Genome wide quantification of ADAR A-to-I RNA editing activity – Supplementary Notes:

Effect of Strand Selection on SNR and region-bias 1
A detailed comparison with alternative methods 2
Global vs. per-gene analysis of ENCODE shRNA data 5
Pseudocode 5
URLs 5

1. Effect of Strand Selection on SNR and region-bias

To quantify the error introduced by our strand-selection scheme, we compared the index calculated for two stranded RNA-seq datasets [1,2] (see Methods) with and without taking into account the correct strand information. The agreement between the two calculations is high, with a mild underestimate of the index (relative index difference of 3.7±0.9% and 2.7±0.7% (mean±sd), for the paired-end and single-end data, respectively) and somewhat lower SNR (15.3±2.5% and 17.1±1.8% for the paired-end and single-end data, respectively) in the unstranded case (Figure 2A and Supp. Figure S1). This should be contrasted with the alternative to the above strand-assignment procedure, in which one averages overall A sites (and the A-to-G mismatches therein) and all T sites (and the T-to-C mismatches therein). This latter approach avoids the errors in assigning all reads in any given region to one strand, but, inevitably, doubles the number of sites considered and the SNR is greatly reduced, as demonstrated in the following table:

	AEI			Both strands, averaged		
Tissue	**Average**	**SD**	**Average SNR**	**Average**	**SD**	**Average SNR**
All	1.58	0.61	27.75	0.87	0.33	19.36
BrainCortex	1.96	0.22	32.36	1.10	0.11	24.06
MuscleSkeletal	0.75	0.06	10.63	0.42	0.03	8.24
WholeBlood	2.02	0.12	40.26	1.08	0.07	25.77

Supplementary Table SN1.1: The strand selection approach in AEI results in higher SNR as compared with averaging over both strands.

Furthermore, we verified that the strand-selection does not bias the editing signal towards intergenic regions. The stranded data includes both mRNA and total RNA samples. The distribution of the editing signal across exonic, intronic and intergenic *Alu* elements observed in the mRNA stranded data is very similar to the one obtained for unstranded data. As expected, this distribution is very different when one studies total-RNA samples that are enriched in introns compared with the poly-A selected mRNA samples (see table SN1.2 below). We thus conclude that the accuracy of our strand-assignment is adequate even in the absence of strand-information, and is superior to averaging over both strands.

Note that the slightly higher signal overestimation in lowly covered regions, mentioned above, is responsible for a very small drift in AEI as a function of coverage, 0.4% relative difference for 10-fold lower coverage (averaged over subsampling realizations), see Figure 2B.

	Stranded Total RNA-Seq		Stranded Poly-A Selected RNA-seq		GTEx Unstranded mRNA	
Genomic Region	**%Total Coverage**	**%Editing Events**	**%Total Coverage**	**%Editing Events**	**%Total Coverage**	**%Editing Events**
Exonic	8.40	8.67	40.80	30.73	20.57 - 41.72	11.35 - 22.88
Intronic	69.29	68.86	40.83	42.12	37.92 - 57.81	43.39 - 62.53
Intergenic	22.31	22.46	18.36	27.14	13.00 - 28.03	22.07 - 36.75

Supplementary Table SN1.2: The distribution of the editing signal across exons, introns and intergenic regions is similar for stranded mRNA-seq and unstranded RNA-seq analyzed through our strand assignment approach, supporting the notion that the strand selection does not introduce a significant bias towards intergenic regions. Total-RNA behaves quite differently.

2. A detailed comparison with alternative methods

Averaging over de-novo detected sites: As an alternative to the AEI, one may envision using the average of all editing levels at adequately covered sites in the samples-of-interest. This approach overcomes some of the biases inherent to the methods that are commonly used is the literature. Although the set of detected sites upon one averages is still coverage-dependent, comparing the average level over the same set of sites does provide a measure of the relative editing strength. However, it suffers from the following disadvantages.

First, current RNA-seq datasets usually include multiple samples. Assume one demands de-novo detection and a minimal coverage of 10 reads (allowing for a rough estimate of the editing level). Applying the above method to a large dataset containing multiple individual samples, results in a very small set of sites that are detected and covered by 10 reads in all samples. For example, taking 10 cortex and 10 muscle samples from our test-set, only 58 *Alu* sites pass this filter for the twenty samples, which brings us back to the undesirable situation of quantifying global editing based on only a few dozen sites is. Therefore, we relax the requirement for detection in all samples, and try to look at sites that are well-covered (≥10 reads) in all samples, but detected as edited in at least one sample. This leads to 4710 sites, still not a very large number compared to the millions of editing sites. The situation is much worse when there are hundreds or thousands of samples originating from multiple tissues, conditions, etc, with varying expression profiles (as is common is current widely use databases, including GTEx [3] which we study).

Second, the method is still biased to the number of samples per condition. Let's say one is interested in comparing the editing level between cortex and muscle. Three groups approach this question. One has ten cortex samples and two muscle samples, the second analyzes ten muscle samples and two cortex samples, and the third has 10 and 10 samples. For simplicity, we use identical samples in all groups (the 10 cortex samples of group 1 are the same as those of group 3, namely our test-set samples).

The following table (SN2.1) compares the results of the three groups, using the AEI or the average over detected sites (sites covered by at least 10 reads in all samples in the cohort, and detected as edited in at least one of these). All three groups overestimate the cortex-to-muscle editing ratio (compared with the ratio estimated by the AEI, which takes into account the full *Alu* signal). In addition, the fluctuations in the estimated ratio are much larger in the Averaged-level method.

		Group1	Group2	Group3
Averaged level	**cortex**	6.72	11.31	6.32
	muscle	1.97	2.36	1.59
	ratio	3.41	4.79	3.97
AEI	**cortex**	1.965	2.179	1.965
	muscle	0.806	0.751	0.751
	ratio	2.44	2.90	2.62

Supplementary Table SN2.1: The averaged editing level over well-covered sites is not as robust as the AEI.

This reveals a deeper problem in the averaging method – this quantification method does not provide a stable per-sample estimator. In other words, it does not lead to a single number that describes the sample, regardless of the other samples under study. The reason for that is clear – the sites upon which one averages, must be covered in all samples, and detected in at least one of the samples. Thus, the resulting list of sites upon which we average depends on the whole cohort, and not just the sample itself. This explains the much larger fluctuations observed in the table above, where (for example) the exact same 10 muscle samples give an averaged level 2.36% in group 2, but 1.59% in group 3. Each group averages over a different set of sites.

Furthermore, this prevents comparison of the analysis results with previous studies. To demonstrate that, we again envision a research effort towards comparing editing in cortex and muscle. First, ten labs analyze one cortex and one muscle RNA-seq sample each, averaging the editing level over the well-covered sites common to the two samples at their hand, and reporting a number per sample. Using our test-set samples, their results average to 19.8% (cortex) and 5.0% (muscle). These numbers are three-fold higher that the levels obtained analyzing the same twenty samples together (group 3 in the table above).

Assume now a third group wants to analyzes a new tissue and compare it with cortex or muscle. They would like to consult the literature, but which number should they use? In fact, they cannot use any of the above numbers, but rather must download all RNA-seq data, re-analyze all samples and get completely different numbers for the same cortex and muscle samples.

In contrast, the AEI measure averages over millions of sites, and provides a well defined number per-sample, independent on the other samples analyzed in the same cohort, which can be quoted for future use.

Averaging over previously reported sites: Alternatively, one may want to average (or even calculate a weighted average, similar to the AEI) over a set of pre-identified sites. For example, one may average or calculate an index over RADAR [4] or REDIportal [5] sites.

Averaging suffers from the same problem discussed above. In order to estimate the editing level, one must require some minimal coverage cut-off, and look at sites that are well-covered (e.g. ≥10 reads) in all samples. This leaves a small number of sites that might not represent well the global editing pattern. To demonstrate this, we re-analyzed the same samples used for batch effect estimation, looking at the average editing level over the well-covered REDIportal sites. Only 3177 and 2707 sites pass the coverage filter (for samples A and B, respectively). Table SN2.2 shows that averaging the editing level in these sites does provide a robust measure (in the sense of low variability across replicates, sequencing centres and protocols), but comparison of the relative editing level between the two samples (A and B) results in a ratio that differs substantially from the one observed using the AEI, which takes into account all *Alu* sites.

			Averaged level over REDI sites			AEI		
seqc_sample	library_id	#samples	CoV	Avg	SD	CoV	Avg	SD
A	1	23	0.0322	0.8050	0.0259	0.0334	1.2654	0.0423
A	2	23	0.0407	0.8024	0.0326	0.0480	1.2785	0.0614
A	3	23	0.0340	0.7959	0.0270	0.0387	1.2861	0.0497
A	4	23	0.0228	0.8093	0.0185	0.0268	1.2745	0.0341
A	5	12	0.0462	0.8223	0.0380	0.0119	1.3287	0.0159
B	1	23	0.0342	0.9498	0.0324	0.0457	2.5308	0.1157
B	2	23	0.0322	0.9602	0.0309	0.0549	2.5391	0.1393
B	3	23	0.0387	0.9598	0.0371	0.0500	2.5511	0.1275
B	4	23	0.0306	0.9538	0.0291	0.0511	2.5483	0.1303
B	5	12	0.0581	0.9705	0.0563	0.0109	2.6071	0.0283

Supplementary Table SN2.2: Averaged editing level over REDIportal sites does provide a robust measure, but the results (and the ratio between the two samples) differ from those obtained from the AEI, which takes into account the full *Alu* editing signal.

Another possibility is to calculate an index over the set of pre-identified sites. This approach is much better, but still has some drawbacks, compared with the AEI. First, it is well known in the community that the abovementioned depositories of editing sites, which integrate results of multiple studies taken at face value, include many false positives (mainly out of *Alu*, though). These are going to introduce noise, inevitably. Second, for many years the A-to-I RNA editing community was focused on RNA editing in the brain. Accordingly, the sites reported in the databases are based, to a large extent, on brain samples. Thus an index based on such lists discriminates against sites expressed or edited more strongly in other tissues. Third, these web-based lists of sites are updated from time to time, as sites found in new studies are added. Therefore, averaging over these sites does not allow comparison between index values calculated by different groups at different time points. Finally, the advantage of the AEI is even clearer considering other species, where the available data on editing sites is scarce and reliable transcriptome-wide datasets of pre-identified editing sites are not available.

Here are the results for index calculated over REDIportal sites, for the batch-effect samples produced by the SEQC consortium[6]. The index is consistently higher than the AEI, since REDIportal selects for sites that are more highly edited. The ratio of editing levels is similar, but not identical to the full signal quantified by the AEI. In addition, the variability across replicates, sequencing centres and protocols is higher than that measured by the AEI.

			Index over REDI sites			AEI		
seqc_sample	library_id	#samples	CoV	Avg	SD	CoV	Avg	SD
A	1	23	0.0452	3.1470	0.1422	0.0334	1.2654	0.0423
A	2	23	0.0711	3.1422	0.2234	0.0480	1.2785	0.0614
A	3	23	0.0615	3.1761	0.1953	0.0387	1.2861	0.0497
A	4	23	0.0419	3.1353	0.1314	0.0268	1.2745	0.0341
A	5	12	0.0164	3.1811	0.0521	0.0119	1.3287	0.0159
B	1	23	0.0616	5.2330	0.3226	0.0457	2.5308	0.1157
B	2	23	0.0716	5.2334	0.3748	0.0549	2.5391	0.1393
B	3	23	0.0713	5.2401	0.3735	0.0500	2.5511	0.1275
B	4	23	0.0736	5.1755	0.3808	0.0511	2.5483	0.1303
B	5	12	0.0183	5.0638	0.0925	0.0109	2.6071	0.0283

Supplementary Table SN2.3: Calculating the index over REDIportal sites overestimates the average over all *Alu* sites given by the AEI.

3. Global vs. per-gene analysis of ENCODE shRNA data

In search for additional putative regulators of editing, we followed Quinones-Valdez et al [7] and analyzed the ENCODE [8] dataset, which includes RNA-seq data following silencing different RNA-binding proteins and gene expression factors. Interestingly, our results (using the AEI) are different from the ones found by Quinones-Valdez et al [7]. The two analyses have somewhat different aims, and thus also differ in the methodology in several aspects. Here we wish to point out just one key difference that might help to shed light on the advantages and disadvantages of using the AEI as an editing measure.

An analysis based on the AEI looks for a global and systematic change in the *Alu* editing pattern, while Quinones-Valdez et al. look for proteins that are associated with differential editing of many individual sites, regardless of the direction of change. Thus, Quinones-Valdez are able to detect proteins that influence different sites in different directions, while our own method is sensitive to a consistent trend even if it's too weak to be detected per site. For example, XPO5, which we identified as one of the prime candidates for ADAR regulation based on its global effect of the AEI, was found by Quinones-Valdez et al to be among the least relevant of all RBPs tested (it caused a statistically significant change in only ~2% of the sites tested, ranked in the lower half of the RBPs tested, in for both cell lines; see their Supp. Figure 2).

4. Pseudocode

AEI Pseudocode:

1. Use as input, a BAM file of uniquely-aligned reads (high-quality alignment is preferred, see Methods).
2. Intersect alignments with region of interest (e.g. *Alu* regions).
3. Count high-quality mismatches, excluding common SNPs (see Methods).
4. If data is unstranded, decide the expressed strand for each region (Figure 2 and Methods). If data is stranded, use the correct strand per read.
5. Calculate the AEI (fraction of A-to-G mismatches to sum of A-to-A and A-to-G calls), and the control index (by default, the C-to-T index: fraction of C-to-T mismatches to sum of C-to-C and C-to-T calls).

5. URLs

RNA-Seq datasets available from https://www.ncbi.nlm.nih.gov/geo/
Stranded mRNA-seq, *GSE113957;* Stranded Total-RNA-seq, *GSE103001;* Batch effect, GSE47774; Human ADARs KO, *GSE99249*; Murine ADAR1 knock-in, *GSE94387*; Storage comparison, GSE113976; accession code for the shRNA data (ENCODE) are detailed in Supp. Table S12
UCSC Data: (Data download commands are also available in the installation script)

Human data:
Genome: hgdownload.soe.ucsc.edu/goldenPath/hg38/bigZips/hg38.fa.gz
Annotations: hgdownload.soe.ucsc.edu/goldenPath/hg38/database, tables rmsk.txt.gz (RepeatsMasker), snp150Common.txt.gz, ncbiRefSeqCurated.txt.gz, and gtexGene.txt.gz

Murine Data:

Genome: hgdownload.soe.ucsc.edu/goldenPath/mm10/bigZips/chromFa.tar.gz
Gene expression: chromosome.sdsc.edu/mouse/download/19-tissues-expr.zip
Annotations: hgdownload.soe.ucsc.edu/goldenPath/mm10/database, tables tables rmsk.txt.gz (RepeatsMasker), snp142Common.txt.gz, and ncbiRefSeqCurated.txt.gz

Analyses Tools:
fastq-sample: https://homes.cs.washington.edu/~dcjones/fastq-tools/fastq-sample.html

Other tools are specified in the README file, Requirements section in GitHub

References

1. Fleischer, J. G. *et al.* Predicting age from the transcriptome of human dermal fibroblasts. *Genome Biol.* **19**, 221 (2018).

2. Wenric, S. *et al.* Transcriptome-wide analysis of natural antisense transcripts shows their potential role in breast cancer. *Sci. Rep.* **7**, 17452 (2017).

3. Lonsdale, J. *et al.* The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* **45**, 580–585 (2013).

4. Ramaswami, G. & Li, J. B. RADAR: A rigorously annotated database of A-to-I RNA editing. *Nucleic Acids Res.* **42**, D109-13 (2014).

5. Picardi, E., D'Erchia, A. M., Lo Giudice, C. & Pesole, G. REDIportal: a comprehensive database of A-to-I RNA editing events in humans. *Nucleic Acids Res.* **45**, D750–D757 (2017).

6. SEQC/MAQC-III Consortium. A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium. *Nat. Biotechnol.* **32**, 903–914 (2014).

7. Quinones-Valdez, G. *et al.* Regulation of RNA editing by RNA-binding proteins in human cells. *Commun. Biol.* **2**, 19 (2019).

8. Sloan, C. A. *et al.* ENCODE data at the ENCODE portal. *Nucleic Acids Res.* **44**, D726-32 (2016).