

SUPPLEMENTARY INFORMATION

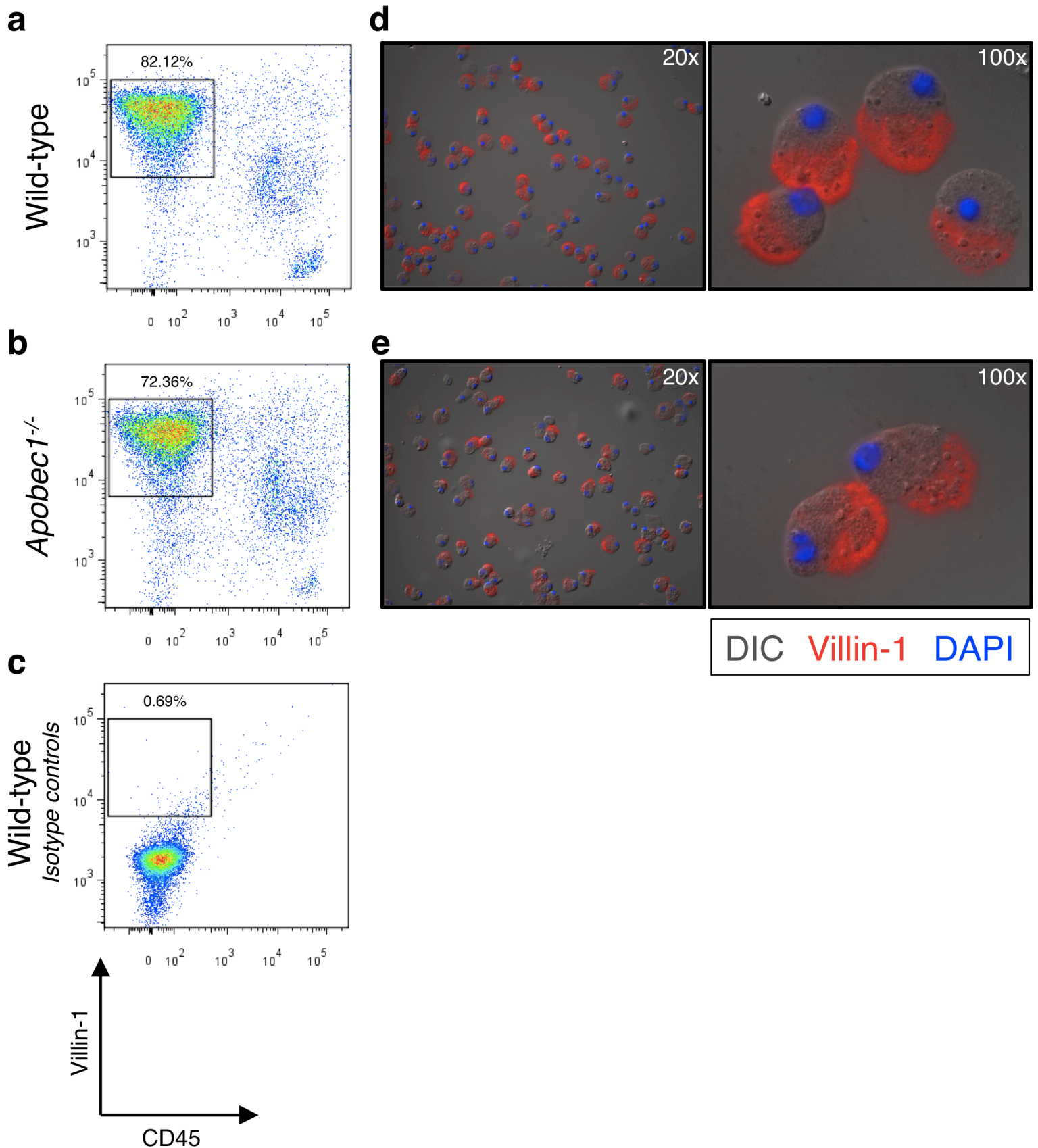
Transcriptome-wide sequencing reveals numerous APOBEC1 mRNA editing targets in transcript 3' UTRs

Brad R. Rosenberg¹, Claire E. Hamilton¹, Michael M. Mwangi², Scott Dewell³ and F. Nina Papavasiliou¹

¹Laboratory of Lymphocyte Biology, The Rockefeller University, 1230 York Avenue, New York, NY 10065, USA.

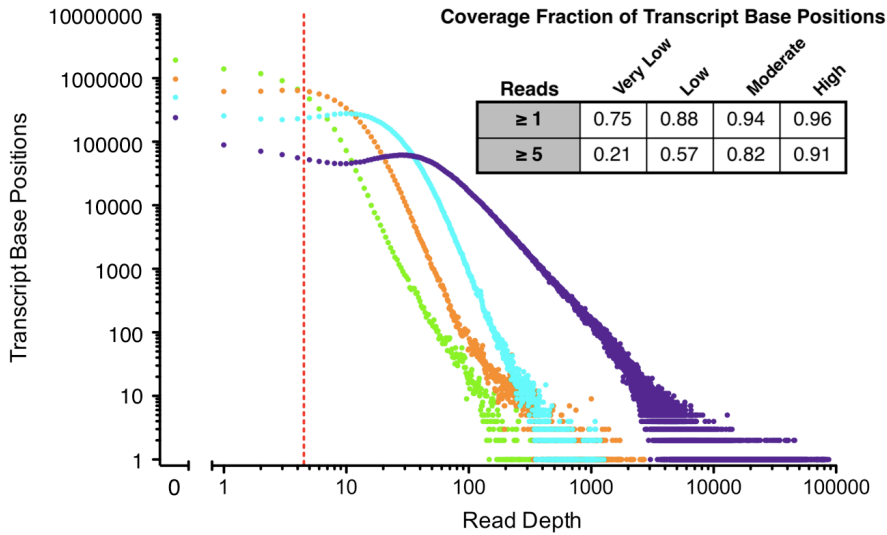
²Laboratory of Microbiology and Infectious Diseases, The Rockefeller University, 1230 York Avenue, New York, NY 10065, USA.

³Genomics Resource Center, The Rockefeller University, 1230 York Avenue, New York, NY 10065, USA.

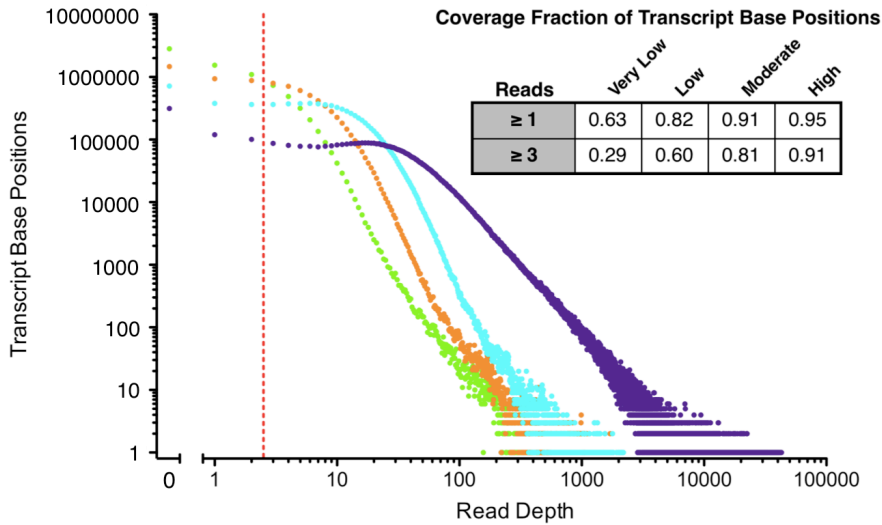


SUPPLEMENTARY FIGURE 1. Isolation of Small Intestine Enterocytes. (a – c) Flow cytometry analysis of enterocyte preps used to generate RNA-Seq libraries. Cells were labeled with antibodies to CD45 (pan-leukocyte marker) and Villin-1 (enterocyte brush border). (a) Wild-type, (b) *Apobec1*^{-/-}, (c) Wild-type cells labeled with isotype control antibodies. (d – e) Immunofluorescence microscopy analysis of enterocyte preps used to generate RNA-Seq libraries. Cells were labeled with antibodies to Villin-1 and stained with DAPI. 20x and 100x magnifications are shown. Villin-1-positive enterocytes with polarized morphology are clearly identifiable. (d) Wild-type, (e) *Apobec1*^{-/-}.

a Wild-type



b *Apobec1*^{-/-}



Gene Expression ● Very Low ● Low ● Moderate ● High

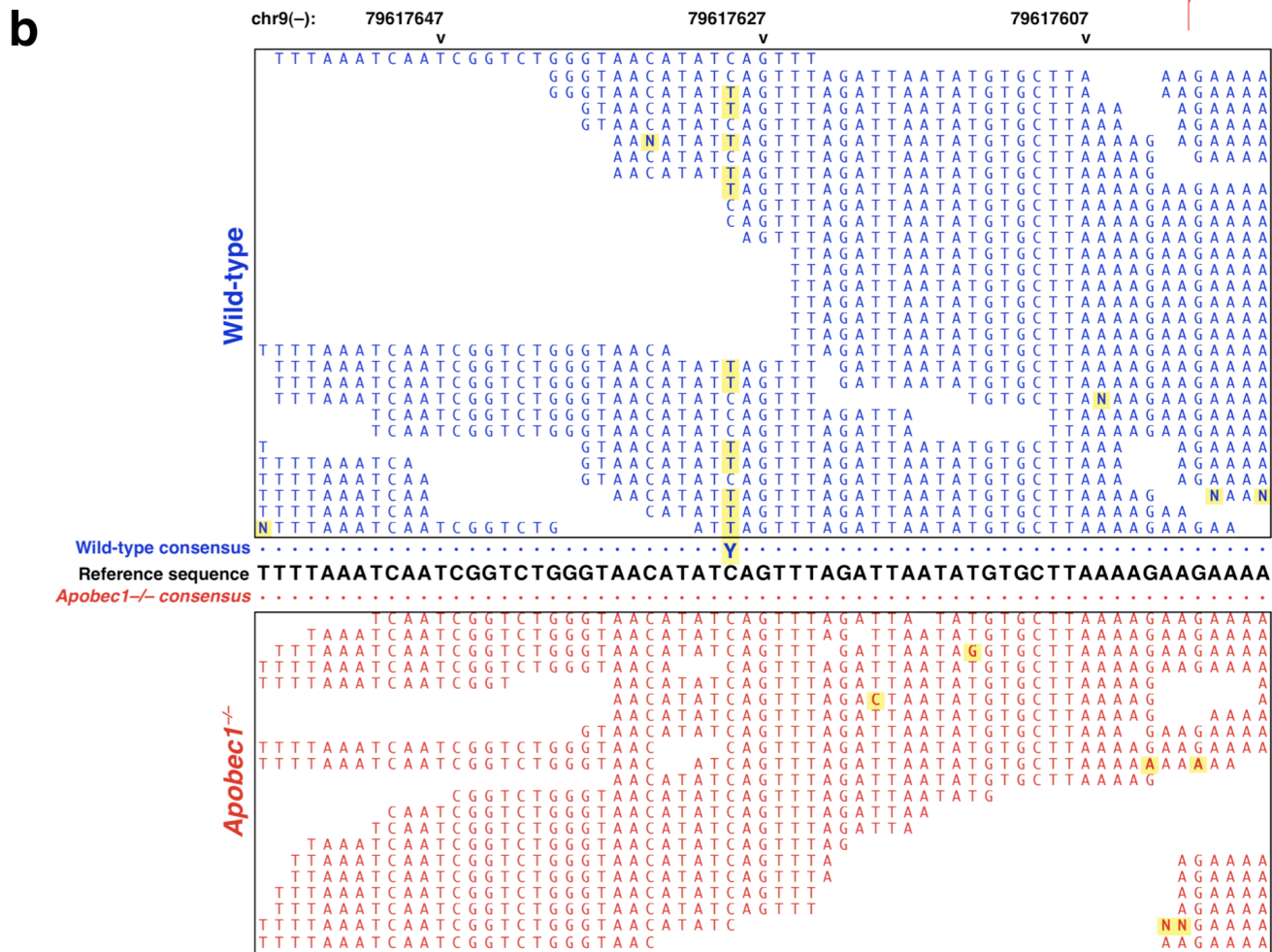
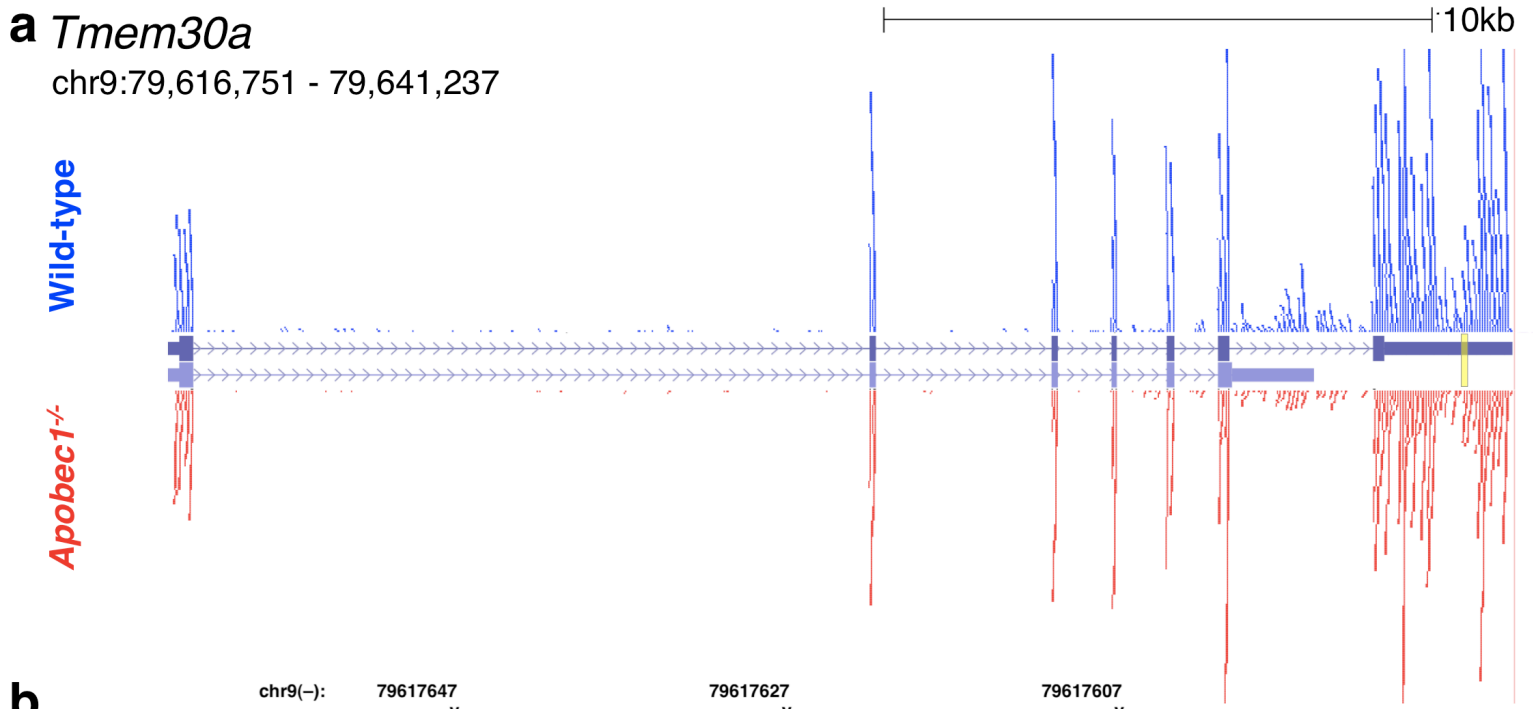
SUPPLEMENTARY FIGURE 2. RNA-Seq Read Coverage Analysis. **(a, b)** Genes expressed in small intestine enterocytes were divided into expression groups (very low, low, moderate, high) by quartile. Plots represent the number of individual base positions of expressed transcripts covered by the indicated number of mapped RNA-Seq reads. Dashed red line indicates the cutoff for inclusion in candidate editing site analyses. Inset tables report the fraction of individual base positions covered by at least 1 and at least the candidate editing site analysis cutoff value of mapped RNA-Seq reads. **(a)** Wild-type, **(b)** *Apobec1*^{-/-}.

SUPPLEMENTARY TABLE 1: RNA-Seq read dataset statistics.

	Raw reads	Uniquely mapped reads
Wild-type	76,766,760	42,770,803 (56%)
<i>Apobec1</i>^{-/-}	50,509,000	28,877,750 (57%)

SUPPLEMENTARY TABLE 2: Candidate APOBEC1 editing site statistics by analysis filter.

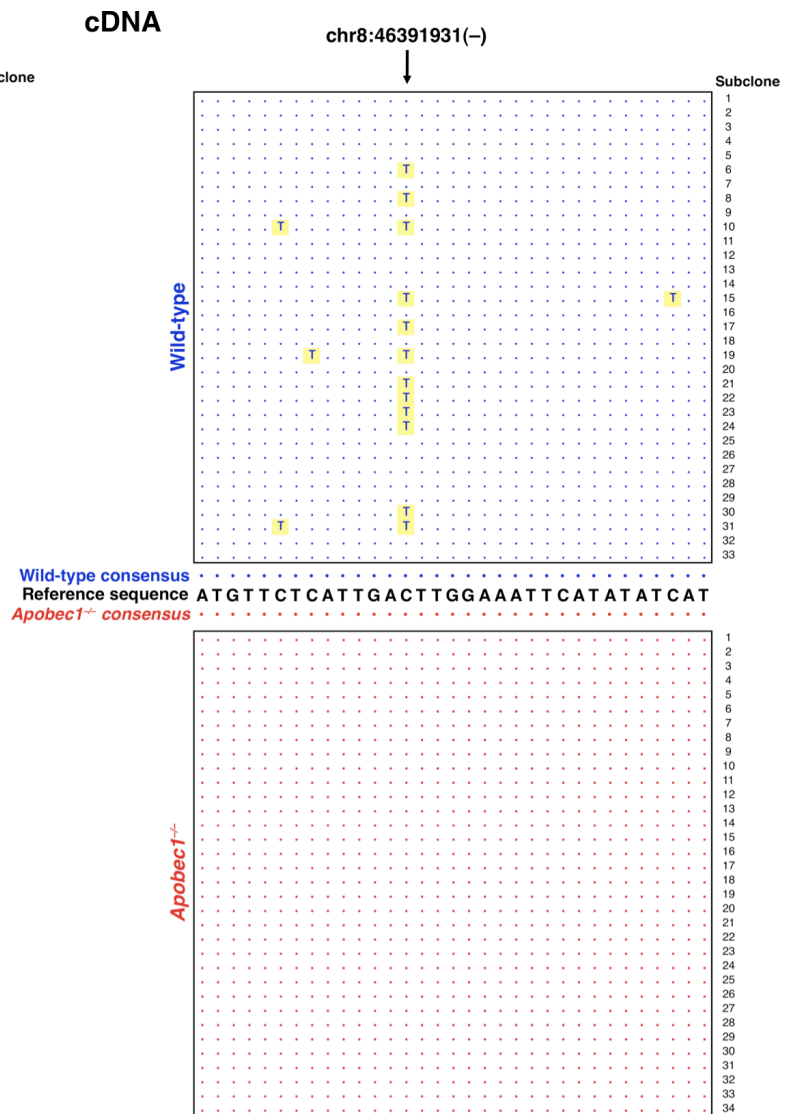
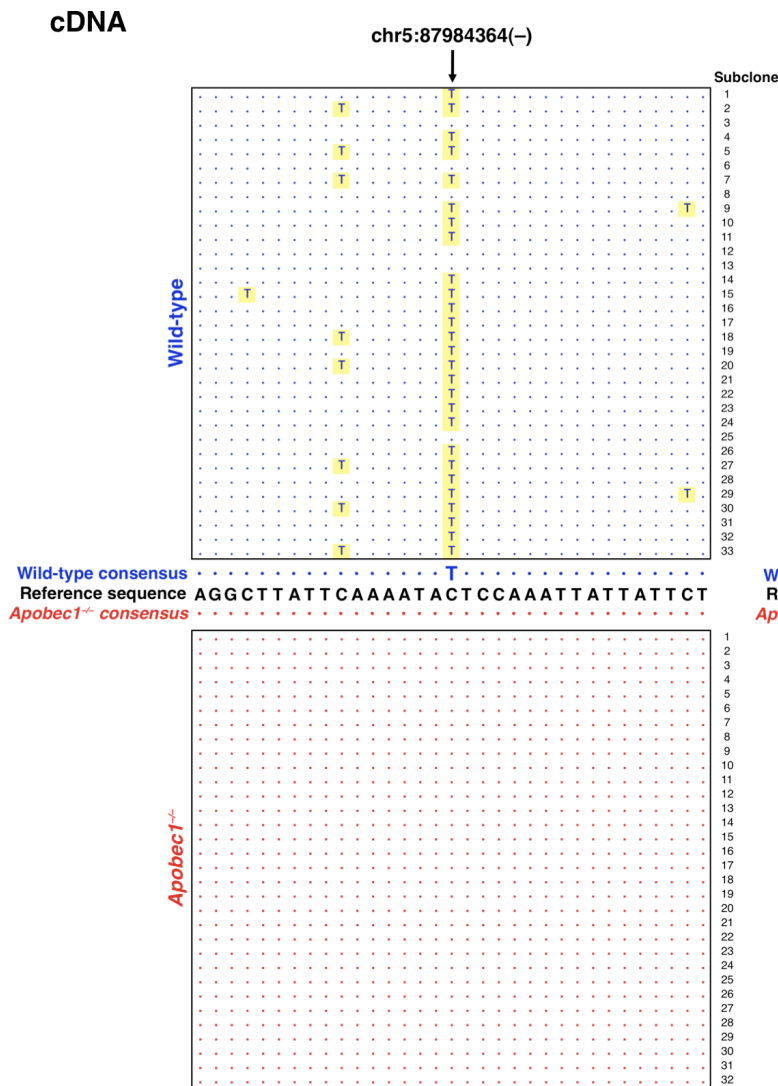
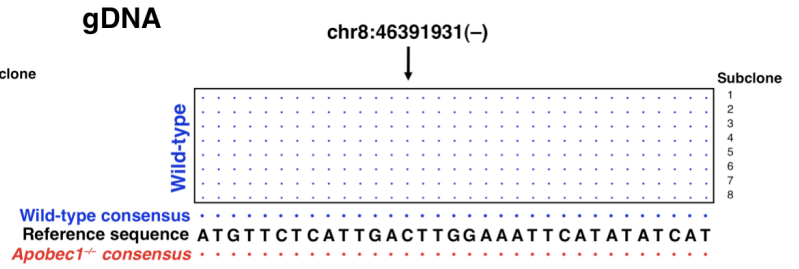
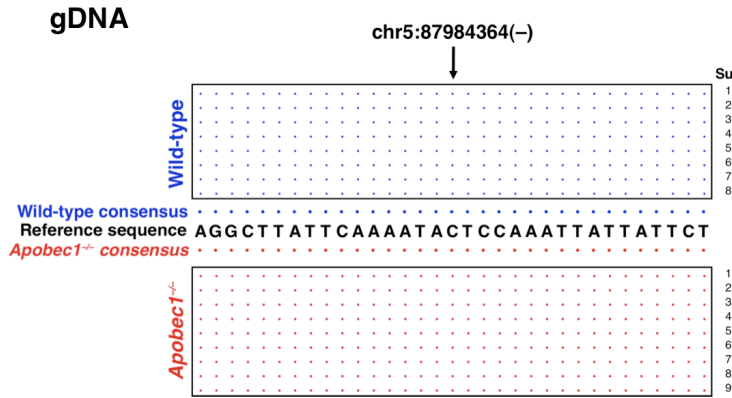
Analysis filter	Candidate edit sites remaining
Wild-type read:reference mismatches (<i>unfiltered</i>)	44,250
Retain sites mapped to RefSeq exons	1,716
Retain reference C / read T mismatches	194
Remove known SNPs	181
Retain APOBEC1-specific mismatches (no mismatch in <i>Apobec1</i> ^{-/-} read set)	93
Remove low read depth / low confidence sites	43
Remove mapping artifacts	39
Validate by Sanger sequencing	33



SUPPLEMENTARY FIGURE 3. Identification of Candidate APOBEC1 Editing Sites By Comparative RNA-Seq Screen: *Tmem30a*. **(a)** Genome annotation for *Tmem30a* displays 2 alternatively-spliced transcript models. Individual RNA-Seq reads (Blue squares, Wild-type; Red squares, *Apobec1*^{-/-}) predominantly map to exons (thick bars). Read distribution suggests that the first transcript (dark blue, upper) is the most abundant transcript isoform in small intestine enterocytes. The screen identified a potential APOBEC1 editing site in the 3' UTR, within the region denoted by the yellow box. **(b)** Detail of region containing potential APOBEC1 editing site (yellow box from *A*). Individual RNA-Seq reads provide overlapping coverage at single-nucleotide resolution. Read mismatches to reference are highlighted yellow. The editing site contains numerous T reads in the wild-type sample, but only C reads in the *Apobec1*^{-/-} sample.

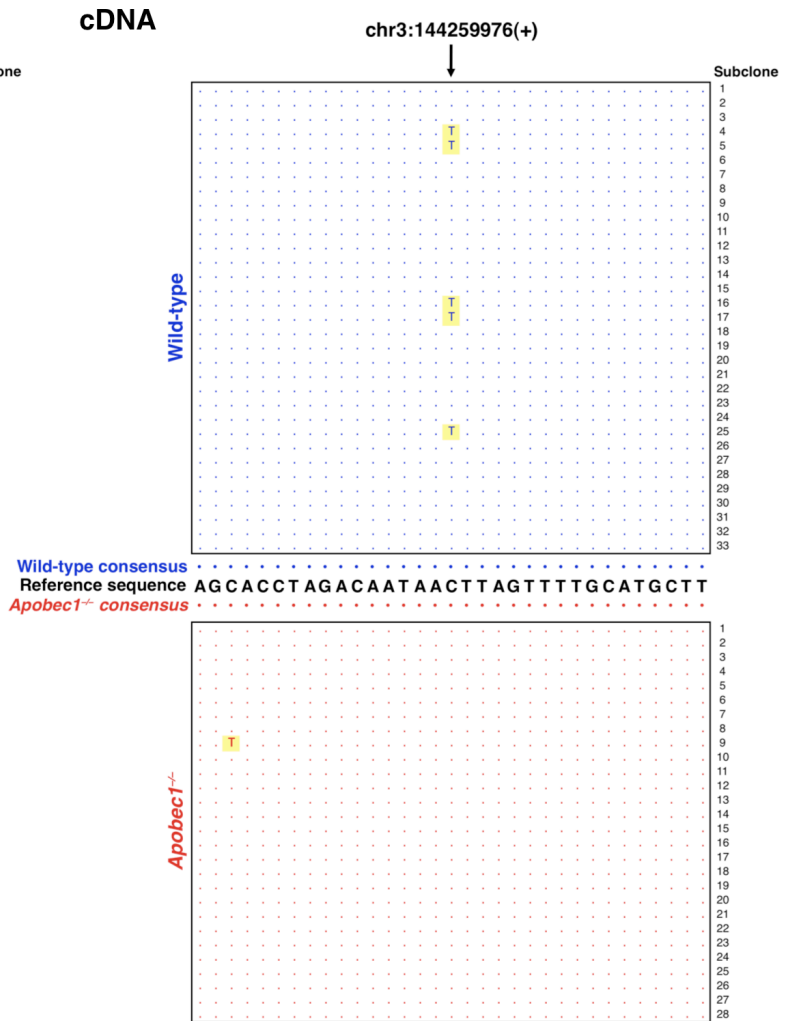
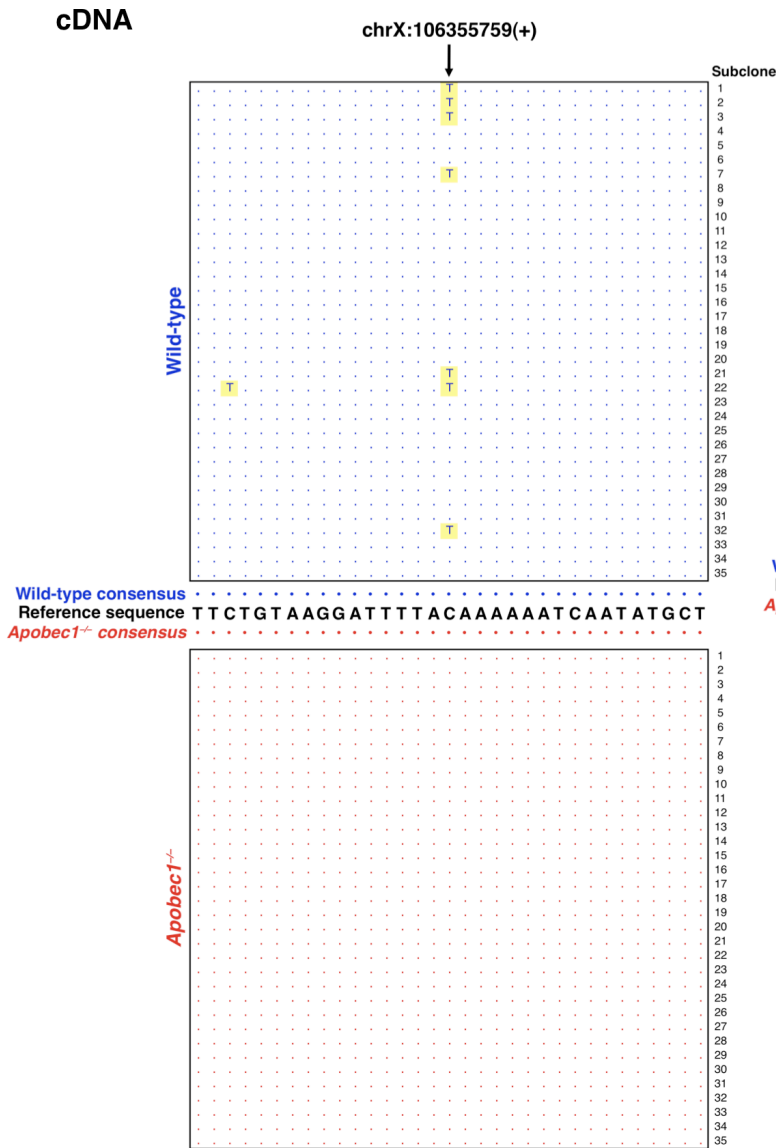
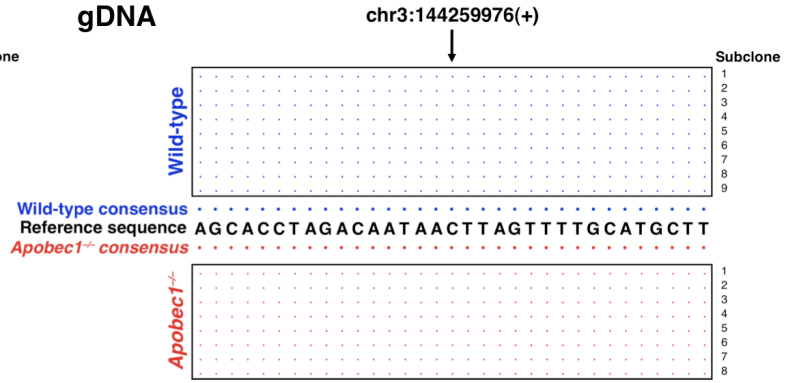
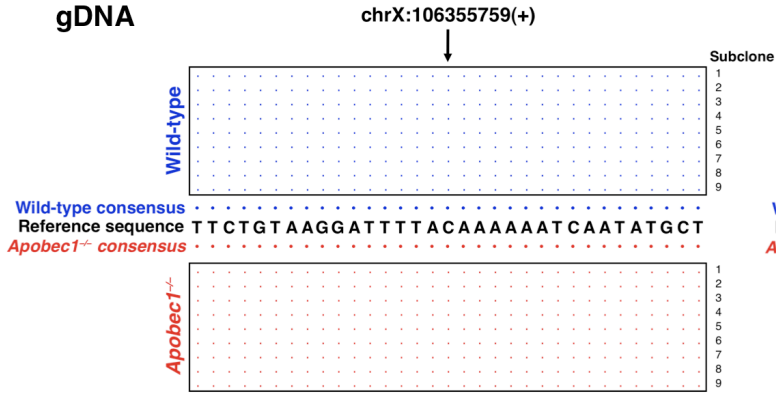
a chr5:87984364(-)
NM_016771
Sult1d1

b chr8:46391931(-)
NM_133969
Cyp4v3



c chrX:106355759(+)
 NM_019989
Sh3bgr1

d chr3:144259976(+)
 NM_053102
Sep15



e

chr5:87984364(-)
NM_016771
Sult1d1

	gDNA		cDNA	
	Wild-type	<i>Apobec1</i> ^{-/-}	Wild-type	<i>Apobec1</i> ^{-/-}
A	0	0	0	0
C	8	9	6	32
G	0	0	0	0
T	0	0	27	0
Subclones sequenced	8	9	33	32

chr8:46391931(-)
NM_133969
Cyp4v3

	gDNA		cDNA	
	Wild-type	<i>Apobec1</i> ^{-/-}	Wild-type	<i>Apobec1</i> ^{-/-}
A	0	0	0	0
C	8	8	21	34
G	0	0	0	0
T	0	0	12	0
Subclones sequenced	8	8	33	34

chrX:106355759(+)
NM_019989
Sh3bgr1

	gDNA		cDNA	
	Wild-type	<i>Apobec1</i> ^{-/-}	Wild-type	<i>Apobec1</i> ^{-/-}
A	0	0	0	0
C	9	9	28	35
G	0	0	0	0
T	0	0	7	0
Subclones sequenced	9	9	35	35

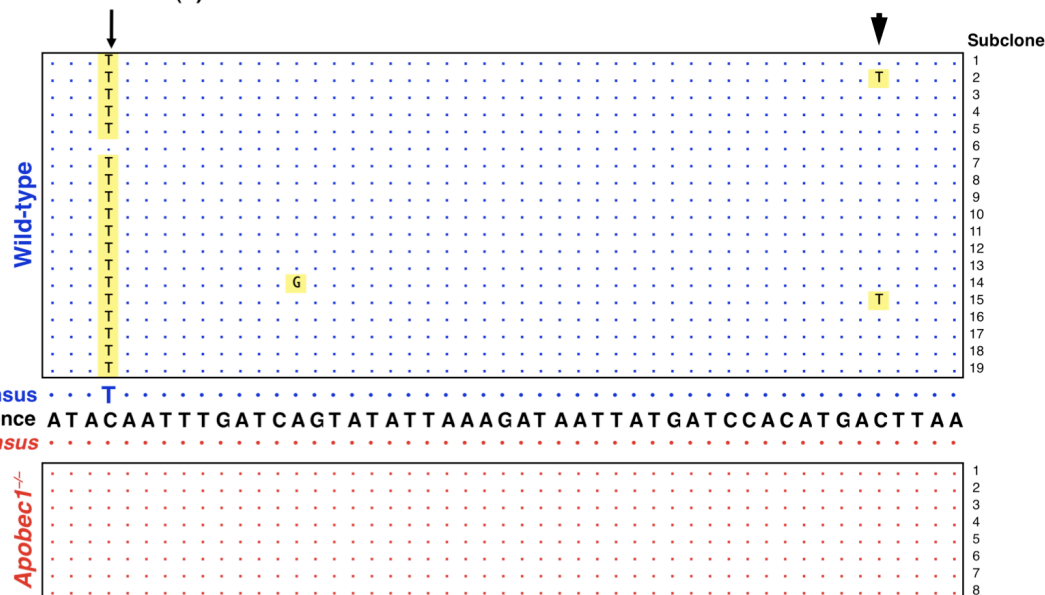
chr3:144259976(+)
NM_053102
Sep15

	gDNA		cDNA	
	Wild-type	<i>Apobec1</i> ^{-/-}	Wild-type	<i>Apobec1</i> ^{-/-}
A	0	0	0	0
C	9	8	28	28
G	0	0	0	0
T	0	0	5	0
Subclones sequenced	9	8	33	28

f

chr12:8014860(+)
NM_009693
Apob

chr12:8014860(+)



SUPPLEMENTARY FIGURE 4. Validation of APOBEC1 mRNA Editing Targets – Subclone Sanger Sequencing. (a – d) Alignments of individual subclone sequences at candidate APOBEC1 editing sites. gDNA and cDNA subclones from wild-type and *Apobec1*^{-/-} enterocytes were sequenced by conventional Sanger techniques and aligned to genome reference. Mismatches to reference are highlighted in yellow. Arrows indicate candidate editing sites identified in RNA-Seq screen. (a) chr5:87984364(-), *Sult1d1*, (b) chr8:46391931(-), *Cyp4v3*, (c) chrX:106355759(+), *Sh3bgr1*, (d) chr3:144259976(+), *Sep15*. (e) Nucleotide frequencies at candidate APOBEC1 editing sites of subclone sequences presented in (a–d). (f) Alignments of individual subclone sequences at *Apob* editing site (indicated by arrow). cDNA subclones from Wild-type and *Apobec1*^{-/-} enterocytes were sequenced by conventional Sanger techniques and aligned to *Apob* reference. “Hyperediting” is apparent in 2 wild-type subclone sequences (arrowhead).

SUPPLEMENTARY TABLE 3. Validated APOBEC1 mRNA Editing Sites.

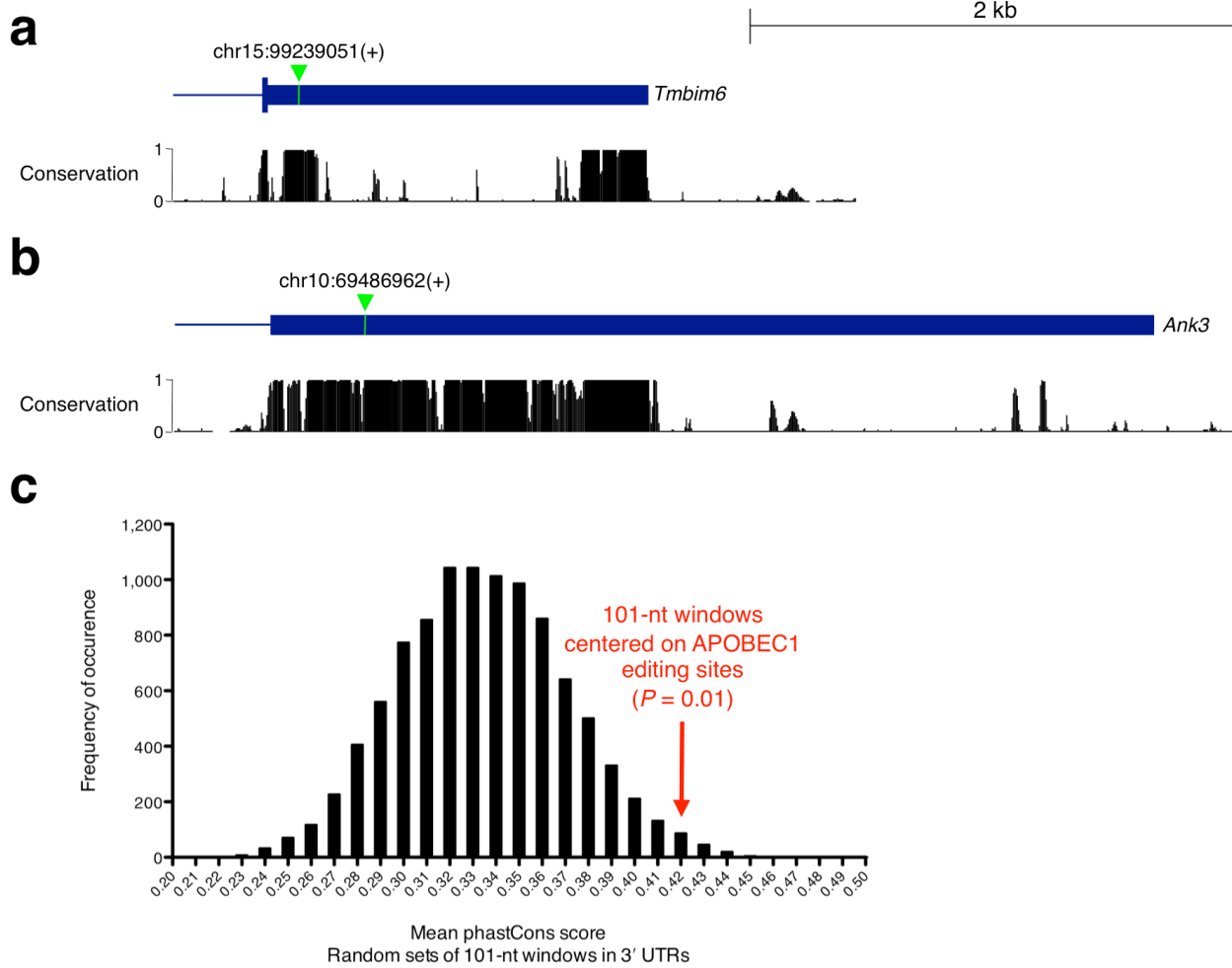
Genome Site	Gene	Type	Ref. Base	Wild-type					<i>Apobec1</i> ^{-/-}				
				Read Cons.	P Cons.	P Mism.	Read Depth	Edit Freq.	Read Cons.	P Cons.	P Mism.	Read Depth	Edit Freq.
chr12:8014860(+)	<i>Apob</i>	CDS	C	T	255	255	204	0.93	C	255	0	128	0.00
chr2:121978638(+)	<i>B2m</i>	3'UTR	C	Y	228	228	2860	0.18	C	255	0	1582	0.00
chrX:109671648(+)	<i>2010106E10Rik</i>	3'UTR	C	Y	228	228	688	0.46	C	255	0	322	0.00
chr8:46391931(-)	<i>Cyp4v3</i>	3'UTR	G	R	228	228	112	0.38	G	117	0	42	0.00
chr3:129616676(+)	<i>Casp6</i>	3'UTR	C	Y	228	228	107	0.50	C	255	0	119	0.00
chr17:44416335(+)	<i>Clic5</i>	3'UTR	C	Y	175	175	186	0.31	C	255	0	92	0.01
chr10:57235791(-)	<i>Serinc1</i>	3'UTR	G	R	77	170	29	0.75	G	39	0	4	0.00
chr5:87984364(-)	<i>Sult1d1</i>	3'UTR	G	R	60	154	28	0.79	G	65	0	20	0.00
chr2:143811725(-)	<i>Rrbp1</i>	3'UTR	G	R	149	149	40	0.38	G	63	0	23	0.00
chr10:7487994(-)	<i>BC013529</i>	3'UTR	G	R	141	141	20	0.45	G	45	0	6	0.00
chr9:79617629(-)	<i>Tmem30a</i>	3'UTR	G	R	129	135	22	0.55	G	87	0	20	0.00
chr1:152208563(-)	<i>BC003331</i>	3'UTR	G	R	54	132	23	0.74	G	48	0	7	0.00
chr4:57203753(-)	<i>Ptpn3</i>	3'UTR	G	R	67	124	15	0.67	G	48	0	7	0.00
chr16:77116537(+)	<i>Usp25</i>	3'UTR	C	Y	116	116	16	0.50	C	45	0	6	0.00
chr3:119135667(+)	<i>Dpyd</i>	3'UTR	C	Y	115	115	26	0.32	C	63	0	12	0.00
chr16:84955113(-)	<i>App</i>	3'UTR	G	R	108	108	563	0.21	G	255	0	357	0.00
chr13:96397289(-)	<i>Iqgap2</i>	3'UTR	G	R	103	103	514	0.23	G	255	0	387	0.00
chr3:144259976(+)	<i>Sep15</i>	3'UTR	C	Y	93	103	13	0.54	C	42	0	5	0.00
chrX:136207009(+)	<i>Rnf128</i>	3'UTR	C	Y	91	91	669	0.20	C	255	0	397	0.00
chrX:106355759(+)	<i>Sh3bgr1</i>	3'UTR	C	Y	89	89	23	0.30	C	75	0	16	0.00
chrX:50374459(+)	<i>Hprt1</i>	3'UTR	C	Y	85	85	55	0.22	C	108	0	27	0.00
chr4:94304303(-)	<i>Lrrc19</i>	3'UTR	G	R	85	85	38	0.26	G	87	0	20	0.00
chr3:119135669(+)	<i>Dpyd</i>	3'UTR	C	Y	84	84	25	0.28	C	60	0	11	0.00
chr14:73595382(-)	<i>Rb1</i>	3'UTR	G	R	83	83	21	0.33	G	30	0	12	0.00
chr12:85772761(-)	<i>Aldh6a1</i>	3'UTR	G	R	64	80	9	0.56	G	42	0	5	0.00
chr2:73654730(-)	<i>Atf2</i>	3'UTR	G	R	73	73	21	0.29	G	54	0	9	0.00
chr16:43981376(-)	<i>Gramd1c</i>	3'UTR	G	R	64	64	17	0.29	G	51	0	8	0.00
chr16:84954758(-)	<i>App</i>	3'UTR	G	R	60	60	293	0.21	G	255	0	118	0.01
chr10:69486962(+)	<i>Ank3</i>	3'UTR	C	Y	56	56	11	0.36	C	36	0	3	0.00
chr13:96397211(-)	<i>Iqgap2</i>	3'UTR	G	R	55	55	124	0.38	G	150	0	41	0.00
chr3:73442586(-)	<i>Bche</i>	3'UTR	G	R	54	54	14	0.36	G	78	0	17	0.00
chr1:192830761(-)	<i>Mfsd7b</i>	3'UTR	G	A	2	48	9	0.78	G	42	0	5	0.00
chr15:99239051(+)	<i>Tmbim6</i>	3'UTR	C	Y	45	45	389	0.20	C	255	0	196	0.00

Validated editing sites are listed with genomic context and RNA-Seq read statistics by sample genotype. *Read Cons*, consensus base call derived from reads; *P Cons*, consensus probability score (defined as the Phred-scaled probability that the read consensus is incorrect); *P Mism*, mismatch probability score (defined as the Phred-scaled probability that the read consensus is identical to the reference base); *Read Depth*, number of reads mapped to given position; *Edit Freq*, editing frequency calculated from read base content.

SUPPLEMENTARY TABLE 4. APOBEC1 Sequence Pattern in RefSeq Transcripts.

MS Excel file: SupplementaryTable4.xls

BOLD indicates transcripts expressed in small intestine enterocytes. **RED** indicates sites with evidence of C-to-U editing in Wild-type RNA-Seq reads.



SUPPLEMENTARY FIGURE 5. Phylogenetic Conservation of Regions Containing APOBEC1 mRNA editing sites. **(a and b)** Examples of APOBEC1 editing sites within well-conserved regions. Blue bars represent transcript 3' UTRs. Conservation plots depict phastCons scores for placental mammal multi-alignments. Editing sites are indicated by green arrows. **(a)** chr15:99239051(+) in the *Tmbim6* transcript, **(b)** chr10:69486962(+) in the *Ank3* transcript. **(c)** Mean phastCons scores for random sets of 101-nt windows within edit site-containing 3' UTRs, as represented by mean phastCons scores for placental mammal multi-alignments. The value for the set of windows centered on the editing sites is to the right of the distribution.

SUPPLEMENTARY TABLE 5. APOBEC1 Editing Sites in miRNA Seed Sequence Matches

Editing site	Transcript accession	Gene	miRNA seed match (C)	miRNA seed match (U)
chr1:152208563 (-)	NM_001077237	<i>BC003331</i>	mmu-miR-669b	
chr1:192830761 (-)	NM_001081259	<i>Mfsd7b</i>		
chr2:73654730 (-)	NM_009715	<i>Atf2</i>	mmu-miR-669n	mmu-miR-297a mmu-miR-297b-5p mmu-miR-297c mmu-miR-539
chr2:121978638 (+)	NM_009735	<i>B2m</i>		
chr2:143811725 (-)	NM_133626	<i>Rrbp1</i>	mmu-miR-539	
chr3:73442586 (-)	NM_009738	<i>Bche</i>	mmu-miR-467e mmu-miR-467h mmu-miR-1970 mmu-miR-599	
chr3:119135667 (+)	NM_170778	<i>Dpyd</i>		
chr3:119135669 (+)	NM_170778	<i>Dpyd</i>		
chr3:129616676 (+)	NM_009811	<i>Casp6</i>	mmu-miR-691	
chr3:144259976 (+)	NM_053102	<i>Sep15</i>		
chr4:57203753 (-)	NM_011207	<i>Ptpn3</i>		mmu-miR-154
chr4:94304303 (-)	NM_175305	<i>Lrrc19</i>		
chr5:87984364 (-)	NM_016771	<i>Sult1d1</i>	mmu-miR-496	
chr8:46391931 (-)	NM_133969	<i>Cyp4v3</i>		
chr9:79617629 (-)	NM_133718	<i>Tmem30a</i>	mmu-miR-190 mmu-miR-190b	
chr10:7487994 (-)	NM_145418	<i>BC013529</i>		mmu-miR-466l
chr10:57235791 (-)	NM_019760	<i>Serinc1</i>		
chr10:69486962 (+)	NM_170729	<i>Ank3</i>		
chr12:85772761 (-)	NM_134042	<i>Aldh6a1</i>		
chr13:96397211 (-)	NM_027711	<i>Iqgap2</i>		
chr13:96397289 (-)	NM_027711	<i>Iqgap2</i>	mmu-miR-370 mmu-miR-683	mmu-miR-323-3p
chr14:73595382 (-)	NM_009029	<i>Rb1</i>		
chr15:99239051 (+)	NM_026669	<i>Tmbim6</i>		
chr16:43981376 (-)	NM_153528	<i>Gramd1c</i>	mmu-miR-1964	
chr16:77116537 (+)	NM_013918	<i>Usp25</i>		
chr16:84954758 (-)	NM_007471	<i>App</i>		
chr16:84955113 (-)	NM_007471	<i>App</i>	mmu-miR-186	
chr17:44416335 (+)	NM_172621	<i>Clic5</i>	mmu-miR-143	
chrX:50374459 (+)	NM_013556	<i>Hprt1</i>		
chrX:106355759 (+)	NM_019989	<i>Sh3bgrl</i>		
chrX:109671648 (+)	NM_026333	<i>2010106E10Rik</i>	mmu-miR-142-3p	
chrX:136207009 (+)	NM_023270	<i>Rnf128</i>		

SUPPLEMENTARY TABLE 6. Primer Sequences for Amplification and Sequencing of Candidate APOBEC1 Editing Sites

Editing site	Transcript accession	Gene	F primer	R primer
chr1:152208563 (-)	NM_001077237	<i>BC003331</i>	CCAGCTAAGGCAACTCAGTCACAT	3' RACE
chr1:192830761 (-)	NM_001081259	<i>Mfsd7b</i>	TTCAGAGAGCCCATGTGTCTCCATTC	TACTTCCTTCTCCTTCTCCTCCT
chr2:73654730 (-)	NM_009715	<i>Atf2</i>	ACCCTCTGCACCCCTCAACATT	GGTACAGAAGTAGTGTGACATCCTGG
chr2:121978638 (+)	NM_009735	<i>B2m</i>	ACAATTTATGCACGCAGAAAGAAATAGCAATG	AAAGCAGAAGTAGCCACAGGGTTG
chr2:143811725 (-)	NM_133626	<i>Rrbp1</i>	GAAGCAACCTGAAGAAGGCATTG	TTTGGGAATAAGGGATACAGCA
chr3:73442586 (-)	NM_009738	<i>Bche</i>	ACACTGTGCTATAGGATGGATCGCAG	CTCCAGGGTGAGGCAGACATTGTTAT
chr3:119135667 (+)	NM_170778	<i>Dpyd</i>	ATGTCTGGTGAATGGCCCACTTTC	CCACAGGTTGTCATTCTCACTTACTTTCT
chr3:119135669 (+)	NM_170778	<i>Dpyd</i>	ATGTCTGGTGAATGGCCCACTTTC	CCACAGGTTGTCATTCTCACTTACTTTCT
chr3:129616676 (+)	NM_009811	<i>Casp6</i>	ATACAAAGGCCAGCTGGTGGGAAGA	ACATGACCAAGTCAAATAGGCCAC
chr3:144259976 (+)	NM_053102	<i>Sep15</i>	TTTGGACGACAACGGGAACATTGC	TGGACTGTGGTGTACTTCAGCTT
chr4:57203753 (-)	NM_011207	<i>Ptpn3</i>	CTGTGGAATTACAAAGATAAATATTACCACCC	3' RACE
chr4:94304303 (-)	NM_175305	<i>Lrrc19</i>	CAAAGTGAGGAACAGGCAGCTTAAAC	TCCTTCAGTAATTTGACCAGTTGCCT
chr5:87984364 (-)	NM_016771	<i>Sult1d1</i>	GTGGCCTCCTAGAGGAAGATTACA	TGACCCTGGTGTGATCCAAA
chr8:46391931 (-)	NM_133969	<i>Cyp4v3</i>	GGTGTGTTTCTTACCAACATGGGTGC	GCACATAACCAGGAAGTTTCTGTGGC
chr9:79617629 (-)	NM_133718	<i>Tmem30a</i>	GCTTCTGCCTTGAAATACCTCAAGC	AATGACAGGAACCAGAGAAGGACAGG
chr10:7487994 (-)	NM_145418	<i>BC013529</i>	ACAGGCTGGCTGTTCAGAAGATGA	ATCCTGGTGCATTCAAGGTAAGGG
chr10:57235791 (-)	NM_019760	<i>Serinc1</i>	ATCCAAACATGAGGCCAGGAGGAT	GGCTGGAACATGAAGATGAACTGC
chr10:69486962 (+)	NM_170729	<i>Ank3</i>	AGTGTACGACACAGGAAGCCATGT	GCCTGTCTTGGATGCATTGTGA
chr12:8014860 (+)	NM_009693	<i>Apob</i>	AGACAAGTAGCTGGTGCCAAGGAA	CTGAATTTGTCTCCTGAGCTGCTG
chr12:85772761 (-)	NM_134042	<i>Aldh6a1</i>	GCCTTCATGTGCCATCTTTGCTCA	TCTGAATTCTGCCAGGGCTGGTTA
chr13:96397211 (-)	NM_027711	<i>Iqgap2</i>	AGTTCTAAGCCCTGTCTTCTGGGA	3' RACE
chr13:96397289 (-)	NM_027711	<i>Iqgap2</i>	AGTTCTAAGCCCTGTCTTCTGGGA	AATAAATGGTGCGGGTGAAGGTGG
chr14:73595382 (-)	NM_009029	<i>Rb1</i>	TGTCCTCAATTTAGTTTCAGTT	3' RACE
chr15:99239051 (+)	NM_026669	<i>Tmbim6</i>	GCACACATCACAGGTGTCGTGTTCTA	ACTCACAAGTCTACACCTCCTCCTCA
chr16:43981376 (-)	NM_153528	<i>Gramd1c</i>	TTACAGTGCCITTCCTTGACTTGGC	CACAAACCACTGGTGTGACACAA
chr16:77116537 (+)	NM_013918	<i>Usp25</i>	ACTGAGTCTTGGACCTAAACA	AATTTACAATAGCCCTTATTCAAGT
chr16:84954758 (-)	NM_007471	<i>App</i>	CTGTACAGATTGCTGCTTCTGCTC	3' RACE
chr16:84955113 (-)	NM_007471	<i>App</i>	GCGAAACCATTGCTTCACTACCCATC	TAATTGGAGACCAGCAGAACACTCCC
chr17:44416335 (+)	NM_172621	<i>Clic5</i>	TAGAAGCAGTAGGTGTCTGGTCGGTA	TATGCAGCTGACCTTGGTCTTCT
chrX:50374459 (+)	NM_013556	<i>Hprt1</i>	GAGGAGTCTGTTGATGTTGCCAGTA	GGAAATCGAGAGCTTCAGACTCGT
chrX:106355759 (+)	NM_019989	<i>Sh3bgrl</i>	TGTTCTGAGTCTTCTCAGCATC	AATGCACTCAACGTGTGTCTGACC
chrX:109671648 (+)	NM_026333	<i>2010106E10Rik</i>	CCACGAGCAGCATATCTCCAAA	TGCATTATGGTCATCAGGAGGG
chrX:136207009 (+)	NM_023270	<i>Rnf128</i>	GTTAACACAGGACTGCCAATCAGGG	CTGTGGGAATTTGCACTGGGAACA

SUPPLEMENTARY METHODS

Immunolabeling, fluorescence microscopy and flow cytometry.

For immunolabeling, enterocyte preparations were pre-incubated with Fc Block (BD Biosciences) and then labeled with PE-Cy7-conjugated antibodies against pan-leukocyte marker CD45 (BD Biosciences). Cells were washed, fixed and permeabilized with Cytofix/Cytoperm solutions (BD Biosciences). After blocking with 5% (v/v) goat serum (Invitrogen), enterocyte preps were labeled with polyclonal antibodies against Villin-1 (Cell Signaling Technology). Secondary labeling was achieved with AlexaFluor 594-conjugated goat anti-rabbit F(ab')₂ fragment.

For flow cytometry, cells were resuspended in acquisition buffer (PBS, 5% FBS v/v) and acquired on an LSR II cell analyzer (BD Biosciences).

For fluorescence microscopy, cells were transferred to slides by Cytospin (Thermo Scientific) centrifugation, labeled with DAPI, and mounted in VECTASHIELD medium (Vector Labs). Images were acquired on an Axioplan 2 fluorescence microscope (Zeiss) and processed with Metamorph software (Molecular Devices).

RNA-Seq Read Coverage Analysis. RNA-Seq read coverage at single-nucleotide resolution was calculated by merging read “pileup” statistics (SAMtools) with transcript models of expressed genes. Expressed genes were defined as RPKM \geq 1.0 by RNA-Seq expression analysis (data not shown). Expressed genes were subdivided into four groups by quartile (RPKM): Very low, low, moderate, and high expression levels. Genomic coordinates for expressed gene exons were derived from RefSeq transcript annotations. These coordinates were merged with SAMtools pileup output to provide the number of reads covering each nucleotide position in expressed transcripts.

APOBEC1 editing site validation subclone sequencing. Sequences containing potential APOBEC1 editing sites were PCR amplified using TurboPfu high-fidelity polymerase (Stratagene). Amplicons were cloned into pSC-B vectors (Stratagene) and transformation colonies were selected by blue/white screening on X-gal. Individual colonies were picked and sequenced as described in *Methods*.

Analysis of APOBEC1 editing site sequence features. Due to alternative splicing, a gene can generate multiple mRNA isoforms, and when the transcribed but untranslated 3' sequence of a gene contains multiple exons, even multiple 3' UTR isoforms may exist. As a result, a single APOBEC1 edit site at the DNA level can appear in different mRNA transcripts and even different 3' UTRs.

Therefore, computations were performed at the DNA level over specific genomic intervals, herein referred to as simply GIs. These intervals are the portions of exons that code for parts of 3' UTR isoforms and are defined in the RefSeq collection. It is important to emphasize that the term "GI" is used very specifically here and does not refer to any exon. There are a total of 26,558 GIs but many more exons. Briefly, the issue of overlapping GIs should be considered. The issue turned out to be a non-factor in the computations and can be safely ignored. None of the GIs that contain an APOBEC1 edit site overlap another GI, and < 5% of the GIs genome-wide overlap another GI.

Furthermore, though the computations were performed at the DNA level, we use the U designation (in place of T) because the results relate to an editing event at the transcript level.

In assessing the AU-richness of the GIs that contain editing sites, the AU content of a set of sequences is defined simply as the number of A- and U-nucleotides divided by

the number of total nucleotides in all the sequences. The AU content of the set of 29 edit-site-containing GIs is 0.63 and is reported in the main text. A random set of 29 GIs was constructed as follows:

- a. For each edit-site-containing GI, a GI was randomly selected from a pool of GIs of comparable length ($\pm 20\%$) according to a uniform distribution over the pool. If the length of the edit-site-containing GI is l , then the length of the new GI was ensured to be in the interval $[0.8l, 1.2l]$. Since the GIs have wildly different lengths, it was felt that it was important to control for length. For example, very long GIs may have sparsely distributed functional elements and so may have long stretches that are not subject to purifying selection. Depending on the edit-site-containing GI, there were a minimum of 262 GIs of comparable length to randomly choose from and a maximum of 1,037, with a median of 949. Therefore, there was ample choice, so the random sets were diverse and rarely contained the same elements.
- b. A random set was allowed to include edit-site-containing GIs.
- c. A random set was not allowed to contain multiple instances of the same GI. If a GI was randomly selected that was selected before, the repeat instance was discarded, and the random selection was redone until a unique GI was obtained.

The AU content was computed for each of 100,000 random sets of 29 GIs. The value was always found to be <0.63 . Hence, a P value of < 0.00001 is reported in the main text.

In assessing the AU-richness of the 101-nt windows centered on the edit sites, the AU content of a set of sequences is defined as above. There are a total of 32 edit sites,

and some of these sites occur in the same GIs. In two cases, a pair of edit sites are separated by < 100-nt, so the 101-nt windows centered on the edit sites overlap (chr3:119135667(+), chr3:119135669(+); chr13:96397211(-), chr13:96397289(-)). To prevent double counting, the edit sites chr3:119135669(+) and chr13:96397289(-) were omitted from the computation. Therefore, only 30 edit sites were actually considered. The AU-content of the resulting 30 edit-site-containing windows is the 0.69 that is reported in the main text.

A random set of 30 windows was constructed as follows:

- a. For each edit-site-containing window, a window of the same length was randomly selected from within the same GI according to a uniform distribution over the GI. The GIs had lengths ranging from 299–4,629 nt, with a median of 1,380-nt. Therefore, there was ample choice, so the random sets were diverse and rarely contained the same elements.
- b. A random set was allowed to include windows that overlapped with and were even centered on edit sites.
- c. A random set was not allowed to contain windows that overlapped with each other – overlapping windows result in double counting and can arise in the case of edit sites in the same GI. If a window was randomly selected that overlapped a previously selected window, the instance was discarded, and the random selection was redone until a non-overlapping window was obtained.

The AU content was computed for each of 100,000 random sets of 30 windows. The value was found to be ≥ 0.69 in only 0.02% of the cases. Hence, a *P* value of 0.0002 is reported in the main text.

To assess the AU skew at sites immediately flanking the edit sites, the binomial test can be used. As described above, the AU content of the edit site-containing windows is $f_{AU} = 0.69$. Consider a column in the multialignment of the transcript segments in mouse encompassing the 32 edit sites. The total number of nucleotides in the column is always $N = 32$. Let k be the number of A- and U-nucleotides in the column. The reported P value is the probability of observing $\geq k$ A- or U-nucleotides under the null hypothesis that A- or U-nucleotides occur with the background frequency f_{AU} . The P value is readily computed from the binomial distribution:

$$P = \sum_{i=k}^N \binom{N}{i} (f_{AU})^i (1 - f_{AU})^{N-i} .$$

Assessment of phylogenetic conservation. PhastCons scores were used to evaluate phylogenetic conservation. In the phastCons analysis, a score is assigned to each nucleotide in mouse in the multialignment of mouse and the 19 other placental mammals. The score is in the interval $[0, 1]$ and reflects the degree of conservation seen in all the mammals ¹. It should be noted that scores are not assigned to indels in mouse. Although phastCons scores are computed across a multialignment of species and such a multialignment can contain indels in any given species, the mouse sequence can be viewed as a stand-alone sequence without indels, with each nucleotide having a unique score. This is the view that should be adopted here.

For a set of windows in the GIs in mouse, the mean phastCons score is computed as follows. First, a sum is performed over the scores of all the nucleotides in all the windows, and then, the result is divided by the total number of nucleotides. The mean phastCons score for the 30 edit-site-containing windows is the 0.42. The random sets of

30 windows were constructed as described above (Analysis of APOBEC1 editing site sequence features). The mean phastCons score was computed for each of 10,000 random sets of 30 windows. The value was found to be ≥ 0.42 in only 1% of the cases. Hence, a P value of 0.01 is reported in the main text.

Computational Analysis Details. Code was written in C++ to perform the sequence feature analysis and conservation computations. The following random number generator was used: mt19937 from the Boost Random Number Library (random.hpp). The generator has a cycle of $2^{19937}-1$ and produces a good uniform distribution in up to 623 dimensions. Each P value was computed 5 separate times with 5 different seeds of the generator to ensure convergence.

Estimation of miRNA Target Sites. In order to estimate if APOBEC1 editing affects miRNA targeting, two sets of sequences were assembled: one set in which 13-nt sequences were centered on the edited cytidine (6-nt upstream, 6-nt downstream) and one set in which 13-nt sequences were centered on the editing site as uridine (6-nt upstream, 6-nt downstream). These sequences were queried against known mature miRNAs (miRBase, <http://www.mirbase.org/>) to determine if any 7-nt substrings is a known miRNA seed.

SUPPLEMENTARY REFERENCES

1. Siepel, A. et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* 15, 1034-50 (2005).