Supplementary Information

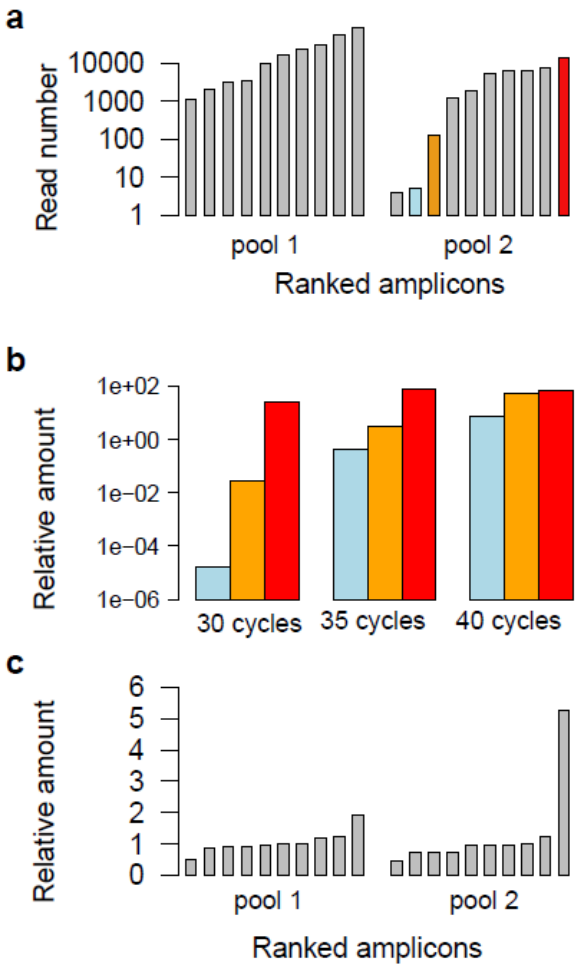
Quantifying RNA allelic ratios by microfluidics-based multiplex PCR and deep sequencing

Rui Zhang[1], Xin Li[2], Gokul Ramaswami[1], Kevin S Smith[2], Gustavo Turecki[3], Stephen B Montgomery[1,2], Jin Billy Li[1]

1. Department of Genetics, Stanford University, Stanford, California, USA.
2. Department of Pathology, Stanford University, Stanford, California, USA.
3. McGill Group for Suicide Studies, Douglas Mental Health University Institute, McGill University, Quebec, Canada.

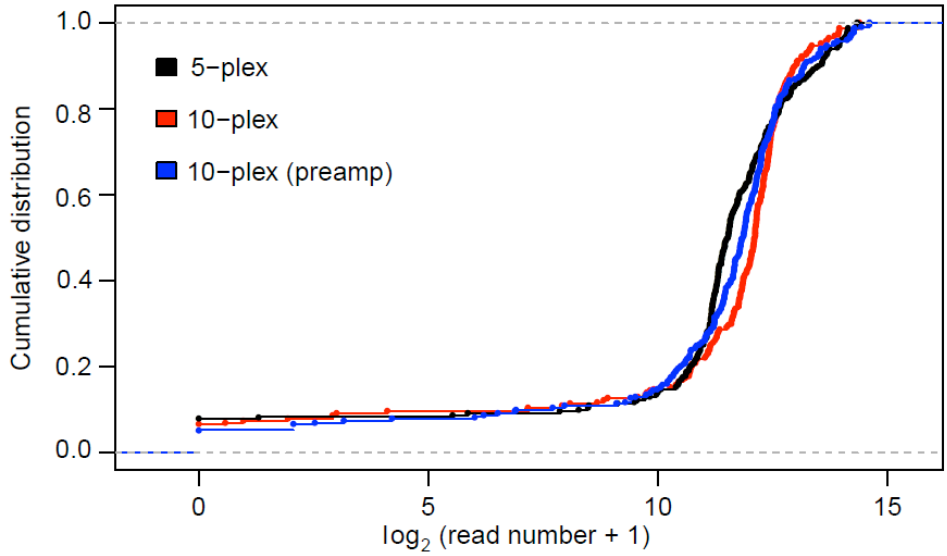
Correspondence should be addressed to J.B.L. (jin.billy.li@stanford.edu) or S.B.M. (smontgom@stanford.edu).

Supplementary Figure 1. The relationship between PCR cycle number and the uniformity of the multiplex PCR products.



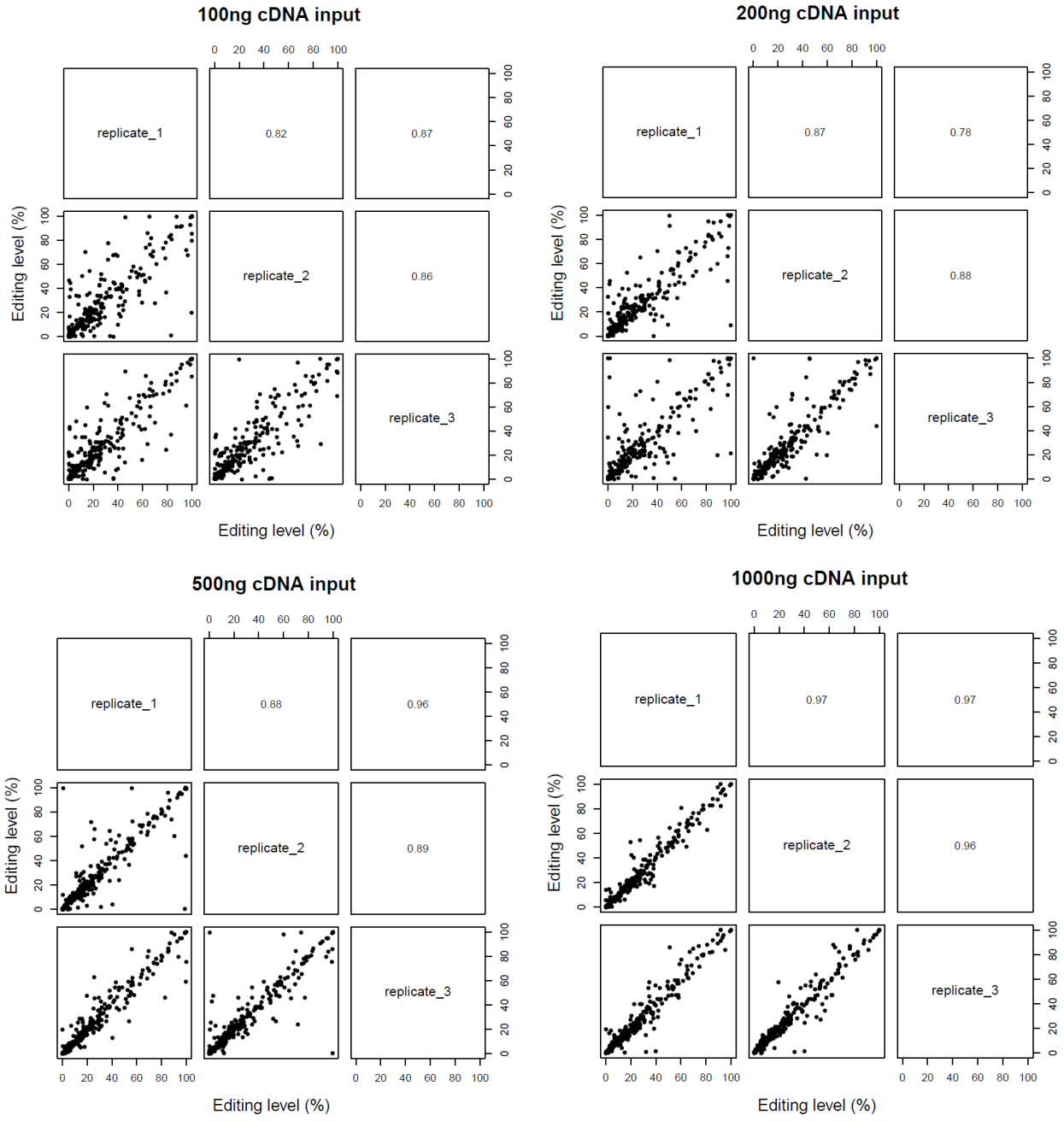
(**a**) The read depth of individual amplicons in a 30 cycle multiplex PCR reaction. The number of reads was measured by deep sequencing of the 20 PCR amplicons using Illumina HiSeq. (**b**) The relative amount of three selected amplicons in PCR reactions with different cycles. The colors in panels **a** and **b** are matched. The quantification of the amount in the multiplex PCR products was described below. In brief, Ct values of the three amplicons (Ct^{a1}, Ct^{b1}, Ct^{c1}) were obtained using multiplex PCR product as the templates. To compare the amount between different amplicons, control Ct values (Ct^{a2}, Ct^{b2}, Ct^{c2}) were obtained using equally mixed individual amplicons as the templates. Finally, the values Ct^{a1}-Ct^{a2}, Ct^{b1}-Ct^{b2}, Ct^{c1}-Ct^{c2} were compared. (**c**) The relative amount of all 20 amplicons in the 40 cycle multiplex PCR. Data were obtained using the same method described in panel **b**. Note that, in panel **c**, the scale of y-axis is linear, unlike the exponential scale in panels **a** and **b**. For pool 1, all amplicons were within 4 fold; for pool 2, all amplicons were within 12 fold.

Supplementary Figure 2. The uniformity of 240 RNA editing amplicons.



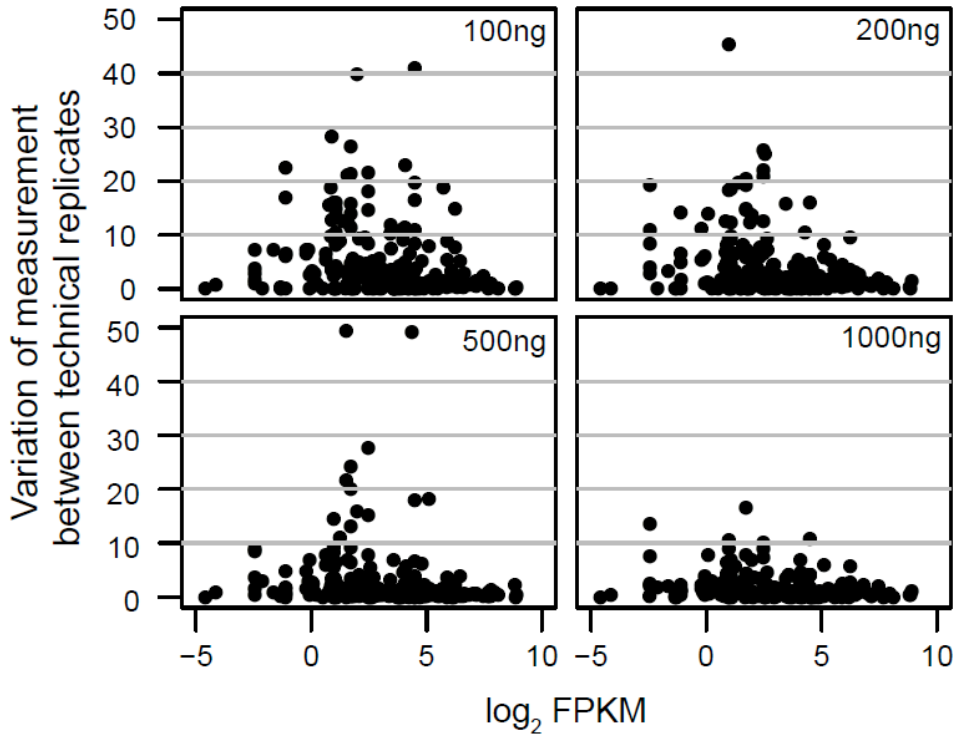
Read numbers are normalized to 1 million reads per sample. For 5-plex and 10-plex results, data from technical replicate 1 of 1 ug cDNA of the HBRR sample were shown here as a representative. For 10-plex preamplification result, data from technical replicate 1 of 50 ng cDNA of the HBRR sample was shown as a representative.

Supplementary Figure 3. The relationship between reproducibility of editing level measurement and the amount of input cDNA.



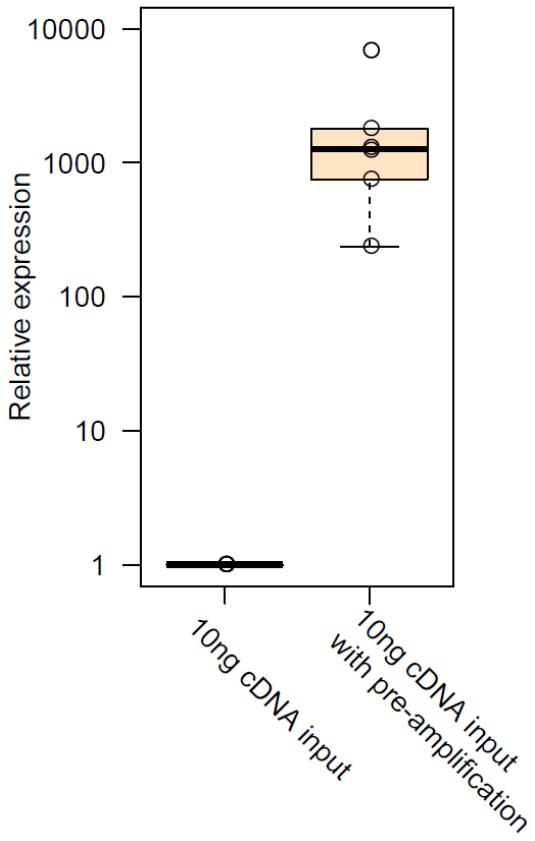
For each amount of input HBRR cDNA (indicated on top), we carried out three technical replicates. Pairwise comparison of editing levels was shown, with the Pearson coefficient also shown above the diagonal.

Supplementary Figure 4. The relationship between gene expression levels and the variation of RNA editing level measurement.



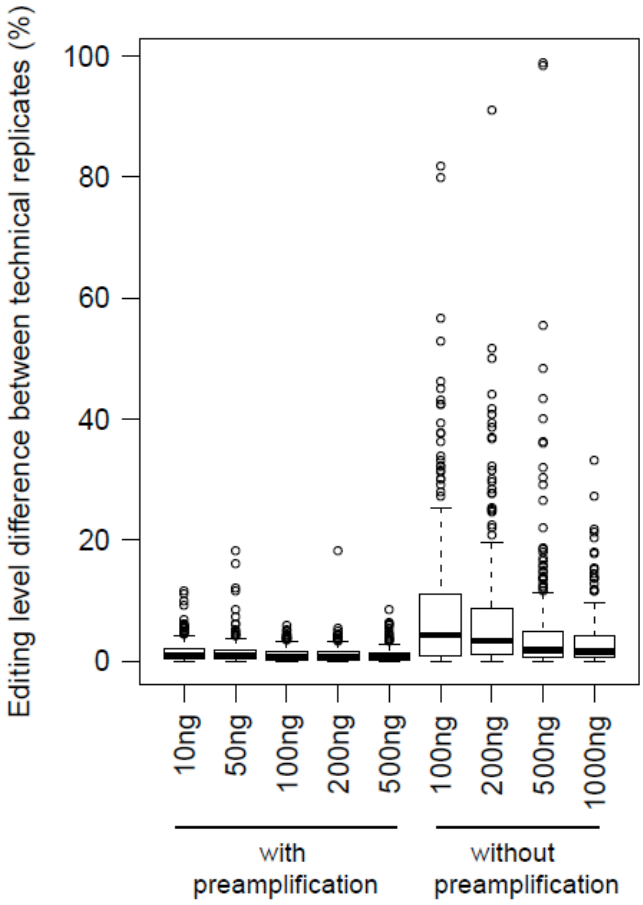
The x-axis is the expression levels of the genes (log2 FPKM) in which the RNA editing sites are located. The y-axis is the variation of measurement which represents by the difference of editing level measurement between two technical replicates shown in **Figure 1d**. Amounts of input cDNA are shown at the top right in each panel.

Supplementary Figure 5. Fold of amplication after the pre-amplification.



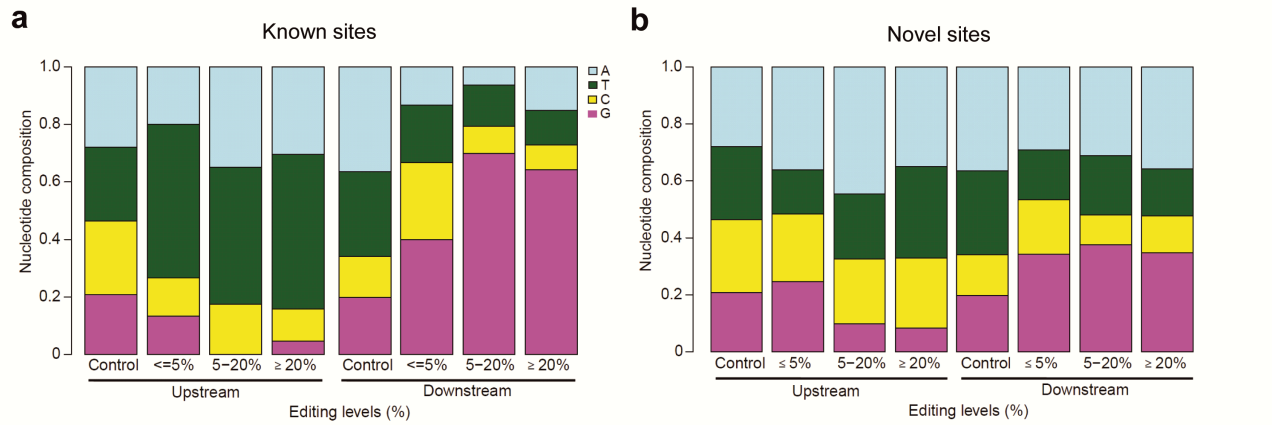
We pre-amplified 10 ng of input cDNA, and quantified the fold of amplification for the desired amplicons. We randomly selected 6 genes and carried out quantitative real time PCR using templates without (left) and with (right) pre-amplification. Each point represents a mean of two duplicate runs.

Supplementary Figure 6. The boxplot of the measurement variances between technical replicates.



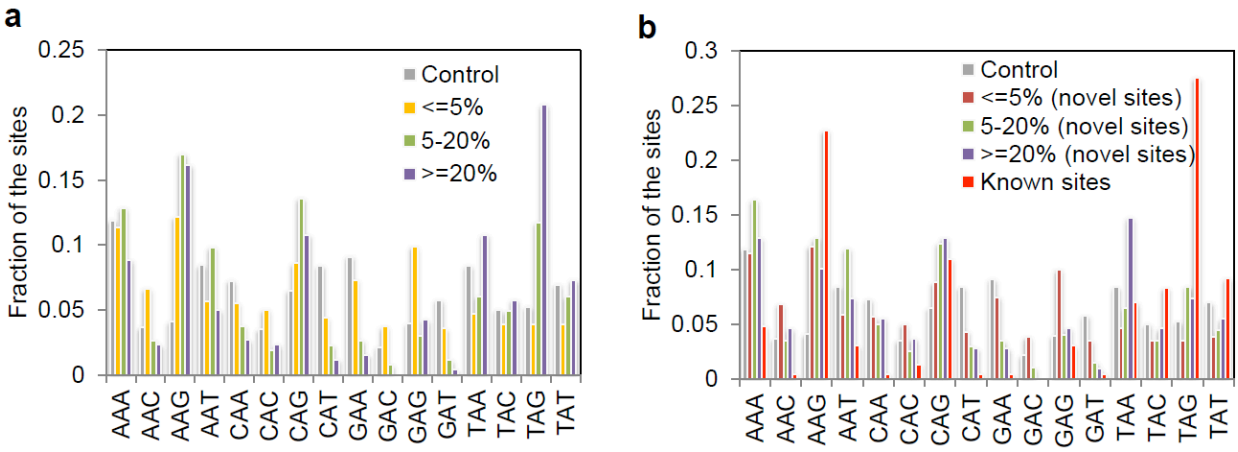
For samples without preamplification, we used the two technical replicates shown in Figure 1d to calculate the variance. The amounts of input cDNA are indicated on the x-axis.

Supplementary Figure 7. Nucleotide composition in positions immediately upstream and downstream of the editing sites.



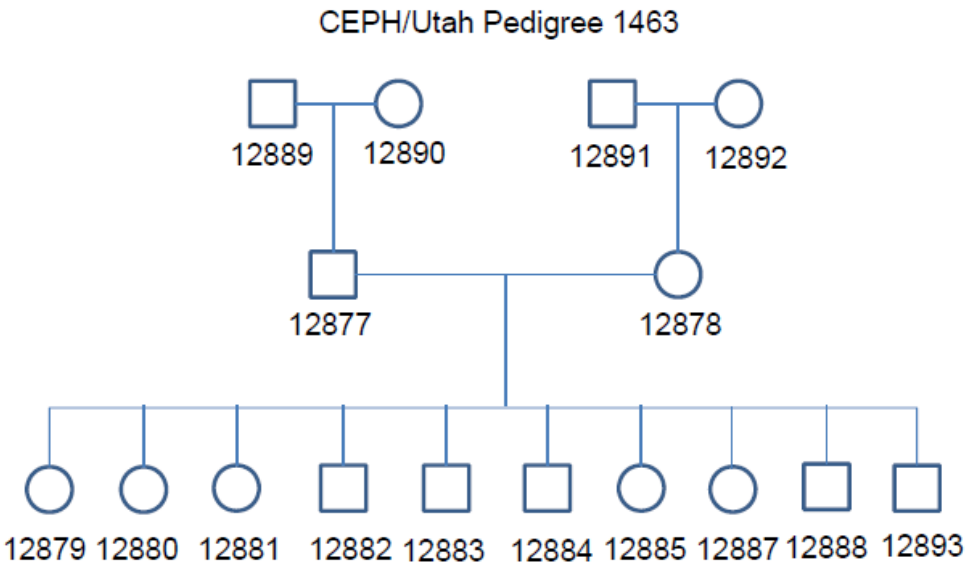
(**a**) known editing sites or (**b**) novel editing sites. The fractions of A, T, C, and G were shown for sites edited at different levels (≤5%, 5-20%, and ≥20%). The control consists of all "A" nucleotides that are covered by mmPCR-seq reads and not edited in any samples tested in this work.

Supplementary Figure 8. Distribution of triplet nucleotide centering the edited adenosine.



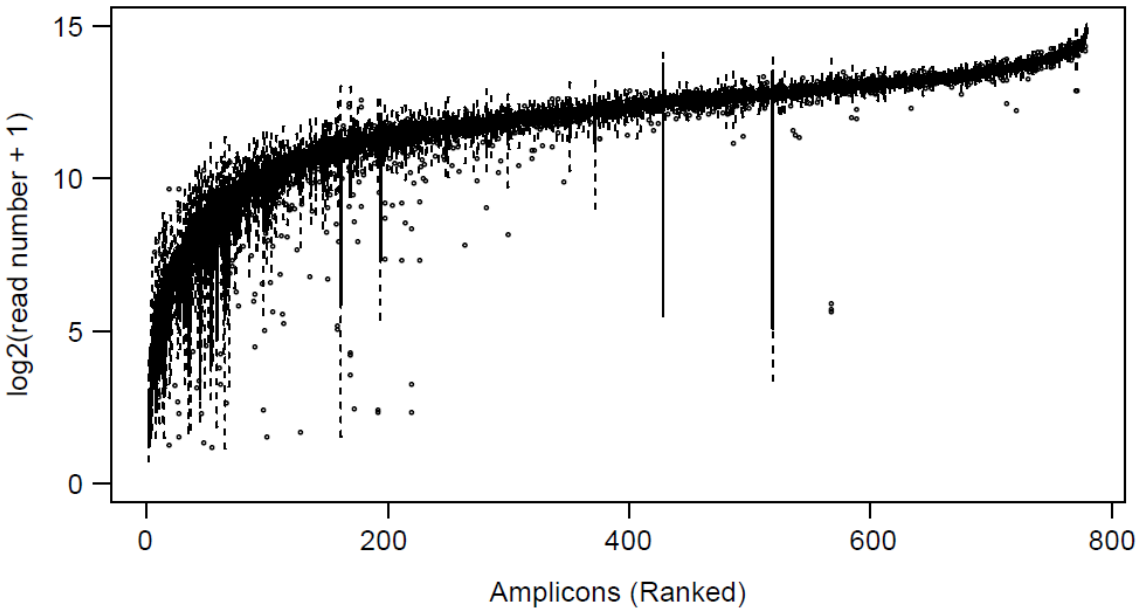
(**a**) All RNA editing sites (known and novel) are separated in three groups based on the editing levels (indicated in different colors). The fractions of sites centered in each of 16 different triplets (NAN) are plotted. (**b**) RNA editing sites are separated into known site and novel site groups. Novel sites are further separated in three groups based on the editing levels. Known sites are not separated based on the editing levels because of the small total number.

Supplementary Figure 9. Family structure of the 16 individuals used in this study.



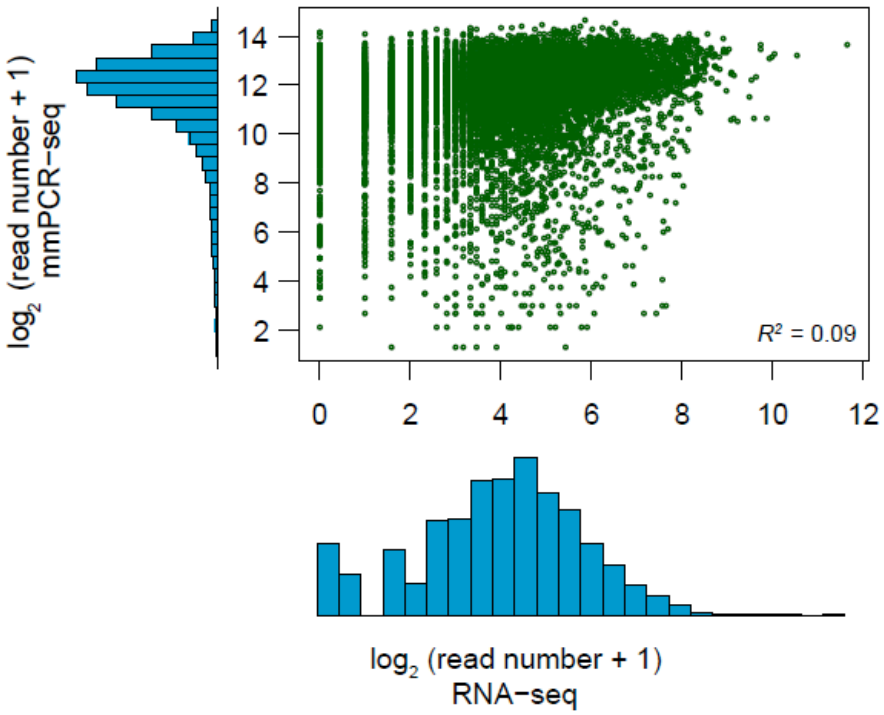
Square symbol represents male and round symbol represents female.

Supplementary Figure 10. Read depth distribution of the ASE amplicons.



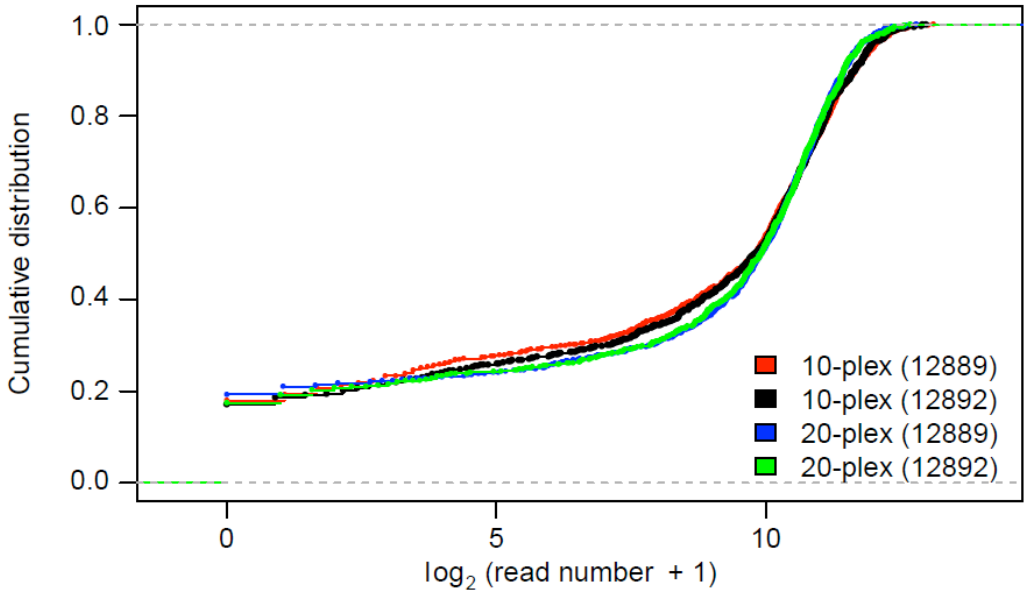
We plotted the read depths of 787 sites detected in all 16 samples using the HiSeq data. Read numbers of amplicons were normalized to 4.7 million mapped reads per sample (the average number of mapped reads per sample).

Supplementary Figure 11. Comparison of the ASE site read depth between mmPCR-seq and RNA-seq.



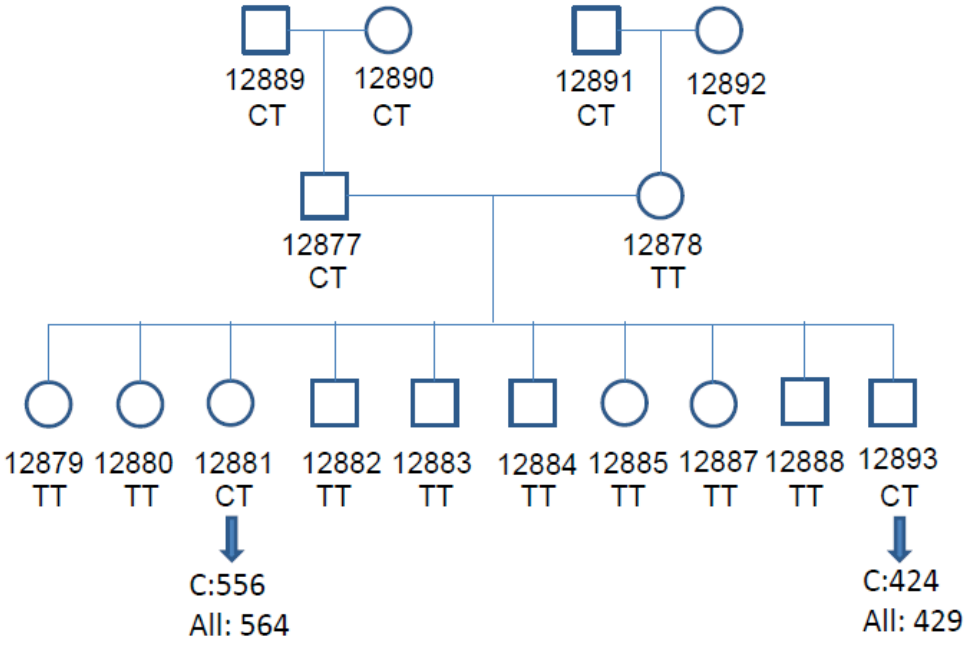
The 787 ASE sites detected in all 16 samples in the mmPCR-seq assay were used.

Supplementary Figure 12. The uniformity of 960 ASE amplicons.



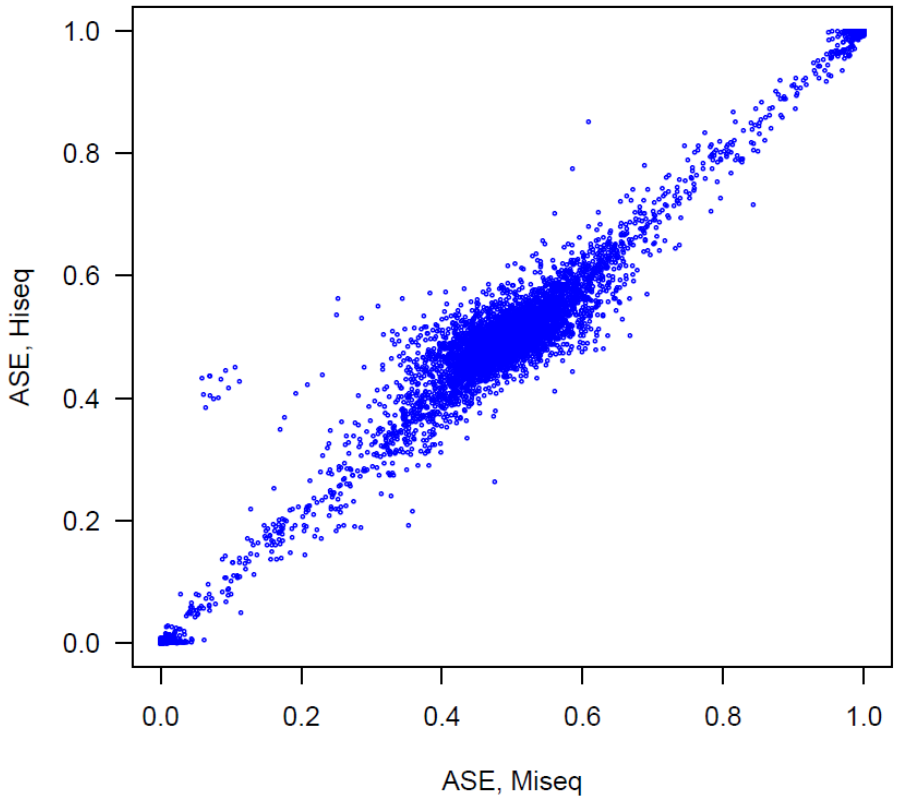
We carried out mmPCR-seq assays for 960 ASE sites with 10- and 20-plex primer sets. Data from individuals 12889 and 12892 are used as representative examples.

Supplementary Figure 13. Monoallelic expression of paternal imprinted gene *ZDBF2*.



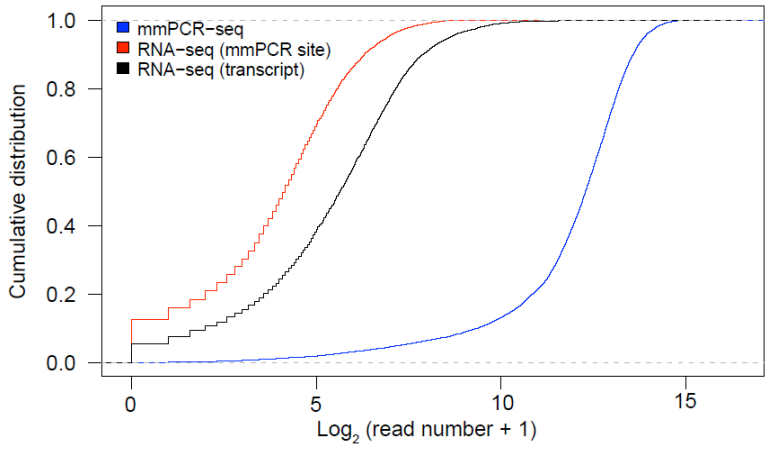
Genotype for SNP rs3732084 of each individual is indicated. For NA12881 and NA12893, the C allele is of paternal origin. The read number of C allele and the number of reads covered this SNP (All) are shown.

Supplementary Figure 14. ASE level comparison between MiSeq and HiSeq data.



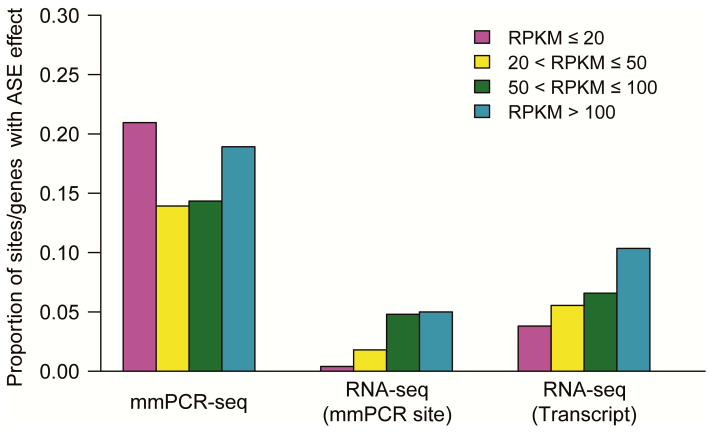
Sites with >=100 reads in both MiSeq and HiSeq are used. The Pearson correlation coefficient (*R*) is 0.995 for all sites and 0.956 for heterozygous sites.

Supplementary Figure 15. The cumulative distribution of read depths from mmPCR-seq or RNA-seq.



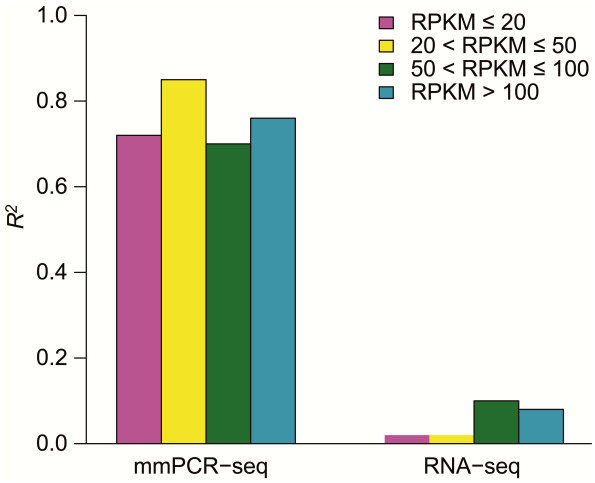
RNA-seq (mmPCR site): the matched sites of mmPCR-seq; RNA-seq (transcript): all heterozygous SNPs in the same gene were combined to count coverage.

Supplementary Figure 16. Proportion of sites/genes with ASE effect among all heterozygous sites using mmPCR-seq or RNA-seq.



RNA-seq (mmPCR site): the matched sites of mmPCR-seq. RNA-seq (Transcript): all heterozygous SNPs in the same gene were used to call ASE.

Supplementary Figure 17. Correlation of ASE effect of IBD siblings among all heterozygous sites for genes with various expression levels.



The matched sites obtained from mmPCR-seq and RNA-seq data were used for analysis**.** Pearson correlation coefficient R^2 reflects degree of correlation between ASE effects among IBD siblings.

Supplementary Table 1. Summary of RNA editing mmPCR-seq experiments

Samples	Primer set (#-plex)	Input (ng)	Number of mapped reads	Percentage of mapped reads
HBRR samples				
HBRR_100ng_1	5	100	1293634	89.97%
HBRR_100ng_2	5	100	915540	93.09%
HBRR_100ng_3	5	100	856707	91.86%
HBRR_200ng_1	5	200	677486	91.94%
HBRR_200ng_2	5	200	750272	90.58%
HBRR_200ng_3	5	200	820813	91.66%
HBRR_500ng_1	5	500	659497	89.47%
HBRR_500ng_2	5	500	715406	91.82%
HBRR_500ng_3	5	500	662805	89.15%
HBRR_1000ng_1	5	1000	850199	90.97%
HBRR_1000ng_2	5	1000	686362	90.31%
HBRR_1000ng_3	5	1000	870546	90.51%
HBRR_100ng_10plex_1	10	100	801112	87.95%
HBRR_100ng_10plex_2	10	100	760515	90.25%
HBRR_200ng_10plex_1	10	200	682542	88.14%
HBRR_200ng_10plex_2	10	200	948807	86.98%
HBRR_500ng_10plex_1	10	500	894307	86.08%
HBRR_500ng_10plex_2	10	500	1045310	86.03%
HBRR_1000ng_10plex_1	10	1000	831820	87.29%
HBRR_1000ng_10plex_2	10	1000	353969	85.71%
HBRR samples with preamplification				
HBRR_preamp_10plex_10ng_1	10	10	334192	39.76%
HBRR_preamp_10plex_10ng_2	10	10	408657	39.75%
HBRR_preamp_10plex_50ng_1	10	50	631706	52.46%
HBRR_preamp_10plex_50ng_2	10	50	505635	52.77%
HBRR_preamp_10plex_100ng_1	10	100	682574	59.90%
HBRR_preamp_10plex_100ng_2	10	100	522964	59.73%
HBRR_preamp_10plex_200ng_1	10	200	596956	54.11%
HBRR_preamp_10plex_200ng_2	10	200	548071	59.18%
HBRR_preamp_10plex_500ng_1	10	500	484869	63.64%
HBRR_preamp_10plex_500ng_2	10	500	720013	65.64%
BA44 region samples				
S6_1	5	1000	513104	86.85%
S6_2	5	1000	480264	81.96%
S7_1	5	1000	735054	88.18%
S7_2	5	1000	734127	88.65%
S6_DNA_1	5	100	943979	96.90%
S6_DNA_2	5	100	725631	97.01%

S7_DNA_1	5	100	959250	96.78%
S7_DNA_2	5	100	988109	97.00%
S1_1	5	1000	692896	91.84%
S1_2	5	1000	713624	90.55%
S2_1	5	1000	739825	91.93%
S2_2	5	1000	791565	91.99%
S3_1	5	1000	566816	85.66%
S3_2	5	1000	1036549	86.04%
S4_1	5	1000	690363	87.85%
S4_2	5	1000	1109034	87.91%
S5_1	5	1000	703714	92.03%
S5_2	5	1000	464735	92.07%
BA44 region samples with preamplification				
S1_preamp_10plex_1	10	200	477381	46.07%
S1_preamp_10plex_2	10	200	378821	48.00%
S2_preamp_10plex_1	10	200	535191	57.17%
S2_preamp_10plex_2	10	200	575386	53.68%
S3_preamp_10plex_1	10	200	393846	40.94%
S3_preamp_10plex_2	10	200	389237	41.79%
S4_preamp_10plex_1	10	200	551721	54.08%
S4_preamp_10plex_2	10	200	533782	56.55%
S5_preamp_10plex_1	10	200	820920	64.58%
S5_preamp_10plex_2	10	200	393666	63.22%

Supplementary Table 2. Accuracy of mmPCR-seq using premixed templates with known allelic frequencies

Expected allelic frequency (%)	Measured allelic frequency (%)[a]					
	Copies of template per PCR reaction[b]					
	50	100	200	1000	5000	25000
1	2.9 (1.8)	2 (1.2)	1.6 (0.83)	1.4 (0.3)	1.5 (0.3)	1.1 (0.6)
2	2.6 (2.1)	3.1 (2.4)	2.3 (0.5)	2.3 (0.8)	2.8 (1.2)	2.5 (0.6)
5	6.3 (1.2)	6.9 (1.1)	6.6 (2.1)	6.1 (1.0)	6.4 (1.1)	6.5 (1.1)
10	11.8 (4.7)	9.3 (5.1)	9.2 (2.9)	8.7 (1.9)	9.2 (1.8)	9(1.8)
20	21 (4.2)	20 (4.2)	18.9 (2.5)	19.1 (2.1)	19 (1.6)	19 (2.1)
30	31 (7.8)	26 (4.2)	27 (4.4)	26.5 (1.6)	26.4 (3.3)	27.8 (3.5)
40	36 (4.5)	39 (4.3)	37.5 (6)	37.1 (3.1)	37.9 (2.8)	38 (4)

[a] A total number of 6 sites are used to calculate the average measured allelic frequency. Standard deviation was shown in parentheses.

[b] The number of template copies is an estimated amount in each of the 48 Fluidigm PCR reactions for each sample. This is based on the estimation that ~20% of the PCR template is loaded to the Fluidigm PCR reactors and the remaining 80% are lost in the microfluidic plumbing system (see **Supplementary Note 2**).

Supplementary Table 3. Summary of the optimized parameters of mmPCR experiments

	High quality RNA sample	Low quality RNA sample*	Low quantity RNA sample*
Optimized cDNA input (ng)	≥1000	≥200	≥10
Preamplification cycle number	Not applicable	15	15
mmPCR cycle number	40	40	40

***** Preamplification was performed in a 10 ul reaction. After preamplification, ~5% of the purified product was used for the microfluidic multiplex PCR.

Supplementary Table 4. Summary of ASE mmPCR-seq experiments

Samples	Number of mapped reads	Percentage of mapped reads
HiSeq data		
S12877_1	1736297	79.84%
S12877_2	1417265	80.31%
S12877_3	1311449	80.19%
S12878_1	1791180	83.28%
S12878_2	1715462	81.42%
S12878_3	2409774	82.81%
S12879_1	1604727	83.91%
S12879_2	1616473	82.84%
S12879_3	1039752	83.57%
S12880_1	2502430	86.93%
S12880_2	2099558	84.95%
S12880_3	1862717	85.60%
S12881_1	2255142	84.08%
S12881_2	1754243	84.35%
S12881_3	2409274	85.42%
S12882_1	935904	81.67%
S12882_2	1770729	85.02%
S12882_3	1238432	83.31%
S12883_1	1290471	84.70%
S12883_2	1344839	84.54%
S12883_3	1303425	86.36%
S12884_1	1793038	86.28%
S12884_2	2069251	83.93%
S12884_3	1112353	85.13%
S12885_1	1640584	82.00%
S12885_2	1468612	82.16%
S12885_3	1275707	82.61%
S12887_1	1383957	81.99%
S12887_2	1519777	84.75%
S12887_3	1051283	82.42%
S12888_1	1785568	80.68%
S12888_2	1279209	83.09%
S12888_3	1203664	81.63%
S12889_1	2477610	87.03%
S12889_2	1121695	89.67%
S12889_3	1304665	87.02%
S12890_1	2079761	85.61%
S12890_2	1989814	87.52%
S12890_3	727607	85.88%
S12891_1	1887668	87.06%

S12891_2	1549227	89.00%
S12891_3	1310679	89.42%
S12892_1	1748633	86.29%
S12892_2	1195593	89.94%
S12892_3	1266617	87.23%
S12893_1	2249913	80.40%
S12893_2	1298565	79.25%
S12893_3	1049085	76.52%
MiSeq data		
S12877_1	200887	76.95%
S12877_2	147457	78.00%
S12877_3	144380	77.48%
S12878_1	212051	80.85%
S12878_2	182966	79.04%
S12878_3	243273	79.53%
S12879_1	157779	80.66%
S12879_2	172628	79.35%
S12879_3	118887	81.13%
S12880_1	288136	84.44%
S12880_2	209345	82.37%
S12880_3	205348	82.67%
S12881_1	260088	80.81%
S12881_2	178158	81.73%
S12881_3	257531	82.53%
S12882_1	110186	78.67%
S12882_2	181771	82.21%
S12882_3	140106	80.32%
S12883_1	146176	82.09%
S12883_2	141450	83.11%
S12883_3	140432	83.58%
S12884_1	205524	83.30%
S12884_2	225633	81.46%
S12884_3	121727	82.05%
S12885_1	187049	80.18%
S12885_2	165257	80.11%
S12885_3	142800	80.08%
S12887_1	162016	79.59%
S12887_2	152056	81.66%
S12887_3	123592	79.56%
S12888_1	184002	78.01%
S12888_2	140713	80.40%
S12888_3	132084	78.29%
S12889_1	286529	83.69%
S12889_2	117567	86.76%

S12889_3	150769	84.16%
S12890_1	237965	82.81%
S12890_2	218135	84.08%
S12890_3	80016	82.88%
S12891_1	213352	84.07%
S12891_2	155201	86.46%
S12891_3	136504	86.25%
S12892_1	202319	83.15%
S12892_2	126366	87.30%
S12892_3	132433	84.09%
S12893_1	258642	77.62%
S12893_2	141286	75.61%
S12893_3	111941	73.63%

Supplementary Table 5. Reproducibility of ASE levels between technical replicates

	R value (All sites)			R value (Heterozygous sites)		
Samples	**Replicate 1 vs Replicate 2**	**Replicate 1 vs Replicate 3**	**Replicate 2 vs Replicate 3**	**Replicate 1 vs Replicate 2**	**Replicate 1 vs Replicate 3**	**Replicate 2 vs Replicate 3**
S12877	0.99	0.99	1.00	0.96	0.96	0.97
S12878	1.00	1.00	0.99	0.96	0.96	0.95
S12879	1.00	1.00	1.00	0.97	0.97	0.97
S12880	1.00	0.99	1.00	0.97	0.94	0.96
S12881	0.99	1.00	0.99	0.93	0.95	0.93
S12882	0.99	0.99	0.99	0.94	0.95	0.95
S12883	0.99	0.99	0.99	0.94	0.93	0.95
S12884	0.99	0.99	0.99	0.94	0.92	0.91
S12885	0.99	1.00	1.00	0.94	0.96	0.96
S12887	1.00	1.00	1.00	0.97	0.95	0.97
S12888	1.00	0.99	1.00	0.97	0.93	0.96
S12889	0.98	0.98	0.98	0.84	0.88	0.89
S12890	0.99	0.99	0.99	0.92	0.89	0.93
S12891	0.99	0.99	0.99	0.91	0.94	0.92
S12892	0.98	0.98	0.98	0.80	0.85	0.85
S12893	0.99	0.99	1.00	0.92	0.93	0.97

The HiSeq data were used in the calculation. The Pearson correlation coefficient (R) is indicated.

Supplementary Table 6. Cost and time comparisons between mmPCR-seq and RNA-seq

mmPCR-seq							
		Library preparation (per sample)			Illumina		
	Primer[a]	cDNA prep[b]	Fluidigm assay[c]	barcode PCR	sequencing per sample[d]	Total cost	Cost per sample
mmPCR-seq (480 loci x 48 samples)	$2,016	$4	$10	$1	$4.5	$2,952	$62
mmPCR-seq (480 loci x 96 samples)	$2,016	$4	$10	$1	$4.5	$3,888	$41
mmPCR-seq (480 loci x 480 samples)	$2,016	$4	$10	$1	$4.5	$11,376	$24
mmPCR-seq (960 loci x 48 samples)	$4,032	$4	$10	$1	$9	$5,184	$108
mmPCR-seq (960 loci x 96 samples)	$4,032	$4	$10	$1	$9	$6,336	$66
mmPCR-seq (960 loci x 480 samples)	$4,032	$4	$10	$1	$9	$15,552	$32
Time of library preparation (96 samples)		2 hours	8 hours	2 hour			
Time of library preparation per sample			7.5 minutes				

RNA-seq					
	Primer	Library preparation (per sample)[e]	Illumina sequencing[f]	Total cost	Cost per sample
RNA-seq (4 samples)	$0	$60	$2,100 (1 lane)	$2,340	$585
RNA-seq (48 samples)	$0	$60	$25,200 (12 lanes)	$28,080	$585
Time of library preparation (4 samples)		12 hours			
Time of library preparation per sample		3 hours			

[a] based on a pair of 35 nt primers for each site, at 6 cents per base (quoted by Invitrogen).

[b] cDNA synthesis using iScript Advanced Kit with higher yield (quoted by Bio-Rad); including the cost of purification.

[c] based on the cost of Fluidigm chip at $450 and reagents at $30 for one chip of 48 samples ($480 total for 48 samples; $10 per sample).

[d] based on Illumina HiSeq 1x101 cycle cost at Stanford sequencing core faciltiy (~150 million reads, $1,400 per lane) to achieve an average of ~1,000 reads/site/sample; pooling barcoded samples from other experiments may be needed.

[e] based on Illumina TruSeq™ RNA Sample Prep Kit v2 (48rxn).

[f] based on Illumina HiSeq paired end 2x101 cycle cost at Stanford sequencing core Facility (~150 million paired end reads, $2,100 per lane); 4 barcoded samples are pooled in the same lane.

Supplementary Table 7. Datasets used for human RNA editing site collection

Resources	Read numbers	Read lengths (bp)	References
ERR030890	64,313,204	75	1
ERR030882	73,513,047	50	2
SRR309262	62,276,114	76	2
SRX083169	54,181,963	76	3
SRX083170	50,989,533	76	3
SRX083171	52,560,671	76	3
SRX083168	47,240,074	76	3
SRX083166	47,806,095	76	3
SRX083167	46,627,463	76	3
SRR085471	15,256,752	36	4
SRR085474	15,772,947	36	4
SRR085725	13,442,077	36	4
SRR087416	14,720,816	36	4
SRR085726	15,228,832	36	4
SRR085473	14,227,702	36	4

1. Human BodyMap 2.0 data from Illumina
2. Cabili, M.N. et al. Genes Dev 25, 1915-27 (2011).
3. Voineagu, I. et al. Nature 474, 380-4 (2011)
4. N. A. Twine, K. Janitz, M. R. Wilkins et al., PLoS One 6 (1), e16266 (2011).

Supplementary Table 8. Primer sequences used for barcode PCR and sequencing

Names	**Sequences**
Barcode forward primer	AATGATACGGCGACCACCGAGATCTACACCCTACACGAGCGTTATCGAGGTC
Barcode reverse primer	CAAGCAGAAGACGGCATACGAGAT(barcode)GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
Custom sequencing primer (Read 1)	CCACCGAGATCTACACCCTACACGAGCGTTATCGAGGTC

Barcode sequences (8 bases long) follow instructions from Fluidigm.

Supplementary Table 9. RNA-seq data for the HBRR sample

Resource	Read number	Read length (bp)	Reference
SRR299026	53,238,798	35	1
SRR299028	58,578,322	100	1
SRR035678	11,712,885	35	2
SRR037439	11,413,794	35	2
SRR037440	11,816,021	35	2
SRR037441	11,244,980	35	2
SRR037442	12,081,324	35	2
SRR037443	11,365,146	35	2
SRR037444	11,616,331	35	2
SRR039628	8,201,143	50	3
SRR039629	8,201,143	50	3
SRR039630	7,760,089	50	3
SRR039631	7,760,089	50	3
SRR039632	7,450,994	50	3
SRR039633	7,450,994	50	3
SRR037452	11,712,885	35	3
SRR037453	11,413,794	35	3
SRR037454	11,816,021	35	3
SRR037455	11,244,980	35	3
SRR037456	12,081,324	35	3
SRR037457	11,365,146	35	3
SRR037458	11,616,331	35	3

1. G. Chen, K. Yin, L. Shi et al., PLoS One 6 (11), e28318 (2011).
2. J. H. Bullard, E. Purdom, K. D. Hansen et al., BMC bioinformatics 11, 94 (2010).
3. K. F. Au, H. Jiang, L. Lin et al., Nucleic acids research 38 (14), 4570 (2010).

Supplementary Note 1. Tuning PCR cycles for multiplex PCR using cDNA templates

We first examined the amplification uniformity using 30-cycle PCR for two pools of 10-plex primers. We chose 30 cycles following the manufacture's recommendation for the multiplex PCR polymerase we used (KAPA 2G, from Kapa Biosystems, Woburn, MA). We then added the adaptor sequences used by the Illumina sequencers to the PCR products via a second round of PCR and performed deep sequencing. An uneven amplification of fragments was observed: the read depth of different amplicons ranges from 1089 to 87513 in pool 1 and ranges from 5 to 97421 in pool 2 (**Supplementary Figure 1a**).

The recommendation from Kapa Biosystems is for using genomic DNA as the PCR template. Compared to genomic DNA samples which have equal starting material for different targeted regions, cDNA samples have a wide range of starting material for different targeted regions due to the wide spectrum of gene expression. Therefore, we examined the uniformity with higher PCR cycle numbers. We performed multiplex PCR with 30, 35, and 40 cycles. We used real-time PCR to quantify the uniformity of different amplicons, using the multiplex PCR product as the template. We first quantified 3 amplicons from pool 2, which had 7, 341 and 97421 reads in the 30-cycle deep sequencing experiment (**Supplementary Figure 1a**). A positive correlation between uniformity and PCR cycles was observed (**Supplementary Figure 1b**). We quantified all 20 amplicons from the 40-cycle PCR products and found more uniform amplification of all target fragments (**Supplementary Figure 1c**).

Supplementary Note 2. **Estimation of the quantity of cDNA template in the Fluidigm microfluidic reactions**

In the Fluidigm Access Array, it is estimated that ~20% of the PCR template is loaded into the 48 PCR reactors for each sample, and the remaining 80% is lost in the microfluidic plumbing system according to Fluidigm technical support. When we start with 1 ug of cDNA input for a sample, ~200 ng will be loaded to the 48 PCR reactors. Therefore, for each PCR reaction, ~4 ng of cDNA template was used.

Assuming that a cell has ~20 pg of total RNA, we would need 200 cells to achieve 4 ng of total RNA (which will be converted to 4 ng of cDNA assuming 100% efficiency in reverse transcription – we typically achieve >70% efficiency using Bio-Rad iScript). With 200 cells, even when a gene is transcribed into a single transcript per cell (which is considered to be extremely lowly expressed), we would have 200 transcripts for the gene.

The estimation above suggests why 1 ug of cDNAs is needed for an accurate measurement of allelic ratios for all sites, located in either highly or lowly expressed genes.

Supplementary Note 3. Determining the threshold of variant frequency to distinguish RNA editing events from sequencing errors

We found a large number of potential RNA editing sites surrounding the known sites using the deep sequencing data from mmPCR-seq. Many of them have low variant frequency, thus making it difficult to distinguish authentic RNA editing events from sequencing errors. To

determine the minimum variant frequency needed to distinguish real RNA editing events from sequencing errors in our platform, we carried out the following three analyses.

First, we analyzed two RNA samples with matched genomic DNA samples on the 240 editing site loci. We compared the density of A-to-G/T-to-C variants (defined as the number of the variants per 10 kb) observed between RNA and matched DNA (enrichment score) with different frequency cutoffs. As expected, the enrichment score increases with higher threshold of variant frequency (**Figure 2a**). With a 1.1% cutoff, we achieved an 11.6 fold enrichment, which would lead to an estimated false discovery rate of 8.6% for A-to-I editing events. Second, 88% of the RNA variants with a minimum of 1.1% level are A-to-G mismatches when annotated by known gene models, indicating A-to-I editing. Third, we examined 694 A-to-G mismatches with ≥1.1% level present in both RNA samples and DNA samples. Almost all (98.6%) of the sites have significantly higher A-to-G frequencies in RNA samples, implying that they are true A-to-I RNA editing events (**Supplementary Data 3, Online Methods**). Taken together, these results suggest that we can distinguish A-to-I RNA editing events from sequencing errors using 1.1% as the frequency cutoff.

We also analyzed the mmPCR-seq data from 960 non-RNA-editing loci from 16 RNA samples (data generated for ASE analysis). With a 1.1% cutoff, we can only find an average of 10 A-to-G/T-to-C variants per 10 kb per sample, which is similar to the density of A-to-G/T-to-C varaints observed in DNA samples. Our results suggest the lack of many extremely low-level editing events in loci that do not contain a known RNA-editing site.

Notably, in the future experiments for the purpose of identifying novel RNA editing sites from mmPCR-seq alone, there will be no need to also sequence DNA of the selected loci. To distinguish novel RNA editing sites from false positives (presumably derived from sequencing and mapping errors), one can determine a minimal variant frequency threshold such that vast majority (often >80-90%) of the identified variants are A-to-G type. For example, requiring 80% of the variants being A-to-G would lead to a FDR at ~3% (using the background of G-to-A mismatches averaged at 2.4%, the false discovery rate at this cutoff is 2.4%/80% = 3%.). In addition, using "spike-in" of a few sites that are known to be unedited (such as a subset of ASE loci) in the Fluidigm experiment may help estimate the FDR rate directly.

Supplementary Note 4. Testing RNA editing coupling and continuous probing hypotheses

It was previously hypothesized that A-to-I RNA editing is initiated by attracting ADARs to a principal site followed by the editing of nearby coupled sites. In this study, we deeply sequenced regions that contain or lack (see below) known RNA editing sites (often with moderate or high editing levels), thus allowing us to detect extremely lowly edited events. Indeed, we observed that a large number of low-level editing events were located around moderately and highly edited events (**Figures 2b and 2c, Supplementary Note 4**), consistent with the coupling hypothesis. In contrast, few editing sites were identified at loci that lack known RNA editing sites (**Supplementary Note 4**). Therefore, our results do not fully support the hypothesis that the continuous probing of the possible secondary structure of pre-mRNAs by ADAR may lead to many low-level A-to-I editing sites across the transcriptome.

Supplementary Note 5. **The reproducibility and uniformity of mmPCR-seq allelotyping assays for 960 ASE sites**

We carried out technical triplicates for each of the 16 samples in the mmPCR-seq assay. We assessed the reproducibility and uniformity of mmPCR-seq alleotyping using the Hiseq data. We found that the ASE levels were highly correlated between technical replicates for all samples measured (**Supplementary Table 5**). Therefore in the following analysis of this work, we combined all reads from technical replicates. A total of 784 sites can be detected in all samples. Of these sites, 92% were covered with 2^{10}~2^{15} reads, within a 32-fold range (**Supplementary Figure 10**).

We next compared the read depth between mmPCR-seq and RNA-seq data. As expected, read depth of mmPCR-seq is not correlated with RNA-seq coverage. Compared to RNA-seq data, read depth from mmPCR-seq is more evenly distributed among sites (**Supplementary Figure 11**). To assess the PCR reaction complexity on uniformity, we also carried out 10-plex PCR by splitting each 20-plex reaction into two and similar uniformity was observed (**Supplementary Figure 12**), suggesting robust design of multiplex primers.

One selected SNP is located in the paternally imprinted gene *ZDBF2* (http://www.geneimprint.com/site/genes-by-species.Homo+sapiens.imprinted-All). As expected, from the mmPCR-seq data, we found that only the paternally allele is expressed (**Supplementary Figure 13**).

Supplementary Note 6. **ASE call in RNA-seq by combining multiple SNPs in the same gene**

One of the benefits of RNA-seq is that reads cover the whole transcript - consequently, when there are multiple SNPs in the same haplotype block, one can use data from all SNPs to call ASE. To investigate how this affects the ASE call, we performed two analyses. We first estimated the number of heterozygous SNPs that could potentially be combined per gene per individual (for each one of the 16 individuals assayed in this study). We found an average of 2.7 heterozygous SNPs per gene per individual, which leads to an average of ~2.7-fold increase of coverage compared to use of any single site (**Supplementary Figure 15**).

We next calculated the proportion of genes with ASE effect by combining all heterozygous SNPs in RNA-seq data. All heterozygous SNPs within a gene were combined to call ASE. We found that, compared to mmPCR-seq, RNA-seq still detected a substantially smaller fraction of ASE effects, especially in lowly or moderately expressed genes (**Supplementary Figure 16**).