Characterization and comparison of human nuclear and cytosolic editomes

Liang Chen¹

Molecular and Computational Biology, Department of Biological Sciences, University of Southern California, Los Angeles, CA 90089

Edited by Peter J. Bickel, University of California, Berkeley, CA, and approved June 13, 2013 (received for review October 29, 2012)

We developed a robust computational statistical framework to identify RNA editing events from RNA-Seq data with high specificity. Our approach handles several outstanding challenges of genome-wide editing analyses, including the effect of editing on read alignment and the utilization of redundant reads. By applying this framework, we characterized the nuclear and cytosolic editomes of seven human cell lines. We found that 93.8–99.2% of the editing events are A-to-G (or A-to-I). Nuclear transcriptomes contain many more editing events than cytosolic transcriptomes. Most of the sites exhibiting nucleus-specific editing are in introns or novel intergenic transcripts that are preferentially localized in the nucleus regardless of their editing status, arguing against the role of editing in nuclear retention. In contrast, many sites that exhibit cytosol-specific editing show comparable nuclear and cytosolic expression, suggesting the differential subcellular compartmentalization of the edited and the unedited alleles. We found that RNA editing is globally associated with the modification of microRNA regulation in 3′ untranslated regions, whereas editing events in coding regions are rare and tend to be synonymous. Interestingly, A-to-G editing at derived alleles in the human lineage tends to result in reversion back to the ancestral forms at the RNA level. This suggests that editing can mediate RNA memory on evolutionary time-scales to maintain ancestral genetic information.

high-throughput sequencing | nucleo-cytoplasmic localization

Although high-throughput sequencing has facilitated the identification of genetic DNA variants, the identification of RNA editing is much more challenging owing to the uneven read distribution in RNA sequencing and the varied RNA editing rates. Errors in base calling and read alignment, the handling of redundant reads, and other systematic sequencing errors all hinder the single-nucleotide-level analysis of RNA editing. Early analyses of high-throughput sequencing data reported more than 10,000 exonic editing events, many of which were non-A-to-G (1). However, recent studies argued that many of these non-A-to-G editing events were false positives resulting from inappropriate bioinformatic analyses (2–4). To deal with multiple sources of potential errors, the existing tools for editing discovery (5–7) usually involve comprehensive pipelines packed with multiple read aligners, read simulations, and various filters. Meanwhile, many critical issues in editing discovery remain to be addressed more properly. For example, the existing methods remove a large number of redundant reads starting from the same positions, although in deep sequencing many redundant reads can result from true expression and are useful for estimating editing levels. Moreover, the effect of editing on read alignment has been commonly overlooked.

Herein, we developed an efficient and robust computational statistical framework for RNA editing discovery. This framework considers the effect of editing on read alignment and handles redundant reads through statistical approaches. Our framework does not require multiple aligners or tedious read simulations. We tested our method on three existing RNA editing studies and achieved improved performance. Our analytic results provide guiding information of the required coverage for editing

detection and explain the reason for an outstanding question in the field: the lack of consistency across replicates in editing discovery.

Although some individual RNA editing events can change coding sequences or are involved in microRNA (miRNA) regulation or nuclear retention (8–11), the subcellular compartmentalization and the functional consequences of global RNA editing are still largely unclear. We applied our method to the nuclear and cytosolic RNA-Seq data of seven human cell lines with a total of 2.8 billion read pairs from the Encyclopedia of DNA Elements (ENCODE) project (12). For each dataset, we identified, at most, 84,316 A-to-G editing events and 2,063 non-A-to-G editing events. The A-to-G percentage was as high as 93.8–99.2%. We then conducted a detailed analysis of the subcellular compartmentalization of editing and the editing landscape of the human genome. This study represents a thorough characterization and comparison of human nuclear and cytosolic editomes, and reveals insights about the functional consequences and the evolutionary features of editing.

Results

An Efficient Computational Statistical Tool for RNA Editing Analysis.

To accurately identify RNA editing sites, we developed an efficient computational statistical framework that properly handles read alignment, redundant reads, and systematic sequencing errors (*SI Methods* and *Fig. S1A*). To consider the effect of editing on read alignment, we adjusted alignment scores and probabilities based on each potential variant allele instead of fixing read alignment based on reference sequences. To handle redundant reads appropriately, we assigned different weights to reads according to their base-call qualities and the adjusted alignment probabilities. The aggregate weight for reads mapped to the same position was set to be one. Then both weighted and

Significance

RNA editing can alter the information content of genome sequences and increase transcriptome diversity. Global analyses of RNA editing remain challenging, largely due to the analytical barriers to separating editing events from errors in sequencing and alignment. Our computational statistical framework addresses these challenges and will also benefit other single-nucleotide level analyses of RNA-Seq data. Our results provide guidance for detecting nucleotide variants in RNA-Seq. We characterized human nuclear and cytosolic editomes and revealed relationships between editing and nucleocytoplasmic compartmentalization, microRNA regulation, and coding capacity. We also discovered that A-to-G editing mediates RNA memory of ancestral DNA alleles.

Author contributions: L.C. designed research, performed research, contributed new reagents/analytic tools, analyzed data, and wrote the paper.

The author declares no conflict of interest.

This article is a PNAS Direct Submission.

¹E-mail: liang.chen@usc.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1218884110/-DCSupplemental.

conventional analysis pipelines. Nuclear RNA contained many more (1.5-fold to 7.2-fold) edited sites than their cytosolic counterparts, although the total numbers of nuclear and cytosolic reads were similar (Fig. 1A and Table S2). The median number of A-to-G editing sites for the cell lines was 52,689 in the nucleus and 10,403 in the cytosol. The median number of non-A-to-G editing sites was 1,993 in the nucleus and 465 in the cytosol.

Both nuclear and cytosolic transcriptomes were enriched for editing events in introns and intergenic regions, despite the small percentage of reads mapped to these regions (Fig. 1B and Fig. S3). Specifically, 72.9% of the nuclear and 45.6% of the cytosolic editing events occurred in introns, although introns contained only 27.0% and 4.8% of the respective mapped reads. The relatively high percentage of intronic reads in nuclear poly(A)⁺ transcriptomes was probably due to the delay in exon splicing (15). Another 17.4% and 27.7% of the editing events were in intergenic regions, whereas only 9.0% of the nuclear and 7.0% of the cytosolic reads were mapped to these regions. Because the samples were generated from >200-base pair (bp) poly(A)⁺ transcripts, these novel intergenic transcripts were likely unannotated long noncoding RNAs. Notably, the prevalence of editing events in introns and intergenic regions was not caused by the high coverage of these sites. Instead, they had lower coverage than the editing sites in other regions (median was 41 vs. 90 for the nucleus and 40 vs. 98 for the cytosol). Additionally, 3' UTR editing events represented 20.2% of the cytosolic editome but only 5.3% of the nuclear editome, although the absolute numbers of the 3' UTR editing events were similar between the cytosol and the nucleus (Fig. 1B).

To compare the chance of being edited for positions in different annotation categories, we normalized the number of editing events by the number of expressed bases. A coverage threshold of 99 was used to declare an expressed base. As shown in Fig. 1C, the median chance of being edited was 0.37% and 0.35% for the expressed positions in introns and intergenic regions, respectively. Expressed coding sequences (CDSs) had the smallest chance to be edited with a median of 0.002%. The median chance was 0.02–0.05% for other regions. All these suggest that introns and intergenic regions are more prone to be edited. As noted earlier, nuclear transcriptomes exhibited 1.9-fold to 17.0-fold more intronic editing sites and 1.2-fold to 4.0-fold more intergenic editing sites than cytosolic transcriptomes across the seven cell lines. However, after the normalization by the number of expressed bases, the chance of being edited in the nucleus was not consistently higher than that in the cytosol. Their differences became smaller and sometimes indistinguishable. Similar conclusions were obtained when the coverage threshold of 38 or 295 was applied.

We further examined the types of editing events. We called positions edited in at least one cell line editable sites. Of the 259,385 editable sites, only seven sites displayed different editing types across different cell lines and different subcellular locations. This result further indicates a low number of false positives in our analysis. Non-A-to-G editing events comprised only 2.9% of all of the editing events. Among them, C-to-T and G-to-A were the most frequent (25.3% and 22.5%, respectively), followed by T-to-C (11.6%). Individual examples of C-to-T, G-to-A, and T-to-C editing were reported previously (16–18). The distributions of the non-A-to-G editing types were similar between the nucleus and the cytosol and between Alu and non-Alu sites (Fig. S4). In contrast to A-to-G events with the vast majority of them in Alu regions, nearly 84.7% of the non-A-to-G editing events occurred in non-Alu regions.

Cell Line Specificity of Human RNA Editing. We next examined the cell-line-specific editing sites. We found that 68.1% of the nuclear editable sites and 70.8% of the cytosolic editable sites were edited in only one cell line. This can be due to the cell-line-

specific regulation of editing, or the cell-line-specific expression of the site, or sampling fluctuations in read coverage. To exclude the latter two possibilities, we considered positions that were covered with ≥ 38 reads in all cell lines and were edited at the level of 0.2–0.8 in at least one cell line. Among these positions, only 20.9% of the nuclear and 32.7% of the cytosolic editable sites were specific to one cell line, indicating cell-line-specific regulation of editing. In contrast to the poor consistency across cell lines or even biological replicates reported in previous studies (14, 19), our results suggest that many editing sites are indeed shared by multiple cell lines if we take into account the expression difference and the sampling fluctuation in read coverage.

At the gene level, 12,205 genes in nuclear transcriptomes and 6,839 genes in cytosolic transcriptomes underwent RNA editing in their exons and/or introns. Approximately 71.7% and 51.1% of these genes, respectively, contained editing events in more than one cell line.

Editing and Nucleus–Cytosol Compartmentalization. At the steady state, there were many cytosol- or nucleus-specific editing sites (Fig. 1B). They can result from the differential compartmentalization of the edited and unedited alleles (e.g., edited alleles are preferentially localized in one compartment), or from the specific subcellular localization of the transcripts independent of the editing status (e.g., both edited and unedited alleles disappear in one compartment). To distinguish the two possibilities, we compared the nuclear and the cytosolic coverage (including both edited and unedited alleles) of these sites (Fig. 2). Across different genome regions, the cytosol-specific editing sites exhibited the median coverage of 29–90 in the nucleus where their editing was not detected (Fig. 2A). Specifically, the sites in introns, exons of noncoding transcripts, CDSs, and 5' UTRs had a comparable or even higher expression level in the nucleus than in the cytosol. Although the cytosol-specific editing sites in intergenic regions and 3' UTRs had a lower nuclear expression level than cytosolic expression, about 41.4% and 60.4% of them had nuclear coverage no less than 38. Thus, the lack of detectable nuclear editing for these sites was not caused by insufficient nuclear coverage.

We further examined the relationship between read coverage and the observed probability of detecting editing in both compartments. With the increased coverage, the chance of detecting editing in both compartments also increased, as shown in Fig. S5. However, at the coverage threshold of 38, the detection probability was reaching the plateau (Fig. S5). Thus, we already had reasonable analytic power to detect the editing events if they existed in both compartments. Taken together, these results suggest the cytosol-specific compartmentalization of the edited alleles. The cytosol-specific editing might result from the differential nucleus–cytosol transport or the differential degradation of the edited and unedited alleles; another possibility is that these editing reactions are catalyzed in the cytosol instead of the nucleus (20).

For the nucleus-specific editing sites, we also examined their nuclear and cytosolic coverage. The nucleus-specific editing sites in introns, intergenic regions, exons of noncoding transcripts, and 5' UTRs had a low cytosolic expression level with a median coverage of 3–9 (Fig. 2B). Thus, these nucleus-specific transcripts or transcript fragments, edited or unedited, are not effectively transported to the cytosol or are degraded rapidly in the cytosol. Although many of the nucleus-specific editing sites in 3' UTRs and CDSs were expressed in the cytosol, they made up only 3.4% of the nucleus-specific editing events. To test whether sequencing depth affected the conclusions on the cytosol- and nucleus-specific editing, we normalized the raw coverage by sequencing depth and obtained the same conclusions (Fig. S6).

Functional Consequences of RNA Editing. The enrichment of editing in 3' UTRs hints at the functional consequences of global edit-

Fig. 2. Nuclear and cytosolic coverage of sites whose editing is detected in only one subcellular location. The read coverage at the site includes both the edited and unedited alleles. (A) Sites with cytosol-specific editing. For the sites in introns, the nuclear coverage was even higher than the cytosolic coverage ($P < 10^{-60}$; paired Wilcoxon tests). For the sites in exons of noncoding transcripts ("noncoding exon"), CDs, and 5' UTRs, the differences between the nuclear and the cytosolic coverage are not significant ($P > 0.1$; paired Wilcoxon tests). (B) Sites with nucleus-specific editing. The sites excluding those in 3' UTRs and CDs have low cytosolic coverage (median coverage 3–9). Outliers are not plotted on the boxplots.

ing. Specifically, we investigated the 3' UTR editing sites for their overlapping with miRNA targets (Fig. 3A). The miRNA targets were predicted by the miRanda algorithm (21) for both unedited alleles and edited alleles. Approximately 16.1% of the edited alleles disrupted the original miRNA targets on the unedited alleles, and another 19.8% created new miRNA targets. Almost 10.2% of the edited alleles switched the miRNA targets by destroying the original ones and creating new ones. For comparison, we performed the same analysis on a control set of 3' UTR sites. Because the majority of the 3' UTR editable sites (11,611 out of 12,718) underwent A-to-G editing, we selected a random set of 12,718 3' UTR positions of nucleotide A and changed them to G. As shown in Fig. 3A, the percentages of affected miRNA targets were significantly lower in the simulated control set ($P < 10^{-60}$; proportion tests). These results suggest that RNA editing events in 3' UTRs are tightly coupled to miRNA regulation.

We also examined to what extent RNA editing can change encoded amino acids. CDS regions contained a small number of editable positions (1,711, or 0.7% of the total editable positions),

and 59.3% of them underwent non-A-to-G editing. As shown in Fig. 3B, 29.3% of the A-to-G changes and 33.5% of the non-A-to-G changes on CDSs were synonymous. If we changed each nucleotide base of all human CDSs to the other three nucleotide bases, only ~23.0% of these changes were synonymous, significantly smaller than the values observed for either A-to-G or non-A-to-G editing ($P = 5.1 \times 10^{-11}$ and 1.0×10^{-23} , respectively; proportion tests). These results show that the likelihood of a CDS position being edited on the RNA level is small, and the edited alleles tend to be synonymous.

Editing Mediates RNA Memory of Ancestral DNA Alleles. Finally, we studied the evolutionary features of the editable sites. The editable positions usually displayed lower conservation scores than their flanking regions, consistent with previous reports (5). Additionally, the lower conservation patterns were consistent across different annotation categories (Fig. S7). The nuclear editable sites were generally less conserved than the cytosolic editable sites in CDs, consistent with the notion that cytosol-specific

Fig. 3. Functional consequences of RNA editing. (A) Effect of RNA editing on miRNA targets in 3' UTRs. A random set of 12,718 positions of nucleotide A in 3' UTRs is changed to G. Compared with the random set, the edited alleles tend to disrupt the original miRNA targets, or create new miRNA targets, or switch miRNA targets ($***P < 10^{-60}$; proportion tests). (B) Effect of RNA editing on coding. Both non-A-to-G and A-to-G events tend to be synonymous, compared with the random situation in which all CDS positions are changed to the other three nucleotides ($**P < 10^{-20}$ for non-A-to-G and $*P < 10^{-10}$ for A-to-G; proportion tests). The errors bars show the 95% confidence intervals.

transcripts are usually more conserved than nucleus-specific transcripts (22). In addition, the three-nucleotide periodic conservation patterns in CDSs was consistent with our earlier observation that editing in CDSs tends to be synonymous and the common notion that the third position of a codon tends to be more mutable and its nucleotide substitution tends to be synonymous. Nevertheless, the CDS editable sites were still much less conserved than the neighboring third positions.

We further examined the DNA evolutionary path of these editable sites. Specifically, the genome sequence of the human ancestor was inferred from the sequence alignments among humans, chimpanzees, and macaques and was downloaded from

the Ensembl database (www.ensembl.org; more details are given in *SI Methods*). We isolated human A-to-G editable sites that were not A (or not T for sites on the minus strand) in the ancestral sequence. Thus, these sites experienced DNA mutation from the ancestor to humans, and they were edited from A-to-G at the RNA level in humans (Fig. 4A). Surprisingly, for 94.2% of the positions, the edited allele matched the ancestral DNA allele. However, if we randomly selected the same number of A or T nucleotide positions from human positions that were mutated during evolution, the expected proportion of positions exhibiting G or C in the ancestral sequence was only 69.6%. This was significantly lower than 94.2% in the case of editing sites ($P = 9.5 \times 10^{-155}$; proportion test; Fig. 4B), suggesting that A-to-G editing in humans tends to reverse the G-to-A transition during human evolution. We called this phenomenon editing-mediated RNA memory of evolution (Fig. 4A).

Although non-A-to-G editing constitutes a small fraction of total human editing events, we tested whether these sites also exhibited editing-mediated RNA memory of evolution. We found that only 77.6% of the non-A-to-G editing reversed the DNA alteration between the human ancestor and humans. This was still significantly higher than the expected proportion 50.6%. However, the P value ($P = 2.8 \times 10^{-7}$; proportion test) was less significant than that of the A-to-G editing (Fig. 4B).

To investigate whether human RNA editing also memorizes the alleles of a more distant ancestor, we inferred the sequence of a distant ancestor for 13 eutherian mammals, including humans and mice (*SI Methods*). As shown in Fig. 4B, among those sites that were mutated from the distant ancestor and were edited from A to G in humans, 78.8% of them were G (or C for sites on the minus strand) in the distant ancestor, significantly higher than that expected by chance ($P = 7.0 \times 10^{-31}$; proportion test). However, this proportion was significantly lower than the 94.2% for the recent human ancestor ($P = 8.4 \times 10^{-72}$; proportion test). It indicates that RNA editing is better at mediating short-term rather than

Fig. 4. Editing-mediated RNA memory of evolution. (A) RNA memory through editing. (B) RNA memory in humans. For positions mutated during evolution and edited at the RNA level in humans, their edited alleles have a significantly higher chance to match the ancestral alleles compared with the control situation (except for non-A-to-G editing in relation to the distant ancestor). Corresponding control sets are the randomly selected human positions of the same nucleotides as the editable sites and were mutated from the ancestors. (C) RNA memory in mice. Only A-to-G editing shows a significantly higher chance to reverse DNA alternations to the recent mouse ancestor sequence. The errors bars show the 95% confidence intervals ($****P < 10^{-150}$, $***P < 10^{-30}$, $**P < 10^{-6}$, and $*P < 10^{-4}$; otherwise, $P > 0.01$; proportion tests).

long-term memory of evolution. Additionally, the study of the non-A-to-G editing sites showed little long-term memory compared with the random case with a P value of 0.04.

To further demonstrate that recent ancestral alleles are more likely to be memorized by RNA editing than distant ancestral alleles, we examined positions that were nucleotide A in humans and G in only one of the two ancestors. For the nonG→G→A (or nonC→C→T) positions along the evolution path of distant ancestor→recent ancestor→humans, 0.029% of them underwent A-to-G editing to memorize the recent ancestral allele (i.e., short-term memory, Fig. 4A). Conversely, among the G→nonG→A (or C→nonC→T) positions, only 0.0027% of them underwent A-to-G editing to maintain the distant ancestral allele (i.e., long-term memory, Fig. 4A). These observations further suggest that A-to-G editing is better at memorizing recent ancestral alleles than at memorizing distant ancestral alleles.

To understand whether editing-mediated RNA memory of evolution is restricted to humans, we applied the same analyses on an existing mouse editing dataset (23). Because the frequency of A-to-G editing in mice is at least an order of magnitude lower than that in humans (24), this dataset contained only about 5,000 editable sites across 15 mouse strains. We inferred the sequence of the mouse ancestor from the alignments among rats, mice, and rabbits (*SI Methods*) and performed similar analyses on the potential editing-mediated RNA memory. As shown in Fig. 4C, about 89.3% of the mouse mutated sites with A-to-G editing reversed the mutations between the mouse ancestor and mice, significantly higher than the proportion expected by chance (57.9%; $P = 2.9 \times 10^{-5}$; proportion test). When we consider the more distant ancestor of the 13 eutherian mammals, the percentage was decreased to 65.5% with a P value of 0.02 compared with the random case. A very small fraction of mouse editable sites were non-A-to-G in the considered dataset. Only two of them reversed the mutations to the recent ancestral alleles, and nine sites reversed the mutations to the distant ancestral alleles, with the expected numbers of 1.2 and 8.3, respectively ($P = 1$ in both cases; proportion test). Overall, the results suggest that A-to-G editing-mediated RNA memory also exists in mice for the recent ancestor, but the signals are not as strong as those in humans.

Discussion

We developed a robust and efficient framework to identify RNA editing events. It obtained a higher A-to-G percentage and a smaller false discovery rate in detecting A-to-G events than other methods. In our analysis of the ENCODE data, only 7 (out of 259,385) sites displayed different editing types across cell lines and subcellular locations, further indicating the low number of false positives. The A-to-G percentage was larger than 93% for all of the nuclear and cytosolic transcriptomes.

Our method handles read alignment comprehensively. An edited nucleotide position can significantly change the alignment score and redistribute the read to different loci. Our model adjusted the alignment scores and the alignment probabilities based on each assumed editing event. This strategy is also valuable for other single-nucleotide-level analyses, such as SNP analysis and allele-specific expression analysis. We used Novoalign to perform the sequence alignment because the hash-based algorithm is generally more sensitive than the algorithms based on Burrows–Wheeler transform data compression (25). We indexed all chromosomes, unplaced contigs, and splice junctions together for the read alignment. Similar to ref. 19, this “non-exon-first” approach avoided the erroneous mapping of real junction reads to intronless pseudogenes.

Our analytic results have provided insights into the coverage requirement for editing discovery. We found that coverage of 38 is usually required to detect sites edited at a level between 0.2 and 0.8. The required coverage increases to 99 or 295 when editing levels are more extreme (Fig. S1C). Due to variable gene

expression levels and strong positional biases in RNA-Seq (26–28), enormous raw reads are required to cover the whole transcriptome at the level of 38×. These results suggest that target-enrichment strategies for weakly expressed genes are urgently needed for editing analysis. The previously reported poor consistency across cell lines or even biological replicates put the biological significance and the importance of global editing under suspicion (14, 19). Our study, however, shows that the poor consistency is mainly due to the sensitivity of editing discovery to sequencing depths.

Owing to the challenges in computational and statistical analyses, the existence and prevalence of non-A-to-G editing events remain controversial. Our analysis identified a relatively small number of non-A-to-G events ranging from 284 to 2,063 in a variety of the ENCODE cell lines. Because we did not use cell-line-specific SNP information for the ENCODE datasets, a proportion of the identified non-A-to-G events may actually be unannotated SNPs. Indeed, the non-A-to-G editing events showed a median editing level of 0.45, close to 0.50, the ratio for SNPs. Although we do not intend to exclude the possible existence of non-A-to-G events, our analyses show that the number of non-A-to-G events is at least an order of magnitude lower than that of A-to-G events.

The comparison of nuclear and cytosolic editomes elucidates the functional consequences of global RNA editing. Many editing events occurred in the introns of nuclear transcriptomes. Intronic editing events are widespread in nascent *Drosophila* RNA (29). Our analyses on human cell lines also show that intronic editing events are prevalent. Because most introns are spliced out when mature mRNA is transported to the cytosol, many of these editing events are not detected in the cytosol. RNA editing was previously hypothesized to mark aberrant transcripts such as novel intergenic transcripts for nuclear retention and/or degradation (30). However, our detailed comparison between the nuclear and cytosolic transcriptomes shows that most of the editable intergenic transcripts are retained in the nucleus regardless of their editing status. Conversely, most of the sites with cytosol-specific editing show comparable expression in the nucleus, but only the edited alleles tend to be compartmentalized in the cytosol.

Most surprisingly, we found that A-to-G editing tends to reverse the DNA mutations to the ancestral alleles at the RNA level. In other words, edited RNAs representing the ancestral sequence status are floating around in the cell. They may revise the DNA to the ancestral genotypes spontaneously in a way similar to the processed pseudogenes. It is unknown how cells use editing to generate RNA memory of evolution. However, such memory might be beneficial during evolution to correct deleterious DNA mutations. Of all of the ancestor–descendant pairs that we examined, the strongest signal of RNA memory came from the comparison between humans and their recent ancestor. It is unclear whether it is a sheer coincidence that RNA editing is more prevalent in humans than other vertebrates that have been examined. Interestingly, in a time scale of evolution, this RNA memory seems to be stronger in a shorter term and extinct in a longer term.

Methods

For reads mapped to a considered position, we adjusted the alignment scores by assuming the potential variant alleles as I_1 and I_2 . Thus, as long as the read nucleotide was the same as I_1 or I_2 , we treated it as a match instead of imposing a mismatch penalty. The probability of the read sequence R_j given the alignment location A_{ji} in the reference genome G : $P(R_j|A_{ji}, G)$ was changed accordingly. As in Novoalign (www.novocraft.com), the posterior alignment probability for location A_{ji} was

$$P(A_{ji}|R_j, G) = P(R_j|A_{ji}, G) / (P(R_j|N, G) + \sum_i P(R_j|A_{ji}, G)),$$

where i' was for all alignment locations of read j , and $P(R_j|N, G)$ was the probability the read resulted from the regions of the reference sequences coded as N s. The likelihood of the read with allele y_j at this position, given l_1 and l_2 , was

$$P(Y_j = y_j | Q_j, l_1)P(A_{ji} | R_j, G, l_1, l_2)(1 - \beta_{l_1 l_2}) + P(Y_j = y_j | Q_j, l_2)P(A_{ji} | R_j, G, l_1, l_2)\beta_{l_1 l_2},$$

where $P(Y_j = y_j | Q_j, l_k)$ was the likelihood of nucleotide y_j given base-call quality Q_j and the underlying nucleotide l_k , and it was equal to $1.0 - 10^{-0.1Q_j}$ when $y_j = l_k$ and $1/3 \times 10^{-0.1Q_j}$ when $y_j \neq l_k$ and $\beta_{l_1 l_2}$ was the editing level for allele l_2 . For simplicity, we estimated $\beta_{l_1 l_2}$ as

$$\frac{\sum_j P(Y_j = l_1 | Q_j, l_1)P(A_{ji} | R_j, G, l_1, l_2)}{\sum_j P(Y_j = l_1 | Q_j)P(A_{ji} | R_j, G, l_1, l_2) + \sum_j P(Y_j = l_2 | Q_j, l_2)P(A_{ji} | R_j, G, l_1, l_2)}. \quad [1]$$

A more accurate estimation could be obtained by the maximum likelihood estimate with increased computational complexity. The posterior probability of allele pair l_1 and l_2 given Y and R was then calculated by assuming the same prior probabilities for different l_1 and l_2 values. The pair of l_1 and l_2 with the largest posterior probability was the two variant alleles for the potential editing sites. We further required that the largest posterior probability should be ≥ 0.99 . Note that the original reads could contain more than two variant alleles because some of alleles may result from sequencing errors. Nevertheless, other methods (6) simply remove all of the positions with ≥ 3 observed alleles. We then retained positions covered by more unique reads than multireads for each allele.

Log-likelihood ratio tests were performed to determine whether $\beta_{l_1 l_2} = 0$ or 1. Thus, we required the expression of both alleles. A weighting scheme was performed to handle reads starting from the same position. Specifically, different weights were assigned to the log-likelihood calculation of different reads. The weighted log-likelihood was

$$\sum_j w_j \log(P(Y_j = y_j | Q_j, l_1)P(A_{ji} | R_j, G, l_1, l_2)(1 - \beta_{l_1 l_2}) + P(Y_j = y_j | Q_j, l_2)P(A_{ji} | R_j, G, l_1, l_2)\beta_{l_1 l_2}).$$

The weight depended on both the base quality Q_j and $P(A_{ji} | R_j, G, l_1, l_2)$, and the sum of the weights for reads starting from the same position was set to 1. Thus,

- Li M, et al. (2011) Widespread RNA and DNA sequence differences in the human transcriptome. *Science* 333(6038):53–58.
- Kleinman CL, Majewski J (2012) Comment on "Widespread RNA and DNA sequence differences in the human transcriptome." *Science* 335(6074):1302, author reply 1302.
- Lin W, Piskol R, Tan MH, Li JB (2012) Comment on "Widespread RNA and DNA sequence differences in the human transcriptome." *Science* 335(6074):1302, author reply 1302.
- Pickrell JK, Gilad Y, Pritchard JK (2012) Comment on "Widespread RNA and DNA sequence differences in the human transcriptome." *Science* 335(6074):1302, author reply 1302.
- Bahn JH, et al. (2012) Accurate identification of A-to-I RNA editing in human by transcriptome sequencing. *Genome Res* 22(1):142–150.
- Peng Z, et al. (2012) Comprehensive analysis of RNA-Seq data reveals extensive RNA editing in a human transcriptome. *Nat Biotechnol* 30(3):253–260.
- Ramaswami G, et al. (2012) Accurate identification of human Alu and non-Alu RNA editing sites. *Nat Methods* 9(6):579–581.
- Pullirsch D, Jantsch MF (2010) Proteome diversification by adenosine to inosine RNA editing. *RNA Biol* 7(2):205–212.
- Prasanth KV, et al. (2005) Regulating gene expression through RNA nuclear retention. *Cell* 123(2):249–263.
- Nishikura K (2010) Functions and regulation of RNA editing by ADAR deaminases. *Annu Rev Biochem* 79:321–349.
- Hundley HA, Bass BL (2010) ADAR editing in double-stranded UTRs and other noncoding RNA sequences. *Trends Biochem Sci* 35(7):377–383.
- Dunham I, et al.; ENCODE Project Consortium (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature* 489(7414):57–74.
- Djebali S, et al. (2012) Landscape of transcription in human cells. *Nature* 489(7414):101–108.
- Park E, Williams B, Wold BJ, Mortazavi A (2012) RNA editing in the human ENCODE RNA-seq data. *Genome Res* 22(9):1626–1633.
- Tilgner H, et al. (2012) Deep sequencing of subcellular RNA fractions shows splicing to be predominantly co-transcriptional in the human genome but inefficient for lncRNAs. *Genome Res* 22(9):1616–1625.

$$w_j = \frac{(1.0 - 10^{-0.1Q_j})P(A_{ji} | R_j, G, l_1, l_2)}{\sum_{j'} (1.0 - 10^{-0.1Q_{j'}})P(A_{j'i} | R_{j'}, G, l_1, l_2)},$$

where j' was for all reads starting from the same position as read j . We required that both the regular log-likelihood ratio test and the weighted log-likelihood ratio test had P values of 10^{-5} based on the χ^2 distribution against both $\beta_{l_1 l_2} = 0$ and 1. We also calculated the weighted $\beta_{l_1 l_2}$ by inserting the weight of each read in Eq. 1, and the weighted β s for the nonreference alleles were reported as the final editing levels.

Our method further removed potential strand-specific errors by requiring at least one read on both strands for each of the two alleles and a P value > 0.1 in the Fisher's exact test of the strand bias between the two alleles. The read count was corrected by the base quality and the alignment probability. Thus, one read was counted as $(1.0 - 10^{-0.1Q_j})P(A_{ji} | R_j, G, l_1, l_2)$. In addition, we required at least one read (also corrected by the base quality and the alignment probability) in which the minimal distance from the editing site to the read end was greater than one-third of the read length. We removed positions with a homopolymer run of ≥ 5 bp upstream or downstream.

Finally, we removed sites that were potential SNPs according to the SNP database (version SNP137, <http://genome.ucsc.edu/>) and sites in simple repeat regions (defined by either the simple repeats track or the repeat masker track in the University of California, Santa Cruz genome browser). We focused exclusively on sites one of whose RNA alleles matched the reference DNA allele. To test our model, we reanalyzed the U87MG dataset (5), the YH poly(A)⁺ dataset (6), and the GM12878 whole-cell poly(A)⁺ dataset (7). The ENCODE poly(A)⁺ RNA-Seq data were for > 200 -nucleotide RNA in the nucleus or the cytosol of seven human cell lines. The seven cell lines, including both cancer cell lines and normal cell lines, were GM12878, HeLa-S3, HepG2, HUVEC, NHEK, K562, and H1-hESC. More details on the data processing and the data analysis can be found in *SI Methods*.

ACKNOWLEDGMENTS. I thank the ENCODE project for generating the RNA-Seq data and making these data available to the public. I thank the anonymous reviewer for providing insightful comments and Sika Zheng for his helpful comments in preparing the manuscript. Research reported in this publication was supported by the National Institute of General Medical Sciences Award R01GM097230 and National Human Genome Research Institute Award P50HG002790.

- Klimek-Tomczak K, et al. (2006) Editing of hnRNP K protein mRNA in colorectal adenocarcinoma and surrounding mucosa. *Br J Cancer* 94(4):586–592.
- Rosenberg BR, Hamilton CE, Mwambi MM, Dewell S, Papavasiliou FN (2011) Transcriptome-wide sequencing reveals numerous APOBEC1 mRNA-editing targets in transcript 3' UTRs. *Nat Struct Mol Biol* 18(2):230–236.
- Sharma PM, Bowman M, Madden SL, Rauscher FJ, 3rd, Sukumar S (1994) RNA editing in the Wilms' tumor susceptibility gene, WT1. *Genes Dev* 8(6):720–731.
- Kleinman CL, Adoue V, Majewski J (2012) RNA editing of protein sequences: A rare event in human transcriptomes. *RNA* 18(9):1586–1596.
- Patterson JB, Samuel CE (1995) Expression and regulation by interferon of a double-stranded-RNA-specific adenosine deaminase from human cells: Evidence for two forms of the deaminase. *Mol Cell Biol* 15(10):5376–5388.
- Enright AJ, et al. (2003) MicroRNA targets in *Drosophila*. *Genome Biol* 5(1):R1.
- Chen L (2010) A global comparison between nuclear and cytosolic transcriptomes reveals differential compartmentalization of alternative transcript isoforms. *Nucleic Acids Res* 38(4):1086–1097.
- Danecek P, et al. (2012) High levels of RNA-editing site conservation amongst 15 laboratory mouse strains. *Genome Biol* 13(4):26.
- Eisenberg E, et al. (2005) Is abundant A-to-I RNA editing primate-specific? *Trends Genet* 21(2):77–81.
- Nielsen R, Paul JS, Albrechtsen A, Song YS (2011) Genotype and SNP calling from next-generation sequencing data. *Nat Rev Genet* 12(6):443–451.
- Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* 5(7):621–628.
- Srivastava S, Chen L (2010) A two-parameter generalized Poisson model to improve the analysis of RNA-seq data. *Nucleic Acids Res* 38(17):e170.
- Wang Z, Gerstein M, Snyder M (2009) RNA-Seq: A revolutionary tool for transcriptomics. *Nat Rev Genet* 10(1):57–63.
- Rodriguez J, Menet JS, Rosbash M (2012) Nascent-seq indicates widespread cotranscriptional RNA editing in *Drosophila*. *Mol Cell* 47(1):27–37.
- Kim DD, et al. (2004) Widespread RNA editing of embedded alu elements in the human transcriptome. *Genome Res* 14(9):1719–1725.