


RESEARCH

Open Access



Human A-to-I RNA editing SNP loci are enriched in GWAS signals for autoimmune diseases and under balancing selection

Hui Zhang^{1,2†}, Qiang Fu^{1†}, Xinrui Shi^{1†}, Ziqing Pan¹, Wenbing Yang¹, Zichao Huang¹, Tian Tang³, Xionglei He¹ and Rui Zhang^{1,4*} 

* Correspondence: zhangrui3@mail.sysu.edu.cn

[†]Hui Zhang, Qiang Fu and Xinrui Shi contributed equally to this work.

¹Key Laboratory of Gene Engineering of the Ministry of Education, State Key Laboratory of Biocontrol, School of Life Sciences, Sun Yat-Sen University, Guangzhou, People's Republic of China
⁴RNA Biomedical Institute, Sun Yat-Sen Memorial Hospital, Sun Yat-Sen University, Guangzhou, People's Republic of China
Full list of author information is available at the end of the article

Abstract

Background: Adenosine-to-inosine (A-to-I) RNA editing plays important roles in diversifying the transcriptome and preventing MDA5 sensing of endogenous dsRNA as nonself. To date, few studies have investigated the population genomic signatures of A-to-I editing due to the lack of editing sites overlapping with SNPs.

Results: In this study, we applied a pipeline to robustly identify SNP editing sites from population transcriptomic data and combined functional genomics, GWAS, and population genomics approaches to study the function and evolution of A-to-I editing. We find that the G allele, which is equivalent to edited I, is overrepresented in editing SNPs. Functionally, A/G editing SNPs are highly enriched in GWAS signals of autoimmune and immune-related diseases. Evolutionarily, derived allele frequency distributions of A/G editing SNPs for both A and G alleles as the ancestral alleles are skewed toward intermediate frequency alleles relative to neutral SNPs, a hallmark of balancing selection, suggesting that both A and G alleles are functionally important. The signal of balancing selection is confirmed by a number of additional population genomic analyses.

Conclusions: We uncovered a hidden layer of A-to-I RNA editing SNP loci as a common target of balancing selection, and we propose that the maintenance of such editing SNP variations may be at least partially due to constraints on the resolution of the balance between immune activity and self-tolerance.

Keywords: A-to-I RNA editing, Autoimmune and immune-related diseases, Transcriptome, Balancing selection

Introduction

RNA editing is a process through which the sequence of an RNA is post-transcriptionally altered from that encoded in the DNA [1, 2]. A-to-I RNA editing is the most common type of RNA editing in metazoans [3]. It is mediated by Adenosine Deaminases Acting on RNA (ADARs), which bind dsRNA regions of protein-coding and non-coding RNAs and deaminate adenosine to inosine [2, 4, 5]. Inosine pairs



© The Author(s). 2020 **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

preferentially with cytidine, as opposed to uridine; therefore, editing alters the sequence and base-pairing properties of both protein-coding and non-coding RNAs. Editing of protein-coding genes may lead to nonsynonymous substitutions, and editing in non-coding RNAs or non-coding parts of protein-coding genes may regulate the splicing and stability of mRNA via multiple mechanisms [6, 7]. Furthermore, it has recently been shown that ADAR1-mediated editing of endogenous dsRNAs, particularly those in the non-coding regions, could disrupt the structures of dsRNA that were potentially bound by MDA5 and block MDA5-mediated immune response; therefore, RNA editing is required to prevent activation of the cytosolic innate immune system. This is most probably the essential function of ADAR1 editing [8–10].

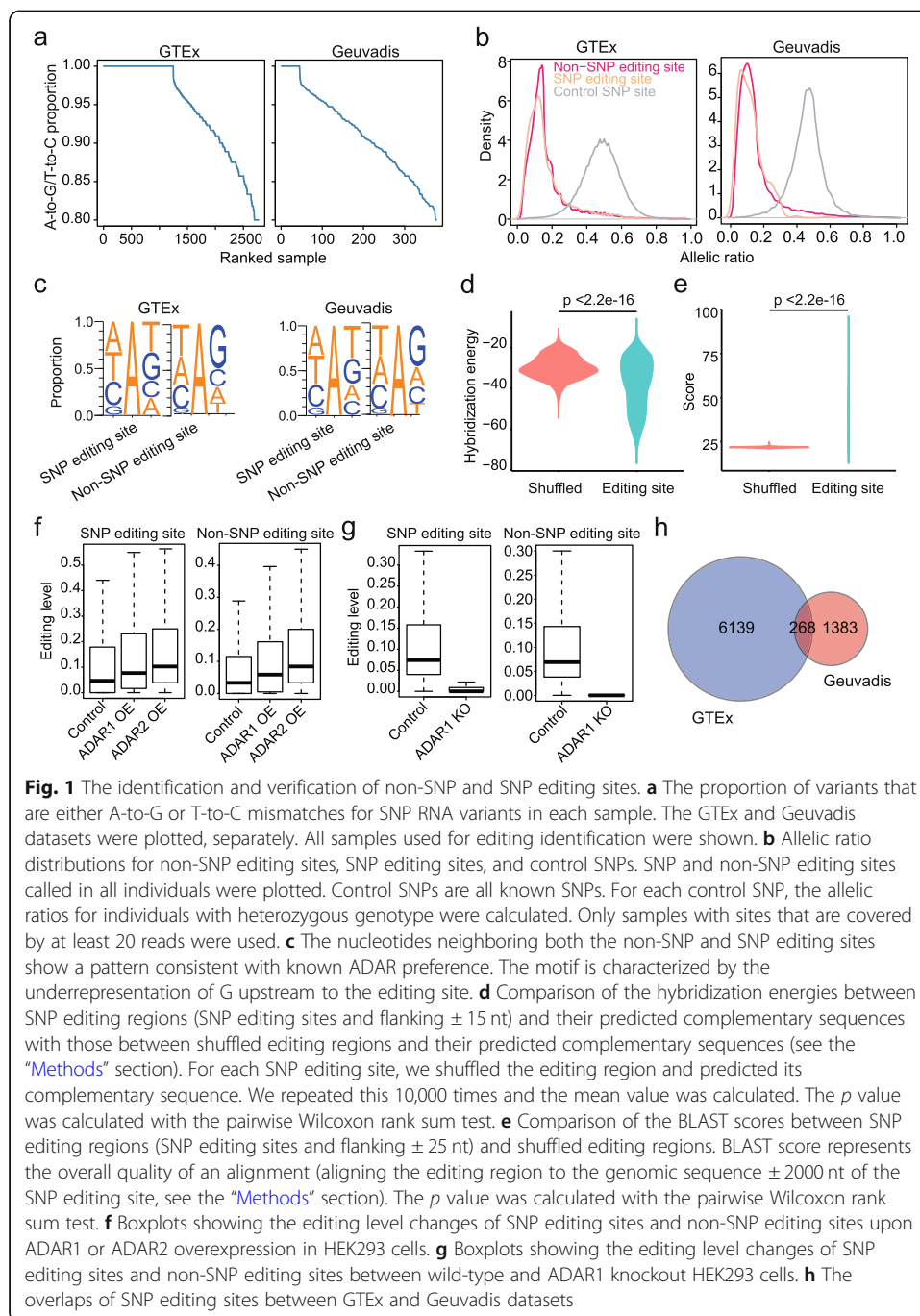
Recent genome-wide searches of editing sites revealed thousands to millions of A-to-I editing events in various species [3]. However, identification of RNA editing sites is still hampered by the difficulty to distinguish true editing sites from A/G genomic variants, especially for samples without matched genomic and transcriptomic data. To minimize false positives arising from genomic variants, the editing identification pipelines that have been recently developed by us and other groups generally discard sites overlapping with known SNPs (e.g., [11–13]). Consequently, previous evolution studies focused on non-SNP editing sites, particularly those that lead to nonsynonymous changes [14–20]. To date, little is known about the population genomic signature of RNA editing. Before the next-generation sequencing era, a pioneer study has identified A-to-I RNA editing sites in the SNP database [21]. In many cases, SNPs overlapped with editing sites are annotated using expressed sequence tags, and thus are RNA editing sites instead of SNPs. However, it is possible that some of these SNPs are real SNPs that can be edited. And such editing SNPs are of importance for functional and evolutionary studies of RNA editing.

Here, we modified our previous approach to achieve robust identification of both non-SNP editing sites and SNP editing sites (editing sites overlapped with SNPs) for samples with matched genomic and transcriptomic data. We applied this approach to samples from Genotype-Tissue Expression (GTEx) and Geuvadis projects and identified SNP editing sites to study the function and evolution of RNA editing in humans.

Results

Identification of SNP editing sites

To generate the list of SNP editing sites, we modified our previous pipeline [11, 22] to retain RNA variants overlapping with known SNPs for editing site calling (Additional file 1: Fig. S1), and applied the modified pipeline to both GTEx and Geuvadis datasets. The GTEx v7 dataset contains a total of 11,688 samples, and we selected the transcriptomes of 3315 human samples (representing 27 tissue types, Additional file 2: Table S1) for analysis. The Geuvadis dataset contains transcriptome data from 464 immortalized B cell line samples. To identify editing sites with high confidence, we only selected samples in which the proportion of A-to-G/T-to-C variants to total variants were at least 80% for editing site call [11, 22]. Editing sites that were overlapped or non-overlapped with SNPs were called separately (Fig. 1a and Additional file 1: Fig. S2a). For the GTEx dataset, we identified 6407 SNP editing sites and 259,462 non-SNP editing sites. For the Geuvadis dataset, we identified 1651 SNP editing sites and 34,419



non-SNP editing sites (Additional file 3: Table S2). The proportions of SNP editing sites in the GTEx and Geuvadis datasets are 0.024 and 0.046, respectively. The number of identified SNP editing sites increased with increasing sample size and remained unsaturated (Additional file 1: Fig. S2b), suggesting the presence of a hidden layer of editing sites that were previously ignored in the human genome. As a control, we applied the same pipeline to call C-to-T/G-to-A editing SNPs in the GTEx and Geuvadis datasets. We identified 11 and 0 sites (including 3 and 0 non-Alu sites in the CDS regions),

which were much less than the 6407 and 1651 A-to-G sites we identified (including 460 and 68 non-Alu sites in the CDS regions).

To validate whether the identified editing sites are bona fide A-to-I editing events, we examined their editing levels in comparison with RNA allelic ratios of known human SNPs. We found that, for both non-SNP and SNP editing sites, the distributions of their editing levels differed from allelic ratios of known heterozygous SNPs, which were centered at 0.5 (Fig. 1b). We also examined RNA editing triplet motifs for non-SNP and SNP editing sites. We found that both were associated with the underrepresentation of guanosines immediately 5' of the edited adenosine (Fig. 1c), consistent with the known ADAR preference [23, 24]. When examining non-Alu CDS and other SNP editing sites separately, we found that non-Alu CDS sites had a slightly weaker ADAR motif than other sites (Additional file 1: Fig. S3). These analyses support that both non-SNP and SNP RNA variants we called were enriched in authentic editing events. In addition, more caution is needed when investigating individual non-Alu CDS sites since they may have a higher false-discovery rate than sites in other genic regions.

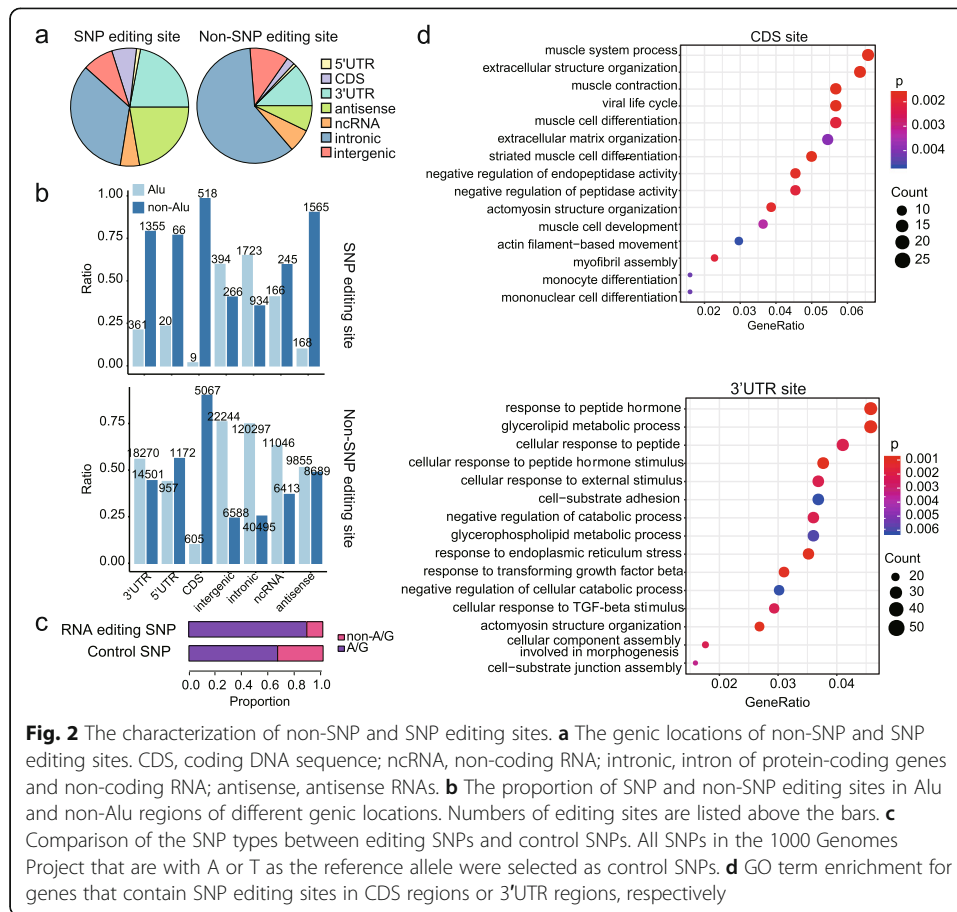
To ask whether SNP editing sites are enriched in regions with dsRNA structures required for ADAR activity, we performed two analyses. First, we predicted the editing complementary sequence (ECS) of each editing region (SNP editing site and flanking ± 15 nt) and the shuffled editing region, as previously described [25]. Next, we compared the hybridization energies between SNP editing regions and their predicted ECSs with those between shuffled editing regions and their predicted ECSs. We found significantly lower hybridization energies of editing regions than the shuffled regions (Fig. 1d). Moreover, about 44% of the SNP editing sites had a statistically significant ECS (see the "Methods" section). Second, we detected the potential dsRNA structures containing SNP editing sites using bl2seq, as previously described [26]. We found that the editing regions formed dsRNA structures with significantly higher alignment scores as compared to the shuffled regions (Fig. 1e). The same analyses were performed for non-Alu SNP editing sites, and the conclusions still held (Additional file 1: Fig. S4a-b).

Finally, to experimentally verify that SNP editing sites are real A-to-I editing events, we examined their editing level changes upon ADAR1 or ADAR2 overexpression in HEK293 cells [27]. We found that both non-SNP and SNP editing sites had increased levels upon overexpression of individual ADARs (Fig. 1f). We also examined ADAR1 knockout HEK293 cells [28] and ADAR1 or ADAR2 knockdown B cells [29]. As expected, both non-SNP and SNP editing sites had decreased editing levels in the knockout or knockdown cells (Fig. 1g and Additional file 1: Fig. S4c-d).

Having proved the validity of SNP editing sites, as many SNP editing sites were not overlapped between the two datasets (Fig. 1h), possibly due to the different origins of tissue types, we merged the two lists and obtained a total of 7790 SNP editing sites, including 278 recoding sites, for analysis.

Characterizing SNP editing sites

A comparison between non-SNP and SNP editing sites revealed the difference in their genic locations and repetitive sequence features. SNP editing sites tended to be in CDS regions and 3'UTR regions, while non-SNP editing sites tended to be in the intergenic regions and intronic regions (Fig. 2a). In addition, compared with non-SNP editing



sites, SNP editing sites tended to be in non-Alu regions in all genic locations studied (Fig. 2b). The densities of editing SNPs varied among different functional classes, and such difference is not due to the difference of background SNP densities in different functional classes (Additional file 1: Fig. S5). These findings suggest that SNP editing sites may be functionally important. In line with this, when examining the types of SNPs in SNP editing sites, we found that editing SNPs were biased toward A/G or T/C (for genes in the forward or reverse strand) genotypes compared with the control SNPs (Fig. 2c). This result suggests that the G allele, which is functionally equivalent to edited I, was selected to be maintained in SNP editing sites.

To ask the possible functional significance of SNP editing sites, we performed Gene Ontology (GO) term analysis. As editing sites in CDS regions and 3'UTR regions may lead to different functional consequences, these two sets of sites were analyzed, separately. We found that genes containing CDS SNP editing sites were highly enriched in muscle-related functions (Fig. 2d). This result is in line with the finding that ADAR2, which is the primary editor of nonrepetitive coding sites [30], had the highest expression level in the artery (Additional file 1: Fig. S6a). For example, a nonsynonymous SNP editing site was identified in gene CASQ2 (muscle contraction GO term). CASQ2 is highly expressed in the heart and artery (Additional file 1: Fig. S6b) and involved in the storage and transport of positively charged calcium atoms [31]. CASQ2 plays an integral role in cardiac regulation, and its mutations have been associated with cardiac

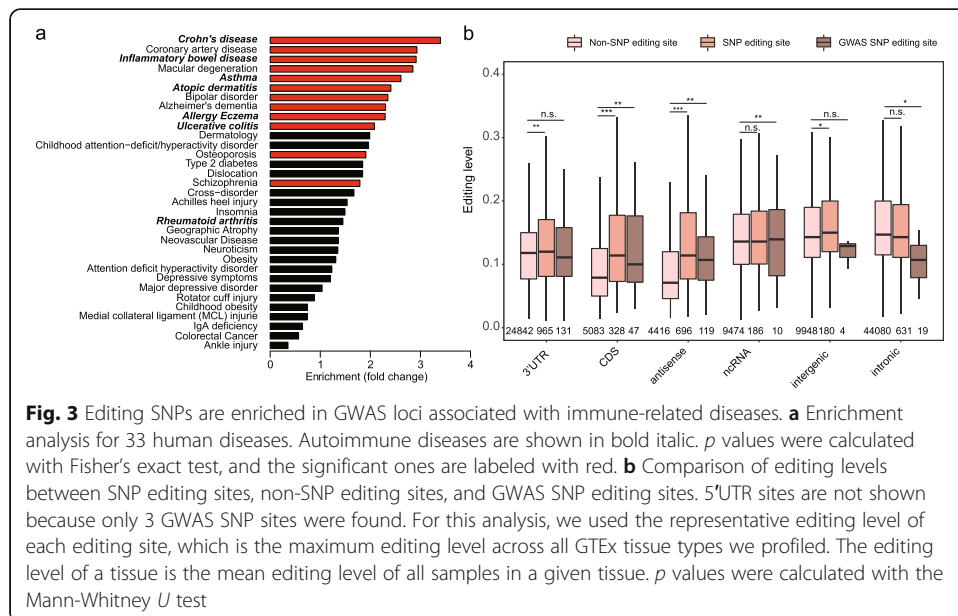
arrhythmia and sudden death [32]. In contrast, genes containing 3'UTR SNP editing sites, which are likely ADAR1 targets, were highly enriched in functional categories such as cellular response to ER stress or external stimulus (Fig. 2d).

Taken together, these results suggest a hidden layer of A-to-I RNA editing events that are likely functionally important.

The link between SNP editing sites and autoimmune and immune-related functions

Genome-wide association studies (GWAS) have led to the identification of thousands of SNPs linked to disease susceptibility in complex human diseases [33]. Based on our finding that in most cases the SNP type of editing SNPs is equivalent to the editing type (A/G SNP vs A-to-I editing), to characterize the potential phenotypic effect of SNP editing sites, we utilized GWAS data and examined the enrichment of editing SNPs in human disease GWAS. Interestingly, we found that editing SNPs are highly enriched in GWAS signals for autoimmune (for example, inflammatory bowel disease (IBD) and Crohn's disease) and immune-related diseases (e.g., coronary artery disease [34]) (Fig. 3a), suggesting that RNA editing may play an important role in autoimmune and immune-related functions. This agrees with the fact that (1) the major biological function of RNA editing is to suppress dsRNA-mediated autoimmunity, and (2) ADAR1 loss-of-function and MDA5 gain-of-function mutations are identified in autoimmune diseases [8–10].

We next compared the editing levels between non-SNP editing sites, SNP editing sites, and GWAS SNP editing sites, which may be informative for assessing their biological significance. We found that SNP and GWAS SNP editing sites in CDS regions had higher editing levels than non-SNP editing sites, while SNP editing sites in intronic or intergenic regions had similar editing levels as compared with non-SNP editing sites (Fig. 3b). Thus, it seems that SNP editing sites in functionally important regions tended to have higher editing levels.



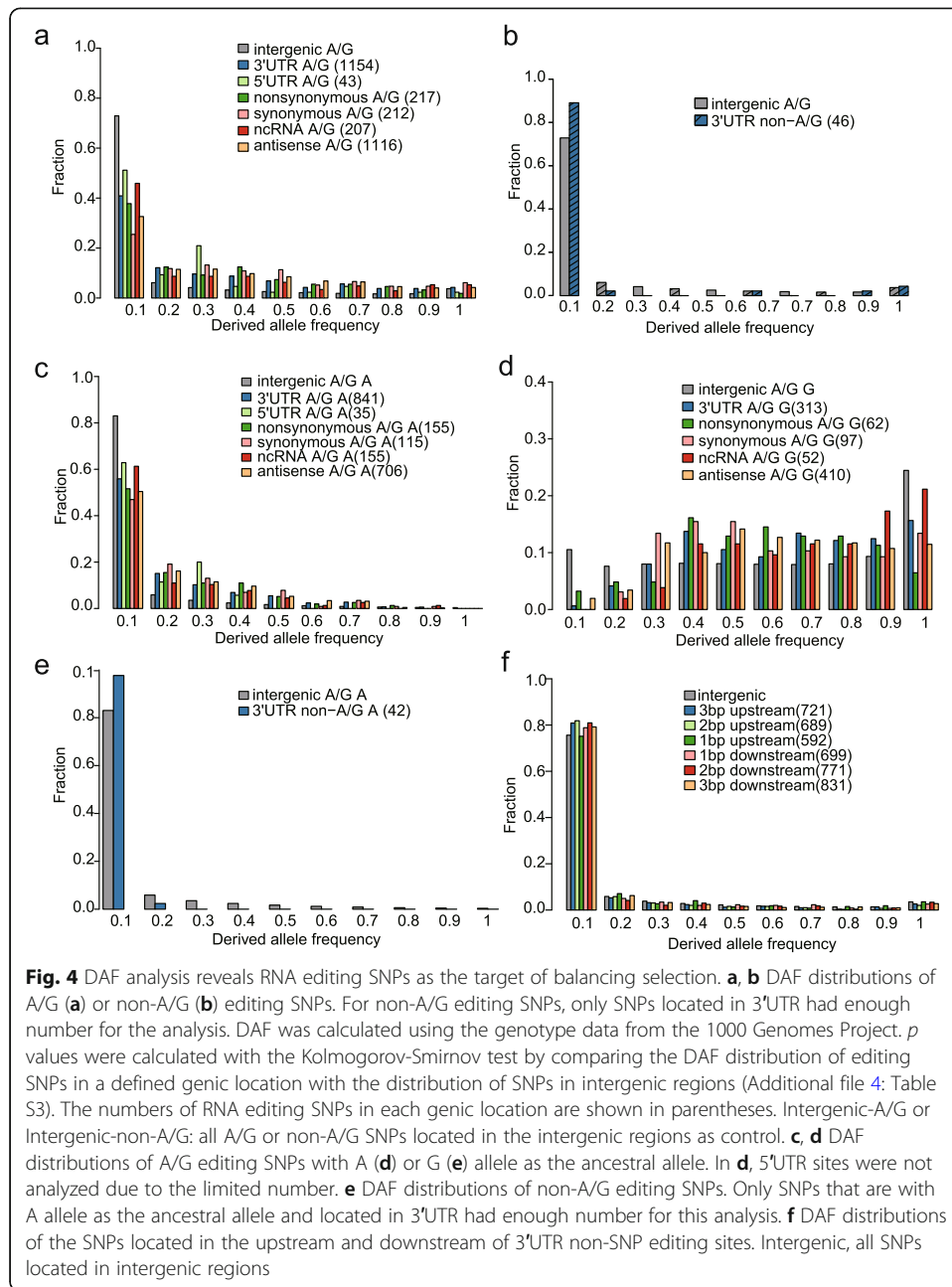
Derived allele frequency (DAF) analysis reveals RNA editing as the target of balancing selection

Balancing selection is the main force shaping the evolution of immunity genes [35–38]. The finding that RNA editing is enriched in loci related to autoimmune and immune-related functions, along with the putative functional difference between the A and I/G alleles, suggested that SNP editing loci may play a role in balancing immune system. If this is the case, evolutionarily, we predicted that a signal of balancing selection in SNP editing loci, particularly for the A/G SNP type that is equivalent to the A-to-I editing, would be found.

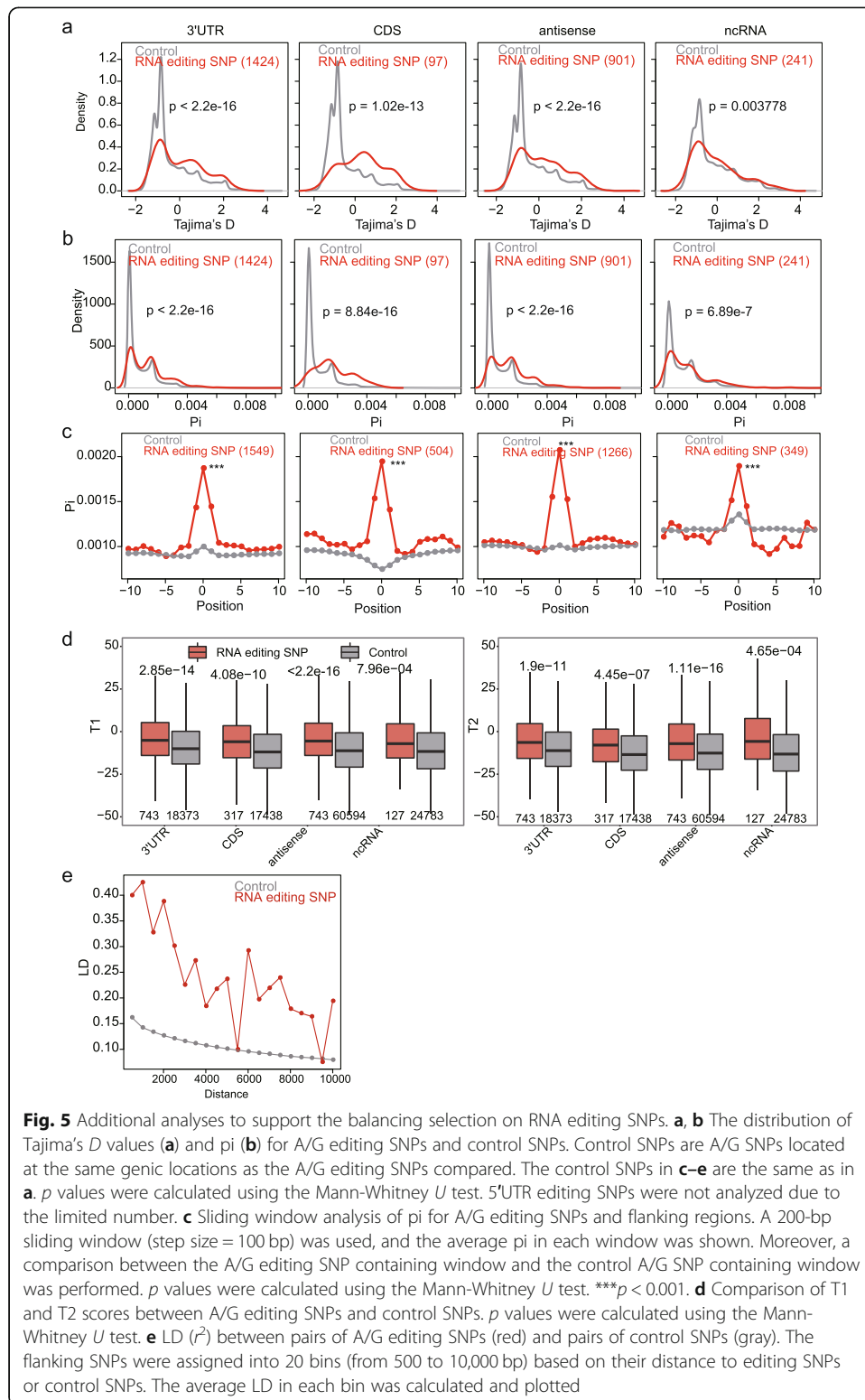
To verify our prediction, we applied population genomic approaches to study the evolution of RNA editing in humans. First, we examined DAF distributions of editing SNPs. A/G and non-A/G editing SNPs were analyzed separately, as they may be subject to different evolutionary constraints. To prevent the ascertainment bias between functional classes, CDS region, UTR of protein-coding genes, and ncRNAs were respectively analyzed. Different DAF distributions were directly compared using the Kolmogorov-Smirnov test, as previously described [39]. Intriguingly, we found that the DAF distribution of A/G editing SNPs was significantly skewed toward intermediate frequency alleles in all functional classes relative to intergenic regions (Fig. 4a, Additional file 1: Fig. S7, *p* values in Additional file 4: Table S3). The excess of intermediate frequency alleles, which is a typical scenario of balancing selection, implies that A/G editing SNPs were under balancing selection. In contrast, a shift in a DAF distribution toward low frequency alleles was observed for non-A/G editing SNPs, which is indicative of negative selection (Fig. 4b, Additional file 1: Fig. S7, *p* values in Additional file 4: Table S3). To ask whether there are different evolutionary patterns for editing SNPs with editable (A allele) or un-editable allele (non-A allele) as the ancestral allele, we examined their DAF distributions separately. In A/G editing SNPs, we found that both were under balancing selection (Fig. 4c, d, Additional file 1: Fig. S7, *p* values in Additional file 4: Table S3). In non-A/G editing SNPs, we found that, for SNPs with editable allele as the ancestral allele, the SNPs were under negative selection (Fig. 4e, Additional file 1: Fig. S7, *p* values in Additional file 4: Table S3). SNPs with un-editable allele as the ancestral allele were not analyzed because only a dozen sites were identified. Finally, we performed the DAF analysis for the upstream and downstream SNPs of non-SNP editing sites. Known editing sites from RADAR2 database [40] and non-SNP editing sites identified in this study were merged for analysis. As expected, no signatures of balancing selection were observed (Fig. 4f, Additional file 1: Fig. S7, *p* values in Additional file 4: Table S3), suggesting that the selection was specific to SNP editing sites. Taken together, these data suggest that the A/G editing SNPs were under balancing selection.

Additional evidence to support the balancing selection on A-to-I RNA editing

In addition to the DAF analysis, we performed additional analyses to confirm A/G editing SNPs as targets of the balancing selection. We found that the sequence variation surrounding the SNP editing sites fulfills two predictions from population genetic theory for genomic regions either directly experiencing long-term balancing selection or genetically linked to them. The first prediction is that we expect editing SNP regions to exhibit an excess of intermediate frequency alleles compared with the expectation



under selective neutrality. This is indicated by positively skewed values of Tajima's *D* (a summary of the mutation site-frequency spectrum [41]) of editing SNP regions, in comparison with the control regions (Fig. 5a). The second prediction is that neutral variants linked to the site under balancing selection are expected to be maintained in a population and generate excess diversity around the target of selection. The excess of high π (nucleotide diversity within species) values of editing SNP regions, compared with the control SNP regions, is consistent with this prediction (Fig. 5b). Moreover, a sliding window analysis revealed that, compared with both the flanking regions and the control A/G SNP regions, editing SNP regions had higher π values (Fig. 5c and Additional file 1: Fig. S8).



Recently, DeGiorgio et al. proposed two model-based summary statistics (T1 and T2, which generate a composite likelihood of a site being under balancing selection) to detect balancing selection [42]. In this method, sites having higher scores are more likely

to be under balancing selection. When comparing these summary statistics between editing SNPs and control SNPs, we found that, as expected, the editing SNPs had higher scores than the control SNPs for both T1 and T2 statistics (Fig. 5d and Additional file 5: Table S4).

In addition to the elevated polymorphism in editing SNP regions, we also found evidence for increased LD (another hallmark of balancing selection). We compared local LD (< 10,000 bp, measured as r^2) between pairs of RNA editing SNPs and pairs of control SNPs. Consistent with balancing selection, we found that pairs of RNA editing SNPs had higher LD than pairs of control SNP sites (Fig. 5e).

To ask whether both Alu and non-Alu editing SNPs are under balancing selection, we repeated the analyses above for the two groups of SNPs, separately. We found that non-Alu editing SNPs were more biased toward A/G genotypes than Alu editing SNPs (Additional file 1: Fig. S9a). DAF distribution of non-Alu editing SNPs was significantly skewed toward intermediate frequency alleles relative to Alu editing SNPs (Additional file 1: Fig. S9b). Consistent with this finding, the comparison of Tajima's D and π values, as well as the T1 and T2 scores, between non-Alu editing SNPs and Alu editing SNPs, all suggested that non-Alu editing SNPs were subject to a stronger balancing selection signal than Alu editing SNPs (Additional file 1: Fig. S9c-e).

DAF analysis supports RNA editing as the target of balancing selection in flies

Finally, to ask whether editing SNPs are under balancing selection in other species, we examined *Drosophila melanogaster* data. The genotype data from the *Drosophila* Genetics Reference Panel Project (DGRP) [43], which consists 205 sequenced inbred lines derived from Raleigh (NC), USA, were used to perform SNP allele type and DAF analysis. We examined all known RNA editing sites from RADAR2 database [40], which were called from three other *D. melanogaster* strains (w1118, Canton-S, and OregonR). A total of 743 sites were found to be overlapped with SNPs in DGRP. Similar to human, fly editing SNPs were biased toward A/G or T/C genotypes as compared with the control SNPs (Additional file 1: Fig. S10a). Moreover, the DAF distribution of A/G editing SNPs was significantly skewed toward intermediate frequency alleles in all functional classes relative to intergenic regions (Additional file 1: Fig. S10b-c, p values in Additional file 4: Table S3). In contrast, a shift in a DAF distribution toward low frequency alleles was observed for non-A/G editing SNPs, which is indicative of negative selection (Additional file 1: Fig. S10d-e, p values in Additional file 4: Table S3). These results support RNA editing as the target of balancing selection in flies.

Discussion

Balancing selection is a mode of adaptation that leads to the maintenance of variation in a species and potentially an important biological force for maintaining advantageous genetic diversity in populations [44–47]. Immunity genes are known as the targets of balancing selection due to the need of immune system genetic plasticity in response to various stimuli, such as different pathogens [35–38]. In this study, we revealed that, unexpectedly, a previously unidentified type of variants (i.e., SNPs in A-to-I RNA editing sites) is the common target of balancing selection in humans. Based on the known functions of ADAR1 and the observed enrichment of editing SNPs in GWAS signals

for autoimmune and immune-related diseases, it is likely that editing SNP variations are maintained at least partially because of constraints on the resolution of the balance between immune activity and self-tolerance.

Several studies have found that editable A's are more likely to be replaced with G's during evolution [15, 48–50]. Some researchers think that this phenomenon suggests that A-to-I RNA editing serves as a safeguard and is beneficial because it reverses harmful G-to-A mutation in RNA transcripts; others argue that this observation suggests that A-to-I RNA editing is non-adaptive because G's are more acceptable at the editable A sites than un-editable A sites. As we found that within the population, A/G alleles at the editing loci were under balancing selection and beneficial, our findings suggest that the previous observation between species was made because the two alleles can be favorable in different conditions, which leads to the fixation of the G allele in some species and the maintenance of the A allele in other species.

Conclusion

In summary, we uncover a hidden layer of human A-to-I editing SNP loci that are of functional importance, enriched in GWAS signals for autoimmune diseases, and subject to balancing selection. Various types of RNA editing, including A-to-I editing, alter sequence relative to the genome at the RNA level, thus providing a rich resource of RNA variants that potentially produce functionally altered genes. For some of the RNA variants that are beneficial under certain conditions, once the same type of mutation occurs at the DNA level, it may be selectively maintained and become the target of balancing selection. Therefore, we hypothesized that RNA editing, as exemplified in this study with A-to-I editing, may be an unrecognized type of the common target of balancing selection in various species.

Methods

RNA-seq data collection

Geuvadis RNA-seq data were downloaded from <https://www.ebi.ac.uk/Tools/geuvadis-das/>. GTEx project data were downloaded from NCBI, and a list of GTEx data sample IDs is shown in Additional file 2: Table S1.

Mapping of RNA-seq reads

We adopted a pipeline that can accurately map RNA-seq reads to the genome [11]. In brief, we used BWA [51] to align RNA-seq reads to a combination of the reference genome and exonic sequences surrounding known splicing junctions from available gene models. We chose the length of the splicing junction regions to be 1 bp shorter than the RNA-seq reads to prevent redundant hits. We obtained gene models from UCSC genome browser: a combination of Gencode, RefSeq, Ensembl, and UCSC Genes. We further used samtools to extract uniquely mapped reads.

Calling RNA variants

RNA variants were called as we described in Additional file 1: Fig. S1. In brief, we detected nucleotide variants between RNA-seq data and the reference genome in each sample. We took variant positions with the mismatch supported by two or more reads

with a base quality score of ≥ 20 and a mapping quality score ≥ 20 . Variants were divided into Alu and non-Alu regions. Non-Alu sites were subject to a more stringent variant call as previously described [11]. Last, to facilitate a fair comparison between non-SNP and SNP editing sites in our analyses, we required a minimum number of reference and altered nucleotides ≥ 3 for both non-SNP and SNP editing site call. We inferred the strand information of the sites based on the strand of the genes.

Notably, for SNP editing site call, we considered a site as an authentic SNP editing site only if the corresponding DNA sample had a homozygous genotype. We collected the genotype information from multiple resources and removed a site if its DNA sample had a conflict genotype information between different data resources.

SNP data

We downloaded all available 1000 Genomes Project and GTEx Project SNP data: (1) Genotype Calls (.vcf) for OMNI SNP Arrays, WES, and WGS of GTEx Project were downloaded from dbGaP database (<https://www.ncbi.nlm.nih.gov/gap/>); (2) WES data of Geuvadis project were downloaded from EMBL-EBI database (<https://www.ebi.ac.uk/arrayexpress/files/E-GEUV-1/genotypes/>); (3) Genotype Calls (.vcf) from OMNI SNP Arrays, Affy SNP Arrays, and WGS of the 1000 Genomes Project were downloaded from ftp website (<http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/>). We discarded all insertion and deletion polymorphisms, SNPs with more than two alleles, SNPs monomorphic (that is, having only one allele) in all populations, and SNPs that did not map uniquely to the human genome (hg19).

Editing level analysis in different cell lines

RNA-seq data of ADAR1 or ADAR2 overexpressed HEK293 cells and control cells were obtained from Song et al. [27]. RNA-seq data of ADAR1 knockout HEK293 cells and control cells were obtained from Song et al. [28]. HEK293 whole genome sequencing data were obtained from NCBI SRA (SRR2123657). RNA-seq data were trimmed with cutadapt (-q 20,20 --trim-n -m 15) and mapped with HISAT2 [52]. DNA-seq data were trimmed with cutadapt (-q 20,20 --trim-n -m 15) and mapped with BWA (aln -n 6 -t 20). Editing levels of SNP and non-SNP editing sites were called from the mapped data. To compare the editing levels between different samples, we required that sites were covered by at least 10 reads in all samples and the editing level was > 0.02 in at least one sample. In addition, for SNP editing sites, we required that the sites were covered by at least 5 reads in DNA-seq data and no G reads were observed.

RNA-seq data of ADAR1 or ADAR2 knockdown B cells and control cells were obtained from Wang et al. [29]. Editing levels were called as above. For SNP editing sites, we required that the genotype was homozygous (AA) based on the corresponding genome sequence data (GM12004 and GM12750) from the 1000 Genomes Project.

RNA secondary structure analysis

Two methods were applied to examine the RNA structure. In the first method [25], to identify putative ECS of a given editing site, we searched for the energetically most favorable hybridization region between the editing region (editing site and flanking ± 15 nt) and the extended surrounding region (± 2500 nt around the editing site) using

RNAplex [53]. We required that the extended surrounding region should be within a gene based on known human gene models. As a control, we shuffled the editing region 10,000 times and calculated the mean value of the lowest hybridization energies. For an ECS of a given editing site, if the hybridization energy between the editing region and the ECS was among the top 100 lowest hybridization energies of the shuffled sequences (i.e., $p < 0.01$), we considered it as an ECS with statistical significance.

In the second method [26], to detect the dsRNA structure formed around a given editing site, we aligned the editing region (editing site and flanking ± 25 nt) to the genomic sequence ± 2000 nt of the SNP editing site. We required that the genomic sequence ± 2000 nt of the SNP editing site should be within a gene based on known human gene models. We used bl2seq, with parameters -F F -W 7 -r 2, to align the sequence, and the best alignment score was obtained. As a control, we shuffled the editing region 10,000 times and calculated the mean value of the best scores.

GO term analysis

GO term enrichment analysis was performed using R package clusterProfiler [54].

GWAS enrichment analysis

A total of 85 GWAS datasets with full GWAS statistics provided in GWAS catalog (<https://www.ebi.ac.uk/gwas/downloads/summary-statistics>) were manually checked, and the ones that are not disease-relevant were excluded. Finally, 45 datasets that represent 33 types of diseases were downloaded. For each GWAS dataset, we examined the percentage of editing SNPs that are overlapped with GWAS SNPs with p value < 0.001 (%editing_SNP). As a control, we examined the percentage of SNPs with p value < 0.001 (%control_SNP). Last, the enrichment score was defined as %editing_SNP/%control_SNP. For a disease with multiple datasets, the dataset with the median enrichment score was shown.

DAF analysis

For each SNP, we extracted the ancestral allele information and DAF from the VCF files of the 1000 Genomes Project. Bi-allelic SNPs were obtained by VCFtools (--min-alleles 2 --max-alleles 2). Only SNPs with a minor allele frequency (MAF) > 0.001 were used for DAF spectrum analysis.

Tajima's D calculation

Tajima's D was calculated by VariScan [55]. In brief, SNP sites and the flanking 300-bp sequences were used for calculation. Only the SNPs with the flanking sequences located in the same functional classes, such as 3'UTR, 5'UTR, CDS, or ncRNA, were selected. RunMode 12 was chosen to calculate Tajima's D .

Nucleotide diversity calculation

We applied two methods to calculate nucleotide diversity (π). In Fig. 5b, VariScan was used to calculate nucleotide diversity of a 300-bp region surrounding the editing SNPs. The parameters are the same as the ones used for Tajima's D calculation.

In Fig. 5c, nucleotide diversity of sliding window analysis was performed using VCFtools. The parameters “--window-pi” and “--window-pi-step” were set to 200 bp and 100 bp, respectively.

Acquisition of T1 and T2 scores

T1 and T2 statistics were obtained from DeGiorgio et al. [42].

LD analysis

The software PopLDdecay [56] was applied to calculate LD (r^2) of pairs of A/G editing SNPs and pairs of control SNPs, using VCF data of the 1000 Genomes Project as the input.

RNA editing SNP analysis in *D. melanogaster*

The genotype data of 205 *D. melanogaster* inbred lines were downloaded from the Drosophila Genetic Reference Panel Project [43] (<http://dgrp2.gnets.ncsu.edu/>). The pairwise *D. melanogaster*/*D. simulans* alignment files were from UCSC (<http://hgdownload.soe.ucsc.edu/goldenPath/dm3/vsDroSim1/>). The list of *D. melanogaster* RNA editing sites was from RADAR database [40]. The ancestral allele of the SNPs was inferred from the homologous *D. simulans* sequence. DAF was calculated based on the genotype information of the 205 inbred lines.

Supplementary Information

Supplementary information accompanies this paper at <https://doi.org/10.1186/s13059-020-02205-x>.

Additional file 1: Fig. S1. The pipeline of non-SNP and SNP editing site identification. **Fig. S2.** The identification of non-SNP and SNP editing sites. **Fig. S3.** Triplet motif analysis of SNP editing sites. **Fig. S4.** The verification of SNP editing sites. **Fig. S5.** Comparison of the ratios of RNA editing SNP and control SNP. **Fig. S6.** ADAR2 and CASQ2 expression in different types of human tissues. **Fig. S7.** The cumulative distribution of DAF. **Fig. S8.** Sliding window analysis of pi for A/G editing SNPs. **Fig. S9.** non-Alu editing SNPs are subject to a stronger balancing selection compared with Alu editing SNPs. **Fig. S10.** DAF analysis reveals RNA editing SNPs as the target of balancing selection in flies.

Additional file 2: Table S1. The list of GTEx data used for analysis.

Additional file 3: Table S2. SNP and non-SNP RNA editing sites.

Additional file 4: Table S3. P values of DAF analysis.

Additional file 5: Table S4. T1,T2 statistics of RNA editing SNPs.

Additional file 6. Review history.

Acknowledgements

We thank Jin Billy Li and Tao Sun for the discussion of the manuscript and Michael DeGiorgio for sharing the T1 and T2 score data. The Genotype-Tissue Expression (GTEx) project was supported by the Common Fund of the Office of the Director of the National Institutes of Health (commonfund.nih.gov/GTEx). Additional funds were provided by the National Cancer Institute (NCI); National Human Genome Research Institute (NHGRI); National Heart, Lung, and Blood Institute (NHLBI); National Institute on Drug Abuse (NIDA); National Institute of Mental Health (NIMH); and National Institute of Neurological Disorders and Stroke (NINDS). Donors were enrolled at Biospecimen Source Sites funded by NCISAIIC-Frederick, Inc. (SAIC-F) subcontracts to the National Disease Research Interchange (10XS170) and Roswell Park Cancer Institute (10XS171). The Laboratory, Data Analysis, and Coordinating Center (LDACC) was funded through a contract (HHSN268201000029C) to The Broad Institute, Inc. Biorepository operations were funded through an SAIC-F subcontract to Van Andel Institute (10ST1035). Additional data repository and project management were provided by SAIC-F (HHSN261200800001E). The Brain Bank was supported by a supplement to University of Miami grant DA006227.

Peer review information

Kevin Pang was the primary editor of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Review history

The review history is available as Additional file 6.

Authors' contributions

H.Z. and R.Z. conceived the project. H.Z., Q.F., and X.R.S. contributed to the computational analyses. Q.F., Z.Q.P., and Z.C.H. contributed to the data collection. H.Z., X.R.S., and R.Z. wrote the paper with input from T.T. and X.L.H.. The authors read and approved the final manuscript.

Funding

This study was supported by grants from National Natural Science Foundation of China (91631108 and 31571341 to R.Z.), Guangdong Innovative and Entrepreneurial Research Team Program (2016ZT065638 to R.Z.), and the State Key Laboratory of Genetic Resources and Evolution (GREKF18-08).

Availability of data and materials

Geuvadis RNA-seq data can be obtained from the Geuvadis consortium (<https://www.internationalgenome.org/data-portal/data-collection/geuvadis>). Geuvadis genotype data can be obtained from the 1000 Genomes Project (<http://www.internationalgenome.org/data/>). GTEx RNA-seq and Genotype data can be obtained from the GTEx consortium (<https://gtexportal.org/home/>). GWAS datasets with full GWAS statistics can be obtained from GWAS catalog (<https://www.ebi.ac.uk/gwas/downloads/summary-statistics>). Fly RNA editing site annotations are available at the RADAR web-site (<http://rnaedit.com/download/>).

Ethics approval and consent to participate

Not applicable

Competing interests

The authors declare no competing financial interests.

Author details

¹Key Laboratory of Gene Engineering of the Ministry of Education, State Key Laboratory of Biocontrol, School of Life Sciences, Sun Yat-Sen University, Guangzhou, People's Republic of China. ²State Key Laboratory of Genetic Resources and Evolution, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming, People's Republic of China. ³State Key Laboratory of Biocontrol, School of Life Sciences, Sun Yat-Sen University, Guangzhou, People's Republic of China. ⁴RNA Biomedical Institute, Sun Yat-Sen Memorial Hospital, Sun Yat-Sen University, Guangzhou, People's Republic of China.

Received: 13 February 2020 Accepted: 16 November 2020

Published online: 30 November 2020

References

- Gott JM, Emeson RB. Functions and mechanisms of RNA editing. *Annu Rev Genet.* 2000;34:499–531.
- Nishikura K. Functions and regulation of RNA editing by ADAR deaminases. *Annu Rev Biochem.* 2010;79:321–49.
- Eisenberg E, Levanon EY. A-to-I RNA editing - immune protector and transcriptome diversifier. *Nat Rev Genet.* 2018;19:473–90.
- Bass BL. RNA editing by adenosine deaminases that act on RNA. *Annu Rev Biochem.* 2002;71:817–46.
- Nishikura K. A-to-I editing of coding and non-coding RNAs by ADARs. *Nat Rev Mol Cell Biol.* 2016;17:83–96.
- Brummer A, Yang Y, Chan TW, Xiao X. Structure-mediated modulation of mRNA abundance by A-to-I editing. *Nat Commun.* 2017;8:1255.
- Hsiao YE, Bahn JH, Yang Y, Lin X, Tran S, Yang EW, Quinones-Valdez G, Xiao X. RNA editing in nascent RNA affects pre-mRNA splicing. *Genome Res.* 2018;28:812–23.
- Mannion NM, Greenwood SM, Young R, Cox S, Brindle J, Read D, Nellaker C, Vesely C, Ponting CP, McLaughlin PJ, et al. The RNA-editing enzyme ADAR1 controls innate immune responses to RNA. *Cell Rep.* 2014;9:1482–94.
- Liddicoat BJ, Piskol R, Chalk AM, Ramaswami G, Higuchi M, Hartner JC, Li JB, Seeburg PH, Walkley CR. RNA editing by ADAR1 prevents MDA5 sensing of endogenous dsRNA as nonself. *Science.* 2015;349:1115–20.
- Pestal K, Funk CC, Snyder JM, Price ND, Treuting PM, Stetson DB. Isoforms of RNA-editing enzyme ADAR1 independently control nucleic acid sensor MDA5-driven autoimmunity and multi-organ development. *Immunity.* 2015;43:933–44.
- Ramaswami G, Lin W, Piskol R, Tan MH, Davis C, Li JB. Accurate identification of human Alu and non-Alu RNA editing sites. *Nat Methods.* 2012;9:579–81.
- Bahn JH, Lee JH, Li G, Greer C, Peng G, Xiao X. Accurate identification of A-to-I RNA editing in human by transcriptome sequencing. *Genome Res.* 2012;22:142–50.
- Peng Z, Cheng Y, Tan BC, Kang L, Tian Z, Zhu Y, Zhang W, Liang Y, Hu X, Tan X, et al. Comprehensive analysis of RNA-Seq data reveals extensive RNA editing in a human transcriptome. *Nat Biotechnol.* 2012;30:253–60.
- Chen JY, Peng Z, Zhang R, Yang XZ, Tan BC, Fang H, Liu CJ, Shi M, Ye ZQ, Zhang YE, et al. RNA editome in rhesus macaque shaped by purifying selection. *PLoS Genet.* 2014;10:e1004274.
- Xu G, Zhang J. Human coding RNA editing is generally nonadaptive. *Proc Natl Acad Sci.* 2014;111:3769–74.
- Yu Y, Zhou H, Kong Y, Pan B, Chen L, Wang H, Hao P, Li X. The landscape of A-to-I RNA editome is shaped by both positive and purifying selection. *PLoS Genet.* 2016;12:e1006191.
- Duan Y, Dou S, Luo S, Zhang H, Lu J. Adaptation of A-to-I RNA editing in *Drosophila*. *PLoS Genet.* 2017;13:e1006648.
- Zhang R, Deng P, Jacobson D, Li JB. Evolutionary analysis reveals regulatory and functional landscape of coding and non-coding RNA editing. *PLoS Genet.* 2017;13:e1006563.
- Liscovitch-Brauer N, Alon S, Porath HT, Elstein B, Unger R, Ziv T, Admon A, Levanon EY, Rosenthal JJC, Eisenberg E. Trade-off between transcriptome plasticity and genome evolution in cephalopods. *Cell.* 2017;169:191–202 e111.
- Yablonovitch AL, Deng P, Jacobson D, Li JB. The evolution and adaptation of A-to-I RNA editing. *PLoS Genet.* 2017;13:e1007064.

21. Eisenberg E, Adamsky K, Cohen L, Amariglio N, Hirshberg A, Rechavi G, Levanon EY. Identification of RNA editing sites in the SNP database. *Nucleic Acids Res.* 2005;33:4612–7.
22. Ramaswami G, Zhang R, Piskol R, Keegan LP, Deng P, O'Connell MA, Li JB. Identifying RNA editing sites using RNA sequencing data alone. *Nat Methods.* 2013;10:128–32.
23. Polson AG, Bass BL. Preferential selection of adenosines for modification by double-stranded RNA adenosine deaminase. *EMBO J.* 1994;13:5701–11.
24. Eggington JM, Greene T, Bass BL. Predicting sites of ADAR editing in double-stranded RNA. *Nat Commun.* 2011;2:319.
25. Licht K, Kapoor U, Amman F, Picardi E, Martin D, Bajad P, Jantsch MF. A high resolution A-to-I editing map in the mouse identifies editing events controlled by pre-mRNA splicing. *Genome Res.* 2019;29:1453–63.
26. Porath HT, Knisbacher BA, Eisenberg E, Levanon EY. Massive A-to-I RNA editing is common across the Metazoa and correlates with dsRNA abundance. *Genome Biol.* 2017;18:185.
27. Song Y, Yang W, Fu Q, Wu L, Zhao X, Zhang Y, Zhang R. irCLASH reveals RNA substrates recognized by human ADARs. *Nat Struct Mol Biol.* 2020;27:351–62.
28. Song Y, Li L, Yang W, Fu Q, Chen W, Fang Z, Li W, Gu N, Zhang R. Sense–antisense miRNA pairs constitute an elaborate reciprocal regulatory circuit. *Genome Res.* 2020;30:661–72.
29. Wang IX, So E, Devlin JL, Zhao Y, Wu M, Cheung VG. ADAR regulates RNA editing, transcript stability, and gene expression. *Cell Rep.* 2013;5:849–60.
30. Tan MH, Li Q, Shanmugam R, Piskol R, Kohler J, Young AN, Liu KJ, Zhang R, Ramaswami G, Ariyoshi K, et al. Dynamic landscape and regulation of RNA editing in mammals. *Nature.* 2017;550:249–54.
31. Yano K, Zarain-Herzberg A. Sarcoplasmic reticulum calsequestrins: structural and functional properties. *Mol Cell Biochem.* 1994;135:61–70.
32. Terentyev D, Viatchenko-Karpinski S, Gyorke I, Volpe P, Williams SC, Gyorke S. Calsequestrin determines the functional size and stability of cardiac intracellular calcium stores: mechanism for hereditary arrhythmia. *Proc Natl Acad Sci U S A.* 2003;100:11759–64.
33. Tam V, Patel N, Turcotte M, Bosse Y, Pare G, Meyre D. Benefits and limitations of genome-wide association studies. *Nat Rev Genet.* 2019;20:467–84.
34. Fioranelli M, Bottaccioli AG, Bottaccioli F, Bianchi M, Rovesti M, Rocca MG. Stress and inflammation in coronary artery disease: a review psychoneuroendocrine-immunology-based. *Front Immunol.* 2018;9:2031.
35. Ferrer-Admetlla A, Bosch E, Sikora M, Marqués-Bonet T, Ramírez-Soriano A, Muntasell A, Navarro A, Lazarus R, Calafell F, Bertranpetit J, Casals F. Balancing selection is the main force shaping the evolution of innate immunity genes. *J Immunol.* 2008;181:1315–22.
36. Barreiro LB, Quintana-Murci L. From evolutionary genetics to human immunology: how selection shapes host defence genes. *Nat Rev Genet.* 2010;11:17–30.
37. Abi-Rached L, Jobin MJ, Kulkarni S, McWhinnie A, Dalva K, Gragert L, Babrzadeh F, Gharizadeh B, Luo M, Plummer FA, et al. The shaping of modern human immune systems by multiregional admixture with archaic humans. *Science.* 2011;334:89–94.
38. Croze M, Zivkovic D, Stephan W, Hutter S. Balancing selection on immunity genes: review of the current literature and new analysis in *Drosophila melanogaster*. *Zoology (Jena).* 2016;119:322–9.
39. Haerty W, Ponting CP. Mutations within lncRNAs are effectively selected against in fruitfly but not in human. *Genome Biol.* 2013;14:R49.
40. Ramaswami G, Li JB. RADAR: a rigorously annotated database of A-to-I RNA editing. *Nucleic Acids Res.* 2014;42:D109–13.
41. Tajima F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics.* 1989;123:585–95.
42. DeGiorgio M, Lohmueller KE, Nielsen R. A model-based approach for identifying signatures of ancient balancing selection in genetic data. *PLoS Genet.* 2014;10:e1004561.
43. Huang W, Massouras A, Inoue Y, Peiffer J, Ràmia M, Tarone AM, Turlapati L, Zichner T, Zhu D, Lyman RF, et al. Natural variation in genome architecture among 205 *Drosophila melanogaster* Genetic Reference Panel lines. *Genome Res.* 2014;24:1193–208.
44. Charlesworth D. Balancing selection and its effects on sequences in nearby genome regions. *PLoS Genet.* 2006;2:e64.
45. Hedrick PW. Balancing selection. *Curr Biol.* 2007;17:R230–1.
46. Leffler EM, Gao Z, Pfeifer S, Ségurel L, Auton A, Venn O, Bowden R, Bontrop R, Wall JD, Sella G, et al. Multiple instances of ancient balancing selection shared between humans and chimpanzees. *Science.* 2013;339:1578–82.
47. Siewert KM, Voight BF. Detecting long-term balancing selection using allele frequency correlation. *Mol Biol Evol.* 2017;34:2996–3005.
48. Tian N, Wu X, Zhang Y, Jin Y. A-to-I editing sites are a genomically encoded G: implications for the evolutionary significance and identification of novel editing sites. *RNA.* 2008;14:211–6.
49. Chen L. Characterization and comparison of human nuclear and cytosolic editomes. *Proc Natl Acad Sci U S A.* 2013;110E2741–7.
50. An NA, Ding W, Yang X-Z, Peng J, He BZ, Shen QS, Lu F, He A, Zhang YE, Tan BC-M, et al. Evolutionarily significant A-to-I RNA editing events originated through G-to-A mutations in primates. *Genome Biol.* 2019;20:24.
51. Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics.* 2010;26:589–95.
52. Kim D, Langmead B, Salzberg SL. HISAT: a fast spliced aligner with low memory requirements. *Nat Methods.* 2015;12:357–60.
53. Tafer H, Amman F, Eggenhofer F, Stadler PF, Hofacker IL. Fast accessibility-based prediction of RNA-RNA interactions. *Bioinformatics.* 2011;27:1934–40.
54. Yu G, Wang L-G, Han Y, He Q-Y. clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS: A Journal of Integrative Biology.* 2012;16:284–7.
55. Vilella AJ, Blanco-García A, Hutter S, Rozas J. VariScan: analysis of evolutionary patterns from large-scale DNA sequence polymorphism data. *Bioinformatics.* 2005;21:2791–3.
56. Zhang C, Dong S-S, Xu J-Y, He W-M, Yang T-L. PopLDdecay: a fast and effective tool for linkage disequilibrium decay analysis based on variant call format files. *Bioinformatics.* 2018;35:1786–8.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.