50_PE_FCB_run_s_6	kidney	80,397,337	54,649,335	67.97	211,458,204	35,777,671	65.47	2,806,832	3.92	181,255	15.42
50_PE_FCB_run_s_7	heart	82,918,784	63,045,326	76.03	234,457,374	36,097,564	57.26	1,832,468	2.54	202,888	17.26
50_PE_FCB_run_s_8	liver	80,048,623	51,627,225	64.49	177,671,526	34,985,456	67.77	1,714,555	2.45	102,194	8.69

method	**tissue**	# reads processed	# aligned reads	% aligned reads	# reported alignments	# unique aligned reads	% unique aligned reads	# reads overlap *Alu*	% reads overlap *Alu*	# Covered *Alu*	% Covered *Alu*
75_FCA_run_s_1	adipose	76,269,225	61,836,980	81.08	125,115,094	42,752,482	69.14	1,073,214	2.51	136,755	11.64
75_FCA_run_s_2	adrenal	76,171,569	63,043,764	82.77	134,037,617	46,536,721	73.82	2,293,585	4.93	293,731	24.99
75_FCA_run_s_3	brain	64,313,204	55,761,432	86.7	107,410,622	40,248,622	72.18	1,025,150	2.55	213,129	18.13
75_FCA_run_s_4	breast	77,195,260	63,751,662	82.58	127,062,025	45,185,850	70.88	1,379,639	3.05	187,092	15.92
75_FCA_run_s_5	colon	80,257,757	67,825,362	84.51	152,597,386	43,604,159	64.29	934,793	2.14	100,327	8.54
75_FCA_run_s_6	kidney	79,772,393	66,587,009	83.47	138,543,901	44,130,602	66.28	1,379,224	3.13	144,183	12.27

75_FCA_run_s_7	heart	76,766,862	66,323,955	86.4	129,582,512	39,525,044	59.59	786,655	1.99	156,056	13.28
75_FCA_run_s_8	liver	77,453,877	63,660,052	82.19	115,122,444	43,618,616	68.52	722,867	1.66	81,189	6.91
75_FCB_run_s_1	lung	81,255,438	66,451,362	81.78	138,557,394	48,850,290	73.51	1,431,640	2.93	98,168	8.35
75_FCB_run_s_2	lymph_node	81,916,460	65,911,770	80.46	149,327,129	45,492,345	69.02	1,693,173	3.72	103,878	8.84
75_FCB_run_s_3	prostate	83,319,902	69,209,226	83.06	151,149,309	50,919,730	73.57	1,484,525	2.92	102,742	8.74
75_FCB_run_s_4	skeletal_muscle	82,864,636	67,451,995	81.4	124,307,888	45,546,002	67.52	922,735	2.03	32,721	2.78
75_FCB_run_s_5	WBC	82,785,673	67,946,971	82.08	157,763,121	48,745,031	71.74	1,360,638	2.79	50,879	4.33
75_FCB_run_s_6	ovary	81,003,052	67,076,854	82.81	150,443,332	48,637,655	72.51	1,535,814	3.16	226,919	19.31
75_FCB_run_s_7	testes	82,044,319	66,446,467	80.99	128,051,731	49,366,312	74.29	1,233,965	2.50	198,515	16.89
75_FCB_run_s_8	thyroid	80,246,657	64,495,919	80.37	123,941,179	48,930,567	75.87	1,423,886	2.91	176,456	15.01
100_FCA_run_s_1	16 Tissues mixture poly-A selected	76,447,153	39,090,660	51.13	77,244,345	27,674,074	70.79	655,329	2.37	161,449	13.74
100_FCA_run_s_2	16 Tissues mixture poly-A selected	78,243,019	39,963,858	51.08	78,833,379	28,326,641	70.88	672,294	2.37	164,697	14.01
100_FCA_run_s_3	16 Tissues mixture poly-A selected	77,229,855	38,733,496	50.15	76,423,922	27,450,027	70.87	649,899	2.37	161,450	13.74
100_FCA_run_s_4	16 Tissues mixture poly-A selected + Normalization	76,274,508	39,811,551	52.2	64,901,830	33,215,772	83.43	616,783	1.86	150,593	12.81
100_FCA_run_s_5	16 Tissues mixture poly-A selected + Normalization	75,929,029	39,253,435	51.7	64,055,853	32,751,559	83.44	603,805	1.84	148,336	12.62
100_FCA_run_s_6	16 Tissues mixture poly-A selected + Normalization	74,756,517	39,203,933	52.44	64,094,790	32,680,774	83.36	602,718	1.84	148,485	12.63
100_FCA_run_s_7	16 Tissues mixture, NEW RiboFree Method	73,420,952	23,149,296	31.53	47,197,070	18,512,162	79.97	543,283	2.93	214,933	18.29
100_FCA_run_s_8	16 Tissues mixture, NEW RiboFree Method	73,520,276	22,705,933	30.88	45,971,302	18,206,044	80.18	538,974	2.96	213,837	18.19

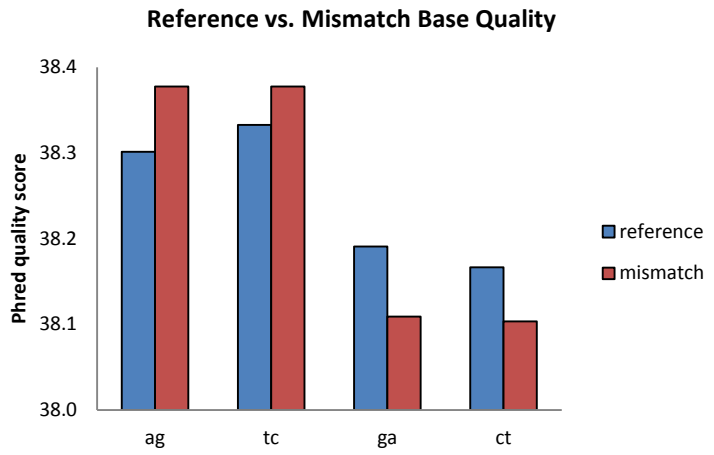
100_FCB_run_s_1	16 Tissues mixture poly-A selected	77,258,890	39,528,201	51.16	78,026,811	28,006,166	70.85	661,503	2.36	163,003	13.87
100_FCB_run_s_2	16 Tissues mixture poly-A selected	75,982,104	38,962,880	51.28	77,411,162	27,503,496	70.59	647,796	2.36	160,655	13.67
100_FCB_run_s_3	16 Tissues mixture poly-A selected + Normalization	74,249,497	39,657,474	53.41	65,467,297	32,938,006	83.06	603,100	1.83	148,725	12.65
100_FCB_run_s_4	16 Tissues mixture poly-A selected + Normalization	73,773,895	40,662,263	55.12	67,522,186	33,700,388	82.88	614,738	1.82	150,371	12.79
100_FCB_run_s_5	16 Tissues mixture, NEW RiboFree Method	72,451,624	24,003,025	33.13	49,271,309	19,156,672	79.81	555,354	2.90	217,809	18.53
100_FCB_run_s_6	16 Tissues mixture, NEW RiboFree Method	70,743,680	24,249,578	34.28	49,725,861	19,357,833	79.83	559,815	2.89	219,276	18.66
100_FCB_run_s_7	16 Tissues mixture, NEW RiboFree Method	71,937,539	23,252,147	32.32	47,052,544	18,648,119	80.2	545,932	2.93	215,602	18.34
100_FCB_run_s_8	16 Tissues mixture, NEW RiboFree Method	72,321,018	22,567,373	31.2	45,540,057	18,122,929	80.31	529,881	2.92	211,369	17.98

(b)

method	**tissue**	**# read pairs processed**	**# aligned pair reads**	**% aligned pair reads**	**# reported pair alignments**	**# unique aligned pair reads**	**% unique aligned pair reads**	**# reads overlap *Alu***	**% reads overlap *Alu***	**# Covered *Alu***	**% Covered *Alu***
100_PE_YH_s_1	lymphoblastoid cell line	13,786,194	8,098,389	58.74	29,811,078	6,009,705	74.21	794,098	6.61	169,973	14.46
100_PE_YH_s_2	lymphoblastoid cell line	17,932,981	12,562,508	70.05	49,598,960	8,986,916	71.54	914,607	5.09	154,707	13.16
75_PE_YH_s_1	lymphoblastoid cell line	18,498,078	9,472,569	51.21	31,307,768	7,849,921	82.87	652,313	4.15	179,184	15.25
75_PE_YH_s_2	lymphoblastoid cell line	17,775,252	10,036,929	56.47	20,581,183	7,193,552	71.67	1,336,820	9.29	277,846	23.64

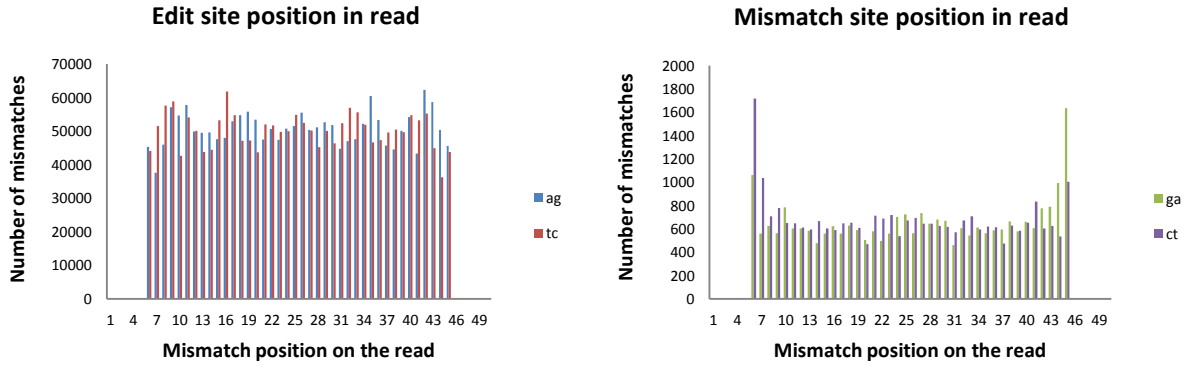
75_PE_YH_s_3	lymphoblastoid cell line	25,618,824	18,559,599	72.45	37,882,133	12,782,323	68.87	2,045,916	8.00	279,233	23.76
75_PE_YH_s_4	lymphoblastoid cell line	25,503,829	16,730,511	65.60	33,442,424	11,574,075	69.18	2,219,134	9.59	319,071	27.15
75_PE_YH_s_5	lymphoblastoid cell line	20,682,027	12,893,039	62.34	19,236,795	11,081,934	85.95	912,597	4.12	197,684	16.82
75_PE_YH_s_6	lymphoblastoid cell line	22,006,296	12,843,982	58.37	18,165,046	11,321,839	88.15	888,263	3.92	206,648	17.58
90_PE_YH_s_1	lymphoblastoid cell line	240,038,801	175,626,280	73.17	478,301,140	150,680,655	85.80	25,711,200	8.53	509,304	43.33
90_PE_YH_s_2	lymphoblastoid cell line	76,468,296	56,160,845	73.44	152,174,940	48,191,588	85.81	8,313,490	8.63	438,580	37.32
90_PE_YH_s_3	lymphoblastoid cell line	105,329,452	73,934,099	70.19	217,563,422	59,434,534	80.39	12,561,426	10.57	492,583	41.91

Supplemental Figure 11: Average Phred quality score at mismatch positions, for read-bp exhibiting the reference genome nucleotide and those exhibiting a mismatch (HBM data). AG/TC mismatch bases have, on average, a higher quality score than the corresponding reference reads. The inverse is true for GA/CT mismatch sites.



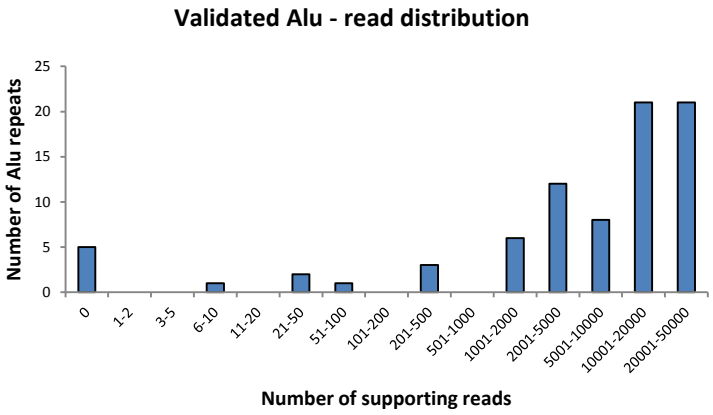
Supplemental Figure 12: Mismatch distribution along the reads (HBM data) (same as Figure 8 in the main text, here for the 50bp reads length data) . (a) 50bp paired-end reads, AG and TC, mismatches; (b) 50bp paired-end reads, GA and CT, mismatches.

(a) (b)

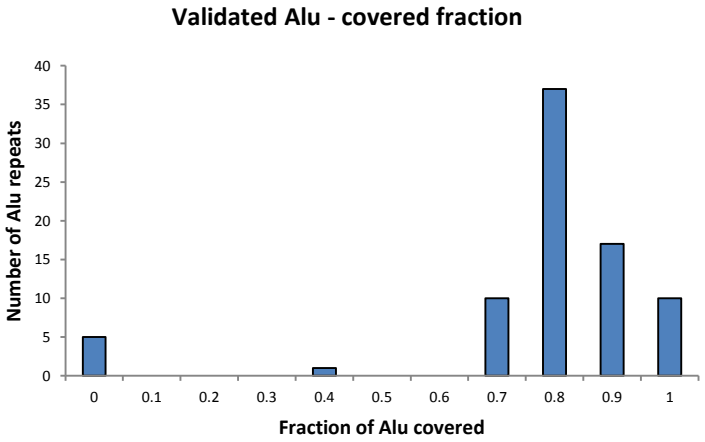


Supplemental Figure 13: (a) Read coverage of the 80 selected *Alu*s. (b) Fraction of the *Alu* that was covered. Most targeted *Alu*s were covered at over 50%, as planned. (MiSeq experiment results).

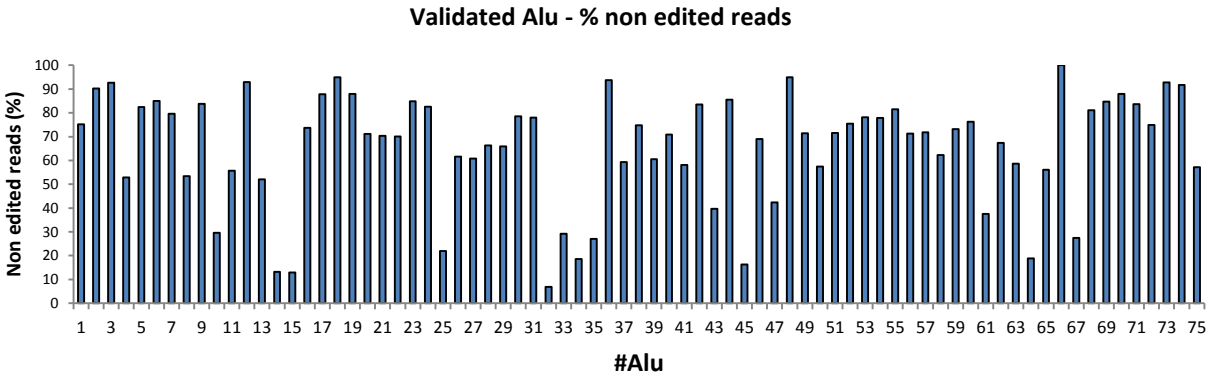
(a)



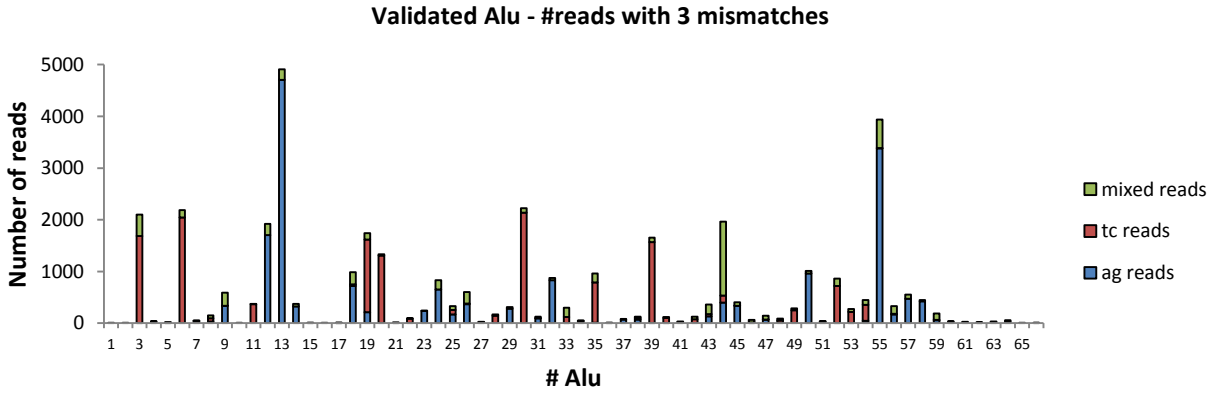
(b)



Supplemental Figure 14: Percentage of non-edited reads in the validated *Alu*s (MiSeq experiment results). For most *Alu*s, the majority of the reads are not edited.

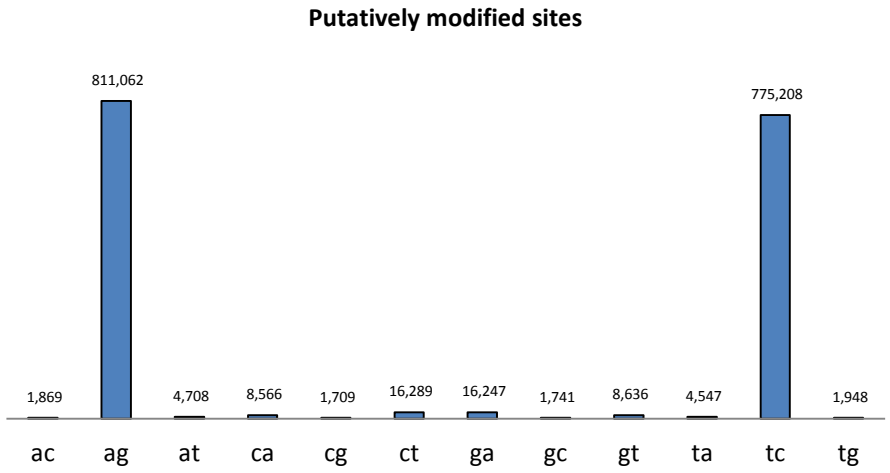


Supplemental Figure 15: *Alu* reads that contain exactly three mismatches. Most *Alu*s present a single dominant type of mismatches coming from the main expressed strand (MiSeq experiment results).

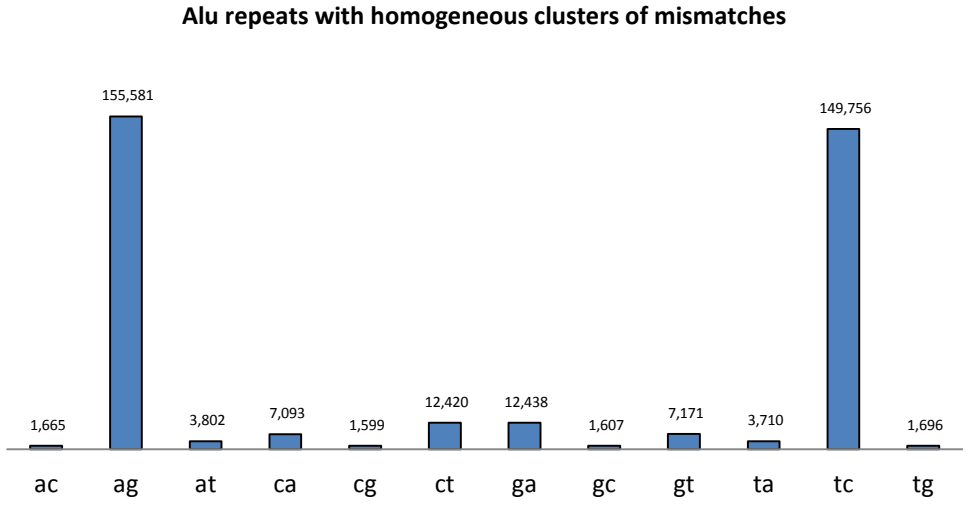


Supplemental Figure16: (a) Mismatch distribution for the putatively modified sites. AG and TC mismatches clearly dominate the distribution. (b) Distribution of *Alu* repeats exhibiting homogeneous clusters of mismatches, by mismatch type (single dominant mismatch type in the *Alu* repeat- number of mismatches of the most common mismatch type is higher than the number of mismatches of all other types combined). (c) Distribution of the number of sites in clusters of mismatches for putative editing sites (AG and TC), and control (GA and CT) sites. While the number of putative editing sites (AG, TC) within a cluster is typically larger than three, almost all clusters of control mismatch types include only one or two mismatches.

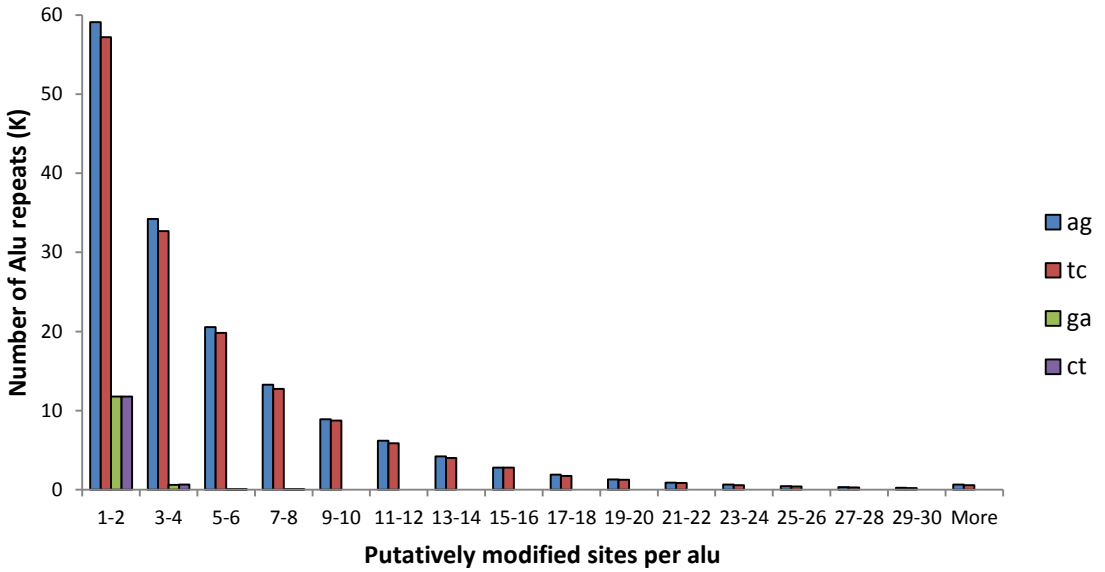
(a)



(b)



(c)



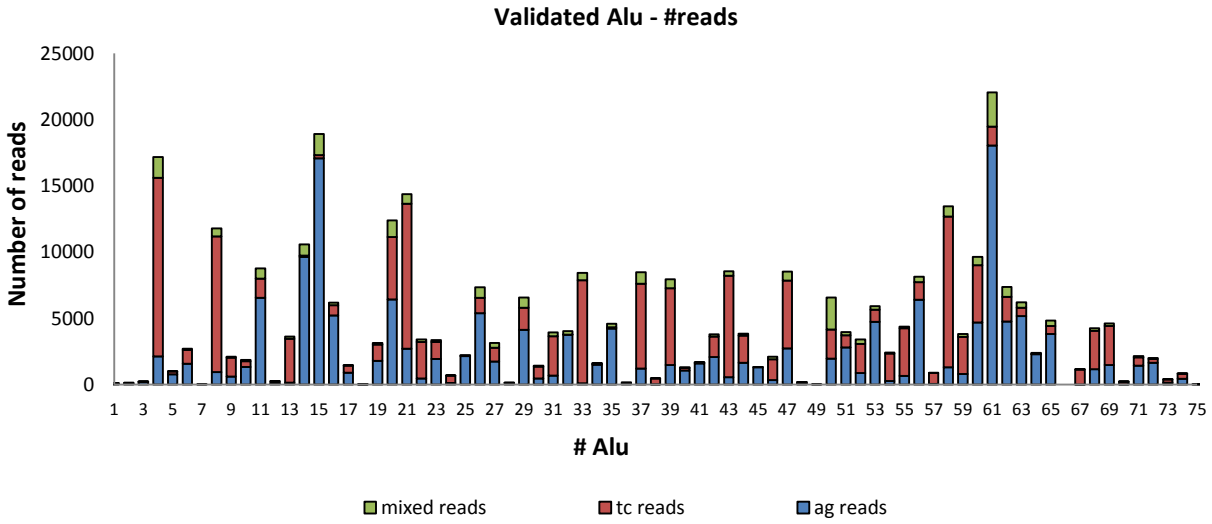
Characterization of editing results

In this section we describe the distribution of edited *Alu* within a given gene. For this purpose, we focused on editing sites that reside within RefSeq genes. Out of the 235,212 *Alu* elements, detected as edited using the HBM dataset, 163,103 overlap RefSeq genes (69%), harboring 694,523 detected sites. The distribution of these *Alu* elements and sites along the gene is presented in Table 4. Most editing within RefSeq genes takes place in *Alu* repeats located in introns; only a few *Alu*s overlap the CDS regions, and even in these cases, most of the *Alu* sequence is typically located in the adjacent introns.

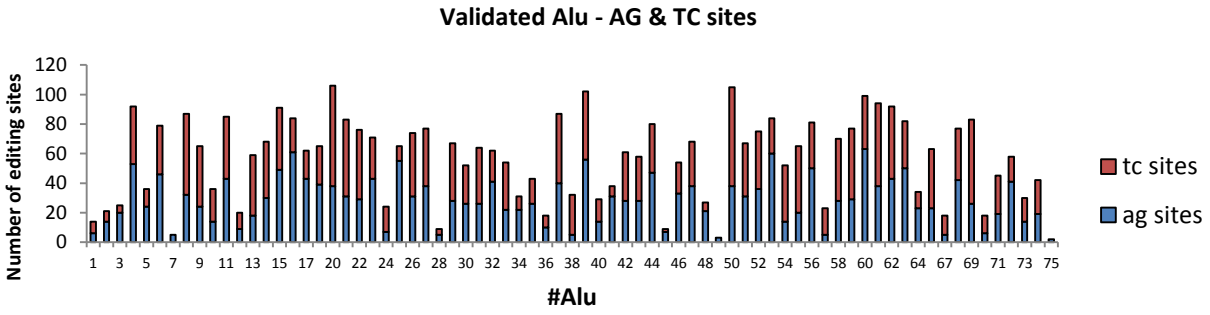
Table 4: Distribution of edited *Alu* genomic location along RefSeq genes.

Intron	160,534
3' UTR	2,835
5' UTR	480
CDS	113

Supplemental Figure17: Number of reads containing mismatches (MiSeq experiment results) . Most reads either contain AG mismatches or TC mismatches. Only a very small proportion is mixed with mismatches of both types. For most *Alu*s the dominant expressed strand can be clearly identified by the dominant editing type.



Supplemental Figure18: Distribution of number of sites, AG and TC, per *Alu* (MiSeq experiment results). Although the number of AG and TC sites per *Alu* is similar in most *Alu*s, data shown in read level (Supplemental Figure17) clearly reflect the dominant expressed strand per *Alu*.



Validation results per group

Table 5: MiSeq results: (full genomic coordinates and primers for each of the *Alu* are given below)

#	Validated group	# *Alu* repeats	# Covered *Alu*	# edited *Alu*	# edited *Alu* both strands
1	HBM edited *Alu*	28	25	25	22
2	Editable *Alu*	52	50	48	48
3	*Alu* in Refseqs				
	KIF1B (NM_015074)	73 (out of 207)	72	68	63
	RILPL1 (NM_178314)	14 (out of 76)	13	11	10
	PSEN1 (NM_000021)	23 (out of 104)	23	22	21
4	lincRNA	7	7	7	5

Primers used for the validation

All forward primers were added with the GCGTTATCGAGGTC adapter for sequencing, and GTGCTCTTCCGATC was used for the reverse primers. Small caps indicate nucleotides masked by Repeatmasker.

Edited *Alus*

Table 6. Set #1 : *Alus* that were found to be edited in the HBM data

All coordinates are given in hg19

	Alu	Product size	Left primer	Right primer
1	chr19:58541893-58542066	244	TTGGGGTAGGCTTGGTGA	ACCGAATGTCACCCCCTTA
2	chr2:198350018-198350247	287	TGAATTCCTAAGTCCTATCACCG	AGACTCTGAAGGTCTAACTGCCA
3	chr17:38316888-38317123	299	TAGAGGCTGCCGCTGTTT	AGGTATTCCTGGAACATATGCAA
4	chr5:55002770-55003018	338	TGAGAATTTCAGGTGACAACAGAT	CCTAAGTCATCAAACTTCGGGA
5	chr11:72561420-72561697	340	caccaggccTAGGACAGAGA	TTGACAGAAACTTCCTTCTTGATTT
6	chr6:11010256-11010554	347	CCCCGGGCAGAGAAAG	TTCCCTTTAAAGATGTGTGCG
7	chr12:27931813-27932099	349	TGAGAGAACATAGGAAACAAGTGG	GAGGACTATTAAGAGAGACCACACG
8	chr2:72905553-72905855	350	TCTGCTTATTAAGGAACTGGGG	TGTTTTAGCTATCAGTCTGCTCCA
9	chr3:40534609-40534914	350	gccCCAAGGAGTTTCCA	AACCATTTCACTTTCCGAAGAA
10	chr7:97805874-97806160	363	AGGCCCAAGAAATAAGGGG	GCATGTGGGAAGCATTCAG
11	chr14:78452189-78452499	364	GCTGGCCAGGAGATAAAGG	TTAAAAATTTGCCCGTTTGC
12	chr21:38335779-38336061	366	GCCCCTCCCTCAGTAAACA	TTCTATGCTTGGATAAAGTGGACA
13	chr1:211982257-211982553	367	TCTCCCCCAAAGTAGCCTG	GCTACCTCAGCTCTCAGTGGTT
14	chr19:9766903-9767199	368	TCTCAATCTTTTCTGAGAGAGGAAA	CAAATGGTAAGATTCATGAGGGA
15	chr19:21212490-21212779	370	CGCTCTGGAGCCTTACCTC	AGCAGCAGGACTCACCAGA
16	chr8:103905551-103905855	380	GAGGTAAACCTTGGGAAGGG	TGCAAACCTCAGTCTGGTCA

17	chr12:69807999-69808307	384	AGCAACGATTGCATGTTCTG	ATTTGTCCAATTTGGGGGA
18	chrX:135243999-135244302	385	TGCTAGATGCTGTGGGGAG	AGAATGGGCGTTCTTTGCT
19	chr20:19944499-19944806	386	GAATGCCACACCATCTTGC	CACCCCTGCTCCAGTGTT
20	chr4:140418328-140418640	391	AGGCGCCAGAAGTGCTATT	CACCACTTCCAGGGCTTTT
21	chr10:24707824-24708128	392	TCCTGGGGCCATGATTATT	GCAAAGAGCAAAGGTCAGAAA
22	chr12:62937378-62937688	392	CACTGCTCATTTATAGCCTTTGC	CAAACAGGTTTTCCCCTTCA
23	chr6:7960049-7960342	392	AGTAGGGCCACGAGAGGAA	CTTCCCCTGCCATTCTCAT
24	chr6:151300626-151300935	395	CAACCTGTATTTTTGGGGGA	TGGGTATAGCCACCAGCAG
25	chr20:5091545-5091845	395	AGAGACAGTGGGTAAGTTTGAAAAA	TTTATGGAGGAAAAGTGGACAAA
26	chr2:208846074-208846368	398	GGGCTAGAGAACTACCATCTCTCTT	GCTTCTGATTTGTCAGTAGCAAAA
27	chr3:170564311-170564620	399	TCAGACAAATTAGAGAGCATTTTCC	GACTCCCCAAATGTCTTCACA
28	chr1:43219311-43219615	399	GGGACTCCAATCATCTGGG	AGTACGCTTCCCACCATCC

Predicted *Alus*

Table 7. Set #2 *Alus* that are predicted to be edited, but with no evidence in the HBM data

	Alu	Product size	Left primer	Right primer
1	chr16:69239922-69240202	331	TTGTTGTCAAAGGCAAGGC	TCGCTAAACAAAAAGCTAAAAGG
2	chr9:127043061-127043337	333	ATTTCACGCTTTGCTTTGC	GCAGCAGGGAAGCTCTGAT
3	chr6:158865865-158866145	339	TGCATTGCTTTGTGTGTATGTG	TCACTGCAGGGCAGTTTCT
4	chr9:127551202-127551503	345	AAGGACCAGATTGTTTGGTGTC	TCAAAAAGCACCAAACTGGA
5	chrX:129309466-129309757	346	TCCAGGACACACAGAACTCAA	TCAAGGTGCAAAAAGTAAACCC
6	chr12:107810455-107810709	348	CCCTGTCGGTGTTTCCAT	CCCACCTTGCATTAAGAGGA
7	chr22:20780940-20781228	349	CACAGGGGTTCCATTATCCTT	GCCTCAGAAGCCTTATCGC
8	chr10:27355739-27356024	349	tGGGTCTGCATATGTTAATTGG	CCTTAAGTTCTATTCTGGGGTCC
9	chr6:135608952-135609262	349	CATAAAAATACCGGGCCGT	TTGTGGGGTAAGGTCCAGG
10	chr16:89945050-89945348	350	AGCTTTTGGGTAGGTAATAAGGTG	CAATCACCGTCAGACAAACCT
11	chr15:41395585-41395888	362	AGATCAGTATTCCCAAATCATCG	TGAGGAACTAATGGCATAATCAGA
12	chr14:59952345-59952633	363	AAATTAATTCACATGTAGACCGCA	CAGTGACATCTGATTCCCTTCA
13	chr2:232633582-232633878	369	TTTGGTCATGGAAACTGGTTTAT	ACGTGGCATATCATCTGCTAAA
14	chr9:99176768-99177057	372	CAATTCAAACCCATGCATCTT	TGGCCTAAAACAGACTTGTCC
15	chr1:8872424-8872710	372	CACTTGCTGACATACTGGCAA	TTTGGTAAAGTGGTCTGTGCC
16	chr15:92478187-92478481	373	AGGCGTCCTTCGTGAATTT	GGCCACTTCTCACCATTGA
17	chr4:99420808-99421110	375	CATGAGCTGTTTGGTTTCCA	TCCCAATGATTCCCTCAAGT
18	chr18:5980429-5980730	379	CTTTTGATTCTTTTGTTGAAAACG	TGCCCAAAAGTGAAATGTGA
19	chr15:44304959-44305263	380	TTGGTTACTGGCAAGTTATTGG	CAGAAATCAATGTGTGCTTTTCA
20	chr14:105907849-105908156	380	AACAGGAAGGTCAGCCAGC	CTTCTGTTCCCTGCACTGG
21	chr14:31549986-31550296	381	CTTCGTCACCCTGCTTCAA	TAGCGCTATCCTGCCTGG
22	chr3:74484338-74484627	382	TCTCTGATTTTGCATTACCATCA	TGTGACAGCTGAATCTGCATT
23	chr3:69144651-69144945	382	GGTCCTGCTGTTTCTCCCT	GGTCTGCTGGGTGGAGTTT

24	chr20:42888132-42888426	384	GGGGCGGTGTTAGGTTGTA	TCCTGAGTCTTGGCCACAT
25	chr9:133620467-133620770	385	GGTGTTATCCCTTCCTGGG	CGCTCTATTATGCAGGCCA
26	chr4:7857976-7858266	386	GGAATCCACTCCTTATTAAAACCA	GGCTATGGCTTCACAACCA
27	chr8:145041249-145041560	386	TGGCTAAAGAAACCACAGACAG	ATGGATGGGTCTCACACCA
28	chr14:31347585-31347860	386	CCAGGCAAATTTTGTTCTTGT	TCAAGAAAATAGCTCATCAGGGA
29	chr8:120894120-120894430	387	GGAGTCAATGGTCAAGAGCC	AGCAATGGCCTTTTAGAAAAGTAA
30	chr14:103426261-103426568	387	CAGGCTTTTCCCCAGAGAG	TGCCCCTGTGGTTTTGAG
31	chr3:170189874-170190168	388	tggtgtGTGCAGGAAAGTG	CCCCAGTCCAACCATGTCT
32	chr10:27297769-27298067	388	GGCCTCCCGCTACATAAAG	TTGCCCTCAGCTGTCATTT
33	chr1:149939942-149940205	389	TCAGGGTCTTGGGAAATACCT	GGAAGAAAGTTCCCAAATGTGT
34	chr21:27371323-27371628	390	TCCCCCGGAATTATGAAAA	CAATGACATAAGGCTCGAGGA
35	chr12:83145978-83146288	391	TGAAGGCTTTCTCGTTGCTT	ATAAACAGCGCCCCTATCG
36	chr5:74160077-74160352	392	CTCCAAATCAAAATCCTTTGTTG	GGCTTCAAATTAGTCTGACCTTTG
37	chr19:9680196-9680508	392	TCCAAGTGCAAAGAGAGATCATT	CACAATATTCGAGCTTTTGACG
38	chr8:19197002-19197311	392	GTTAGCTCACTTTCCTCTTCCCT	CCTTGGGATCCTCTCCTCA
39	chr5:159380086-159380388	393	GGAACGTTCTTTGAAAGTGACA	TGGAGAGATGTGTGGTGGG
40	chr1:90123019-90123319	394	GACCATTGTGGTTTAGAAGATGC	TGAAGAATAGGTCCCCTGAAAA
41	chr19:13437709-13437994	394	GCAAATACCACATAGCTCCCA	TGGCAAGGAATCACATTTAGC
42	chr2:206871124-206871423	395	TCAGTTGAAGTACAGGGTGGG	ATTTTTGAACAAAGACATCTCGAA
43	chr9:80926467-80926773	396	CAAGGGTTACCTTGAAGAGCA	CATTGGGTGAGCAATGATTTT
44	chr19:9869995-9870303	397	ggcattagccacAATACATCTG	CATCCACAAACATACAACTCAGC
45	chr4:113564721-113565035	398	ACTGGGATGAGAAGGCCAC	AGGGGCAGAATTCAACTGG
46	chr8:68211899-68212221	398	CAGGCCTTGGAGAACCTGT	AACCTTGTTTATGTGGAAAGCAA
47	chr10:18328374-18328668	398	AGGAAGGTGGTGGCCTATG	ATGGGAATTTTCCAATACCTTG
48	chr15:43337374-43337669	399	GGAAATGATGACAGGTCCATTT	GGGCCAGACATTCCTGTTT
49	chr20:36684506-36684815	399	TGCAGATGTTCCTGTCTCCTT	TGAAGTGAAAGCCAGGTCAA
50	chr21:34920011-34920328	399	GCTTTGGCCTATGTCAAGTGT	GCTCATCAGAAATCTGAGCCA
51	chr12:122417459-122417771	395	GAGGGGCTCAACACACTGA	GGGTCGACTTAGGTTTCCG
52	chr15:44836938-44837214	333	aaaaaGGTGGGAGGGGC	TCACCTTAAATTGCCTGTAGTGG

Comprehensive set of *Alu* of single genes

Table 8. Set #3 *Alu*s within selected genes

	Alu	Product size	Left primer	Right primer
1	chr1:10283594-10283901	374	GAAGATGGTTGCAAGGTGG	TTTTTCTGATTTTACTGTCTCATGG
2	chr1:10293769-10293999	345	GAAAAAGAGATGCAAAAGTCAAca	TCCAGTTTATAGTTGAGGAAACCA
3	chr1:10294934-10295247	422	cactgtacccagccTCCCT	AACACTTCGATCACTTTTTATCCTG
4	chr1:10298156-10298459	393	ATGGAGGAAGGGCAAAATG	tcacacctggccTAAATGC
5	chr1:10300264-10300565	345	caccacaccctgccAAAC	TCACTATGGAATTAAAAATGGTTCA

6	chr1:10310938-10311246	391	tctggggaaaacaaaTGGA	CTTATATCTTCAAAACCAAGCTGC
7	chr1:10312092-10312397	373	CAGTCCTCCTAGTTCCCACTTTT	cCCCACAAAAACAAAGAACC
8	chr1:10312779-10313082	384	tcaccgagccaggtcATAA	TATGTGGCCACCCCACTAA
9	chr1:10325620-10325928	394	accatgcccagccTTTAAC	TTGCCTTTTTCTTTCATTCAGTC
10	chr1:10329290-10329587	447	GCAATTATAAAAGATCCTTTGCCT	ccatgcccagTAGGAGTGT
11	chr1:10335967-10336219	450	cactgttcttcattcACTTCTCCT	CGAGAATTGCCACCTAAAAAG
12	chr1:10336856-10337156	400	TTGTCGTTATAGAAATGCTGAAGTC	gcacctggcTATGTTGGTAA
13	chr1:10337388-10337686	349	aaaaCAATTACCAACAGAATGTTCA	AAAAATACAGGCCTCGGGA
14	chr1:10373968-10374239	383	tgcttggccATTAGGATTTC	AAAAACTAATTACGGAAGGCACA
15	chr1:10379161-10379458	449	TGGAAAAATGAAGGATACATCTGA	AGGTAACCAAAAGCTCATAGTGC
16	chr1:10393201-10393444	320	caccactcccagccAATAA	gcctggccTATGGGTCTT
17	chr1:10417523-10417782	298	CCTCTCAAGAATTGCCAGTGT	cactgcgcctggccTA
18	chr1:10429708-10430015	346	TCTTAATGATAACGGACTGGGG	gccCATTATTTTGGGGG
19	chr1:10273912-10274192	344	AAAACCTGCAAAAGGTGCC	GGCCCGTTAACCAAACAGT
20	chr1:10275231-10275533	432	CAAGCTTGTAAATGTTTGATGGA	TGGTTGAATGATATGCTTTGGT
21	chr1:10275608-10275906	395	TACCCTCTACCATTGCCCG	CATCTGCCAAGATGATGTCAA
22	chr1:10278483-10278778	390	TTTCTGGCTACTCACATGGCT	ACCAAACGACACAGCAAGC
23	chr1:10283029-10283344	449	GGCTTTCAAAACTTGGAAGATG	TGAAATCAAGTTTAAAATCCAGCA
24	chr1:10284488-10284781	387	cccagccTGTTTTGTGTGT	GGTTGAATTCCTATTTAACTGGATG
25	chr1:10286744-10287053	385	CCCAGCCTATCTTCAGACTACC	TGCGTTATTTATCAACCAAATACAA
26	chr1:10287090-10287390	349	TGGTTGATAAATAACGCAAGTCTC	GGTTTTATAACCCTGGGCTTG
27	chr1:10290485-10290793	445	TCCTTTGTTTGGTTATTCTTGAAA	CAAAGCCAACACTACTGTGTGAA
28	chr1:10290989-10291292	368	TCAGTGGCATATCTGAGGGA	CAGTGATTTCCTTTAACTTCTGGG
29	chr1:10291436-10291744	375	TGTATGGAGAGTTCAGTGCCC	TTCCACGTGCAAACCACTT
30	chr1:10297785-10298092	348	CCACTATTGAAAATTTGGGGG	CCCTTCCTCCATCCCAA
31	chr1:10298989-10299259	343	CCACAGCATACTGCTGCTTT	ttGCTTGTGTCTGTATTCATCAAA
32	chr1:10301187-10301461	350	TGTGGTTAGTTTTGTGTGTGTGTG	TGCCTTTAAACTTATATTCCCACTG
33	chr1:10311510-10311804	366	AAAAGGCTGAGTGAAATATGGC	CACCATAAGACAGGGCCAA
34	chr1:10314652-10314958	393	TGAATGCTTCATTACCAAATGC	TCCAACCACTATTTGAAAGCAG
35	chr1:10315001-10315305	394	CTGCTTTCAAATAGTGGTTGGA	TCAAGGCACCCGTTAAGC
36	chr1:10315365-10315668	349	CGGGTGCCTTGAAATAGTAATG	CAAAATTTGGAAATGAACGACA
37	chr1:10315867-10316166	400	TCTGGTTTTTCATGTTCCATGT	CACTAAATCTGAATAGGGCAATCA
38	chr1:10316777-10317068	378	CCAACATTATTGCCCATTGA	GCAATTTTAATTTGGCATTTCC
39	chr1:10317721-10318034	397	TCATCTCTTTGTTACATGTTTTTGG	TCAGTCCCTTCCCTTGACA
40	chr1:10322223-10322506	397	GGCTATTTCTGGGTTTTGGG	AACTGCAAATAAATGCACAGACTT
41	chr1:10323077-10323355	337	TGTGCTTAAAGAAGCTGTTTTGG	GGAGGAAGAGAAAGAAACAGCA
42	chr1:10333329-10333506	246	TGCTTTGGAGAAGGAATTAAGAA	ATCCTTGAGGCCTAAAACTTCA
43	chr1:10333989-10334295	387	AGCAGTGAGACCCTGGCTT	ATGCTGGCTGACTACTGCG
44	chr1:10339747-10340060	441	GGGCCTCGGAATGATTAGA	GGAGAGAATAATGAGGTTCATGTG
45	chr1:10340283-10340591	398	TCTTGATTTTGGGGCTAATTTT	GCCAAGCCTCATTACATGC
46	chr1:10342982-10343289	380	TGTGTAGAGCCATGCTTCTGTT	CCTGGAACAGCTACTCAAATGTT

47	chr1:10346103-10346397	400	GGCTTTGTTCACTGAATAGCCT	TGTGGCTACAGCTGCATGTT
48	chr1:10351486-10351781	408	TGCTAGGGATTCATAAAAATGGA	TTTGTCTCAACTGACTTTGGGA
49	chr1:10354644-10354920	348	CCAGAACTCATGTTTTAAAATGGAA	CAAACGGAACCAAGTGCAT
50	chr1:10374465-10374777	385	TCAATTGTGGTCAAAATGCAG	TGAAAAGATTTCCCATAGCCC
51	chr1:10378497-10378788	342	TTGGGTCACTGCTAGGGAA	TACTGCCAGGGTTTCCTCC
52	chr1:10380773-10381058	449	CTTGGCAATCTCAAGGCTTC	CTCCTGACCTCTGTTACATTGTATG
53	chr1:10383406-10383678	384	GCATTTTGTGGTCGACCTTT	ATCTTCTTGAGGGGAGGCA
54	chr1:10389510-10389805	415	GATTGGATTTCTCCTATTGGCA	CAGAGGCAGTCATGGCTAAA
55	chr1:10392114-10392423	397	AGGAGCAAACGCAAACCTT	GAGGAGACTCTGAATGGGCA
56	chr1:10393930-10394227	380	TTGGGAAAAATGAAACAAAACC	GAGGAGCCCATTCAAGAGC
57	chr1:10396212-10396504	400	TGGAGACAGTCAGGATTTCTAGG	TTATCAACTTCCTCAGCTCAAATG
58	chr1:10403533-10403844	432	CCACATGTATTCCTGCTCCTC	AAGGCCCCCTATTGTTTCTTA
59	chr1:10406416-10406721	384	GCCTTCTCGTTAGCAACAGC	AAAACCCTTACAACAAGGTGCT
60	chr1:10409214-10409514	350	GAGGGCATAAAAGAAATCTCTGC	AAATGCAAGCTCACAATTAGCA
61	chr1:10410872-10411151	340	TGGACCAGATTAAGAACCTTGG	TGCCCCAGACAAGAGAGAA
62	chr1:10416143-10416432	431	GGGAAGAAAATTGAAGGAAGTG	AATGGGGTCCAAATTGAGG
63	chr1:10420460-10420760	400	CAGAGGCAGGCTTGTCAAA	TAGGTCCCCATCCTCTTGC
64	chr1:10422850-10423137	450	TGTTTCCCTCATATCGTCTCTCT	GAGTCCGCTTTCATCTCAATTT
65	chr1:10423467-10423772	444	GCTCTTGTGGATTGAGGAGG	TCCGCCTAGCTAGAGGAACA
66	chr1:10424258-10424561	416	TGTGTGTGTATTCCCTTCCG	TGATGTTAGGTGAAAATGCAGAA
67	chr1:10424782-10425082	395	CAAAATCCCATATATCTTCAGACCA	TTCCAAGTTTAGGTCAAAACGAA
68	chr1:10426748-10427034	372	ATGGGAGCAAGGCAGAAAT	TCTCTTTATTGCTGCCAGAGAA
69	chr1:10427157-10427436	396	AGAATGGATTGATGGTGAAGAGA	CTTTTATCTGGCTGGGCGT
70	chr1:10429404-10429674	346	tggaaggtgcACCACTGTAA	cCCCCAGTCCGTTATCATT
71	chr1:10435838-10436159	444	TCTTTGGCTTTTGGCTTCA	ACACTCCTGTCCTTCCTCTCTTT
72	chr1:10437714-10438008	388	GGCGGGAGTGTTTCACTTT	CACAAGGTCACTACCCAGCA
73	chr1:10440483-10440777	350	GATTGTTAACCTGCCTCTTGAAA	TTCTGGGAGCAGCAGCAT
74	chr12:123967155-123967449	383	ccagcAAAGGGCATCATTT	CAGGTTTGTCACTGCTGGG
75	chr12:123975297-123975596	387	gtctggccCCAAAGTCCT	CTCTCGAGAAGCTGGAGCC
76	chr12:123984821-123985131	400	cccggccTCATGTATTACC	gtgcagcgtcatctttcGT
77	chr12:124000299-124000607	343	caccgagccAAACACCA	ctcaatcatctcggggcA
78	chr12:124012631-124012924	386	GCCTCCATAATACTGTATGACATTG	ggctcaaaacagcagaaatG
79	chr12:123961876-123962176	390	CCTTTGATGTGCTTCACAGAGT	ACCATCCAGGAGGTTCTGC
80	chr12:123962534-123962847	383	CCTGGGACACCTGGAGTG	TGAGCACAGCAAGACACAGA
81	chr12:123964702-123965007	399	GCTGGAGGCTCTGAATGC	CCATTTGTTCCCAGGTCGT
82	chr12:123972703-123973021	382	ACAGACGTCCTTCCCAACC	TCTCCCTGTGCTGCTATGG
83	chr12:123978329-123978631	347	AAGATCTGTCCTGAGGCCC	TGAAGTCAAGAGGCAGGGA
84	chr12:123984481-123984776	378	TTCTCTCCATCCCTGACCC	aaagaaaGGGACACATAGGAACTG
85	chr12:123985369-123985687	390	tgtcaacctaaatgaTGAACAGAGA	TGCGTGGCCATCTTTTATT
86	chr12:124001959-124002262	346	CAGAGAGAGGGCACAGCAA	GAGCTCAACGGTGCTCTCA
87	chr12:124004823-124005137	357	tTCGGACATGAAAAGGCAA	CTTCAGCCCAAGAAGGAGG

88	chr14:73636372-73636634	350	TGCCACTCAATCAAGTTTGC	catataatcaagtaacaGGCCAGG
89	chr14:73666981-73667263	450	ACTGGCTTGGTATAATCTTTTTCAT	AGGGGTAAGTACTTGATAAGGCAT
90	chr14:73604509-73604791	341	TTTAGAGCAGCGCTAGCCA	ctccAAAAGAAAAACAAAACAGG
91	chr14:73611594-73611900	400	TGCGTGGTAATCACTTCTGC	GGGTAATGCAAGCAAAAATAAAA
92	chr14:73612892-73613136	392	ctcaaaataaaGACAGGAACAAAGG	TGAGATCAGCGAAATGCAA
93	chr14:73613768-73613990	296	AGAGGTCTGAGGCTGGGAG	ctggCAAGAAATCGCAGAC
94	chr14:73615358-73615691	446	TCGGACAAAGAGATTCCTCAC	CACAGGACATCCATGCTCTC
95	chr14:73619036-73619331	344	GAACAGCTGATACAAAAGAAACCA	TGCTACCGAGCTTTTGTGG
96	chr14:73619469-73619764	388	TGCCCTTGTCATCAATAATAGC	GCTCATAGGTTTGCTTTTTGTCT
97	chr14:73621024-73621323	417	GCATTTGGGGTGTCGTAAA	GTTCATGGCAGTTTTTGCC
98	chr14:73631664-73631963	399	TGTCTCTAACCAACACTTCTGTCAC	TGAGACACCATCTTTAAGGAAAAA
99	chr14:73646689-73646992	395	TTTTGTCGCATTTGTACATCTTTT	AAATGTGGAGATCTGGCTGAA
100	chr14:73648473-73648781	370	CATGTCTGAAAGGGAAACGG	AGGCCAGTGACAACCAAAA
101	chr14:73648921-73649225	441	TTGGTTGTCACTGGCCTTT	ACCTTCTAGGAGCCCCCAA
102	chr14:73649269-73649579	382	TGGGGGCTCCTAGAAGGTA	TTGGCAAAGGAAGAAGGAGA
103	chr14:73652808-73653103	381	TGTGGAAGAACACCAGAAACA	TGGTACTCCTTGGACATTTGAA
104	chr14:73653202-73653501	382	TGGCCTTTGCCAAATTGTA	AACAGGTCCCACAACCCTT
105	chr14:73656913-73657213	433	CTGACAGCTTAGAGGGGTAAAAA	TCATGTACATCCTTCCAGAAAGTC
106	chr14:73659736-73660025	399	CGTGTACATCCCATAACTCTTCA	ttggccagCTTGTCAAAAC
107	chr14:73660059-73660286	272	CAAGCTTTTTGAGGTTTTGACA	TGTCTACCACTCTGCTTGTGC
108	chr14:73665482-73665782	364	GCTTGAAGGAAAGGTGATGAA	CAGATCCTTCACACAAGAAAACA
109	chr14:73669784-73670095	450	TGTCTGGTGAAGATTTGAGGATT	GCATTGCGTTGTGGTTTTC
110	chr14:73671902-73672209	400	TGTGATAAAAAGCCAGAGAGACA	AGCAGCAAGCTGAGTCCAA

LincRNA's *Alu*

Table 9. Set #4 selected *Alu*s within LincRNA

	Alu	Product size	Left primer	Right primer
1	chr1:93800272-93800425 (TCONS_00001027)	295	TTCCCTCCAAACAATTGGC	AAAATTCAATGTTATTCAGACCCAA
2	chr17:30734889-30735179 (TCONS_00025349)	336	TGAATGAAAGGTATCTCATGGGT	TGTTTCTATCGAAATGTGTCTGG
3	chr1:150584891-150585168 (TCONS_00000640)	345	GACTAAGTGGGCCCAGACC	TGCTTAAGAAAGGGGCTCTTG
4	chr18:74520240-74520520 (TCONS_00026576)	347	AATAGTTTTGTTTGGTTTTATGCCA	TGAGGGTGATGGGAGATCA
5	chr13:79997704-79997917 (TCONS_00021499, TCONS_00021848, TCONS_00021849, TCONS_00021851)	350	CAATGCATCAGAGATTATGTGTCA	CAGAGCCCAAGTAGAGTGCAG
6	chr6:25998921-25999229 (TCONS_00012437, TCONS_00012113, TCONS_00012114, TCONS_00011743)	382	GCATTGTGGTTGGGTATGG	GGGCACACCAGAGAGCATA
7	chr11:82785550-82785847 (TCONS_00019409)	400	GGTAAAGCTTTTCATCTCTCAAAAA	GCTCTAACAGCAATTGGCTTC

Table 10: Primers used in the Sanger screen

	Alu	Product size	Left primer	Right primer
1	chr4:39955679-39955993	377	GGCCCCAACTGAGGTACTAA	tgagatcgtggcaTGttttg
2	chr8:100,134,899-100,135,346	448	ttttcaatgccagctgtgtc	acagagtttcgctcttgttgc
3	chr14:102682546+102682945	400	AAGCGGAGGGAACATTCAG	GCTATTTTTAAAAGACCAAAATGTCA
4	chr14:23371998+23372366	369	cggagtcccactctgtcact	GCTGTCAGCACTAGGGTGTG
5	chr12:117649466+117649964	499	TCCCAATTCCTGGACAACTC	TTGCTTCTCCAGTGTGTGGA
6	chr2:206982070+206982493	424	TGTTAGCTTCTTGAGGACATTAGC	GCTCTTCTGATGCCTGGTTC
7	chr9:80496258+80496656	399	agctcactggtgaataaatgaaAA	TTGGGCAGAACACAGATTGA
8	chr19:23523933+23524468	536	accactgactcccaggttca	tcatttgcagttgccttttg
9	chr3:15113548+15114046	499	GATCCTGGGTTGTGCTTCTT	GCTTTTGCAGAACATCACCA
10	chr1:1751166-1751455	290	ATTTGCAGTGAGCAGTTTGC	GATGGGGATCTCGTCATGTT
11	chr19:44643474-44643861	388	AACCACACAATGCCACACAT	TCGTTCTGTCACCCAGATTG
12	chr17:6826988+6827284	297	GTTCAAATCCCATCCCATCA	AATGGCTTGATCTTGGCTCA

Table 11: Comarison of editing activity in Alu repeats with that in recoding sites

A list of well-characterized recoding editing sites was given in table S3 of (Li et al. 2009) Although many more recoding sites are reported since, most of them are weakly edited and not conserved across species. Here we compare the total amount of editing activity in these sites with that in Alu repeats. We use the number of AG (or TC) mismatches as a proxy for the number of editing events, assuming, as usual, that the number of reads aligned to a genomic region is proportional to its expression level.

		thyroid		testes		ovary		WBC		skeletal_muscle		prostate		lymph_node		lung		adipose		adrenal		brain		breast		colon		kindey		heart		liver		lymphoblastoid cell line	
		A	G	A	G	A	G	A	G	A	G	A	G	A	G	A	G	A	G	A	G	A	G	A	G	A	G	A	G	A	G	A	G	A	G
GRIA4	chr11:105804693-105804694	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
KNCA1	chr12:5021741-5021742	19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	22	10	0	0	0	0	0	0	0	0	0	0	2	0
GRIA2	chr4:158257874-158257875	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	7	77	0	0	0	0	0	0	0	0	0	0	0	0
GRIA2	chr4:158281293-158281294	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	11	0	0	0	0	0	0	0	0	0	0	0	0
CYFIP2	chr5:156736807-156736808	5	1	9	2	15	0	108	0	0	0	8	0	6	0	0	0	3	0	17	2	18	42	3	0	0	0	57	1	8	0	4	0	39	1
GluR6	chr6:102337688-102337689	0	0	0	0	2	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	2	1	2	0	4	0	0	0	10	0	0	0	0	0
GluR6	chr6:102337701-102337702	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	1	0	0	0	0	0	0	6	0	0	0	0	0
GluR6	chr6:102372588-102372589	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	16	0	0	0	0	0	0	2	0	0	0	0	0
HTR2C	chrX:114082681-114082682	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	0	0	0	0	0	0	0	0	0	0	0	0
HTR2C	chrX:114082683-114082684	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0
HTR2C	chrX:114082687-114082688	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0
HTR2C	chrX:114082688-114082689	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0
HTR2C	chrX:114082693-114082694	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	3	0	0	0	0	0	0	0	0	0	0	0	0
GRIA3	chrX:122598961-122598962	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	2	12	0	0	0	0	0	9	0	0	0	0	0	0
GABRA3	chrX:151358318-151358319	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0
FLNA	chrX:153579949-153579950	31	2	116	42	263	7	126	1	17	1	102	417	126	34	59	72	98	25	232	11	3	2	66	5	291	122	67	6	10	6	1	0	129	0
BLCAP	chr20:36147532-36147533	245	1	142	3	111	0	79	8	427	0	131	8	92	3	72	4	247	0	126	5	111	10	230	5	82	2	65	3	92	2	72	4	105	1
BLCAP	chr20:36147562-36147563	49	8	41	10	47	19	61	11	271	7	76	17	69	16	40	20	291	15	164	5	109	12	300	8	101	6	77	15	65	4	72	2	66	19
BLCAP	chr20:36147571-36147572	261	16	179	19	152	26	111	12	410	7	117	24	79	36	63	21	292	25	151	14	115	13	301	9	94	6	68	19	53	4	67	3	50	21
GluR5	chr21:30953749-30953750	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0
IGFBP7	chr4:57976233-57976234	501	842	1293	1860	326	192	65	3	96	147	676	2082	1677	3671	1074	2160	319	415	943	1381	127	197	177	281	260	1185	2279	1767	743	310	293	232	0	0
IGFBP7	chr4:57976285-57976286	411	91	1225	112	142	7	8	0	81	34	605	348	1714	329	889	403	312	65	957	205	90	7	208	30	607	178	1764	113	469	41	190	7	0	0
	All 22 known sites	**1522**	**961**	**3006**	**2048**	**1061**	**254**	**558**	**35**	**1302**	**196**	**1715**	**2896**	**3764**	**4089**	**2197**	**2680**	**1562**	**545**	**2590**	**1626**	**622**	**426**	**1287**	**338**	**1439**	**1499**	**4377**	**1933**	**1458**	**367**	**699**	**248**	**391**	**42**
	Alu's	**4E+06**	**527042**	**2896263**	**629186**	3617223	**587226**	7632519	**432217**	9872574	**194193**	9009574	**883873**	3146553	**1045305**	5319343	**862568**	**5061058**	**682861**	**3969798**	**1000551**	2811567	**617672**	4552598	**518937**	4676603	**604873**	3213258	**553799**	1273607	**189449**	3459594	**257087**	**1.2E+08**	**8469675**
	Inosine fraction (known/alu)		**0.182**		**0.325**		**0.043**		**0.008**		**0.101**		**0.328**		**0.391**		**0.311**		**0.080**		**0.163**		**0.069**		**0.065**		**0.248**		**0.349**		**0.194**		**0.096**		**0.000**

References

Athanasiadis A. 2004. Widespread A-to-I RNA editing of Alu-containing mRNAs in the human transcriptome. *PLoS Biol 2, e391*.

Bahn JH, Lee J-H, Li G, Greer C, Peng G, Xiao X. 2012. Accurate identification of A-to-I RNA editing in human by transcriptome sequencing. *Genome Res* **22**: 142–50.

Blow M, Futreal PA, Wooster R, Stratton MR. 2004. A survey of RNA editing in human brain. *Genome Res* **14**: 2379–2387.

Carmi S, Borukhov I, Levanon EY. 2011. Identification of widespread ultra-edited human RNAs. *PLoS Genet* **7**: e1002317.

Gong C, Maquat LE. 2011. lncRNAs transactivate STAU1-mediated mRNA decay by duplexing with 3' UTRs via Alu elements. *Nature* **470**: 284–8.

Ju YS, Kim J-I, Kim S, Hong D, Park H, Shin J-Y, Lee S, Lee W-C, Kim S, Yu S-B, et al. 2011. Extensive genomic and transcriptional diversity identified through massively parallel DNA and RNA sequencing of eighteen Korean individuals. *Nat Genet* **43**: 745–52.

Kim DD. 2004. Widespread RNA editing of embedded alu elements in the human transcriptome. *Genome Res* **14**: 1719–1725.

Kiran A, Baranov P V. 2010. DARNED. A DAtabase of RNa EDiting in humans. *Bioinformatics* 1–5.

Levanon EY. 2004. Systematic identification of abundant A-to-I editing sites in the human transcriptome. *Nat Biotechnol* **22**: 1001–1005.

Li JB, Levanon EY, Yoon J-K, Aach J, Xie B, Leproust E, Zhang K, Gao Y, Church GM. 2009. Genome-wide identification of human RNA editing sites by parallel DNA capturing and sequencing. *Science* **324**: 1210–3.

Park E, Williams B, Wold BJ, Mortazavi A. 2012. RNA editing in the human ENCODE RNA-seq data. *Genome Res* **22**: 1626–33.

Peng Z, Cheng Y, Tan BC-M, Kang L, Tian Z, Zhu Y, Zhang W, Liang Y, Hu X, Tan X, et al. 2012. Comprehensive analysis of RNA-Seq data reveals extensive RNA editing in a human transcriptome. *Nat Biotechnol* **30**: 253–60.

Ramaswami G, Lin W, Piskol R, Tan MH, Davis C, Li JB. 2012. Accurate identification of human Alu and non-Alu RNA editing sites. *Nat Methods* **9**: 579–81.