Supplementary Table 1. Number of identified editing sites in the ENCODE data sets (cytoplasmic PolyA+ RNA)

	Alu sites		Non-Alu sites					
Cell line	**Editing events**	**AG, %**	**Editing events**	**AG, %**	**Total editing events**	**Total non-AG**	**Total AG, %**	**Total raw reads pairs (in millions)**
H1hESC	19781	99.9	772	93.0	20553	80	99.6	97.2
HeLa-S3	22803	99.7	750	90.8	23553	140	99.4	225.6
HepG2	8708	99.7	339	86.4	9047	72	99.2	224.4
HUVEC	7833	99.8	389	82.3	8222	83	99.0	230
K562	15636	99.4	425	81.2	16061	180	98.9	213.3
NHEK	7927	99.7	467	90.1	8394	66	99.2	222.9

Supplementary Table 2: Performance of GIREMI in different types of regions (accuracy measured as 1-% SNPs among predicted editing sites in each category)

Data	Region	Location	Genome-aware		GIREMI				Multiple data sets method (Overlap of two data sets)*				GIREMI (union of results)***			Multiple data sets method (Pooled read alignments)**		
			Number of sites	%AG	Number of sites	%AG	Accuracy	Overlap rel. to Genome-aware	Number of sites	%AG	Accuracy	Overlap rel. to Genome-aware	Number of sites	%AG	Accuracy	Number of sites	%AG	Accuracy
GM12878 (30% SNPs assumed unknown)	All		41027	98.8%	37591	98.6%	98.1%	90.0%	8307	90.2%	85.0%	18.5%	51082	97.1%	97.8%	35026	86.4%	97.0%
	Alu	Non-synonymous	119	100.0%	107	98.2%	100.0%	89.9%	31	100.0%	83.9%	24.3%	126	97.7%	100.0%	103	94.5%	100.0%
		Synonymous	58	100.0%	53	98.1%	100.0%	91.4%	15	100.0%	86.7%	24.5%	67	97.1%	100.0%	51	96.2%	100.0%
		UTR	7056	99.7%	6814	99.1%	99.6%	96.2%	2792	99.1%	86.7%	35.5%	8690	98.7%	99.7%	5994	96.8%	99.3%
		noncoding	2555	99.4%	2375	98.8%	99.2%	92.3%	688	97.0%	82.0%	23.7%	3154	98.0%	99.5%	2262	94.9%	98.9%
		Intronic	22017	99.8%	19752	99.0%	99.2%	89.0%	4135	98.3%	89.0%	18.6%	28353	97.7%	99.5%	23994	95.8%	99.4%
		Intergenic	7952	99.6%	7030	99.1%	99.4%	87.9%	136	100.0%	68.4%	1.3%	7344	99.1%	99.5%	812	97.0%	99.4%
	Repetitive non-Alu	Non-synonymous	NA	NA	1	NA	NA	NA	NA	NA	NA	NA	1	NA	NA	NA	NA	NA
		UTR	20	71.4%	20	76.9%	85.0%	85.0%	8	66.7%	75.0%	30.0%	41	83.7%	87.8%	15	46.9%	80.0%
		noncoding	10	52.6%	11	64.7%	81.8%	90.0%	3	50.0%	66.7%	18.2%	35	74.5%	80.0%	8	47.1%	25.0%
		Intronic	134	90.5%	149	81.4%	79.2%	88.1%	15	68.2%	60.0%	6.0%	695	88.5%	85.6%	53	41.7%	77.4%
		Intergenic	96	93.2%	86	92.5%	94.2%	84.4%	NA	NA	NA	NA	109	93.2%	94.5%	NA	NA	NA
	Non-repetitive	Non-synonymous	13	18.1%	35	72.9%	34.3%	92.3%	25	18.9%	4.0%	2.9%	59	68.6%	39.0%	91	16.5%	13.2%
		Synonymous	7	22.6%	31	73.8%	22.6%	100.0%	15	17.2%	6.7%	3.2%	51	77.3%	17.6%	65	19.4%	9.2%
		UTR	205	64.5%	257	82.4%	71.6%	89.8%	166	36.2%	57.8%	37.4%	416	78.0%	72.8%	407	24.9%	52.8%
		noncoding	100	46.1%	137	77.0%	65.0%	89.0%	45	35.4%	51.1%	16.8%	225	71.0%	68.0%	166	26.1%	57.8%
		Intronic	418	81.6%	467	83.4%	77.1%	86.1%	229	49.4%	62.9%	30.8%	1404	83.0%	76.9%	959	37.1%	69.1%
		Intergenic	267	82.2%	266	87.5%	85.7%	85.4%	4	44.4%	100.0%	1.5%	312	88.6%	87.2%	46	38.0%	69.6%
GM12878 (50% SNPs assumed unknown)	All		41027	98.8%	37956	97.5%	97.2%	89.9%	8445	85.3%	83.5%	17.2%	51985	96.3%	96.5%	35470	82.4%	95.8%
	Alu	Non-synonymous	119	100.0%	108	99.1%	99.1%	89.9%	31	96.9%	83.9%	21.8%	127	97.7%	99.2%	103	93.6%	100.0%
		Synonymous	58	100.0%	53	98.1%	100.0%	91.4%	15	100.0%	86.7%	22.4%	67	97.1%	100.0%	52	96.3%	98.1%
		UTR	7056	99.7%	6868	98.5%	99.2%	96.5%	2798	99.0%	86.5%	34.3%	8736	98.2%	99.4%	6011	96.3%	99.1%
		noncoding	2555	99.4%	2393	98.0%	99.0%	92.7%	691	96.6%	81.6%	22.1%	3179	97.4%	99.2%	2267	93.9%	98.7%
		Intronic	22017	99.8%	19834	98.3%	98.7%	88.9%	4145	98.2%	88.8%	16.7%	28630	97.1%	99.1%	24072	95.0%	99.1%
		Intergenic	7952	99.6%	7054	98.4%	99.0%	87.9%	136	100.0%	68.4%	1.2%	7369	98.3%	99.1%	817	96.5%	98.8%
	Repetitive non-Alu	UTR	20	71.4%	23	67.6%	73.9%	85.0%	9	52.9%	66.7%	30.0%	49	72.1%	73.5%	16	37.2%	75.0%
		noncoding	10	52.6%	13	65.0%	53.8%	70.0%	2	28.6%	100.0%	20.0%	43	76.8%	62.8%	9	36.0%	22.2%
		Intronic	134	90.5%	167	77.7%	70.1%	87.3%	14	43.8%	64.3%	6.7%	780	86.9%	76.8%	62	37.1%	66.1%
		Intergenic	96	93.2%	92	91.1%	88.0%	84.4%	NA	NA	NA	NA	118	92.9%	89.0%	2	40.0%	0.0%
	Non-repetitive	Non-synonymous	13	18.1%	47	68.1%	25.5%	92.3%	36	16.7%	2.8%	7.7%	89	71.8%	25.8%	128	16.0%	9.4%
		Synonymous	7	22.6%	38	79.2%	18.4%	100.0%	26	18.1%	3.8%	14.3%	71	81.6%	12.7%	84	17.2%	7.1%
		UTR	205	64.5%	310	76.9%	59.0%	89.3%	210	32.1%	45.7%	46.8%	519	74.5%	58.0%	498	22.6%	43.2%
		noncoding	100	46.1%	150	71.4%	58.7%	88.0%	52	26.5%	44.2%	23.0%	257	69.1%	59.5%	190	23.0%	50.5%
		Intronic	418	81.6%	522	78.4%	69.5%	86.8%	270	39.3%	53.3%	34.4%	1611	81.5%	67.5%	1103	32.1%	60.1%
		Intergenic	267	82.2%	284	82.8%	80.3%	85.4%	10	45.5%	40.0%	1.5%	340	84.6%	80.0%	56	36.8%	57.1%
Averages																		
All				76.3%		85.4%	77.7%	89.1%		62.5%	63.2%	19.3%		86.6%	78.4%		56.2%	65.7%
30% SNPs unknown	Repeitive non-Alu	Coding		NA		NA	NA	NA		NA	NA	NA		NA	NA		NA	NA
		all non-coding		77.0%		78.9%	85.0%	86.9%		61.6%	67.2%	18.1%		85.0%	87.0%		45.2%	60.8%
	Non-repetitive	Coding		20.3%		73.4%	28.4%	96.2%		18.1%	5.3%	3.0%		72.9%	28.3%		17.9%	11.2%
		all non-coding		68.6%		82.6%	74.8%	87.6%		41.4%	68.0%	21.6%		80.2%	76.2%		31.5%	62.3%
50% SNPs unknwon	Repeitive non-Alu	Coding		NA		NA	NA	NA		NA	NA	NA		NA	NA		NA	NA
		all non-coding		77.0%		75.4%	71.5%	81.7%		41.8%	77.0%	18.9%		82.2%	75.5%		37.6%	40.8%
	Non-repetitive	Coding		20.3%		73.6%	22.0%	96.2%		17.4%	3.3%	11.0%		76.7%	19.3%		16.6%	8.3%
		all non-coding		68.6%		77.4%	66.9%	87.4%		35.8%	45.8%	26.4%		77.4%	66.3%		28.6%	52.7%

*Multiple data set methods - Overlap of two data sets: editing sites identified in GM12878 and YH RNA-Seq data separately (see Supplementary Note 3), then GM12878 editing sites were called by requiring their presence in YH.

**Multiple data set methods - pooled read alignments: GM12878 and YH mapped reads were pooled together, then editing sites were identified using the pooled reads. Thus the results shown here were derived from two data sets.

*** GIREMI (union of results): results of GIREMI for GM12878 and YH data (analyzed separately) were combined.

NOTE: The number of editing sites shown for GIREMI is slightly different from those in Fig. 1 because only one of the 9 randomized trials for SNP exclusion was used (see Fig. 1 legend).

Supplementary Table 3: Comparison of GIREMI and the "mutliple data sets" methods on a set of primary human brain tissue RNA-Seq data.
(sample information in Supplementary Table 5)

Data	Region	Location	GIREMI		Multiple data sets method (Overlap: 2/3 data sets)*		Overlap with GIREMI		Multiple data sets method (Overlap: 2/17 data sets)**		Overlap with GIREMI	
			Number of sites	%AG	Number of sites	%AG	Number of sites	%AG	Number of sites	%AG	Number of sites	%AG
SRR627451	All		6351	97.5%	900	74.4%	754	95.2%	3023	69.5%	2549	97.0%
	Alu	Non-synonymous	9	100.0%	3	100.0%	3	100.0%	9	100.0%	9	100.0%
		Synonymous	6	100.0%	3	100.0%	3	100.0%	5	100.0%	5	100.0%
		UTR	243	96.4%	198	97.1%	173	99.4%	244	93.1%	213	99.1%
		noncoding	170	98.3%	78	88.6%	67	97.1%	132	93.0%	114	98.3%
		Intronic	3856	98.2%	414	91.0%	367	94.6%	1618	95.4%	1467	97.5%
		Intergenic	146	100.0%	28	100.0%	25	100.0%	104	99.0%	93	100.0%
	Repetitive non-Alu	Non-synonymous	NA	NA	NA	NA	NA	NA	1	100.0%	NA	NA
		UTR	3	100.0%	1	50.0%	NA	NA	2	33.3%	1	100.0%
		noncoding	3	100.0%	NA	NA	NA	NA	1	25.0%	1	100.0%
		Intronic	100	99.0%	18	90.0%	16	100.0%	44	73.3%	37	100.0%
		Intergenic	1	100.0%	1	100.0%	NA	NA	2	100.0%	1	100.0%
	Non-repetitive	Non-synonymous	11	61.1%	20	31.7%	6	54.5%	51	21.6%	10	62.5%
		Synonymous	2	66.7%	7	35.0%	NA	NA	22	19.1%	2	66.7%
		UTR	33	86.8%	33	21.9%	15	93.8%	86	18.2%	24	88.9%
		noncoding	36	92.3%	13	40.6%	7	87.5%	38	33.0%	14	87.5%
		Intronic	1704	96.7%	69	55.6%	61	89.7%	641	59.1%	541	95.9%
		Intergenic	28	80.0%	14	87.5%	11	100.0%	23	65.7%	17	85.0%
SRR663681	All		12436	88.7%	6591	81.6%	6044	88.9%	9461	73.7%	8412	89.4%
	Alu	Non-synonymous	42	97.7%	21	95.5%	20	100.0%	34	91.9%	30	96.8%
		Synonymous	33	97.1%	26	100.0%	25	100.0%	32	100.0%	30	100.0%
		UTR	2931	94.5%	1865	94.7%	1743	95.7%	2645	93.1%	2445	95.8%
		noncoding	1418	84.8%	843	85.2%	798	85.9%	1122	84.0%	1043	85.4%
		Intronic	7148	87.6%	3342	84.5%	3101	86.4%	4614	85.0%	4263	87.3%
		Intergenic	264	94.0%	134	89.3%	114	92.7%	206	89.6%	178	93.2%
	Repetitive non-Alu	Non-synonymous	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
		UTR	12	100.0%	3	42.9%	2	100.0%	8	34.8%	7	100.0%
		noncoding	3	75.0%	1	50.0%	1	100.0%	4	33.3%	3	75.0%
		Intronic	35	100.0%	26	76.5%	19	100.0%	39	63.9%	27	100.0%
		Intergenic	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
	Non-repetitive	Non-synonymous	18	62.1%	29	24.4%	8	53.3%	82	17.9%	15	60.0%
		Synonymous	9	56.3%	11	21.2%	3	50.0%	28	13.1%	6	50.0%
		UTR	110	82.7%	87	26.7%	45	83.3%	257	22.9%	89	84.8%
		noncoding	83	87.4%	41	38.3%	27	79.4%	102	32.1%	60	85.7%
		Intronic	300	82.9%	146	49.2%	124	83.8%	270	38.8%	200	86.2%
		Intergenic	30	90.9%	16	76.2%	14	93.3%	18	45.0%	16	94.1%
SRR815232	All		14065	87.0%	6559	81.3%	5947	88.9%	10080	72.7%	8876	89.2%
	Alu	Non-synonymous	52	98.1%	21	95.5%	19	100.0%	39	95.1%	36	97.3%
		Synonymous	49	94.2%	27	100.0%	26	100.0%	37	100.0%	34	100.0%
		UTR	3415	93.6%	1868	94.4%	1745	95.4%	2952	92.8%	2735	95.3%
		noncoding	1514	85.3%	853	85.3%	797	87.1%	1167	84.6%	1076	87.1%
		Intronic	8212	85.0%	3315	84.3%	3016	86.1%	4905	84.4%	4439	86.8%
		Intergenic	354	91.5%	140	89.7%	127	92.0%	248	90.2%	219	92.0%
	Repetitive non-Alu	Non-synonymous	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
		UTR	2	100.0%	2	40.0%	2	100.0%	6	30.0%	2	100.0%
		noncoding	2	100.0%	1	100.0%	1	100.0%	4	66.7%	1	100.0%
		Intronic	33	100.0%	22	75.9%	18	100.0%	30	53.6%	24	100.0%
		Intergenic	NA	NA	1	100.0%	NA	NA	1	100.0%	NA	NA
	Non-repetitive	Non-synonymous	16	53.3%	25	21.0%	6	66.7%	75	14.9%	12	57.1%
		Synonymous	7	46.7%	12	22.6%	5	83.3%	34	14.2%	7	53.8%
		UTR	97	82.9%	85	25.0%	39	84.8%	240	19.9%	83	84.7%
		noncoding	51	68.9%	41	39.4%	23	79.3%	79	24.8%	36	70.6%
		Intronic	230	82.4%	127	46.0%	105	81.4%	239	32.8%	150	81.5%
		Intergenic	31	91.2%	19	73.1%	18	90.0%	24	46.2%	22	88.0%
			GIREMI (union of results)		**GIREMI (Pooled 3 samples)**		**Multiple data sets method (Pooled 3 samples)**		**Overlap (Pooled 3 samples)**			
	All		26715	89.6%	43547	95.4%	55932	80.2%	43547	95.4%		
All 3 data sets ***	Alu	Non-synonymous	82	97.6%	113	98.3%	144	92.3%	113	98.3%		
		Synonymous	61	93.8%	77	97.5%	88	91.7%	77	97.5%		
		UTR	4809	93.9%	5862	97.6%	7353	91.8%	5862	97.6%		
		noncoding	2307	85.4%	3124	93.8%	3905	84.0%	3124	93.8%		
		Intronic	16129	88.7%	30136	95.3%	38319	86.2%	30136	95.3%		
		Intergenic	635	93.9%	1045	97.4%	1334	90.3%	1045	97.4%		
	Repetitive non-Alu	UTR	15	100.0%	24	96.0%	29	50.9%	24	96.0%		
		noncoding	7	87.5%	6	85.7%	9	40.9%	6	85.7%		
		Intronic	141	99.3%	156	99.4%	224	70.9%	156	99.4%		
	Non-repetitive	Non-synonymous	34	56.7%	46	65.7%	84	15.1%	46	65.7%		
		Synonymous	16	51.6%	13	52.0%	30	11.7%	13	52.0%		
		UTR	188	81.7%	246	77.6%	400	18.2%	246	77.6%		
		noncoding	141	81.5%	206	91.6%	295	40.5%	206	91.6%		
		Intronic	2084	93.3%	2385	95.6%	3567	58.1%	2385	95.6%		
		Intergenic	66	84.6%	108	90.8%	151	25.7%	108	90.8%		

*Multiple data set methods - Overlap of 2 out of 3 data sets: editing sites identified in the 3 RNA-Seq data separately), then for each data set, editing sites were called by requiring their presence in 2 out 3 samples
**Multiple data set methods - Overlap of 2 out of 17 data sets: editing sites identified in the 3 RNA-Seq data separately), then for each data set, editing sites were called by requiring their presence in 2 out 17 samples.
*** GIREMI (union of results): results of GIREMI for the 3 data sets (separately) were combined. Pooled 3 samples - mapped reads of the three data sets were pooled together, then editing sites were identified using the pooled reads with GIREMI or the Multiple data sets methods.

Supplementary Table 4. Identification of recoding sites by GIREMI or mutual information (MI) alone (sample IDs same as in Supplementary Table 5).

Samples	No. of sites predicted by GIREMI	No. of sites as input to GIREMI*	GIREMI Sensitivity (GIREMI_ predicted /input)	No. of sites predicted by MI	No. of sites as input to MI**	MI_predicted/ GIREMI_ Predicted
SRR595926	9	13	69.2%	4	4	44.4%
SRR607679	5	7	71.4%	2	3	40.0%
SRR607839	5	8	62.5%	4	4	80.0%
SRR608456	7	9	77.8%	0	0	0.0%
SRR613627	6	8	75.0%	0	0	0.0%
SRR613747	6	10	60.0%	0	0	0.0%
SRR627449	7	9	77.8%	3	3	42.9%
SRR627451	11	13	84.6%	6	9	54.5%
SRR627455	11	13	84.6%	4	4	36.4%
SRR627462	10	11	90.9%	6	6	60.0%
SRR658573	6	8	75.0%	4	4	66.7%
SRR660933	8	13	61.5%	4	6	50.0%
SRR660969	3	5	60.0%	0	0	0.0%
SRR662162	3	4	75.0%	1	1	33.3%
SRR662233	2	8	25.0%	0	0	0.0%
SRR663681	11	15	73.3%	4	4	36.4%
SRR810319	7	10	70.0%	0	0	0.0%
SRR815232	6	8	75.0%	1	1	16.7%
SRR817751	14	18	77.8%	6	6	42.9%
SRR818033	7	9	77.8%	2	2	28.6%
SRR821690	3	4	75.0%	0	0	0.0%
Total sensitivity	**43**	**47**	**91.5%**	**32**	**34**	**74.4%**
Average per sample sensitivity			**71.4%**			**30.1%**

*"No. of sites as input to GIREMI" refers to the total number of SNVs (required to have ≥5 reads) that were tested by GIREMI (MI followed by GLM) to predict editing sites.

**"No. of sites as input to MI" refers to the number of SNVs that were testable by the MI step, required to have ≥5 reads harboring two SNVs.

Supplementary Table 5. RNA-Seq data obtained from the GTEx project. Sample IDs are SRR followed by the numeric ID shown. Samples in red were excluded from further analysis because of low sequencing coverage/quality.

Subject ID	Cerebellum	Cortex	Frontal Cortex	Hippocampus	Lung	Thyroid	Heart	Skeletal Muscle
N7MS	627451	627455	595926	608456,600724	607839	607679	608096	612839
NPJ8	627462	627449	613627	821690	627457	602951	598148	601695
RU72	613747	NA	612563	NA	614948	614743	612875	615044
T6MN	663681	660933	662233	660969	662162	658573	659637	661639
WWYW	815232	810319	818033	817751	NA	808886	815517	816226

Supplementary Table 6. Number of identified editing sites in the GTEx data sets. (Samples in red were excluded from further analysis because of low sequencing coverage/quality.)

		Alu sites		Non-Alu sites					
Tissue	Data ID	Editing sites	AG, %	Editing sites	AG, %	Total sites	Total AG, %	Raw read pairs (x million)	Uniquely mapped read pairs (x million)
Cerebellum	SRR613747	4578	90.7	260	85	4838	90.4	56	20.1
Cerebellum	SRR627451	4430	98.1	1921	95.9	6351	97.5	34.5	15.6
Cerebellum	SRR627462	7452	97.6	1974	95.3	9426	97.1	36.5	15.4
Cerebellum	SRR663681	11836	89	600	83.4	12436	88.7	48.8	31.1
Cerebellum	SRR815232	13596	87.3	469	80	14065	87	69.1	41.4
Cortex	SRR627449	2848	97.2	1037	95.5	3885	96.8	32.2	15.5
Cortex	SRR627455	3577	97.7	1311	96.5	4888	97.4	43.5	15.7
Cortex	SRR660933	5067	91.9	622	93.6	5689	92.1	54.9	29.9
Cortex	SRR810319	4174	91.1	316	84.2	4490	90.6	65.1	35.4
Frontal cortex	SRR595926	3358	89.8	433	90.1	3791	89.8	61.8	31.4
Frontal cortex	SRR612563	185	95.7	22	90.9	207	95.2	58	5.7
Frontal cortex	SRR613627	4685	90.5	320	78.4	5005	89.8	50.8	26.1
Frontal cortex	SRR662233	4255	91	413	94.4	4668	91.3	72.8	40.8
Frontal cortex	SRR818033	6655	90.6	371	85.7	7026	90.3	65.9	39.3
Hippocampus	SRR600724	390	87.7	49	95.9	439	88.6	104.5	8.1
Hippocampus	SRR608456	3543	89	273	91.9	3816	89.3	56.1	26
Hippocampus	SRR660969	1391	90.7	118	97.5	1509	91.3	58.1	30.9
Hippocampus	SRR817751	7270	90.2	508	87	7778	90	68.5	39.4
Hippocampus	SRR821690	2284	92.6	130	86.2	2414	92.3	94.4	35.7
Heart	SRR598148	1337	95.7	100	88	1437	95.1	53.7	19.4
Heart	SRR608096	2337	96.6	206	89.3	2543	96	58.5	24.2
Heart	SRR612875	1449	97.8	154	87.7	1603	96.8	53.7	17.9
Heart	SRR659637	3190	96.8	356	91	3546	96.2	50.2	26.6
Heart	SRR815517	4136	96.7	295	88.1	4431	96.1	69.1	28.5
Lung	SRR607839	9499	97.6	1346	95.5	10845	97.3	49.8	23.8
Lung	SRR614948	3112	97.5	391	95.4	3503	97.3	38.1	12.3
Lung	SRR627457	6350	99.1	720	97.5	7070	99	42.9	13.2
Lung	SRR662162	6753	96.5	779	95.5	7532	96.4	40.8	23.2
Skeletal muscle	SRR601695	295	91.9	36	66.7	331	89.1	96.2	26.5
Skeletal muscle	SRR612839	965	93.4	111	85.6	1076	92.6	52	23.2
Skeletal muscle	SRR615044	131	83.2	25	52	156	78.2	50.2	15.9
Skeletal muscle	SRR661639	67	88.1	17	58.8	84	82.1	39	9.8
Skeletal muscle	SRR816226	1416	96.4	110	82.7	1526	95.4	60.7	33.3
Thyroid	SRR602951	3282	93.6	271	91.1	3553	93.4	100.3	24.4
Thyroid	SRR607679	4563	95.2	660	92.7	5223	94.8	80.8	29.3
Thyroid	SRR614743	3615	96.5	406	92.1	4021	96	48.9	15.7
Thyroid	SRR658573	18335	94.8	1384	92.2	19719	94.6	55.5	35.7
Thyroid	SRR808886	10878	95.8	840	88.2	11718	95.3	50	32.1

Supplementary Table 7. Gene ontology analysis of genes with tissue-specific editing (TSE). All GO categories shown here were associated with a p value less than 0.0001.

GO ID	GO Description	Tissues with TSE
GO:0045454	cell redox homeostasis	Cortex,Frontal cortex,Thyroid
GO:0070469	respiratory chain	Frontal cortex,Hippocampus
GO:0007059	chromosome segregation	Frontal cortex,Hippocampus
GO:0008635	activation of caspase activity by cytochrome c	Cortex,Frontal cortex
GO:0006309	DNA fragmentation involved in apoptotic nuclear change	Cortex,Frontal cortex
GO:0006366	transcription from RNA polymerase II promoter	Lung,Thyroid
GO:0008233	peptidase activity	Lung,Thyroid
GO:0006465	signal peptide processing	Lung,Thyroid
GO:0004437	inositol or phosphatidylinositol phosphatase activity	Hippocampus
GO:0006120	mitochondrial electron transport, NADH to ubiquinone	Hippocampus
GO:0005747	mitochondrial respiratory chain complex I	Hippocampus
GO:0044237	cellular metabolic process	Hippocampus
GO:0022900	electron transport chain	Hippocampus
GO:0008137	NADH dehydrogenase (ubiquinone) activity	Hippocampus
GO:0009055	electron carrier activity	Cortex
GO:0006916	anti-apoptosis	Cortex
GO:0006479	protein amino acid methylation	Frontal cortex
GO:0045333	cellular respiration	Frontal cortex
GO:0006364	rRNA processing	Thyroid
GO:0006954	inflammatory response	Lung
GO:0007507	heart development	Lung
GO:0008047	enzyme activator activity	Lung
GO:0009615	response to virus	Lung
GO:0006383	transcription from RNA polymerase III promoter	Lung

Supplementary Notes:

Supplementary Note 1 – Predicting RNA editing via mutual information (MI) in GIREMI

The calculation of mutual information (MI) in GIREMI utilizes (pairs of) reads that harbor two SNVs (SNPs, editing or unknown type) and determines their degree of allelic linkage. As shown in Fig. 1b, MI distributions for combinations of SNPs and editing sites (defined using the genome sequencing data) are readily distinguishable. Thus, MI is an effective measure to discriminate editing sites from SNPs. Fig. 1b shows the MI data for SNP pairs, which is the distribution used in GIREMI for predicting editing sites. In contrast, Supplementary Fig. 1a (upper panel) shows the MI values of SNPs relative to any other SNVs in their neighborhood, some of which may be editing sites. Consistently, we observed a minor peak at lower MI range indicating SNP-editing site pairs. It should be noted that this distribution was not used in GIREMI. We also examined the MI values for editing sites and SNPs located in different types of regions (*Alu*, repetitive non-*Alu*, non-repetitive). It can be appreciated that the MI values are generally similar for different types of editing sites, but those in non-*Alu* regions tend to have a small fraction of sites with higher MI values (Supplementary Fig. 1a, lower panel), suggesting that these editing sites may indeed be SNPs. Similarly, the MI distribution of SNPs also has a pronounced (*Alu* SNPs) or minor (non-*Alu* SNPs) peak near the lower MI range, indicating that they are likely paired with editing sites.

It should be noted that the MI step alone does not render an advantage for predicting SNPs (which is not the goal of GIREMI as an RNA editing predictor). In the GM12878 data, if 50% SNPs were assumed to be unknown, 12,359 SNPs were excluded from the MI calculation due to lack of neighboring SNVs or inadequate reads (< 5) covering the SNP and its neighboring SNVs. A total of 4,815 SNPs were included for MI calculation, among which 1,323 SNPs were incorrectly predicted as RNA editing sites based on their MI values. Thus, 3,492 (72.5%) of the 4,815 SNPs remained to be SNPs. If 30% SNPs were assumed to be unknown, 11,298 SNPs were excluded from the MI calculation due to lack of neighboring SNVs or inadequate reads (<5) covering the SNP and its neighboring SNVs. A total of 2,903 SNPs were included for MI calculation, among which 740 SNPs were incorrectly predicted as RNA editing sites based on their MI values. Thus, 2,163 (74.5%) of the 2,903 SNPs remained to be SNPs. Overall, SNP prediction using MI alone is obviously not sensitive, nor very accurate.

In general, the requirement of multiple SNV-containing reads in MI calculation suggests that this step alone may have limited sensitivity in pinpointing editing sites that are in isolation from other editing sites or SNPs. Although the vast majority of A-to-I editing sites are located in *Alu* elements and in close proximity with other editing sites, a relatively small number of editing sites, especially those in coding regions, are located in isolation from others. To complement the MI step, GIREMI includes a second-step based on GLM (Online Methods). Importantly, the predicted editing sites by MI were used as training data to drive the GLM parameter estimation. Thus, the GLM is data set-specific and does not rely on pre-parameterization of the model. Overall, GIREMI has better sensitivity and accuracy than another genome-independent editing prediction method (see Supplementary Note 3).

To better understand the sensitivity of GIREMI for isolated editing sites, we examined its prediction of known recoding sites[1]. The analysis is fully described in Supplementary Note 5. Here, we only elaborate on the identification of these sites by the MI method. Overall, GIREMI (combining MI and GLM steps) has a high sensitivity in predicting these recoding sites (Supplementary Note 5). We then examined how many of the recoding sites were identified by the MI step. Among all GIREMI-predicted recoding sites, 74.4% were identifiable in the MI step in at least one sample (Supplementary Table 4). For each sample, an average of 30% of the identified recoding sites were resulted from the MI step, with the rest predicted by the GLM step. It is expected that the sensitivity to recoding sites in the MI step alone is highly dependent on the genetic background and editome profile of a specific sample. We thus examined the type of mismatches paired with the MI-identified recoding sites (i.e., those harbored in the same pairs of reads as the MI-identified recoding sites). As shown in Supplementary Fig. 1e, the recoding sites often had neighboring editing sites (all located in non-repetitive regions), SNPs or un-determined SNVs to enable MI calculation. Thus, some recoding sites are identifiable by MI due to their proximity to other SNVs. Given the limited length of mRNAs (after intron removal by splicing), it is expected that the sensitivity using the MI calculation alone would be further improved once longer read length and insert size of RNA-Seq libraries become available in the near future.

Supplementary Note 2 – Read mapping and variant calling methods to generate input files for GIREMI

RNA-Seq read mapping is an important first step to generate the necessary input files for GIREMI (i.e., lists of single-nucleotide variants (SNVs) in the reads). For this purpose, different mapping methods can be adopted. For results presented in this paper, we used our previously published mapping strategy that facilitates accurate mapping of reads harboring SNVs[2]. This mapping method reduces errors due to existence of homologous regions in the genome and minimizes mapping bias for the alternative alleles of SNVs in the reads[3]. Importantly, we showed that this stringent mapping approach enables more accurate quantification of editing levels compared to those resulted from nominal mapping methods[3]. To call SNVs from mapped reads, we followed the procedures described in our previous work[2] and implemented a few quality filters that emerged in recent literature of RNA editing analysis[3] (see Online Methods).

GIREMI can also be applied to SNVs identified using alternative read mapping and variant calling methods. As an example, we used another read mapping strategy (BWA[4]) that is often applied in RNA-Seq analysis. In addition, we adopted a popular variant calling method (the GATK tool[5]) to analyze these data. The specific procedures we used were very similar to those described in[6]. As shown in Supplementary Fig. 2, we observed that the false positive rate using the nominal mapping strategy and GATK variant calling was somewhat higher than using our previous method described above, although the difference is not large. This more relaxed mapping and variant calling strategy led to prediction of higher numbers of editing sites.

Overall, it is highly recommended that stringency in read mapping and variant calling is practiced for any methods that predict RNA editing sites in RNA-Seq data, including GIREMI.

Supplementary Note 3 – Comparison of results from GIREMI and other methods

We compared GIREMI-predicted editing sites with those resulted from two other approaches (Supplementary Table 2). The first is the nominal method that utilizes whole-genome sequencing data to distinguish RNA editing from SNPs (the "genome-aware" method). The second approach calls RNA editing using RNA-Seq data from multiple samples[6] (the "multiple data sets" method). It does not necessitate genome sequence data and essentially requires a predicted editing site be present in multiple data sets. To conduct a fair comparison, we used the same read mapping and artifact-filtering procedures as used for GIREMI (Online Methods). In addition, to apply the "multiple data sets" method to the GM12878 data, at least one other RNA-Seq data set must be included. For this purpose, we used another deeply sequenced lymphoblast RNA-Seq data set (YH data) that also has matched genome sequencing data[7]. Two alternative implementations of the "multiple data sets" method were carried out as proposed in the original paper[6]. The first is to call GM12878 editing sites by requiring their presence in the YH data. The second is to pool the reads from the two samples and predict RNA editing sites using the same method as for individual samples. To compare with this data-pooling mode in a fair manner, we simply combined the results of the 2 data sets predicted by GIREMI (note these results were still identified from individual data set). In all analyses, either 30% or 50% of SNPs of the corresponding cell line was assumed to be unknown since almost all SNPs in both cell lines are already included in dbSNP. This procedure enables an unbiased performance evaluation resembling realistic cases where genome data were not available. As shown in Supplementary Table 2, editing sites predicted by GIREMI overlapped considerably with those from the genome-aware method. Results from GIREMI had higher % overlap (relative to genome-aware), %AG and accuracy (calculated as 1 - %SNPs among predicted editing sites) compared with the "multiple data sets" method, especially for editing sites in non-*Alu* regions. Thus, the above evaluations showed that GIREMI outperforms the existing genome-independent method.

As another performance evaluation, we analyzed a panel of primary human brain tissue RNA-Seq data (Supplementary Table 3) using GIREMI and the "multiple data sets" methods. These data sets were obtained from the GTEx database, with their IDs shown in Supplementary Tables 5 and 6. We analyzed 3 cerebellum data sets as an example, which mimics typical individual lab-based projects where a small number of samples were collected at modest depth. Each of the 3 data sets was analyzed separately by GIREMI. However, since the "multiple data sets" method needs more than 1 data set, we conducted three types of comparisons. First, each of the 3 data sets was also analyzed separately by this method. Then, editing sites for each data set were called by requiring their presence in at least one of the other two RNA-Seq data sets. Second, since this method will obviously benefit from availability of a large number of data sets for comparison, we also expanded the number of comparison data sets to 17 (all data sets from sub-regions of brain, Supplementary Table 6). Overall, it can be appreciated that GIREMI largely outperforms the "multiple data sets" method in sensitivity and %AG, although it only uses one data set whereas the latter method uses 3 or 17 data sets. While GIREMI favors A-to-G changes by nature, the facts that the "multiple data sets" method yielded very low %AG in certain regions (e.g., non-repetitive) and sites common to both methods had much higher %AG suggest that the "multiple data sets" method produced limited results. In the third type of evaluation, the 3 data sets were pooled together (a mode of the "multiple data sets" method) to increase statistical power for editing prediction. To enable a fair comparison, GIREMI's results on the 3 data sets were combined. In addition, to be equivalent to the "multiple data sets" method, GIREMI was also applied to the pooled data sets. The latter method ignores the distinction of individual data sets and treats them as a whole. It applies to biological replicates of the same experiment, but is disadvantageous if comparisons of editomes across samples are desirable. In general, the sensitivity of any method should increase with deeper sequencing data, which was observed for both methods. In this case, the inputs to GIREMI were essentially the output of the "multiple data sets" method since they were applied to the same pooled data set (mapped and filtered in exactly the same way) and both excluded public dbSNPs. Thus, GIREMI outputs a subset of the input, but dramatically increases the accuracy of the results (much improved %AG especially for non-*Alu* regions) (Supplementary Table 3). Since the 3 data sets used here had very low sequencing depth (Supplementary Table 6), statistical power was a limiting factor and data pooling rendered an overriding advantage. In contrast, in Supplementary Table 2, the "union of results" of GIREMI already outperformed the data-pooling mode of the "multiple data set" method, largely due to the very high sequencing depth of the GM12878 and YH data sets.

In summary, GIREMI represents a substantial improvement over the existing method, given its higher accuracy, sensitivity and its advantageous applicability to single data set. We recommend applying GIREMI to individual data sets if given high sequencing depth or, even with low sequencing depth, if the data sets represent different types of samples to be compared against each other. In addition, GIREMI can be easily applied to pooled data sets or comparative analysis across multiple samples, to achieve higher sensitivity. For example, if multiple biological replicates are available and sequencing depth is modest, we recommend using GIREMI on pooled data sets combining the replicates.

Supplementary Note 4 – GIREMI performance in different types of genomic regions

We evaluated the accuracy of predicted editing sites in different types of regions: *Alu*, non-*Alu*-repetitive, non-repetitive, synonymous, non-synonymous etc (Supplementary Tables 2 and 3), as it is known that existing methods have very different performance for different types of regions with non-*Alu* sites most challenging to predict[6]. In Supplementary Table 2, we used (1 - % known SNPs among predicted editing sites) to represent accuracy since the genome for GM12878 is known. The accuracy of GIREMI for non-*Alu* sites is lower than *Alu* sites in general. For example, when assuming 30% of GM12878 genomic SNPs were unknown to dbSNP, we had nearly perfect accuracy for *Alu* sites (both coding and non-coding). For non-*Alu* repetitive and non-repetitive regions, GIREMI had an average accuracy of 85% and 75% respectively for non-coding editing sites, but both higher than those of the "multiple data sets" method. The accuracy of GIREMI is reduced if a large fraction (e.g. 50% in Supplementary Table 2) of SNPs of the specific sample is unknown. However, given the rapidly expanding public SNP databases owing to large-scale genome sequencing efforts, it is highly likely that only a minor fraction of SNPs is unknown for a particular sample. The performance of GIREMI in non-*Alu* ***coding*** regions is discussed below.

Supplementary Note 5 – GIREMI performance for coding sites

Current methods for editing identification suffer from low sensitivity and low accuracy for coding sites. On an initial analysis, the accuracy of GIREMI is also low with the average being 28% for synonymous and non-synonymous coding sites in non-repetitive regions (Supplementary Table 2). Nevertheless, this level of accuracy for non-repetitive coding sites is better than that of the "multiple data sets" method.

To further evaluate our method, we asked whether GIREMI could identify known coding sites with high sensitivity. For this purpose, we used a list of high quality recoding sites reported in previous work (Table S1 and S2 of Pinto et al 2014[1]). We focused on the GTEx data sets (human primary tissue RNA-seq, Supplementary Table 5) for this evaluation due to tissue-specificity of the recoding sites. As shown in Supplementary Table 4, we identified a total of 43 (91.5% sensitivity) out of the 47 recoding sites that had at least 5 total reads in ≥ 1 sample. Note that not all recoding sites in[1] were testable in this analysis since they did not have adequate read coverage (≥ 5 as required), possibly due to the relatively low sequencing depth of the GTEx RNA-seq data and/or tissue-specificity of the related genes. For each sample, the sensitivity in detecting recoding sites varies, with an average sensitivity of 71.4%.

Since the sensitivity of GIREMI in detecting known coding sites is high, we examined whether applying an additional filter on non-*Alu* coding sites to retain only known sites could increase the accuracy without compromising sensitivity greatly. For example, among the 66 coding sites in non-repetitive regions predicted by GIREMI in GM12878 data (Supplementary Table 2, 30% of SNPs assumed unknown), 12 were included in databases of known or predicted editing sites[8, 9]. Among these 12 sites, 4 were genomic SNPs. Thus, with a filter to retain only sites in editing databases, the accuracy of GIREMI for non-repetitive coding sites is increased to 67%. If 50% of genomic SNPs were assumed unknown, this accuracy is 80%. The presence of SNPs in predicted sites after applying such a filter is possibly due to false positives existing in the public editing databases themselves. Alternatively, it suggests an interesting observation that SNPs in one sample could be editing sites in another. With this filter, the sensitivity is somewhat reduced. In the above example, 19 of the initially predicted 66 sites were not SNPs, thus likely true editing sites, 8 of which were retained after filtering. Overall, since the number of non-repetitive coding sites is relatively small and their inclusion in public databases seems to be saturating, we recommend directly using GIREMI results for this type of sites if a high sensitivity is desired, but filtering to retain for sites in editing databases to achieve higher accuracy.

Supplementary Note 6 – Variation of editomes across human individuals

As shown in the main text, editing sites common to many individuals were associated with relatively high editing levels (Supplementary Fig. 10b). These data argue against the possibility that these sites are randomly occurring transcriptome innovations. Rather, common editing sites should be associated with certain advantage such that evolution has preserved their prevalence in the population. To this end, two possibilities exist with the first being that these editing sites themselves are adaptive changes in the RNA that render functional fitness. Alternatively, editing at these positions is the consequence (or by-products) of an adaptive function executed, for example, by the ADAR enzymes that are known to have additional roles beyond RNA editing[10]. The level of sequence conservation of regions immediately flanking common editing sites may provide a clue to distinguish the two hypotheses. Higher conservation in such regions is more likely associated with functional significance of the editing sites themselves. In contrast, if ADAR's non-editing function is adaptive and evolutionarily preserved, sequence conservation of the immediate neighborhood of the editing sites themselves may not be high, given the predominant specificity of ADAR enzymes to dsRNA structures instead of RNA sequences. Indeed, the conservation profile of common editing sites is similar to that of non-TSE sites and lower than that of TSEs (Fig. 2b vs. 2d). Thus, it is highly likely that common editing sites in general are not enriched with functional editing sites, although it is important to note that a subset of functional sites, such as the TSEs, does exist. Overall, our data support the hypothesis that many common RNA editing sites are likely by-products of the RNA editing machinery carrying out functions to mediate other aspects of gene expression. Evolutionary selection to preserve the other regulatory functions led to an apparent preservation of the RNA editing sites across the population.

Supplementary References:

1. Pinto, Y., Cohen, H.Y. & Levanon, E.Y. Mammalian conserved ADAR targets comprise only a small fragment of the human editosome. *Genome Biol* **15**, R5 (2014).
2. Bahn, J.H. et al. Accurate identification of A-to-I RNA editing in human by transcriptome sequencing. *Genome Res* **22**, 142-150 (2012).
3. Lee, J.H., Ang, J.K. & Xiao, X. Analysis and design of RNA sequencing experiments for identifying RNA editing and other single-nucleotide variants. *RNA* **19**, 725-732 (2013).
4. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-1760 (2009).
5. McKenna, A. et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20**, 1297-1303 (2010).
6. Ramaswami, G. et al. Identifying RNA editing sites using RNA sequencing data alone. *Nat Methods* **10**, 128-132 (2013).
7. Peng, Z. et al. Comprehensive analysis of RNA-Seq data reveals extensive RNA editing in a human transcriptome. *Nat Biotechnol* **30**, 253-260 (2012).
8. Kiran, A. & Baranov, P.V. DARNED: a DAtabase of RNa EDiting in humans. *Bioinformatics* **26**, 1772-1776 (2010).
9. Ramaswami, G. & Li, J.B. RADAR: a rigorously annotated database of A-to-I RNA editing. *Nucleic Acids Res* **42**, D109-113 (2014).
10. Wang, I.X. et al. ADAR regulates RNA editing, transcript stability, and gene expression. *Cell Rep* **5**, 849-860 (2013).