Supplementary Information

Nature Biotechnology: doi:10.1038/nbt.2140

Supplementary Discussion

Additional discussion on editing sites calling using the poly(A)$^-$ sequence data

For both types of RNA library preparations, we initially set out to call editing sites located in the annotated gene sequences as well as the intergenic regions. For the poly(A)$^+$ data, we identified editing sites (16,905, before strand annotation) in annotated gene regions. However, because these libraries were not done in the strand-specific mode, they were not suitable for identification of editing sites in the intergenic genes (which include both noncoding and novel coding genes) due the lack of concrete strand annotation. This part of the data thus was not included in the manuscript. Conversely, editing sites called from the strand-specific poly(A)$^-$ data could be distributed in, in addition to coding and noncoding transcripts, the processed intron RNA. There were a total of 62,866 editing sites identified from the poly(A)$^-$ data, of which 51,655 resided in the annotated gene regions, and 11,211 in the intergenic regions. However, we focused primarily on the intergenic editing sites in the manuscript. This adjustment was made owing to a high representation of intronic reads in the poly(A)$^-$ libraries (Supplementary Table 3) as well as the intronic editing sites (the total number is 50,863, or 98.47% of all sites with annotation). However, because of their over-abundance and as 97.06% of these sites were found on the same strand as their annotated genes, these site were likely present in processed introns or premature transcripts in the poly(A)$^-$ RNA preparation. Therefore, only the intergenic editing sites were included in our final analysis to depict a novel and more biologically relevant editome dataset. Analysis of these sites in poly(A)$^-$ RNAs showed that they exhibit sequence attributes characteristic of A-to-I(G) editing (Table 1).

Additional discussion on the clustering extent of our candidate editing sites

RNA editing sites inferred from this study largely agree with known features of RNA editing (Fig. 2), again illustrating the validity and reliability of our calling strategy. However, one editing sequence feature that our pipeline failed to recapitulate completely was the clustering of sites. Interestingly, the degree of editing site clustering was also limited in another deep sequencing dataset that comprised both DNA and RNA sequence information [2] (22.36% of the editing sites are in clusters of 3 sites per 100 bp). The similar observations of the seemingly low level of concurrent editing may have several explanations. First, a technical challenge likely lies in the current read mapping algorithms, which normally set a prior restriction on number of mismatch allowed between the genome and cDNA sequences. Interestingly, by partially relaxing this criterion but employing stringent filters at subsequent steps, our pipeline was able to specifically capture the majority of the known clustered sites in a simulated sequence read set. Given the results shown by the simulation studies, which yielded 80% sensitivity, and the extent of clustering for sites identified by our approach (34.80%), we estimated the true clustering rate of the YH editome to be about 44%. Second, underrepresentation of hyperedited transcripts in these datasets may also be attributable to the non-brain tissue sources of the sequencing samples under study, as RNA editing is enriched in the brain. Finally, the information archived in DARNED originated from multiple sources of individuals and/or tissues,

Nature Biotechnology: doi:10.1038/nbt.2140

representing a composite collection that might lead to overestimation of the site clustering extent.

Additional discussion on pipeline stringency

Similar to the recent reevaluation of the Li et al approach [1], our methodology has also addressed two important technical issues – paralogues and genomic variant sequences, by incorporating the "BLAT", "YH genomic variants", and "Strand" filters. When we examined the Sanger validation results for sites that were identified in a version of the pipeline without these filters vs. those remained after applying the newer version, the number of called sites expectedly dropped – by 5,397 for the poly(A)$^+$ data and by 6,878 for poly(A)$^-$ (for these sites, 78.77% of poly(A)$^+$ and 45% of poly(A)$^-$ were removed by the "BLAT" filter alone). Intriguingly, after installing these filters, the false discovery rate remained approximately the same: for A-to-G sites, the FDR is 7.59% for the previous dataset and 6.74% for the new one; for non-A-to-G sites, the rate is 37.00% vs. 48.98%. Based on these observations, it is therefore likely that, by incorporating the aforementioned criteria, both false positives and true positives were removed to the same extent. One possible explanation for the exclusion of true positives may be the stringency in read alignment imposed by BLAT, as we found that ~73% of the SOAP uniquely aligned reads that were filtered actually represented the first best matches in the BLAT analysis, and thus could be regarded as supporting evidence for edits calling. Nevertheless, to ensure the accurate identification of RNA edits, a complete and robust workflow was employed to carry out our analyses. It is noteworthy that the criterion for duplication polymorphisms used in the Schrider et al analysis, which was based on the variant regions archived in the Database of Genomic Variants, may be overcritical. Our own CNV analysis (see Methods) revealed that ~10% of the YH genome exhibits variation in copy number (as opposed to >30% by referencing against the Database of Genomic Variants) and that <500 putative sites (<2%) fall in these regions (as opposed to ~30% of our dataset that overlap with the genomic variant regions annotated in the database). This observation thus underscores the importance of the genome-specific information in proper detection of potential artifacts in this type of study.

Nature Biotechnology: doi:10.1038/nbt.2140

Supplementary Methods

Construction of transcriptome libraries and sequencing

Total RNA was isolated from viable lymphoblastoid cell line (LCL) of an anonymous male Han Chinese using Trizol (Invitrogen, Cat #15596-018) according to the manufacturer's instructions, and subsequently treated with RNase-free DNase I for 30 min at 37℃ to remove residual DNA. Libraries were prepared according to the Illumina's protocol (Preparing Samples for Sequencing of mRNA, Part #1004898, Rev. A). poly(A)$^+$ RNA was isolated using the oligo(dT) beads (Dynabeads mRNA Purification Kit; Invitrogen, Cat. #610-06). Upon chemical fragmentation, double-stranded cDNA was synthesized from these RNA samples using random hexamer-primer and reverse transcriptase (Invitrogen). Following the synthesis of 2nd strand, end repair, addition of a single A base, and adaptor ligation, cDNA was further size-selected on agarose gels (~200 bp). These cDNA templates were enriched by PCR amplification and size-selected again on agarose gels.

For some of the libraries, samples were subjected to Duplex-Specific thermostable nuclease (DSN) normalization prior to cluster generation, using the DSN Trimmer Kit (Evrogen, Cat. #NK001). The procedure was done according to the manufacturer's instructions as well as the sample preparation guideline provided by Illumina (Part #15014673, Rev. B). Briefly, the sample library mixed with Hyb buffer was denatured at 98℃ for 2 min and incubated at 68℃ for 5 h. DSN buffer and 2 ml of the DSN enzyme were added to the mixture and incubated at 68℃ for 25 min followed by the addition of stop solution. After purification of the DSN-treated library using SPRI beads, the library was enriched by PCR. The library construction was completed by final purification of the PCR product using SPRI beads.

Isolation of the nonribosomal poly(A)$^-$ RNA [3] (termed accordingly) was first done by removing rRNA from total RNA using the RiboMinus Human/Mouse Transcriptome Isolation kit (Invitrogen, Cat. #K1550-01), and further enriched by removing poly(A)$^+$ RNA with the oligo(dT) beads. Sequencing libraries for strand-specific transcriptome analysis was carried out according to a previous report [4]. Briefly, the first cDNA strand was synthesized from fragmented RNA with random hexamer primers. After purification with the G-50 gel filtration spin-column (GE Healthcare Life Sciences) to remove dNTPs, second-strand synthesis was performed by incubation with RNase H, DNA polymerase, and dNTPs that contain dUTP in place of dTTP (Promega). A single 3' 'A' base was added using Klenow exo$^-$ and dATP to the end-repaired cDNA. Upon ligation with the Illumina PE adaptors, the products were gel-recovered and subsequently digested with N-Glycosylase (UNG; Applied Biosystems) to remove the second-strand cDNA. Samples were then amplified by 15 cycles of PCR with Phusion polymerase and PCR primers with barcode sequence.

The concentration of each library was measured by real-time PCR. Agilent 2100 Bioanalyzer was used for profiling the distribution of insert size. The fragment size of RNA depended on the chemical condition during the fragmentation process, and thus

Nature Biotechnology: doi:10.1038/nbt.2140

may vary between libraries (see Supplementary Table 1).

The poly(A)$^+$ libraries were sequenced by the Illumina Genome Analyzer IIx (GAIIx) platform (CS: Illumina Paired End Cluster Generation Kit v4; SBS: Illumina Sequencing Kit v4 36-Cycle Run), while the poly(A)$^-$ libraries were sequenced by Illumina HiSeq^{TM} 2000 [CS: HiSeq^{TM} Paired End Cluster Generation Kit; SBS: Illumina HiSeq^{TM} Sequencing Kit (200 cycles)]; both types of experiment were done based on the manufacturer's instructions (Illumina Inc., USA). Eight lanes of a GAIIx's flow cell were applied to the poly(A)$^+$ RNA libraries, which were sequenced for 75 or 100 cycles . And five lanes of HiSeq^{TM} 2000's were used for the poly(A)$^-$ RNA libraries, which were sequenced for 90 cycles.

The Illumina Sequence Control Software (SCS v2.5) with Real Time Analysis (RTA v1.6) was used to provide the management and execution of the Genome Analyzer II experiment runs, while HiSeq Control Software (HCS v1.1.37) with RTA (v1.7.45) was equipped for HiSeq^{TM} 2000.

Validation with Sanger sequencing

To confirm whether the putative sites are edited, we analyzed a selection of targets by region-specific PCR amplification of gDNA and cDNA (Supplementary Table 8). To verify the existence of editing sites with low degree of variation (<20%), we performed TA cloning for six of A-to-G amplicons followed by Sanger sequencing validation (Supplementary Table 8). In total, 127 sites were amplified and sequenced successfully using AB 3730xl. The genotypes were called manually from the trace files. Primer sequences are listed in Supplementary Table 17.

Small RNA library preparation and sequencing

Small RNA library preparation was performed according to the manufacturer's instructions (Preparing Samples for Analysis of Small RNA, Part # 11251913, Rev. A). Briefly, sRNA ranging from 18 to 30 nt were gel-purified and ligated to the Illumina 3' adaptor and 5' adaptor. Ligation products were gel-purified, reverse transcribed, and amplified using Illumina's sRNA primer set. Libraries were sequenced on an Illumina HiSeq^{TM} 2000 platform [CS: cBot Single Read Cluster Plate; SBS: Illumina HiSeq^{TM} Sequencing Kit (50 cycles)].

Nature Biotechnology: doi:10.1038/nbt.2140

Supplementary Table 1. Overview of sequencing samples

	Library	Total reads	Reads length	Insert size (nt)	Q20[2]	GC	% of reads aligned to genome	
							Pair	Single[3]
poly(A)+	HUMpsfTDRAAPEI	12,331,915	100	200	90.33%	46.91%	71.01%	3.11%
	HUMpsfTDRACPEI	12,116,155	100	300	89.29%	47.36%	67.46%	5.21%
	HUMpsfTARAAPE[1]	15,379,917	75	500	85.48%	45.69%	69.45%	6.74%
	HUMpsfTBRAAPE[1]	15,395,115	75	500	83.31%	47.92%	69.19%	7.14%
	HUMpsfTCRAAPE[1]	15,970,605	75	500	91.40%	47.78%	81.42%	3.89%
	HUMpsfTCRACPE	15,393,303	75	200	90.52%	47.77%	76.99%	5.25%
	HUMpsfTCRBBPE	15,517,982	75	200	90.16%	49.18%	79.77%	3.43%
	HUMpsfTCRBCPE	15,679,166	75	300	84.71%	46.02%	73.04%	4.86%
	total (8 lanes)	**117,784,158**	**18,890,027,200[4]**		**88.18%**	**47.33%**	**73.82%**	**4.99%**
poly(A)-	HUMwktTBRAAPE_L7	73,615,826	90	200	95.10%	47.90%	76.49%	2.00%
	HUMwktTBRAAPE_L1	73,692,795	90	200	92.62%	47.83%	73.19%	2.38%
	HUMwktTBRAAPE_L2	76,465,541	90	200	92.76%	47.82%	73.40%	2.37%
	HUMwktTBRAAPE_L3	77,143,736	90	200	92.68%	47.85%	73.18%	2.38%
	HUMwktTBRBAPE	101,089,779	90	200	93.60%	47.26%	73.22%	2.38%
	total (5 lanes)	**402,007,677**	**72,361,381,860[5]**		**93.36%**	**47.70%**	**73.84%**	**2.31%**
	Total	**519,791,835**	**91,251,409,060[6]**		**92.29%**	**47.62%**	**767,581,884 (73.84%)**	**30,285,351 (2.91%)**

[1]These libraries were normalized by DSN treatment prior to cluster generation (Methods).

[2]% of nucleotides with sequencing quality ≥20.

[3]% of reads of which only one read mate of a particular pair could be successfully aligned to genome while the other failed in this regard.

[4]Total reads length (nt) of poly(A)$^+$

[5]Total reads length (nt) of poly(A)$^-$

[6]Total reads length (nt) of poly(A)$^+$ and poly(A)$^-$

Supplementary Table 2. Sequence coverage and depth values for genes identified by RNA-Seq

(Please see separate MS Excel file.)

Supplementary Table 3. Genome coverage of each sequencing library by regions

(Please see separate MS Excel file.)

Supplementary Table 4. Results of pipeline simulation on random-sites data

(random-sites dataset, total simulated SNV sites = 8,213, with parameters m=15, n=2)

Depth	SOAPsnp			Basic filter			Read parameter filter			MES filter		
	Sites	PS[1]	PR[2]	Sites	PS	PR	Sites	PS	PR	Sites	PS	PR
50X	50,356	7,903	15.69%	12,655	7,712	60.94%	6,952	6,816	98.04%	6,866	6,816	99.27%
20X	26,709	7,716	28.89%	8,764	7,179	81.91%	6,553	6,454	98.49%	6,489	6,454	99.46%
10X	15,530	6,574	42.33%	5,631	5,176	91.92%	5,048	4,975	98.55%	4,998	4,975	99.54%

[1]Positive sites.

[2]Positive call rate = PS/sites.

Nature Biotechnology: doi:10.1038/nbt.2140

Supplementary Table 5. Results of pipeline simulation on DARNED-sites dataset

(50X simulation data, with parameters m=15, n=2)

SOAPsnp			Basic filter			Read parameter filter			MES filter		
Sites	PS[1]	PR[2]	Sites	PS	PR	Sites	PS	PR	Sites	PS	PR
45,189	612	1.35%	5,943	595	10.01%	672	544	80.95%	588	544	92.52%

[1]Positive sites.

[2]Positive call rate.

Supplementary Table 6. List of all editing sites identified in the study

(Please see separate MS Excel file.)

Supplementary Table 7. Editing site validation by Sanger sequencing

Type		Total validation	True editing	False editing
A-to-G	No. of sites	74	69	5
	% of total	100%	93.24%	6.74%
Non-A-to-G transitions	No. of sites	25	15	10
	% of total	100%	60.00%	40.00%
Transversions	No. of sites	24	10	14
	% of total	100%	41.67%	58.33%

Supplementary Table 8. List of editing sites subjected to validation by Sanger sequencing

(Please see separate MS Excel file.)

Supplementary Table 9. List of editing sites that alter codon usage in transcripts

(Please see separate MS Excel file.)

Nature Biotechnology: doi:10.1038/nbt.2140

Supplementary Table 10. Known features (for all editing types)

Type of feature		poly(A)$^+$	poly(A)$^-$
# of editing sites	Total	11,467	11,221
Site distribution (counts)	Intergenic	-	11,221
	5-UTR	18	-
	CDS	80	-
	Intron	9,362	-
	3-UTR	1,905	-
	Unknown	102	-
dsRNA structure	Counts (%)	4,791 (41.78%)	5,324 (47.45%)
Overlap with Alu	Counts (%)	9,660 (84.24%)	9,766 (87.03%)
Site clusters	≥3 sites in 100 bps	3,354 (29.25%)	3,634 (32.39%)
	≥3 sites in 50 bps	2,088 (18.21%)	2,259 (20.13%)
Codon changes	Synonymous (%)	40 (50.00%)	-
	Non-synonymous (%)	40 (50.00%)	-
Comparison with other data	Genes count	3,077	-
	Genes overlap with DARNED	938	-
	Sites overlap with DARNED	1,098	351
	Genes overlap with cancer	473	-
	Sites overlap with cancer	334	22
Sequence conservation	Counts (%)	216 (1.88%)	85 (0.76%)

Supplementary Table 11. Overrepresented KEGG pathways in the highly edited gene sets (≥10 variants per gene)

KEGG pathway term	Count	*P*-Value	FDR (%)	Genes
p53 signaling pathway	5	7.21×10^{-3}	7.479	*DDB2, MDM2, ATR, MDM4, PTEN*
B cell receptor signaling pathway	5	1.01×10^{-2}	10.376	*CARD11, LYN, SOS1, PLCG2, CD22*
Glioma	4	3.45×10^{-2}	31.411	*SOS1, PLCG2, MDM2, PTEN*
Pathways in cancer	8	7.75×10^{-2}	57.926	*HIF1A, XIAP, SOS1, PLCG2, MDM2, STK4, PTEN, TRAF3*
Protein export	2	8.77×10^{-2}	62.685	*SRP54, SRP9*

Supplementary Table 12. List of editing sites with conserved flanking sequences

(Please see separate MS Excel file.)

Supplementary Table 13. Statistics and features of editing sites identified in two individuals of the Li et al study by the new pipeline

(Please see separate MS Excel file.)

Nature Biotechnology: doi:10.1038/nbt.2140

Supplementary Table 14. Overview of small RNA-Seq

Total reads	63,516,056
Total reads length (bp)	1,453,159,137
# of reads mapped to human genome	52,896,655
% of reads mapped to human genome	83.28%
# of reads mapped to human miRNA[1]	31,400,627
% of reads mapped to human miRNA[1]	49.44%

[1]Reference sequences based on miRBase 16

Supplementary Table 15. List of editing sites identified in miRNA

(Please see separate MS Excel file.)

Supplementary Table 16. miRNAs with A-to-G variant sites

miRNA ID	Location (mature)	Location (hairpin)	Sequence	Variant read count	Total read count	Variant degree (%)
hsa-let-7c	17	27	UAU	56	534	10.49
hsa-miR-1260b	9	18	CAC	5	58	8.62
hsa-miR-1273	9	84	AAA	6	9	66.67
hsa-miR-200b	5	61	UAC	15	139	10.79
hsa-miR-301b	20	64	AAA	49	542	9.04
hsa-miR-378c	21	31	GAG	30	509	5.89
hsa-miR-381[a]	4	52	UAC	15	20	75
hsa-miR-422a	10	19	UAG	30	35	85.71
hsa-miR-625*	7	58	UAG	136	2194	6.2

[a]Chiang et al. 2010

Supplementary Table 17. List of primers used for Sanger sequencing validation of called sites

(Please see separate MS Excel file.)

Nature Biotechnology: doi:10.1038/nbt.2140

Legends to Supplemental Figures

Supplementary Figure 1. Overall gene coverage and sequencing depth of the RNA-Seq data set.

Distribution of the extent of (**a**) gene coverage and (**b**) sequencing depth for genes with at least one corresponding read. Notably, overwhelming majority of the identified transcripts was covered significantly (> 90%) at considerable depth (2×) by aligned RNA-Seq reads.

Supplementary Figure 2. Editing site calling accuracy and sensitivity as a function of the read parameter criterion. (**a**) A two-parameter filter was applied to simulated reads (harboring random mismatches at arbitrary positions of mRNAs encoded by chromosome 1; see Methods) at the indicated sequencing depths, and the performance of the approach was evaluated: accuracy is defined as the false discovery rate (FDR; dotted lines) while sensitivity (SN; gray bars) equals positive calling rate of the simulated editing sites. A particular set of read parameters (m = 15, n = 2; highlighted by yellow shade) was selected and incorporated into the final pipeline. (**b**) Accuracy and sensitivity of the pipeline for each given filter stage. As successive filters were applied to the simulated reads (random-sites dataset; see Methods), the performance of the approach was evaluated as in (**a**).

Supplementary Figure 3. Sequencing chromatogram traces from additional validated gene loci are shown. Type and genomic location (and strand annotation) of each editing site are denoted on top. The editing positions are highlighted by yellow shades. Top trace is genomic DNA (gDNA), bottom trace cDNA. A complete list of validated sites is in Supplementary Table 8.

Supplementary Figure 4. Additional examples of flanking sequence features for: *Alu*-associated and non-*Alu*-associated A-to-G sites in (**a**) poly(A)$^+$ RNA and (**b**) poly(A)$^-$ RNA; (**c**) C-to-T editing type in poly(A)$^+$ RNA; and (**d**) T-to-C editing type in poly(A)$^-$ RNA (see also Fig. 2, g & h).

Supplementary Figure 5. Distinct overlap between datasets reveals conservation of RNA editing at the gene level.

Extent of overlap in editing sites between data sets in terms of nucleotide position ("site") and corresponding gene ("gene"). The DARNED data were compared with those of a breast cancer RNA-Seq study. Proportions of sites and genes that are unique or common between data sets are shown. See also Figure 3a.

Supplementary Figure 6. Another example of highly edited gene (*SPN*) that undergoes editing at both novel and known sites (see also Fig. 3, b & c).

Supplementary Figure 7. Distribution of identified editing sites along the length of miRNAs (nucleotide position based on miRBase reference sequence).

Nature Biotechnology: doi:10.1038/nbt.2140

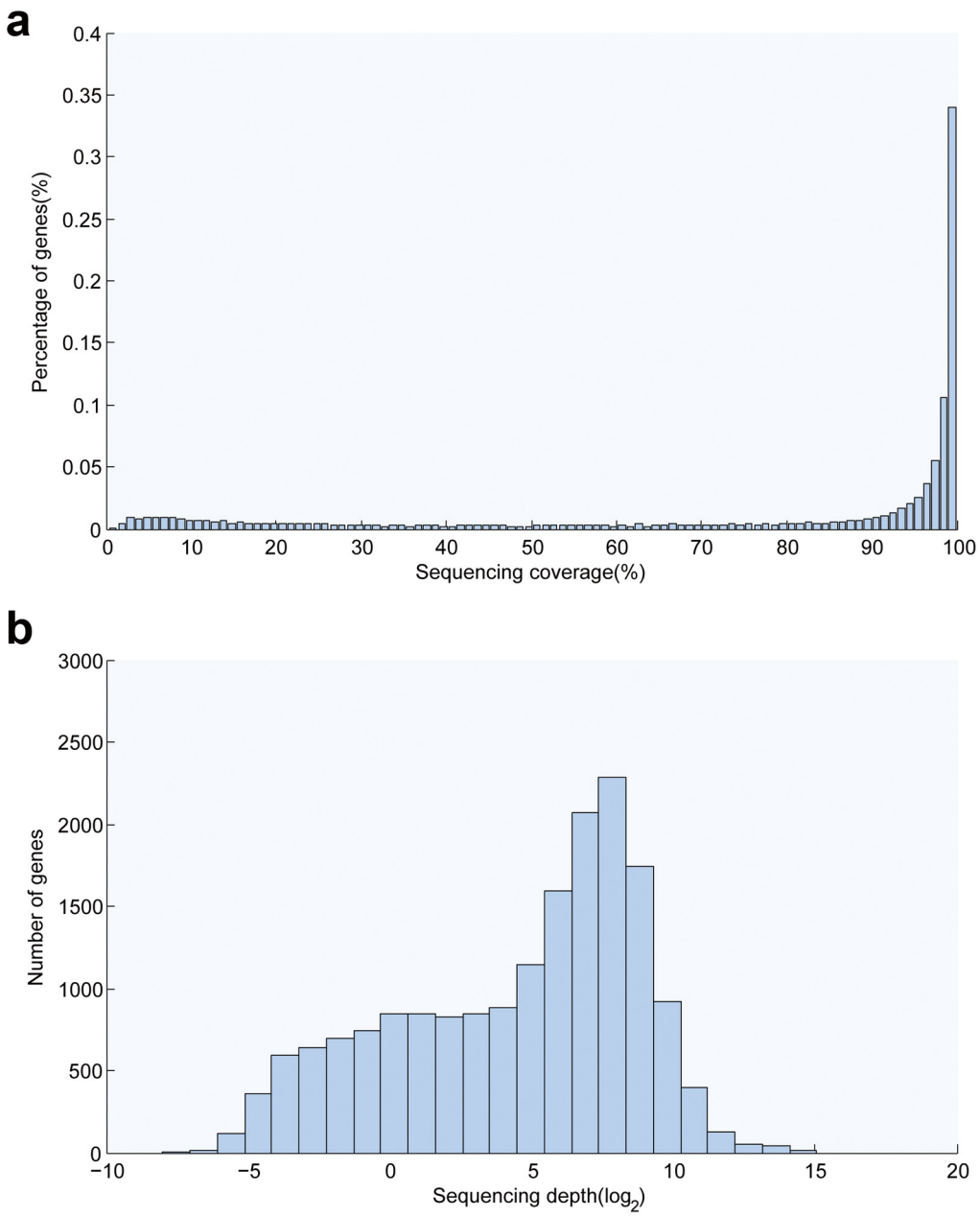
Supplementary Figure 8. Distribution of editing levels for sites in the MES (see Methods).

References

1. Schrider, D.R., Gout, J.F. & Hahn, M.W. Very few RNA and DNA sequence differences in the human transcriptome. *PLoS One* **6**, e25842 (2011).
2. Shah, S.P. et al. Mutational evolution in a lobular breast tumour profiled at single nucleotide resolution. *Nature* **461**, 809-813 (2009).
3. Morin, R. et al. Profiling the HeLa S3 transcriptome using randomly primed cDNA and massively parallel short-read sequencing. *Biotechniques* **45**, 81-94 (2008).
4. Parkhomchuk, D. et al. Transcriptome analysis by strand-specific sequencing of complementary DNA. *Nucleic Acids Res* **37**, e123 (2009).

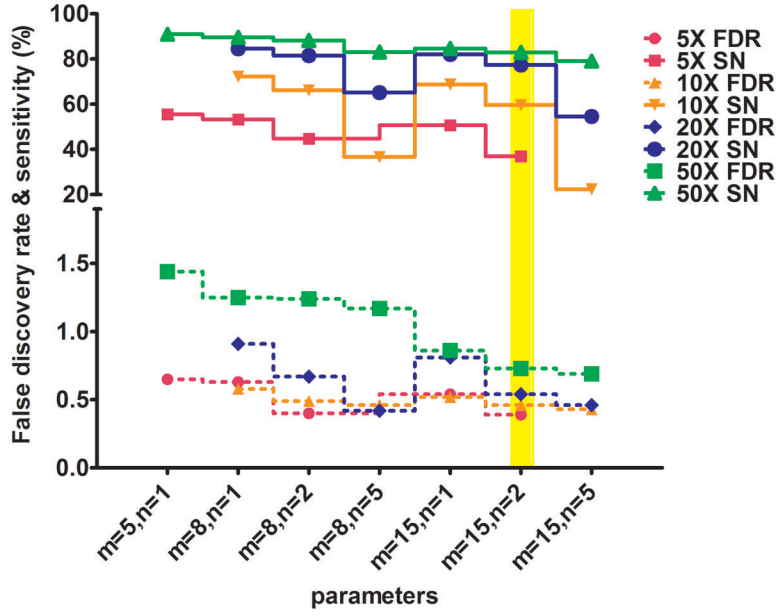
Nature Biotechnology: doi:10.1038/nbt.2140

Supplementary Figure 1

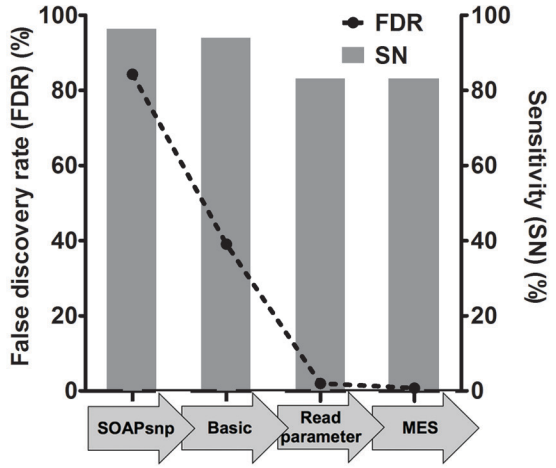


Supplementary Figure 2

a

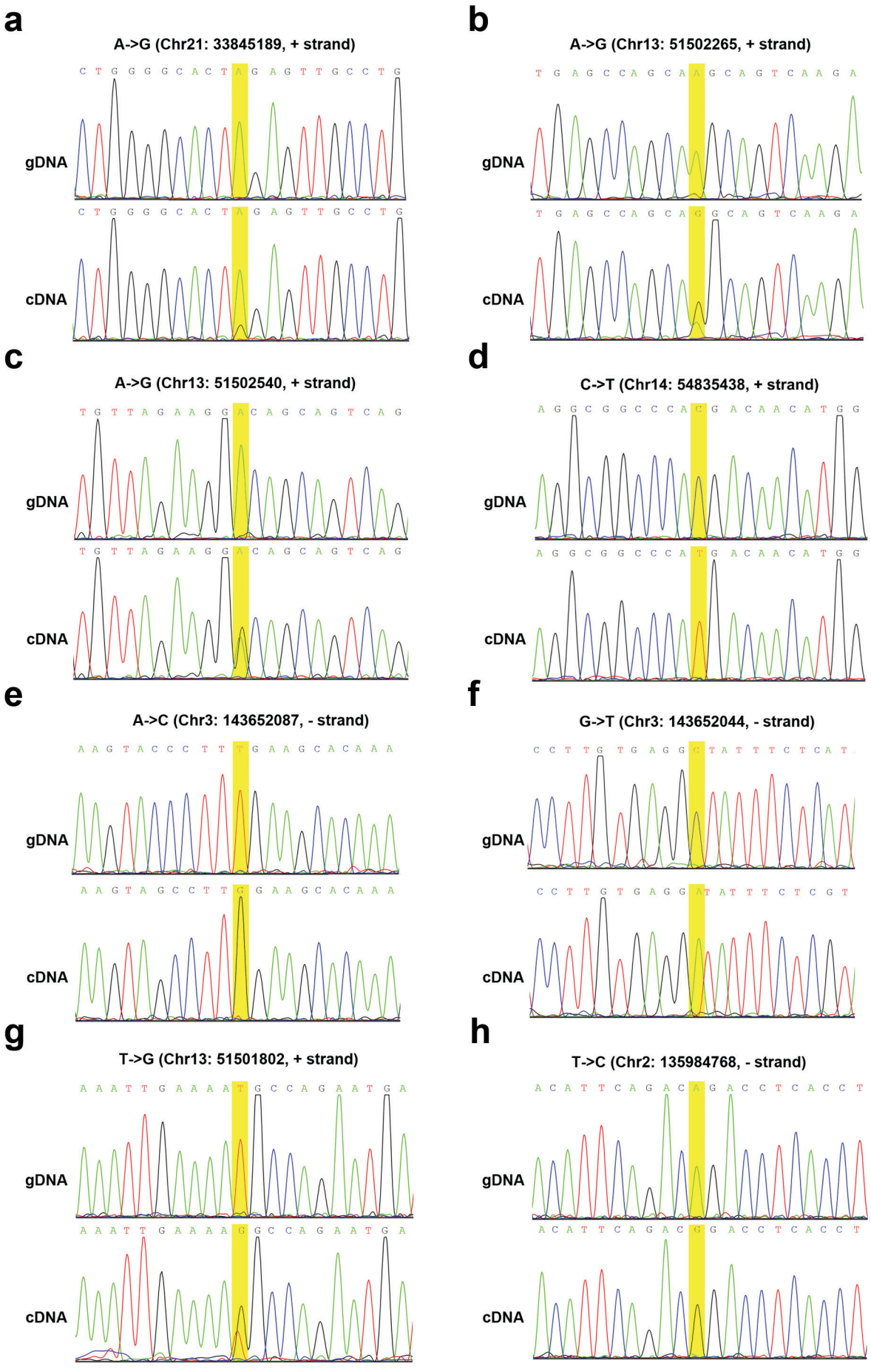


b

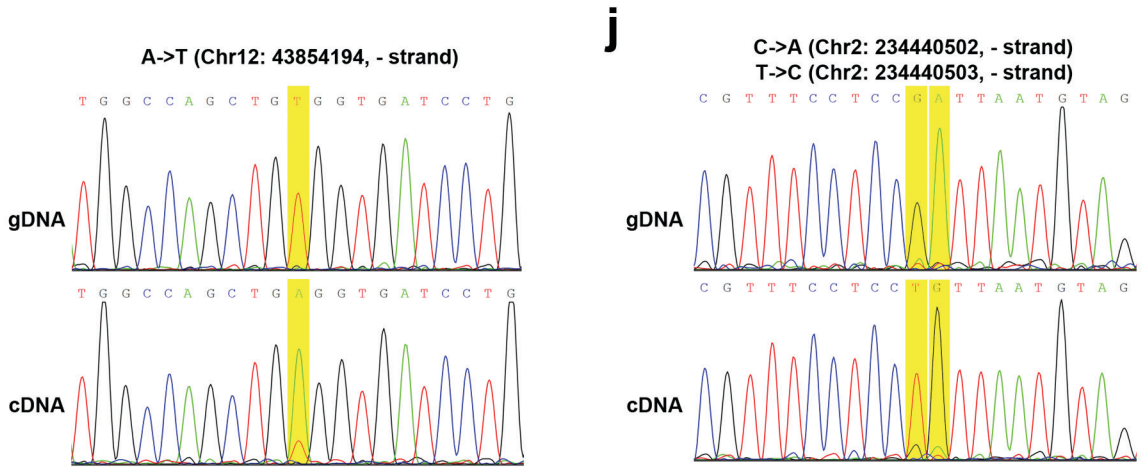


Nature Biotechnology: doi:10.1038/nbt.2140

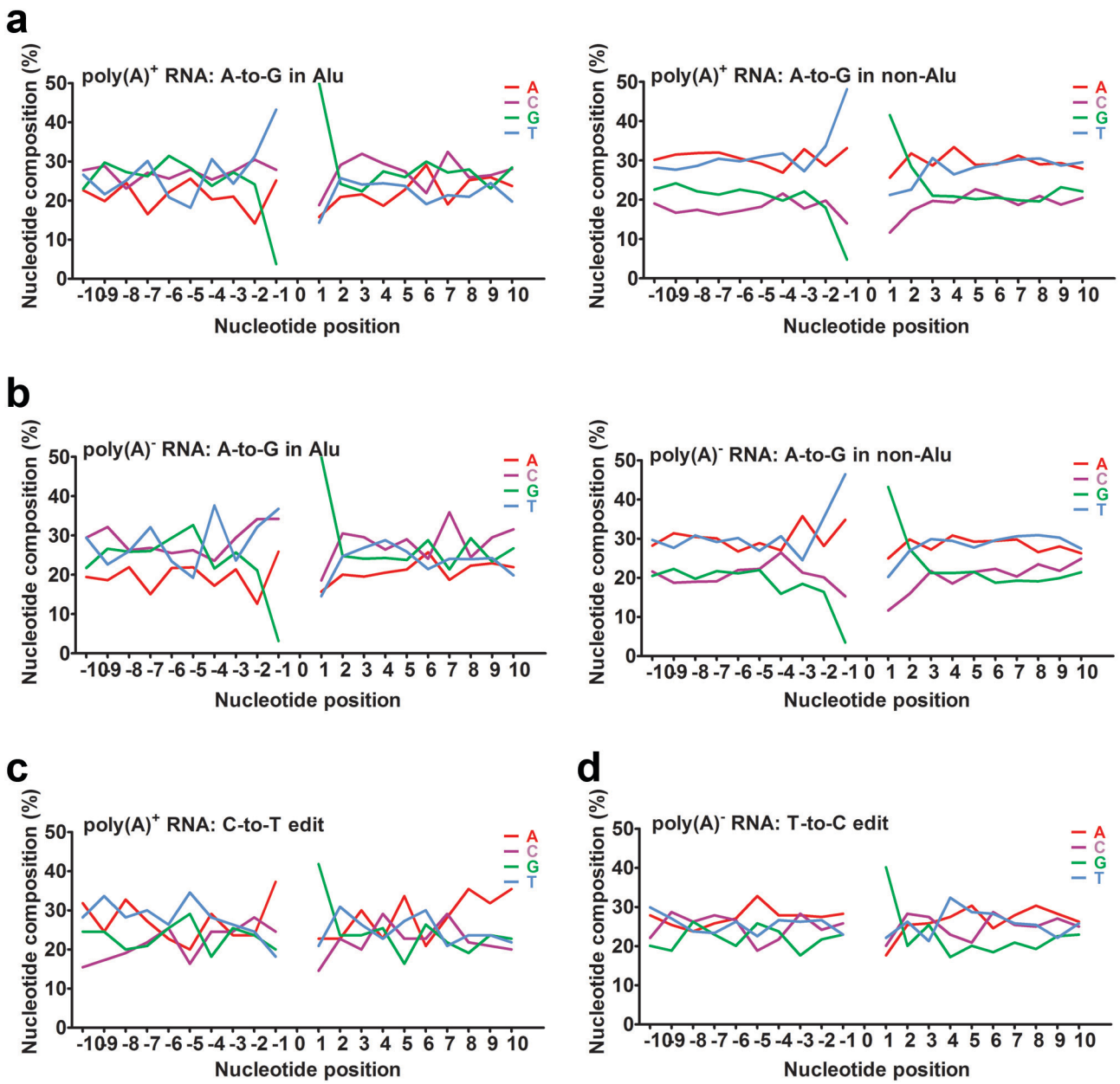
Supplementary Figure 3



Supplementary Figure 3

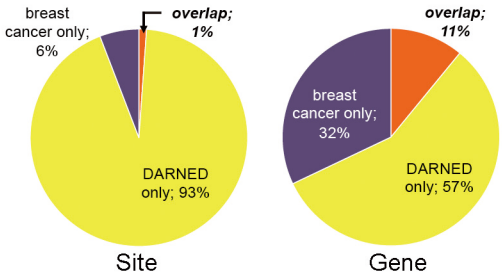


Supplementary Figure 4



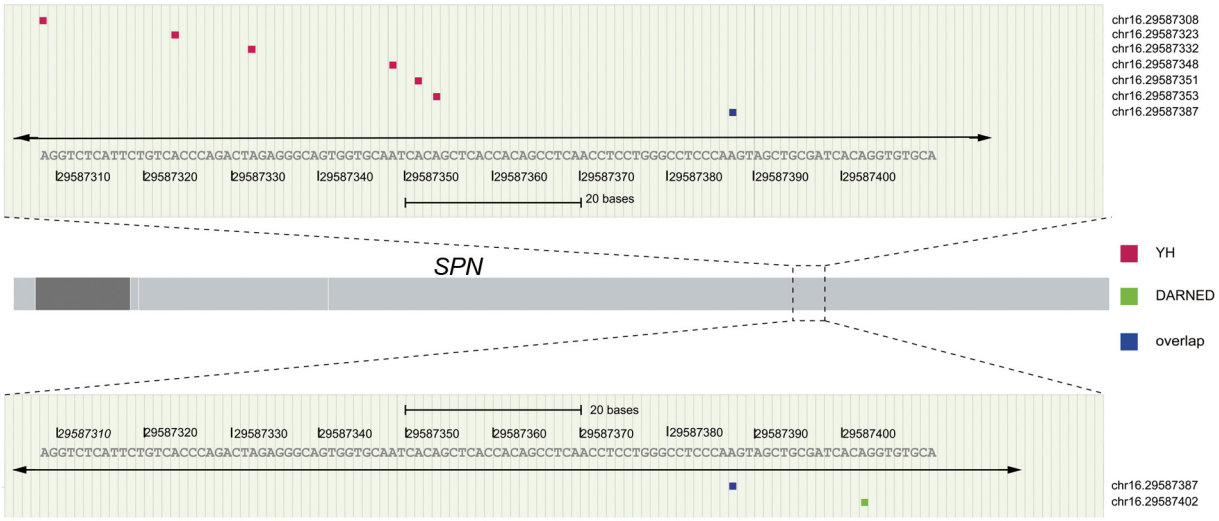
Nature Biotechnology: doi:10.1038/nbt.2140

Supplementary Figure 5

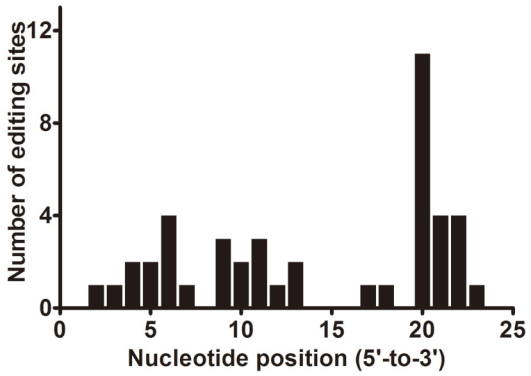


Nature Biotechnology: doi:10.1038/nbt.2140

Supplementary Figure 6



Supplementary Figure 7



Supplementary Figure 8