Genome sequence–independent identification of RNA editing sites

Qing Zhang[1] & Xinshu Xiao[1–3]

RNA editing generates post-transcriptional sequence changes that can be deduced from RNA-seq data, but detection typically requires matched genomic sequence or multiple related expression data sets. We developed the GIREMI tool (genome-independent identification of RNA editing by mutual information; https://www.ibp.ucla.edu/research/xiao/GIREMI.html) to predict adenosine-to-inosine editing accurately and sensitively from a single RNA-seq data set of modest sequencing depth. Using GIREMI on existing data, we observed tissue-specific and evolutionary patterns in editing sites in the human population.

Accurate identification of the RNA 'editome' is needed to better understand the diversity of gene expression and its functional implications[1–3]. Many tools have been developed recently to identify RNA editing sites from RNA-seq data (summarized in ref. 4). However, challenges still exist, including the requirement for genome sequence data from the same individual in order to discriminate RNA editing sites from genomic single-nucleotide polymorphisms (SNPs)[4]. Even with matched whole-genome sequence, some SNPs still escape identification, possibly owing to nonuniformity in sequencing coverage or other issues. Other methods use multiple RNA-seq data sets to increase the confidence of finding individual sites, but this precludes analysis of single data sets and may miss unique changes[5]. Here we report a method to accurately identify the RNA editome independently of genome sequence using a single RNA-seq data set of modest sequencing depth.

Our GIREMI method uses allelic linkage between single-nucleotide variants (SNVs) to detect candidate editing sites and improves the predictive power with generalized linear models (GLMs). In a typical RNA-seq data set, many reads contain SNVs that may correspond to genomic SNPs, RNA editing sites or experimental errors. A pair of SNPs in the same read (or read pair, in paired-end sequencing) maintains the same haplotype in the RNA as in reference genomic DNA (**Fig. 1a**). In contrast, a SNP and an RNA editing site exhibit variable allelic linkage because RNA editing occurs post-transcriptionally to either allele randomly (unless allele-specific editing exists, which is presumably rare). Similarly, the allelic linkage for a pair of RNA editing sites may also appear random, although processive editing does exist[6] that may lead to allelic bias of multiple editing sites.

To examine whether allelic linkage may enable the discrimination of RNA editing sites from SNPs, we calculated the mutual information (MI) associated with SNPs or RNA editing sites in RNA-seq reads (Online Methods). Indeed, MI values associated with the two types of variants demonstrated a striking difference (**Fig. 1b**, **Supplementary Fig. 1a** and **Supplementary Note 1**), reflecting the discriminative power of this approach. On the basis of this rationale, the GIREMI method calculates the MI of publicly available SNPs (such as from the dbSNP database) and uncharacterized RNA variants in a given RNA-seq data set and uses this to predict RNA editing sites and further parameterize a GLM for enhanced performance (**Supplementary Fig. 1**, Online Methods, **Supplementary Software** and https://www.ibp.ucla.edu/research/xiao/GIREMI.html).

As a proof of concept, we first applied GIREMI to a deeply sequenced Encyclopedia of DNA Elements (ENCODE) RNA-seq data set derived from the GM12878 human lymphoblastoid cell line, which has associated genome sequencing data[7]. The MI step predicted 31,660 RNA editing sites (99.6% being the A-to-G type), and the GLM found 5,117 additional putative A-to-G editing sites. Because the genome of GM12878 has been well studied, most of the SNPs in this cell line are already included in dbSNP, which afforded an advantage in predicting RNA editing sites. To evaluate the performance of GIREMI, we assumed that 10–90% of the GM12878 SNPs were unknown (**Fig. 1c**). Strikingly, the false discovery rate (FDR, % GM12878 SNPs in predicted editing sites) was only 3% when 30% of GM12878 SNPs were assumed to be unknown (**Fig. 1c**). The FDR increased to only 7.6% if 90% of SNPs were unknown, which is an extreme overestimate of the fraction of unknown SNPs in a common human sample given the recent expansion of dbSNP. Performance did not change substantially when a different read-mapping method was used (**Supplementary Fig. 2** and **Supplementary Note 2**). It should be noted that the FDR defined here assumes that SNPs are the only source of error; other possible artifacts, for example, those due to alignment mistakes, are not accounted for. Applied to other data sets, GIREMI also outperformed previous methods[8] in sensitivity and accuracy (**Supplementary Fig. 3a**,**b** and **Supplementary Table 1**).

The identification of RNA editing sites depends closely on sequencing depth[4,8], and prediction accuracy may deteriorate at lower depths. To examine this relationship, we repeated the analysis with downsampled GM12878 data (**Fig. 1d** and **Supplementary Fig. 3c**,**d**). As expected, the number of

[1]Department of Integrative Biology and Physiology, University of California, Los Angeles, Los Angeles, USA. [2]Bioinformatics Interdepartmental Program, University of California, Los Angeles, Los Angeles, USA. [3]Molecular Biology Institute, University of California, Los Angeles, Los Angeles, USA. Correspondence should be addressed to X.X. (gxxiao@ucla.edu)

RECEIVED 15 JULY 2014; ACCEPTED 23 JANUARY 2015; PUBLISHED ONLINE 2 MARCH 2015; DOI:10.1038/NMETH.3314

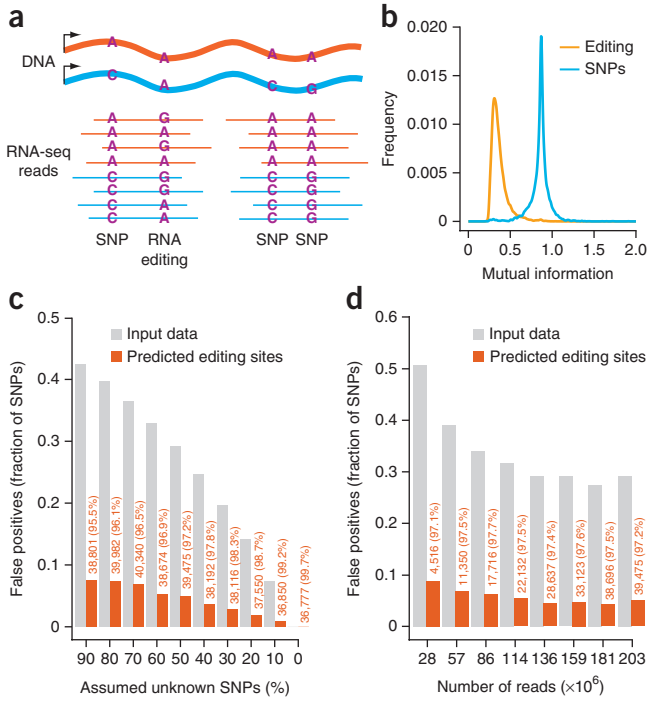

Figure 1 | The GIREMI method. (**a**) The allelic combinations of two SNPs in the same RNA-seq reads are the same as their DNA haplotypes, whereas a SNP and an RNA editing site (or a pair of RNA editing sites) exhibit variable allelic linkage. (**b**) Distributions of mutual information associated with SNPs and RNA editing sites, estimated using GM12878 RNA-seq data (ENCODE, cytosolic, poly(A)$^+$) and its associated genome sequencing data. (**c**) RNA editing sites predicted by GIREMI in the GM12878 data. Different fractions of genomic SNPs of GM12878 were assumed as unknown by excluding them from dbSNP (mean of 9 trials of randomized SNP selection shown for each). Gray bars show the fraction of GM12878 SNPs among all single-nucleotide mismatches in the mapped RNA-seq reads after filtering for artifacts (Online Methods). Orange bars show the fraction of false positives (GM12878 SNPs) among all predicted editing sites (i.e., FDR). The number of predicted editing sites and percent A-to-G editing are listed in orange. (**d**) Performance of GIREMI at different sequencing depths (downsampled GM12878 data). Number of mapped reads (singletons) is shown along the *x* axis. Fifty percent of the GM12878 SNPs were assumed to be unknown. Labels are as in **c**.

RNA editing sites dropped as sequencing depth decreased. Remarkably, unlike previous methods, the accuracy of GIREMI was not affected much by sequencing depth, with low FDR (8.8%) even at very low sequencing depth (<30 million singleton reads or 15 million pairs). Similar performance was observed for single-end data (**Supplementary Fig. 4**).

To further evaluate performance, we compared GIREMI-predicted editing sites to those from a "genome-aware" method that utilizes SNPs identified in whole-genome sequencing data[9] (**Table 1**, **Supplementary Table 2** and **Supplementary Note 3**). In addition, we included results of the genome-independent "multiple data sets" method that calls RNA editing sites using RNA-seq data from multiple samples[5]. For two levels of assumed unknown SNPs (30% and 50%), GIREMI consistently predicted more editing sites at higher accuracy (measured as 1 – % SNPs among predicted editing sites), with greater overlap with the genome-aware method and a higher percentage of A-to-G sites (%AG) than the multiple data sets method (**Supplementary Note 3**). Thus, GIREMI exhibited superior performance despite requiring only a single RNA-seq data set.

Recent studies identified a large number of editing sites in *Alu* regions with high confidence[10,11]. In contrast, accuracy of predicted non-*Alu* editing sites was relatively low, especially for those in coding regions[5]. GIREMI also demonstrated variable accuracy for different types of regions (**Supplementary Table 2** and **Supplementary Notes 3** and **4**). Overall, the sensitivity and accuracy of GIREMI are both high compared to the existing genome-independent method in pinpointing *Alu* and noncoding editing sites of non-*Alu* regions. We obtained similar results for human brain RNA-seq data (**Supplementary Table 3**) reflecting typical individual lab-based projects in which a small number of samples are collected, either with or without biological replicates.

Compared to noncoding sites, editing sites in coding regions (recoding sites) are much less prevalent, and existing methods suffer from low sensitivity and accuracy in pinpointing non-*Alu* coding editing events[5]. On an initial examination, the accuracy of GIREMI was also low (~28% on average) for these sites in nonrepetitive regions but still higher than that of the multiple data sets method (5.3% on average; **Supplementary Table 2**). For a detailed evaluation, we examined whether GIREMI could identify previously reported recoding sites that are conserved between human and mouse[12]. As most recoding sites are highly tissue specific, we used RNA-seq data sets derived from a panel of primary human tissues (**Supplementary Note 5**). Among the 47 recoding sites with adequate read coverage (≥5) in at least one sample, 43 were correctly identified by GIREMI, yielding an overall sensitivity of 91.5% and an average per-sample sensitivity of 71.4% (**Supplementary Table 4**). Given the high sensitivity and the expected small number of non-*Alu* coding sites, we can leverage the rapidly expanding sets of known coding sites to improve accuracy. For the GM12878 data, the accuracy in predicting nonrepetitive coding sites was 67–80% if only known sites were considered (**Supplementary Note 5**).

Table 1 | Performance of GIREMI compared with other methods applied to the GM12878 data (cytosolic, poly(A)$^+$ RNA-seq)

	Genome-aware method[9]		GIREMI				Multiple data sets method[5,a]			
Region	No. of sites	%AG	No. of sites	%AG	Accuracy[b] (%)	Overlap[c] (%)	No. of sites	%AG	Accuracy[b] (%)	Overlap[c] (%)
All	41,027	98.8	37,591	98.6	98.1	90.0	8,307	90.2	85.0	18.5
Alu	39,757	99.7	36,131	99.0	99.4	90.4	7,797	98.5	87.1	24.9
Repetitive non-*Alu*	260	88.6	267	83.7	84.3	86.4	26	65.6	65.4	14.8
Nonrepetitive	1,010	73.5	1,193	82.8	73.8	87.6	484	41.0	55.6	29.2

[a]Results were derived using two data sets (GM12878 and YH RNA-seq; **Supplementary Note 3**). Editing sites were identified in the two data sets separately, and final GM12878 editing sites were called by requiring their presence in YH results. Results of another mode of the multiple data sets method (pooled samples) are included in **Supplementary Table 2**. [b]Accuracy was defined as (1 – % SNPs among predicted editing sites in each category); 30% of GM12878 SNPs were assumed to be unknown in applying the GIREMI and multiple data sets methods. [c]Overlap was calculated relative to the results of the genome-aware method. %AG, percentage of A-to-G sites.

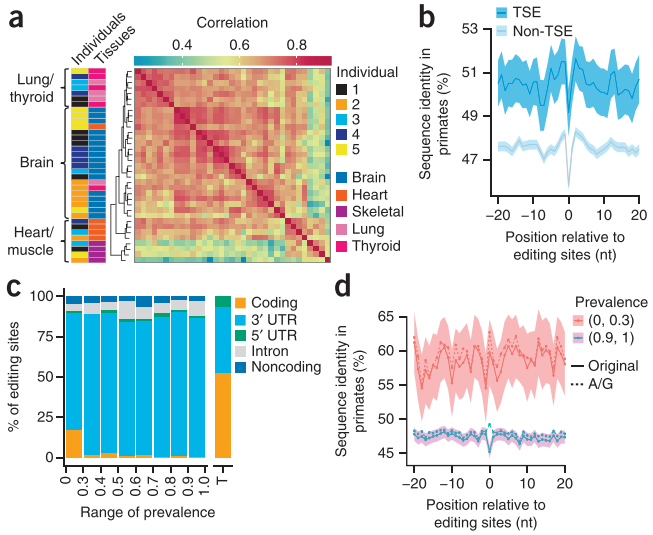

Figure 2 | RNA editomes of human tissues and individuals. (**a**) Comparison of RNA editing sites across human tissues by hierarchical clustering of Pearson correlation coefficients (calculated for editing ratios of all editing sites that are present in 35 samples). Different brain regions are represented in the same color given their highly similar editing profiles. (**b**) Conservation of the immediate neighborhood of tissue-specific editing (TSE) sites in 3′ UTRs. Sequence conservation (percentage of sequence identity in primates) of each position flanking editing sites (position 0) is shown. Shaded regions represent 95% confidence intervals. A similar plot for non-TSE sites is included for comparison. (**c**) Distribution of editing sites in different types of intragenic regions for 93 humans. Editing sites were grouped according to their prevalence in this population. "Noncoding" refers to noncoding genes or noncoding transcripts of coding genes. Regional distribution of nucleotides in the entire transcriptome of coding genes (without introns) is shown as a reference (T). (**d**) Conservation of 3′ UTR regions flanking two groups of editing sites with different prevalence levels (solid lines), similarly as in **b**. Dashed lines correspond to the sequence identity if Gs in other genomes were assumed as a conserved base given a reference nucleotide A in human[9].

Owing to the genome-independent nature of GIREMI, it can be applied to any RNA-seq data set without restrictions. We first examined the variation of editomes across human tissues, a fundamental question not yet addressed on a genome-wide scale. We used a panel of 38 Genotype-Tissue Expression (GTEx) RNA-seq data sets from five human subjects and eight primary tissue types (four brain regions, heart, skeletal muscle, thyroid and lung)[13]. The samples were chosen such that each individual had data from nearly all eight tissues types (**Supplementary Table 5**). When clustered according to how RNA editing ratios correlate in pairwise comparisons, the samples segregated largely by tissue instead of by individual (**Fig. 2a**). Three major tissue groups were observed, encompassing (i) lung and thyroid, (ii) brain regions and (iii) muscle (heart and skeletal). Different brain regions were barely distinguishable on the basis of their editing profiles. This tissue-dominated clustering pattern is especially striking given that the number of predicted editing sites varied greatly across samples largely owing to sequencing depth variation (**Supplementary Fig. 5** and **Supplementary Table 6**). This result is unlikely to be a byproduct of the expected tissue-dominated clustering of overall gene expression, as the editing ratios are not correlated with gene expression levels (**Supplementary Fig. 6**). Thus, our observation supports the existence of tissue-specific regulation of RNA editing. In addition, our result is consistent with a recent report of tissue-dominant clustering of editing sites in rhesus macaque[14]. Notably, in contrast to the previous study, our study included only shallowly sequenced RNA-seq data (12.3–41.1 million mapped read pairs) without specific genomic data of the samples.

In examining the patterns of tissue-specific editing (TSE), we observed the largest difference in RNA editing between brain and muscle-related tissues, with up to 24% of editing sites being specific to brain tissues (**Supplementary Fig. 7a**). In addition, muscle demonstrated a considerably smaller number of editing sites and lower editing levels compared to thyroid or lung (**Supplementary Fig. 7b**). The mRNA expression levels of the RNA-editing enzyme ADAR1 (**Supplementary Fig. 7a**) explained approximately 77% of the variability in editing levels across tissues (**Supplementary Fig. 8a**). Similarly notable concordance was not observed for ADAR2 (**Supplementary Fig. 8b**).

Overall, TSE sites were highly enriched in 3′ UTR regions compared to all editing sites ($P < 2.2 \times 10^{-16}$, Fisher's exact test; **Supplementary Fig. 9a**). Interestingly, higher sequence conservation was observed in 3′ UTR regions harboring TSE sites than in those flanking non-TSE sites (**Fig. 2b**), supporting existence of selection pressure in TSE regions. We observed a number of distinctive genomic features of 3′ UTR TSE sites and their associated genes (**Supplementary Fig. 9**). In addition, brain-specific editing sites were often in genes related to energy, cellular metabolism and apoptosis, whereas lung- or thyroid-specific editing sites were found in genes related to signal peptide processing and response to stimuli (viral or inflammatory) (**Supplementary Table 7**).

We next examined the level of variability in editomes across human individuals, a fundamental question that has not been addressed on a global scale. To this end, we analyzed RNA-seq data of lymphoblastoid cells of 93 people in the 1000 Genomes Project (GBR population)[15] and identified a total of 22,715 editing sites. For each editing site covered by at least ten total reads in ≥50% of individuals, we calculated the fraction of these individuals expressing the edited nucleotide. We used this value to represent the prevalence of an editing site in the population and observed that the majority of editing sites (88%) had a prevalence of at least 50% (**Supplementary Fig. 10a**). Levels of RNA editing varied considerably across the prevalence groups, with an overall trend of enhanced editing as prevalence increased (**Supplementary Fig. 10b**).

All prevalence groups consisted of editing sites enriched in 3′ UTRs relative to the general composition of the human transcriptome (**Fig. 2c**). Notably, a smaller percentage of intronic editing sites was observed here than in the GTEx data set (**Supplementary Fig. 9a**), possibly owing to differences in RNA-seq protocols. Intriguingly, the group of rare editing sites showed a considerably higher enrichment in coding regions than other groups. In addition, rare editing sites were associated with more highly conserved 3′ UTR regions, whereas common editing sites were found in less conserved regions (**Fig. 2d** and **Supplementary Fig. 11**). Although located in functionally important regions (i.e., coding and highly conserved 3′ UTRs), rare editing sites are probably not functionally significant given their low editing levels. Possibly, these editing sites represent random innovations of the transcriptome of few

individuals that have not yet undergone long-term selection. Purifying selection may exist to prevent these sites from gaining higher editing levels or higher prevalence in the population.

In contrast, common editing sites were associated with relatively high editing levels. This observation argues against the possibility that these sites are randomly occurring transcriptome innovations. Rather, common editing sites should be associated with certain advantage such that evolution has preserved their prevalence. Because these sites are less conserved than TSE sites, as with non-TSE sites (**Fig. 2b,d**), it is unlikely that most of the common editing sites are functionally critical. An alternative hypothesis is that many common RNA editing sites are byproducts of the RNA-editing machinery carrying out functions to mediate other aspects of gene expression; perhaps this machinery being under selection has led to an apparent preservation of the RNA editing sites across populations (**Supplementary Note 6**).

We have presented a method for the identification of RNA editing independent of sample-specific genome sequences with high accuracy, even given low sequencing depth. Applying GIREMI yielded novel insights about tissue-specific editing and evolutionary implications of RNA editing, and we expect that the tool will enable many new discoveries from routine RNA-sequencing studies.

METHODS

Methods and any associated references are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We thank members of the Xiao laboratory for comments on this work and for helping with RNA-seq read mapping. We thank the ENCODE, GTEx and the 1000 Genomes Project for generating the data and making their data available to the public. This work was supported in part by US National Institutes of Health grants R01HG006264 and U01HG007013 and by US National Science Foundation grant 1262134.

AUTHOR CONTRIBUTIONS

Q.Z. implemented and developed the GIREMI method and conducted bioinformatic analyses; X.X. conceived the idea, designed and conducted bioinformatic analyses, and wrote the paper with input from Q.Z.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Bass, B.L. *Annu. Rev. Biochem.* **71**, 817–846 (2002).
2. Nishikura, K. *Annu. Rev. Biochem.* **79**, 321–349 (2010).
3. Farajollahi, S. & Maas, S. *Trends Genet.* **26**, 221–230 (2010).
4. Lee, J.H., Ang, J.K. & Xiao, X. *RNA* **19**, 725–732 (2013).
5. Ramaswami, G. *et al. Nat. Methods* **10**, 128–132 (2013).
6. Ensterö, M., Daniel, C., Wahlstedt, H., Major, F. & Ohman, M. *Nucleic Acids Res.* **37**, 6916–6926 (2009).
7. Djebali, S. *et al. Nature* **489**, 101–108 (2012).
8. Chen, L. *Proc. Natl. Acad. Sci. USA* **110**, E2741–E2747 (2013).
9. Bahn, J.H. *et al. Genome Res.* **22**, 142–150 (2012).
10. Bazak, L. *et al. Genome Res.* **24**, 365–376 (2014).
11. Bazak, L., Levanon, E.Y. & Eisenberg, E. *Nucleic Acids Res.* **42**, 6876–6884 (2014).
12. Pinto, Y., Cohen, H.Y. & Levanon, E.Y. *Genome Biol.* **15**, R5 (2014).
13. The GTEx Consortium. *Nat. Genet.* **45**, 580–585 (2013).
14. Chen, J.Y. *et al. PLoS Genet.* **10**, e1004274 (2014).
15. Abecasis, G.R. *et al. Nature* **491**, 56–65 (2012).

ONLINE METHODS

Mapping of RNA-seq reads. RNA-seq reads were mapped to the reference human genome (hg19) and transcriptome (Ensembl release 71) using our previously published method[9,16]. The method was designed to enable unbiased mapping of alternative RNA alleles corresponding to RNA editing or expressed single-nucleotide polymorphisms (SNPs). Briefly, Bowtie[17] and Blat[18] were used to align all reads to the reference genome, and Bowtie was used to align all reads to the transcriptome. Results from the three parallel mapping procedures were merged into a union. Final mapped reads were required to satisfy a dual-filtering scheme such that a read (or a pair of read in paired-end data) maps uniquely with up to n_1 mismatches (per read) and does not map to any other regions with up to n_2 mismatches (per read) ($n_2 > n_1$). For all data sets, n_1 and n_2 values were set to be about 5% and 12% of the read length, respectively. We previously showed that this mapping method effectively reduced the mapping bias to alternative alleles[9,16] and facilitated relatively accurate quantification of allelic ratios compared to other methods[4].

All data sets from ENCODE cell lines, U87MG cells and GTEx human tissues were mapped in the same way as described above. For the 1000 Genomes data sets, we downloaded mapped reads (.bam files) directly. However, we implemented an additional filtering step using Blat to remove possible ambiguous mapping (such as those due to existence of pseudogenes or homologs), similarly to in refs. 19,20.

Preprocessing to identify and filter mismatches in RNA-seq reads. The RNA-seq reads were piled up to identify mismatches relative to the reference human genome (hg19). All duplicate reads were removed within each RNA-seq library except the one with the highest-quality score at the mismatch position. Duplicate reads were defined here as (pairs of) reads mapped to exactly the same genomic locations. For each mismatch position, a total read coverage of ≥5 was required and the variant allele was required to be present in at least three reads. According to previous literature[4,8,19–23], a number of filters were desirable to remove potential artifacts resulted from sequencing or mapping bias. We thus imposed additional procedures as described in our previous work[4] to discard the following types of mismatches: those located in simple repeats regions or homopolymer runs of ≥5 nt, those associated with reads substantially biased toward one strand, those with extreme variant allele frequencies (>95% or <10%) and those located within 4 nt of a known spliced junction. To further reduce the impact of sequencing errors, we calculated a log-likelihood ratio (LLR) to examine the likelihood of a mismatch being a sequencing error, as described in ref. 9. We only retained mismatches passing an LLR cutoff of 2.

The same procedures as described above (read mapping and mismatch filtering) were applied for all methods included in this study, i.e., the GIREMI, genome-aware and multiple data sets methods. In addition, known SNPs (in dbSNP) were excluded from predicted editing sites by the multiple data sets method.

GIREMI. The GIREMI method combines statistical inference of MI between pairs of single-nucleotide variants (SNVs) in RNA-seq reads with machine learning to predict RNA editing sites. The input to GIREMI includes a list of SNVs (mismatches) derived from an RNA-seq data set and known SNPs in public databases such as dbSNP. The output is a collection of predicted RNA editing sites and their editing levels. Except public SNP information, GIREMI carries out all analyses using one RNA-seq data set of interest and does not rely on any other genomic or RNA-seq data sets.

Mutual information (MI) of SNVs and RNA editing prediction. As the first step of GIREMI, we identify known SNPs (from dbSNP) in the list of SNVs derived from the RNA-seq reads. We then extract all RNA-seq reads that harbor the known SNPs and the subset of reads (or read pairs in paired-end RNA-seq; required ≥5 such reads) that cover more than one SNP. SNP pairs located in the same (pairs of) reads were retained for MI calculation. As an example, in the GM12878 data set, a total of 5,306 SNPs (out of 37,775 SNPs covered by ≥5 RNA-seq reads) were involved in this calculation. In another less deeply sequenced RNA-seq data set (GTEx SRR595926, 31M mapped reads), 884 SNPs out of a total of 10,590 RNA-seq–covered SNPs were used for this step. Although the percentage of SNPs used for the calculation of MI is not high, it is adequate to generate the reference MI distribution (such as that in **Fig. 1b**) for further prediction of RNA editing sites. The number of RNA editing sites suitable for MI calculation is much larger than that of SNPs. For example, 32,548 editing sites were used to generate the example distribution of MI (**Fig. 1b**), where our previous genome-dependent method was applied to predict RNA editing sites[9].

For each SNV s_i, we consider all possible nucleotides A, C, G and T as the four possible states of the variable s_i. Thus, for a joint variable representing a pair of mismatches (s_i, s_j), a total of 16 states are possible. Although it is unlikely that all 16 states are present in one RNA-seq data set, we use this scheme because it is general and can accommodate possible existence of sequencing errors or other complexity. The probabilities of observing each state of s_i, s_j or (s_i, s_j) were calculated using the maximum-likelihood method. A value of 0.01 was assumed for states that were not observed in the actual data considering existence of sequencing errors of all possible nucleotides and accounting for low sequencing depth in realistic data sets. Incorporation of this pseudo-value led to an increase of MI of about 0.2 for both SNPs and editing sites (**Fig. 1b**), but the final editing predictions with or without this pseudo-value are very similar (data not shown). The MI of (s_i, s_j) is thus

$$I(s_i,s_j) = \sum_{n_i \in N} \sum_{n_j \in N} p(n_i,n_j) \log\left(\frac{p(n_i,n_j)}{p(n_i)p(n_j)}\right)$$

where $N = \{A, C, G, T\}$ and n_i and n_j represent the states of s_i and s_j, respectively. We used natural log for the above formula.

Then, the MI of a SNP s_i is defined as

$$I(s_i) = \frac{\sum_{s_j \in S} I(s_i,s_j)}{T}$$

where S is the collection of other SNPs paired with s_i, and T is the total number of pairs in S.

As an example, the distribution of $I(s_i)$ values for SNP pairs detected in the GM12878 data set is shown (**Fig. 1b**). In theory, the maximum MI should be log(2) = 0.7 for SNP pairs.

However, in practice, larger values were sometimes observed, owing to limited read coverage at each site and the numerical difference between joint probability and marginal probability of the states. The marginal probability was estimated using all reads covering the particular SNV, whereas the joint probability was estimated using reads covering both SNVs. Thus, the number of joint reads is often smaller than that of the marginal reads and the joint probability is less accurately estimated than the marginal ones. This discordance sometimes led to MI values larger than the theoretical upper bound.

For each RNA-seq data set, the MI of SNPs is calculated independently. Thus, a data set–specific distribution of $I(s_i, s_j)$ is derived. Subsequently, for a SNV s_x that is not a known SNP, an $I(s_x)$ value is calculated similarly as described above by examining its relationship with other SNVs (either known SNPs or otherwise). On the basis of the distribution of $I(s_i)$ of known SNPs, a P value is calculated for s_x to test the null hypothesis that $I(s_x)$ is not different from the distribution of $I(s_i)$ for SNPs. A P-value cutoff of 0.05 was used to call RNA editing sites. Correction of P values for multiple testing was not applied owing to the discovery nature of this test. As an example, we predicted 31,660 RNA editing sites (99.6% being the A-to-G type) in this step for the GM12878 data set.

Generalized linear model (GLM) for the prediction of RNA editing. As the second step of GIREMI, RNA editing sites identified by the MI approach are used to train a GLM to predict additional editing sites. The GLM incorporates two types of features that have discriminative power for SNPs and RNA editing sites. The first feature quantifies the deviation of the allelic ratio of the unknown SNV from an expected allelic ratio reflecting the allelic expression of the respective gene. The second type of feature represents the sequence preferences of the neighborhoods of RNA editing sites (mainly A-to-G). It should be noted that the GLM step only analyzes A-to-G mismatches as candidate RNA editing sites, without including other types of SNVs.

To estimate the expected allelic ratio r of a gene g, we extract all expressed heterozygous known SNPs (S) (dbSNP) in gene g (read coverage ≥5). The allelic ratio r is calculated by maximizing the log-likelihood function $\log L(r \mid D)$, where D refers to the RNA-seq data for gene g. We assume reads covering a specific SNP s_j in gene g follow a binomial distribution. Thus, the estimated allelic ratio $\hat{r}$ that maximizes $\log L(r \mid D)$ of gene g is

$$\hat{r} = \frac{\sum_{s_j \in S} m_{s_j}}{\sum_{s_j \in S} (m_{s_j} + n_{s_j})}$$

where m and n refer to the number of reads with alternative and reference alleles, respectively. In practice, the haplotype information for SNPs in S is not known. Thus, we arbitrarily assign m_{s_j} as the read count for the major allele and n_{s_j} as that for the minor allele in the RNA-seq data. This assumption may cause a biased allelic ratio larger than the actual value. Nevertheless, the same directional bias exists in the allelic ratio for a specific SNV that is to be compared to $\hat{r}$. Thus, the impact of this bias will be largely canceled out. In cases where no SNP is available in gene g, an expected ratio of 0.5 is used assuming the gene has no allelic expression bias.

A heterozygous SNP is expected to have an allelic ratio that is largely consistent as the allelic ratio of the gene. In contrast, RNA editing sites may have allelic ratios that substantially deviate from that of the gene. Thus, we use the absolute difference (d) between the allelic ratio of the unknown SNVs and the estimated $\hat{r}$ of the gene as one feature in the GLM. This feature has the discriminative power for SNPs and RNA editing sites (**Supplementary Fig. 1c**), but exceptions do exist. For example, it cannot identify editing sites with editing levels similar to allelic ratios of genetic SNPs in the same gene. In addition, a minor fraction of SNPs may have allelic ratios largely different from that of the entire gene if allele-specific splicing or other local RNA processing events affect the allelic expression of the SNPs[16].

To increase the discriminative power, we incorporated sequence-based features into the GLM. Importantly, these features are not based on sequence motifs built from a priori knowledge regarding RNA editing. Instead, they were derived using editing sites predicted by the MI step of GIREMI. Thus, the features are specific to the data set of interest without any a priori assumptions. To this end, we generate a positional weight matrix (PWM) for the sequence neighborhood of the predicted editing sites (**Supplementary Fig. 1d**). For an unknown SNV, a composite sequence score was calculated using its −1 and +1 nucleotides according to the PWM. It should be noted that putative editing sites predicted by the MI-based approach are mostly (>97%) of the A-to-G type. Thus, the sequence features derived here largely reflect those of A-to-I editing that is known to demonstrate nucleotide preferences at the −1 and +1 positions[9].

Together, for each unknown SNV of the A-to-G type, the GLM estimation is

$$Y = g^{-1}(\beta_0 + \beta_d d + \beta_c c)$$

where d represents the difference between the allelic ratio of the SNV and the estimated $\hat{r}$ of the gene and c denotes the composite score for the sequence features. β_0, β_d and β_c are the respective coefficients of the GLM, which are solved using a binomial link function.

The GLM of each RNA-seq data set was trained using the putative editing sites predicted by the MI approach. In addition, a leave-one-out scheme was applied where the genetic allelic ratio was estimated using all expressed heterozygous SNPs except one per gene. These randomly excluded SNPs were used as training data together with the putative editing sites. The training data were then separated into two random subsets of the same size. The first subset was used to parameterize the GLM model. The recall and precision of the predictive model were evaluated using the second subset. To reach a trade-off between the recall and precision, an F measure was calculated as follows:

$$F_\beta = \frac{(1+\beta^2) \times \text{precision} \times \text{recall}}{\beta^2 \times \text{precision} + \text{recall}}$$

In the above F measure, we set β to be 0.5, which puts more emphasis on precision than recall. Finally, a cutoff for the predicted probability of a site being an RNA editing site was chosen to achieve an F measure of 0.75.

doi:10.1038/nmeth.3314

It should be noted that, although GLM was designed to predict A-to-G sites only, the MI method was not restricted to identification of A-to-G sites alone. Thus, other types of RNA-DNA mismatches do exist in the final results, but with the vast majority being A-to-G. The biological credibility of the other types of RNA-DNA mismatches is still under debate, which is not a focus of this work.

The two steps in GIREMI demonstrate different efficacies for different types of editing sites. The MI step is most effective for editing sites in close proximity with other editing sites or SNPs (such as A-to-I editing in *Alu* regions that are known to cluster together). Its sensitivity is lower in predicting editing sites in isolation. In contrast, the GLM step, although contributing a relatively small number of additional sites overall as a second step of GIREMI, is an important procedure to ensure high sensitivity in identifying recoding sites. Thus, both steps are essential for our method.

Code availability. GIREMI was implemented using a combination of R, Perl and C codes. The package is available at https://www.ibp.ucla.edu/research/xiao/GIREMI.html.

RNA-seq and SNP data sets. ENCODE RNA-seq data sets were downloaded from the UCSC Genome Browser (http://genome.ucsc.edu/ENCODE/). In this study, we used the data sets derived from cytosolic polyadenylated RNA. U87MG RNA-seq data (wild type and *ADAR1* knock down) were obtained from our previous study[9]. GTEx RNA-seq data sets were downloaded from dbGAP with permission. The 1000 Genomes RNA-seq data were downloaded from the Geuvadis project (http://www.geuvadis.org/). SNPs derived from genome sequencing data were obtained from the 1000 Genomes Project for GM12878 and a genome sequencing project for U87MG cells[24]. Public SNP data were obtained from dbSNP (version 137).

Tissue-specific editing (TSE). In the analysis of the GTEx data, we compared the editomes of any pair of tissues included in this study. Each editing site was required to have a read coverage of at least ten reads in ≥75% of samples (i.e., individuals) in either tissue under comparison. A moderated *t*-test[25] was applied to determine whether the editing levels were significantly different across the two tissues (using samples that meet with the read coverage cutoff; FDR <5%). Editing sites that passed this test were defined as TSEs.

The heat map of editomes of different tissues (**Fig. 2a**) was generated on the basis of Pearson correlation of pairs of samples. For each sample pair, only RNA editing sites with adequate read coverage (at least ten reads) in both samples were included. Hierarchical clustering was used to generate the clusters.

Conservation analysis of regions flanking editing sites. The same method as in our previous work[9] was used to evaluate the conservation level of each editing site and their flanking regions. Briefly, with the 46-way multiz alignments from the UCSC browser[26], we focused on the ten primates, including human, chimp, gorilla, orangutan, rhesus, baboon, marmoset, tarsier, mouse lemur and bush baby. On the basis of the multiple sequence alignments, the percent identity at each nucleotide position of interest was calculated, together with a 95% confidence interval.

Gene Ontology (GO) analysis. GO analysis was conducted similarly as in ref. 27. Briefly, the GO terms of each gene were obtained from Ensembl. To identify GO categories that are enriched in a specific set of genes, we compared the number of genes in the set with a particular GO term to that in a control gene set. The control gene set was constructed so that the randomly picked controls and the test genes have one-to-one matched transcript length and G+C content. On the basis of 10,000 randomly selected control sets, a *P* value for enrichment of each GO category in the test gene set was calculated as the fraction of times that F_{test} was lower than or equal to $F_{control}$, where F_{test} and $F_{control}$ denote, respectively, the fraction of genes in the test set and a random control set associated with the current GO category. A *P*-value cutoff (the smaller of 1/10,000 or 1/total number of GO terms considered) was applied to choose significantly enriched GO terms.

Comparison of TSEs with binding sites of RNA-binding proteins (RBPs). Publicly available CLIP-Seq data were collected for hnRNP A1, A2/B1, F, M, U (ref. 28), hnRNP H (ref. 29), hnRNP C (ref. 30), AGO2, IGF2BP1, QKI, PUM2 (ref. 31), DGCR8 (ref. 32), ELAVL1 (ref. 33), EWSR1, FUS, TAF15 (ref. 34), LIN28 (ref. 35), MOV10 (ref. 36), PTB (ref. 37), SFRS1 (ref. 38), TDP43 (ref. 39), TIA1 and TIAL1 (ref. 40). CLIP tag clusters were directly downloaded from the above publications or generated using our in-house pipeline. TSEs in 3′ UTRs were then examined for their overlap with CLIP clusters of the above proteins collectively, similarly for non-TSEs.

16. Li, G. *et al. Nucleic Acids Res.* **40**, e104 (2012).
17. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S.L. *Genome Biol.* **10**, R25 (2009).
18. Kent, W.J. *Genome Res.* **12**, 656–664 (2002).
19. Peng, Z. *et al. Nat. Biotechnol.* **30**, 253–260 (2012).
20. Ramaswami, G. *et al. Nat. Methods* **9**, 579–581 (2012).
21. Kleinman, C.L. & Majewski, J. *Science* **335**, 1302 (2012).
22. Lin, W., Piskol, R., Tan, M.H. & Li, J.B. *Science* **335**, 1302 (2012).
23. Pickrell, J.K., Gilad, Y. & Pritchard, J.K. *Science* **335**, 1302 (2012).
24. Clark, M.J. *et al. PLoS Genet.* **6**, e1000832 (2010).
25. Smyth, G.K. in *Bioinformatics and Computational Biology Solutions Using R and Bioconductor* (eds. Gentleman, R., Carey, V., Dudoit, S., Irizarry, R. & Huber, W.) 397–420 (Springer, 2005).
26. Dreszer, T.R. *et al. Nucleic Acids Res.* **40**, D918–D923 (2012).
27. Lee, J.H. *et al. Circ. Res.* **109**, 1332–1341 (2011).
28. Huelga, S.C. *et al. Cell Rep.* **1**, 167–178 (2012).
29. Katz, Y., Wang, E.T., Airoldi, E.M. & Burge, C.B. *Nat. Methods* **7**, 1009–1015 (2010).
30. König, J. *et al. Nat. Struct. Mol. Biol.* **17**, 909–915 (2010).
31. Hafner, M. *et al. Cell* **141**, 129–141 (2010).
32. Macias, S. *et al. Nat. Struct. Mol. Biol.* **19**, 760–766 (2012).
33. Mukherjee, N. *et al. Mol. Cell* **43**, 327–339 (2011).
34. Hoell, J.I. *et al. Nat. Struct. Mol. Biol.* **18**, 1428–1431 (2011).
35. Wilbert, M.L. *et al. Mol. Cell* **48**, 195–206 (2012).
36. Sievers, C., Schlumpf, T., Sawarkar, R., Comoglio, F. & Paro, R. *Nucleic Acids Res.* **40**, e160 (2012).
37. Xue, Y. *et al. Mol. Cell* **36**, 996–1006 (2009).
38. Sanford, J.R. *et al. Genome Res.* **19**, 381–394 (2009).
39. Tollervey, J.R. *et al. Nat. Neurosci.* **14**, 452–458 (2011).
40. Wang, Z. *et al. PLoS Biol.* **8**, e1000530 (2010).