European Journal of Human Genetics (2020) 28:1753–1762
https://doi.org/10.1038/s41431-020-0688-7

ARTICLE

# Identification of genetic variants controlling RNA editing and their effect on RNA structure stabilization

Aziz Belkadi[1] · Gaurav Thareja[1] · Anna Halama[1] · Yasmin Mahmoud[1] · Danielle Jones[1] · Sam Agnew[1] · Joel Malek[1] · Karsten Suhre[1]

Received: 21 December 2019 / Revised: 31 May 2020 / Accepted: 11 June 2020 / Published online: 10 July 2020


## Abstract

Post-transcriptional modification of RNA (RNA editing, RNAe) results in differences between the RNA transcript and the genomic DNA sequence (RDD). Enzymatic modification of adenosine to inosine (A2I) by ADAR is the most studied type of RNAe. However, few genetic association studies with A2I RNAe events have been conducted. Some studies have analyzed the inter-population RNAe-QTL diversity in humans, but the sample size of these studies was limited. Other types of RNA and DNA differences have been reported but are largely understudied. Here, we report a comprehensive analysis of all types of RDD, based on two independent datasets. We found that A2I was by far the most observed type of RDD. Moreover, manual curation suggests that A2I is likely the only enzymatically driven RNAe type observed in blood derived DNA, all other non-A2I RDD could either be attributed to sequencing and processing artifacts, or are a result of somatic DNA rearrangements. We then conducted an in-cis genetic association study and identified 472 genetic associations (RNAe-QTL), that were replicated in both datasets. We confirm the potential effect of the RNAe-QTL on RNA structure by showing that allele specific RNAe occurs in heterozygotes. Although the generally assumed function of RNAe is to destabilize double stranded RNA structure, we found clear evidence for the potential additional involvement of RNAe in maintaining RNA hairpin that has been altered by the RNAe-QTL. Our study confirms, in two independent datasets, the potential role of RNAe in maintaining RNA structure in the presence of genetic variation.

## Introduction

RNA editing (RNAe) is a post-transcriptional process consisting of the enzymatic modification of a nucleotide in the RNA molecule, resulting in a nucleotide in the transcriptomic sequence other than what is encoded by the genome. The conversion of adenosine (A) nucleotides to inosines (I) (A2I) – which will be read as a Guanine (G) using standard RNA-sequencing techniques -constitutes the most common type of RNAe [1]. A2I RNAe is performed by the adenosine deaminase acting on the RNA (ADAR) family of proteins, which is found in many Metazoans and mammals [2, 3]. There are three mammalian ADAR proteins: ADAR (ADAR1), ADARB1 (ADAR2), and ADARB2 (ADAR3). ADAR and ADARB1 have A2I RNAe activity while ADARB2 appears to be catalytically inactive [4].

Taking advantage of the rapid progress in RNA-sequencing technologies, millions of A2I RNAe sites have been reported in humans [5, 6]. RNAe occurs predominantly in alu repeats of the 3′ untranslated and intronic regions, and their proposed function in vertebrates is to edit endogenous long double stranded RNAs – the favorite target of ADAR enzymes – in order to prevent the immune system activation caused by the presence of double stranded RNA, which is also a marker of viral infection [2]. In addition, RNAe has been shown to play a role in RNA stability [7], RNA nuclear retention [8] and in transcript diversification by acting on splicing [9] and micro RNA target sequence and abundance [10].

**Supplementary information** The online version of this article (https://doi.org/10.1038/s41431-020-0688-7) contains supplementary material, which is available to authorized users.

✉ Aziz Belkadi
abb2013@qatar-med.cornell.edu

✉ Karsten Suhre
kas2049@qatar-med.cornell.edu

[1] Department of Physiology and Biophysics, Weill Cornell Medicine-Qatar, Education City, Doha, Qatar

Variants in ADAR1 are directly associated with dyschromatosis symmetrica herediataria, an autosomal dominant pigmentary genodermatosis characterized by hyperpigmented and hypopigmented macules on the extremities [11] and Aicardi-Goutières syndrome, a genetically determined inflammatory disorder particularly affecting the brain and skin, which is associated with increased production of the antiviral cytokine interferon α [12]. Reduction of RNAe due to loss or suppression of ADAR2 is associated with a range of neuronal disorders such as amyotrophic lateral sclerosis [13]. Alterations of ADAR2-mediated RNAe levels has been linked to a brain cancer [14]. Increase of overall RNAe and ADAR1 expression has been related to different cancer types [15, 16] and reduced RNAe efficiency has been proposed to be a potential innovative therapeutic target of those tumors [17].

Three genome wide association studies in humans and two target gene association studies in *Drosophila melanogaster* have been conducted to identify genetic variants associated with variability in A2I events (RNAe-QTLs). These studies have shown that RNAe events are mostly regulated by *in cis* genetic effects and RNAe-QTL overlapped with QTL associated with blood metabolite levels, obesity related traits, and blood protein levels. Nonetheless, none of the A2I genetic association studies has been replicated in an independent dataset [18–22]. In light of the growing concern about publication bias and the dearth of compelling replications, a follow-up analysis is a necessary step in validating these findings.

Here, we assessed RNA-to-DNA sequence differences (RDD) in two independent cohorts. We performed RNA sequencing for transcriptome quantification (RNA-seq) using RNA extracted from white blood cells of 320 individuals from the multi-ethnic QMDiab study [23]. The mean read depth was 16 million reads per sample. We analyzed publicly available RNA-seq data from 421 European and African individuals from the Geuvadis RNA-seq project [24]. We identified 2599 RDD sites in common to both datasets. After manual curation, we found that all the non-A2I RDD were technical limitations rather than biological RNAe events. We then reported a genetic association study (*in cis*) with 2099 high quality A2I sites. We detected 1257 *cis*-RNAe-QTLs in Geuvadis and replicated 472 of these RNAe-QTLs in QMDiab. Furthermore, RNA structure prediction and local sequence alignment indicate that RNAe-QTL might regulate RNAe by acting on RNA secondary structure. Finally, we found clear evidence of RNAe-QTL allele specific RNAe (ASE) suggesting that RNAe-QTL is involved in ASE by influencing the RNA secondary structure.

## Material and methods

### Sample collection

The Geuvadis RNA-seq data is a collection of 462 mRNA sequencing on lymphoblastoid cell line samples from five populations: the CEPH (CEU), Finns (FIN), British (GBR), Toscani (TSI), and Yoruba (YRI) [24]. Of these, 421 are in the 1000 Genomes Phase 1 data set. QMDiab is a cross-sectional case-control study that was conducted between February and June 2012 at the Dermatology Department of HMC in Doha, Qatar. QMDiab has been described previously and comprises male and female participants in near equal proportions, aged between 23 and 71 years, mainly of Arab, South Asian and Filipino descent [23]. The initial study was approved by the Institutional Review Boards of HMC and Weill Cornell Medicine—Qatar (research protocol #11131/11). Written informed consent was obtained from all participants. RNA-seq was performed on whole blood cells in QMDiab sample.

### RDD sites detection

The pipeline for detecting RDD site is described in Fig. 1. Briefly, 421 and 320 individuals from Geuvadis and QMDiab projects respectively were included. RNA sequence reads from both studies were aligned to the Genome Reference Consortium human genome reference build 37 (GRCh37) using STAR aligner and data were processed following The Genome Analysis ToolKit (GATK) best practices for RNA sequence variant calling [25, 26]. Insertions and deletions were removed from the analysis. We filtered out poorly covered calls (≤10 reads) and we kept only common RDD sites present in >10% of individuals in both datasets. Polymorphic sites of the genome reported with a frequency ≥1% in at least one of the four public databases, 1000 Genomes, Exac, Exome Variant Server, and GnomAD, were removed leaving 10,865 and 29,166 high quality RDD sites in the Geuvadis and QMDiab dataset, respectively.

We defined RDD level by the fraction of the edited to total number of reads covering the RDD position. We removed 250 RDD with <10% of difference in the RNAe level between the 90th and the 10th percentile in either Geuvadis or QMDiab. We removed 70 RDD sites matching a position covered by <5 reads in DNA-seq for >90% of Geuvadis individuals. We then removed 272 RDD sites for which the fraction of the alternative allele in DNA-seq was ≥10% in more than 1% of Geuvadis individuals. RDD site annotation was carried out using Annovar [27]. R was used for data organization, plotting, and statistical analyses [28].

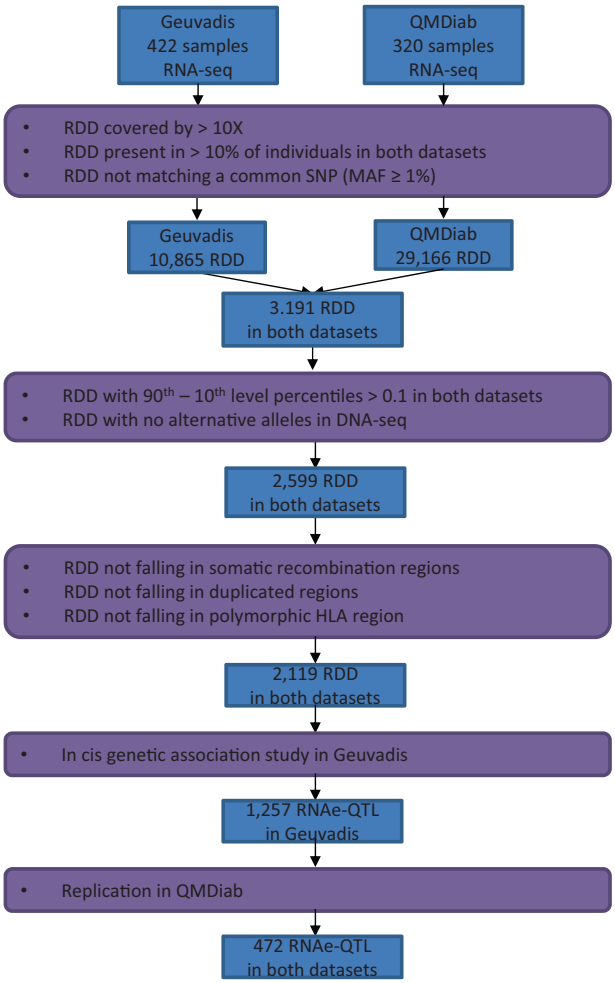


**Fig. 1 Schematic view of the study design.** A workflow for detecting RDD in two independent datasets.

## DNA genotyping and imputation

Whole genome sequencing imputed variant calling files for the 1,000 genome project phase 3 were downloaded from the University of California, Santa Cruz website (http://hgdownload.cse.ucsc.edu/gbdb/hg19/1000Genomes/phase3/). A total of 38,187,570 autosomal variants were called for 462 individuals. Forty one individuals with 10% of missing genotypes (Plink software [29] option –mind 0.1) were excluded. In all, 32,079,946 SNVs were removed due to missing genotypes, Hardy Weinberg Equilibrium, or allele frequency filters (Plink options --geno 0.02 --hwe 1E-6 --maf 0.05). A total of 6,107,624 remaining variants were included in the analysis. DNA was extracted from 320 samples from QMDiab and genotyped by the WCM-Q genomics core facility using the Illumina Omni 2.5 array (version 8). We used Shapeit [30] for phasing and impute 2 [31] for imputation using the 1000 genomes as reference. A total of 18,829,416 high quality imputed autosomal SNPs were obtained for 320 samples. No sample was excluded due to a low overall call rate (<90%). In all, 14,052,475 variants were removed due to missing genotype, Hardy–Weinberg equilibrium, or minor allele threshold (PLINK option --geno 0.02 --hwe 1E-6 --maf 0.05), leaving 4,776,941 autosomal variants.

## In cis association discovery

The RNAe data from the Geuvadis project was used for discovery. We used Plink to fit linear models to RNAe levels, using the three first principal components (PCs) computed on SNPs and the expression level of the gene carrying the RNAe site as covariates (Fig. 1). Twenty-six RNAe sites that do not belong to annotated transcripts were excluded. The three first PCs explain 41.8% of the variability. As RNAe is mainly regulated *in cis* [32], we targeted SNPs within 10 Mb upstream or downstream of the RNAe site. A total of 6,973 RNAe-QTL associations covering 4,823 SNPs that passed the Bonferroni multi-testing correction ($P < 0.05/N/2{,}099$ where N is the number of SNPs within 10 Mb upstream or downstream the RNAe site) were retained. For each RNAe-QTL, we grouped all the associations that are in LD >0.8 into one RNAe-QTL locus. The sentinel is then defined for each group by its strongest association. A total of 1257 genetic associations representing one or more sentinels associated with one or more RNAe sites that passed the Bonferroni multi-testing correction were identified.

## Replication

We used QMDiab data for replication. Plink software was used to fill the linear model with the three first PCs computed on SNPs (explaining 20.8% of the variation), gender, age, diabetes, and the expression level of the gene carrying the RNAe site as covariates. The tag SNP was defined by the strongest association in QMDiab from all the highly correlated SNPs with the tag SNP in Geuvadis (LD >0.8). A total of 827 RNAe-QTL (65.8%) from Geuvadis have a tag SNP in QMDiab. In all, 472 (57.1% of RNAe-QTL with a tag SNP in QMDiab) passed the Bonferroni correction ($P < 0.05/1{,}257$). All replicated RNAe-QTL with $P < 0.05$ in QMDiab and minor allele frequency <30% in Geuvadis and QMDiab showed the same trend in both datasets (Table S1). As the sample size of Geuvadis (421) and QMDiab (320) was comparable, no statistical power for replication was estimated.

## Multiple alignment and RNA secondary structure prediction

Alignment of reverse sequence of ±30 bp surrounding the RNAe site and the ±30 bp surrounding the RNAe-QTL was

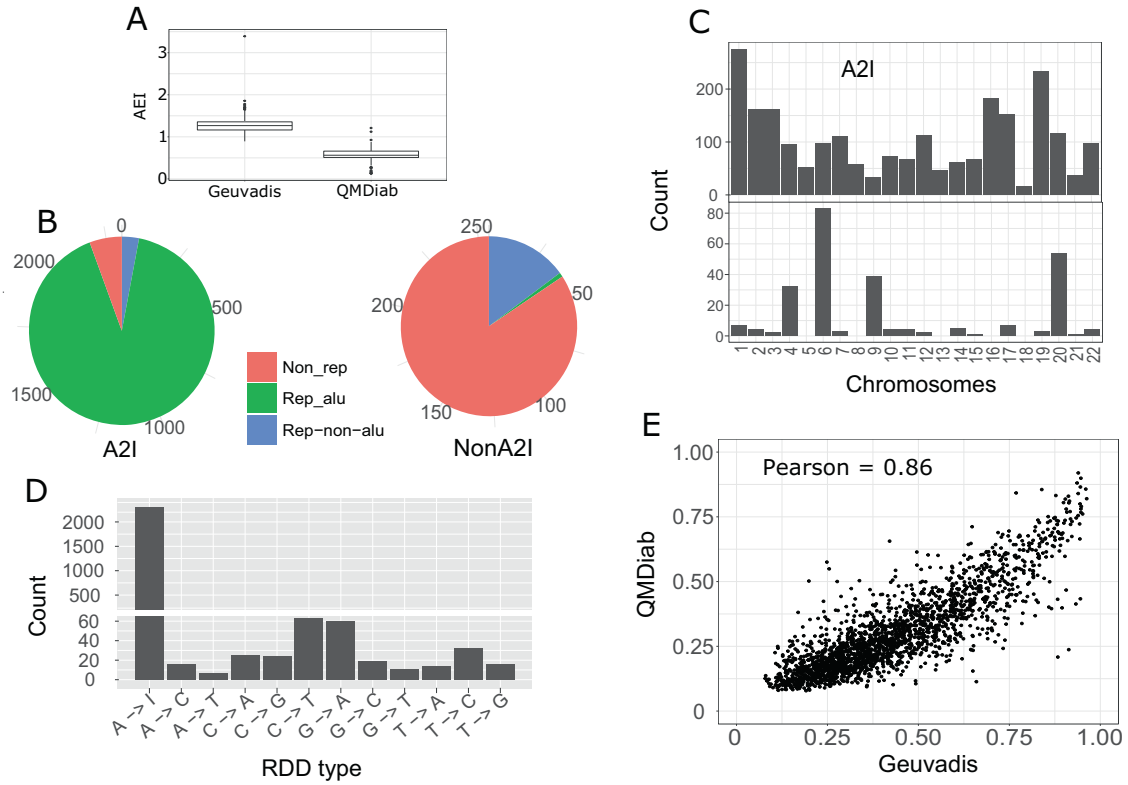


**Fig. 2 Characteristics of detected A2I and non-A2I. a** Alu editing index [35] shows a higher RNAe activity in Geuvadis than in QMDiab. **b** Pie chart describing the percentage of A2I and nonA2I RDDs falling in Alu repetitive regions, non-Alu repetitive regions, and nonrepetitive regions. **c** Distribution of A2I and non-A2I RDDs across the human chromosomes. A2I RDDs are distributed evenly among regions with sequencing reads, while non-A2I RDDs are enriched in specific genomic regions (**d**) Types of RDD detected. (**e**). Correlation between overall (mean) RNAe level of 2,119 A2I RNAe sites between two independent datasets.

performed using Clustal [33]. Alignment free energy was estimated using the RNAalifold function of Vienna RNA package 2 [34]. We used RNAfold program implemented in the Vienna RNA package 2 [34] with default parameters to predict the RNA secondary.

## Results

### RNA-DNA-differences (RDD)

We identified RDD sites by calling Single Nucleotide Polymorphisms (SNPs) from RNA-sequencing in two independent datasets (Fig. 1 and Methods). A total of 10,865 and 29,166 RDD sites were detected in Geuvadis and QMDiab, respectively. In total, 24,570 (95%) RNAe sites detected exclusively in QMDiab were excluded from Geuvadis because they did not reach the required coverage (>10×) in 10% of Geuvadis individuals. We identified 5,145 RNAe sites exclusively in Geuvadis, most likely due to the higher RNAe activity in lymphocyte cell lines used in Geuvadis comparing to whole blood cells used in QMDiab [35] (Fig. 2a). A total of 2,599 RDD sites were present in both datasets. A2I which is sequenced as an A to G modification is the most common type of RDD. A total of 2312 (89%) of the RDD sites we identified were A2I RDD. The remaining 287 were non-A2I RDDs. Interestingly, 2237 (96.8%) of A2I were already reported in at least one of the three public RNAe databases: Radar [5], Darned [6], or GTEx [3]. Furthermore, 1893 (84.6%) of these A2I sites were present in all three databases and only 46 (2.1%) were exclusive to one database (Fig S1). Ninety one percent of the A2I sites located to repetitive *alu* regions (Fig. 2b). A2I sites were found across all chromosomes and the number of A2I sites per chromosome correlates with the number of coding genes (Pearson = 0.81), where chromosomes 1 and 19 consequently have the largest number of A2I RDDs (Fig. 2c). Unlike A2I, non-A2I sites were mainly localized on HLA regions of the chromosome 6, and chromosomes 4, 9, and 20 (Fig. 2c). All possible types of non-A2I base modifications were observed at a different level. The most common non-A2I RDD types were C2T and T2G substitutions, representing 43% of the total non-A2I RDDs (Fig. 2d).

The difference between the RNA and the DNA that we identified here as RDD can be explained by different

processes: editing of the RNA (RNAe); editing of the cellular DNA i.e., rearrangement of the genomic DNA within the immunoglobulin genes in cells of the B-lymphocyte lineage (somatic recombination) or hypermorphic HLA regions; APOBEC-mediated DNA editing; the presence of unannotated SNPs or alignment and calling errors. We investigated whether an observed RDD is due to RNAe, somatic recombination or processing artifact. Starting with non-A2I RDD, we first considered the difference between the DNA that is the predecessor of the RNA and the DNA sequence in the reference genome used, as both datasets used lymphocytes for RNA-seq (Lymphoblastoid cells in Geuvadis and whole blood cells in QMDiab). A total of 97 non-A2I RDDs belong to the HLA region on chromosome 6, five belong to the immunoglobulin heavy locus on chromosome 14, two belong to the immunoglobulin kappa locus on chromosome 2, and three belong to the immunoglobulin lambda locus on chromosome 22.

We then assessed the effect of sequence alignment ambiguity due to sequence duplications not reflected in the reference genome. For example, 30 non-A2I sites locate to the FRG1 gene region on chromosome 4, and 54 non-A2I sites locate to the homologous pseudogene FRG1B on chromosome 20, and 45 non-A2I sites locate to a region on chromosome 9. Sequence comparison analysis showed 93% sequence identity between these three regions. Forty-seven further non-A2I RDD fall in regions which also have high sequence homology with another region of the genome (>90% identity). These RDDs are thus likely artifacts and the result of ambiguity in sequence read alignment.

Only four non-A2I RDDs remained after filtering for possible misalignment and RDD located in V(D)J recombination regions. Read alignment visualization using the Integrative Genomics Viewer [36] of these four non-A2I sites shows that all the reads seem to start or end at these sites. Thus, the absence of reads that actually span these non-A2I sites suggest these RDDs to be an artifact (Fig S2). Taken together, these observations suggest that none of the observed non-A2I sites are due to RNAe, at least not in the blood circulating immune cells analyzed here.

We then assessed the number of A2I sites falling in duplicated or V(D)J recombination regions. Based on these observations, we decided to also remove 193 A2I sites that fall into HLA, somatic recombination or duplicated genome regions. A total of 2119 A2I sites were kept for further analysis. Importantly, all these A2I sites were reported in at least one public database (Table S1). In addition, the overall (mean) RNAe level of the 2,119 A2I sites was strongly correlated between the two datasets (Pearson = 0.86. Figure 2e). Together, these observations suggest that these A2I RDD are real RNAe events and unlikely due to any technical limitations.

## Effect of RNAe-QTL on RNA secondary structure

Genetic association analysis showed that most of the RNAe sites were associated with more than a single SNP (Table S1). For each RNAe site, we aimed to identify the causal RNAe-QTL over all the associated SNPs. RNAe-QTL could be localized on the same or the opposite strand of the RNAe site in a folded double-stranded RNA molecule. To characterize the potential effect of RNAe-QTL controlling RNAe event on RNA secondary structure, we focused on RNAe-QTL opposite to the RNAe site within a potential RNA hairpin structure. We performed local alignment of the RNA sequence in the direct vicinity of the RNAe site against the reverse sequences surrounding all the associated SNPs ($P < 1e–6$). We identified the causal RNAe-QTL by pairwise sequence alignment and then estimated the minimal estimated free energy.

To identify the potential causal RNAe-QTL, we targeted RNAe sites filling the two conditions: (i) the strongest association was replicated in QMDiab, and (ii) both RNAe site and the associated SNPs were located on the same transcript and within 5000 bp distance. We then performed local alignment of the ±30 base pairs (bp) surrounding the RNAe site to the inverse + /- 30 bp surrounding all the associated SNPs. We chose ±30 bp sequence for a reliable pairwise sequence alignment in repeated alu elements where most of RNAe sites and their associated SNPs were localized. We found 65 RNAe sites filling conditions (i) and (ii) (Table S2). In 31 cases, the potential causal RNAe-QTL (the alignment with the minimal free energy) was the strongest association (Fig S7–S10 for examples). We identified 20 other cases where the potential causal RNAe-QTL was not the strongest associations. Nonetheless, the potential causal RNAe-QTL for these 20 cases was highly correlated (R2 > 0.9) with the strongest association (Table S2). For the 14 remaining RNAe sites, the RNAe-QTL with the lowest free energy alignment was not correlated to the strongest association (R2 < 0.9).

To confirm the causal RNAe-QTL identification, we predicted the RNA secondary structure of sequence in the vicinity of the 65 RNAe sites. Interestingly, the predicted RNA structure of the transcripts harboring 4 RNAe sites showed the potential causal RNAe-QTL was localized on the exact opposite of the RNAe site within the double stranded RNA segment (hairpin). Moreover, the potential causal RNAe-QTL for these four cases was the strongest association over all the associated SNPs (Table S2). Importantly, the potential causal RNAe-QTL alleles were Uracil (U) to Cytosine (C) where U was associated with a lower RNAe level and C with a higher RNAe level [37].

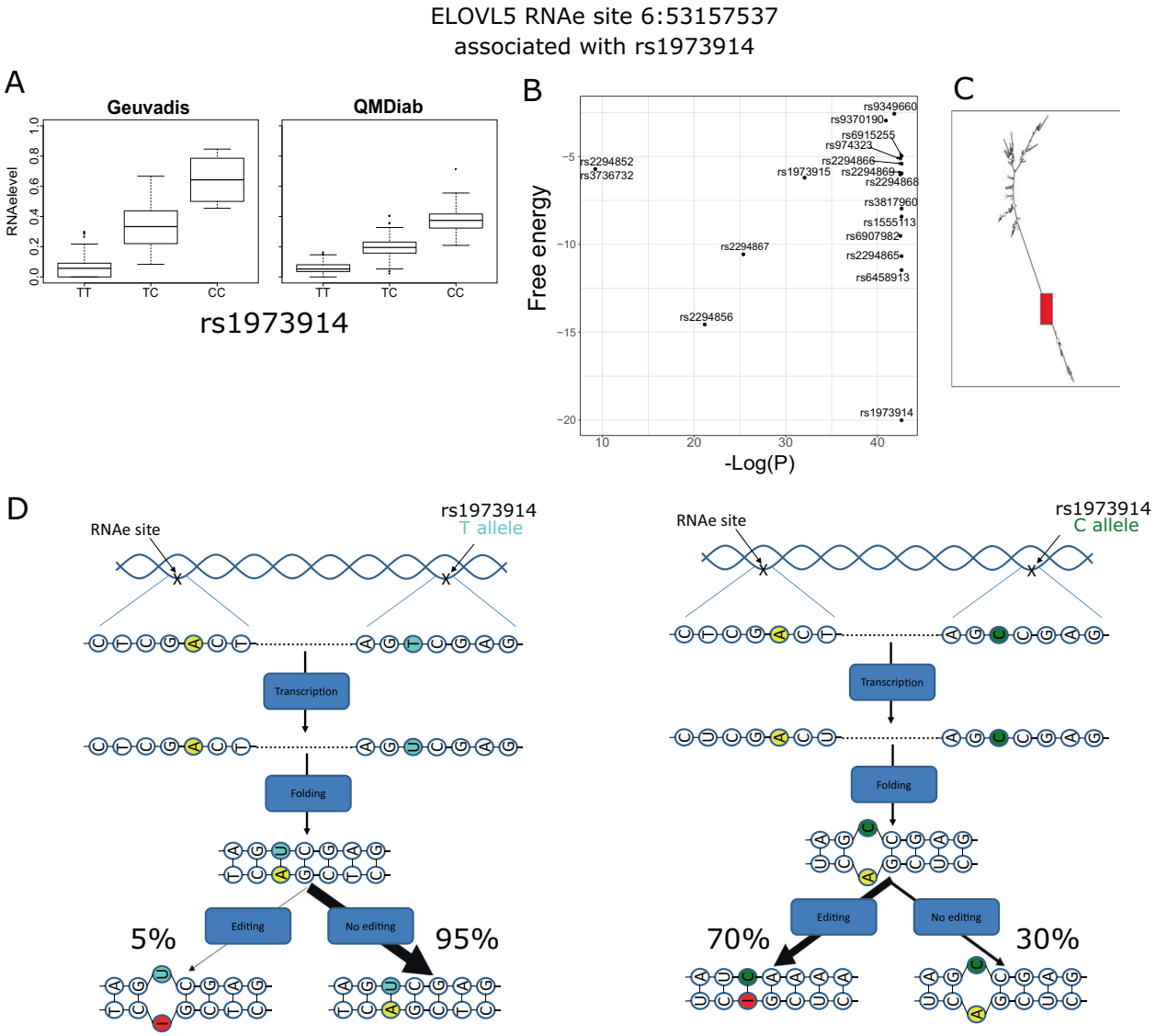


**Fig. 3 Effect of the RNAe-QTL on the RNA secondary structure. Three other cases are provided in Fig S4–S6. a** Boxplot showing the significant association of SNP rs1973914 with the RNAe level of the RNAe site chr6:53157537 in Geuvadis and QMDiab. **b** Distribution of the RNAe/associated SNPs pairwise alignment estimated free energy as function of -log P-value of RNAe level/SNP association. The strongest association (rs1973914) shows the alignment with the lowest estimated free energy. **c** The complete RNA predicted secondary structure. The red rectangle indicates the localization of the hairpin carrying the RNAe and the RNAe-QTL on opposite strand. **d** Cartoon representation of the RNAe and the associated RNAe-QTL. All the four possible cases are represented for both alleles (T and C) of the RNAe-QTL rs1973914 in a nonedited and edited RNA. Colored in yellow and red is the RNAe site in nonedited and edited RNA, respectively. Colored in green is the C allele of rs1973914 and in magenta is the T allele.

This observation indicates that depending on the RNAe state, each allele has a potential opposite effect on the RNAe event by enhancing as in the case of C in nonedited RNA or U in edited RNA, or reducing as in the case of U in nonedited RNA or C in edited RNA the generation of a loop in the double stranded structure (Fig. 3 for an example of one case and Fig S4–S6 for the remaining three cases). These results suggest the involvement of the potential causal RNAe-QTL in regulating RNAe events by acting on the RNA structure.

## Allele specific RNA editing

Next, we assessed allele specific RNA editing (ASE) for the two alleles of the RNAe-QTL. ASE is the determination whether RNAe displays RNAe-QTL allele specificity (Fig. 4a). All the RNAe-QTLs were bi-allelic SNPs with only two alleles: reference and alternative (multi-allelic SNPs and indels were filtered out). ASE was estimated on heterozygous individuals as they carry both alleles. Both datasets were sequenced using paired end sequencing

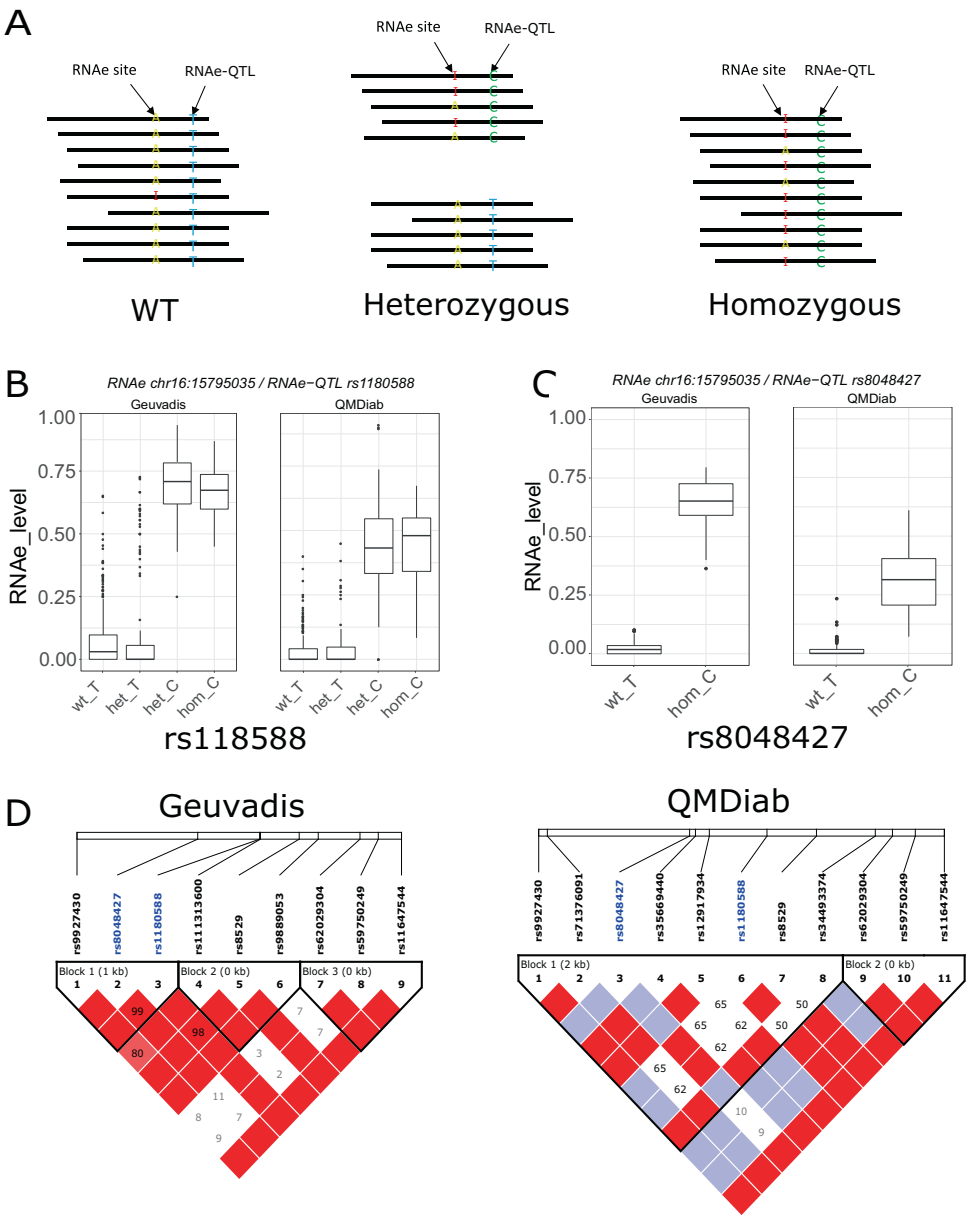


**Fig. 4 Allele specific RNA editing (ASE). a** Schematic diagram of ASE. Pairs of RNAe sites and associated RNAe-QTL falling on the same reads offer the opportunity to compute ASE on heterozygous. **b** ASE level (RNAe displaying RNAe-QTL allele-specificity) for the association RNAe site chr16:15795035 and the ANP rs1180588 showing a similar ASE level for the reference allele (T) in wild type homozygous (wt_T) and heterozygous genotypes (het_T), and a similar ASE level associated with the alternative allele (**c**) in homozygous for the alternative allele (hom_C) and heterozygous genotypes (het_C) in Geuvadis (left) and in QMDiab replication (right). In this case, both RNAe site and the associated SNP (rs1180588) are close enough to be on the same reads. C. ASE level for the strongest association with the RNAe site chr16:15795035; rs8048427. The RNAe site and the RNAe-QTL were too distant to fall on the same reads. Represented here are only ASE level for reference allele in homozygous wild type (T) and ASE level for the alternative allele for homozygous for the alternative allele (**c**). **d** Linkage disequilibrium plot for the region covering the two RNAe-QTL rs1180588 and rs8048427. The value within each diamond represents the pairwise correlation between SNPs defined by the upper left and the upper right sides of the diamond. Shading represents the magnitude and significance of pairwise LD, with a red-to-white gradient reflecting higher to lower LD values. In both datasets, the two RNAe-QTL (colored in blue) belong to the same LD block suggesting that these two RNAe-QTLs represent a single signal.

technology. We involved only sequencing reads or pair of reads (mates) that carry both the RNAe site and the RNAe-QTL. We counted for each allele of the RNAe-QTL, the number of edited and nonedited reads. For each RNAe site, we enumerated individuals having at least five reads or mates covering both the RNAe site and the RNAe-QTL and kept only sites with more than ten individuals in both projects. Finally, the significance of the ASE was

estimated using Fisher's exact and chi squared tests in both datasets.

Out of 18 identified cases, both statistical tests show clear evidence for ASE in 17 cases in both datasets (Table S3). For example, the RNAe site chr16:15795035 was associated with the RNAe-QTL rs1180588 in Geuvadis ($P = 1.67e{-}104$) and the RNAe-QTL association was replicated in QMDiab ($P = 6.17e{-}60$). Both alleles T (reference) and C (alternative) show RNAe level specificity in heterozygous individuals (Table S3). To confirm the ASE at the individual level, we estimated the ASE level associated with each allele. Interestingly, the ASE level was higher for allele C (associated with high RNAe level) compared with allele T (associated with low RNAe level) in both datasets (Fig. 4b).

Furthermore, another SNP (rs8048427) represents the strongest association for this RNAe site chr16:15795035. In addition, rs8048427 is the potential causal RNAe-QTL (Fig S5). Following these observations, we investigated the ASE for the RNAe-QTL rs8048427. The two RNAe-QTL rs1180588 and rs8048427 for the RNAe site chr16:15795035 were correlated ($R2 = 0.64$ and 0.58 in Geuvadis and QMDiab, respectively) indicating these two RNAe-QTL might be a single signal (Fig. 4d). The RNAe-QTL rs8048427 was too distant from the RNAe site chr16:15795035 to fall on the same reads or mates (650 bp distance). We assessed the ASE level for rs8048427 in homozygous wild type and homozygous for the alternative allele, for the reference and the alternative allele, respectively (Fig. 4c). We found that ASE level of rs8048427 was similar to rs1180588.

## Discussion

In 2011, Li et al. published the first study on RDD in a population of 27 individuals [38]. All types of modification have been reported where the A2I RDD was the most common. This study has been widely discussed by the scientific community [39]. These discussions noted that more than 82% of the detected RDD were sequencing artifacts, read alignment errors, or incorrect genotypes in poorly covered regions. Since then, to our best knowledge, no comprehensive analysis of RDD quantity and origin has been performed. Our study confirms that enzymatic-driven RNAe is limited to A2I, at least in the cell types analyzed here. In addition, DNA duplication and rearrangement (somatic recombination and hyper-morphic MHC regions) are the most common source of non-A2I RDD.

We are aware of some technical limitations to this study. A conversion of cytosine to uracil is another reported type of RNAe mediated by Apobec-1. But due to the stringent RNAe filtering strategy adopted here, this RNAe type is probably excluded because its RNAe level is ~ 15–20% [40]. Various other interesting findings are probably missed due to the different ethnicity structure between the discovery and the replication cohorts. The material similarity used in both RNA-seq data (Lymphoblastoid cells in Geuvadis and whole blood cells in QMDiab) did not emphasize the tissue intervariability as has been shown in the multi-tissue RNA-seq data [3] (Fig S3). We confirmed by our analysis (Fig. 2e) that the overall RNAe level was correlated between the two cell types (lymphocyte cell lines and whole blood). Lastly, the conservative approach used to detect common RDD may omit the identification of rare non-A2I RDD.

RNAe is mainly controlled by *in cis* genetic variations [32]. We performed an in cis genetic association study of 2,119 RNAe sites in Geuvadis to detect 1,257 genetic associations and replicated 472 in QMDiab with Bonferroni multi testing correction. Two studies identified the genetic variants controlling the RNAe level in the Geuvadis datasets [18, 22]. In total, both studies identified 754 RNAe sites associated with a variant in at least one of the 5 Geuvadis populations: 393 in [18] and 416 in [22]. In our analysis, the majority of these RNAe sites (528, 70%) were filtered out because they were not identified in QMDiab. A total of 80 RNAe-QTLs were replicated in our analysis. The remaining 146 RNAe-QTLs were not confirmed in our analysis due to the different strategy adopted here. Unlike the two cited studies, we did not focus on the RNAe population intervariability.

RNA secondary structure is commonly predicted by computational methods based on a free energy model. However, the prediction accuracy is limited due to errors from thermodynamic parameters, kinetic barriers, the existence of multiple structural conformations, and protein interactions [41]. Here, we applied a strategy based on aligning the region in the vicinity of the RNAe site to the region in the vicinity all associated SNPs to identify the causal RNAe-QTL. Out of 65 tested cases, the potential causal RNAe-QTL for 50 cases was the strongest association or a variant highly correlated with the strongest association. We replicated the RNAe-QTL rs8048427 association with the RNAe chr16:15795035 identified by Park et al. [18].

We demonstrated clear evidence of a specific RNAe level associated with the two alleles of the RNAe-QTL. Recently, ASE has been proposed to be a potential mechanism for explaining how UTR and nonsynonymous SNPs impact phenotypes and diseases by controlling non-synonymous RNAe sites [42]. In addition, ASE confirms the effect of RNAe-QTL on the RNA structure [42].

In summary, our study reviews, through a comprehensive analysis the origins of all types of RDD in white blood cells. A2I is the only likely RDD resulting from a

RNAe process. Moreover, it is widely accepted now that RNAe plays a role in double stranded RNA structure weakening. As we have shown in the examples, RNAe might also be involved in RNA structure maintaining. Last, it will be relevant in the next steps to experimentally validate both hypotheses stating that RNAe destabilizes (maintains) RNA structure by abrogating (supporting) the formation of long perfectly matched double stranded RNA structure.

## Data availability

Complete summary statistics for genetic association with RNAe level in both datasets, Gauvadis and QMDiab, are available at https://figshare.com/projects/RNAe/77310.

**Acknowledgements** We thank the staff of the HMC dermatology department and of WCM-Q for their contribution to QMDiab. We thank Laurent Abel from the human genetics of infectious diseases lab for the scientific discussions. Finally, we are grateful to all study participants of QMDiab for their invaluable contributions to this study.

**Funding** This work was supported by the Biomedical Research Program at Weill Cornell Medicine in Qatar, a program funded by the Qatar Foundation. The statements made herein are solely the responsibility of the authors.

## Compliance with ethical standards

**Conflict of interest** The authors declare that they have no conflict of interest.

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

## References

1. Nishikura K. Functions and regulation of RNA editing by ADAR deaminases. Annu Rev Biochem. 2010;79:321–49.
2. Liddicoat BJ, Piskol R, Chalk AM, Ramaswami G, Higuchi M, Hartner JC, et al. RNA editing by ADAR1 prevents MDA5 sensing of endogenous dsRNA as nonself. Science. 2015;349:1115–20.
3. Tan MH, Li Q, Shanmugam R, Piskol R, Kohler J, Young AN, et al. Dynamic landscape and regulation of RNA editing in mammals. Nature. 2017;550:249–54.
4. Jacobs MM, Fogg RL, Emeson RB, Stanwood GD. ADAR1 and ADAR2 expression and editing activity during forebrain development. Dev Neurosci. 2009;31:223–37.
5. Ramaswami G, Li JB. RADAR: a rigorously annotated database of A-to-I RNA editing. Nucleic Acids Res. 2014;42 Database issue:D109-113.
6. Kiran AM, O'Mahony JJ, Sanjeev K, Baranov PV. Darned in 2013: inclusion of model organisms and linking with Wikipedia. Nucleic Acids Res. 2013;41 Database issue:D258-261.
7. Stellos K, Gatsiou A, Stamatelopoulos K, Perisic Matic L, John D, Lunella FF, et al. Adenosine-to-inosine RNA editing controls cathepsin S expression in atherosclerosis by enabling HuR-mediated post-transcriptional regulation. Nat Med. 2016;22:1140–50.
8. Prasanth KV, Prasanth SG, Xuan Z, Hearn S, Freier SM, Bennett CF, et al. Regulating gene expression through RNA nuclear retention. Cell. 2005;123:249–63.
9. Rueter SM, Dawson TR, Emeson RB. Regulation of alternative splicing by RNA editing. Nature. 1999;399:75–80.
10. Brümmer A, Yang Y, Chan TW, Xiao X. Structure-mediated modulation of mRNA abundance by A-to-I editing. Nat Commun. 2017;8:1255.
11. Miyamura Y, Suzuki T, Kono M, Inagaki K, Ito S, Suzuki N, et al. Mutations of the RNA-specific adenosine deaminase gene (DSRAD) are involved in dyschromatosis symmetrica hereditaria. Am J Hum Genet. 2003;73:693–9.
12. Rice GI, Kasher PR, Forte GMA, Mannion NM, Greenwood SM, Szynkiewicz M, et al. Mutations in ADAR1 cause Aicardi-Goutières syndrome associated with a type I interferon signature. Nat Genet. 2012;44:1243–8.
13. Gaisler-Salomon I, Kravitz E, Feiler Y, Safran M, Biegon A, Amariglio N, et al. Hippocampus-specific deficiency in RNA editing of GluA2 in Alzheimer's disease. Neurobiol Aging. 2014;35:1785–91.
14. Galeano F, Rossetti C, Tomaselli S, Cifaldi L, Lezzerini M, Pezzullo M, et al. ADAR2-editing activity inhibits glioblastoma growth through the modulation of the CDC14B/Skp2/p21/p27 axis. Oncogene. 2013;32:998–1009.
15. Chen L, Li Y, Lin CH, Chan THM, Chow RKK, Song Y, et al. Recoding RNA editing of AZIN1 predisposes to hepatocellular carcinoma. Nat Med. 2013;19:209–16.
16. Shah SP, Morin RD, Khattra J, Prentice L, Pugh T, Burleigh A, et al. Mutational evolution in a lobular breast tumour profiled at single nucleotide resolution. Nature. 2009;461:809–13.
17. Zipeto MA, Court AC, Sadarangani A, Delos Santos NP, Balaian L, Chun H-J, et al. ADAR1 Activation Drives Leukemia Stem Cell Self-Renewal by Impairing Let-7 Biogenesis. Cell Stem Cell. 2016;19:177–91.
18. Park E, Guo J, Shen S, Demirdjian L, Wu YN, Lin L, et al. Population and allelic variation of A-to-I RNA editing in human transcriptomes. Genome Biol. 2017;18:143.
19. Ramaswami G, Deng P, Zhang R, Anna Carbone M, Mackay TFC, Li JB. Genetic mapping uncovers cis-regulatory landscape of RNA editing. Nat Commun. 2015;6:8194.
20. Kurmangaliyev YZ, Ali S, Nuzhdin SV. Genetic determinants of RNA editing levels of ADAR targets in drosophila melanogaster. G3. 2015;6:391–6.
21. Franzén O, Ermel R, Sukhavasi K, Jain R, Jain A, Betsholtz C, et al. Global analysis of A-to-I RNA editing reveals association with common disease variants. PeerJ. 2018;6:e4466.
22. Ouyang Z, Ren C, Liu F, An G, Bo X, Shu W. The landscape of the A-to-I RNA editome from 462 human genomes. Sci Rep. 2018;8:12069.
23. Mook-Kanamori DO, MME-D Selim, Takiddin AH, Al-Homsi H, KAS Al-Mahmoud, Al-Obaidli A, et al. 1,5-Anhydroglucitol in saliva is a noninvasive marker of short-term glycemic control. J Clin Endocrinol Metab. 2014;99:E479–483.
24. Lappalainen T, Sammeth M, Friedländer MR, t Hoen PAC, Monlong J, Rivas MA, et al. Transcriptome and genome sequencing uncovers functional variation in humans. Nature. 2013;501:506–11.
25. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinforma Oxf Engl. 2013;29:15–21.
26. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010;20:1297–303.
27. Yang H, Wang K. Genomic variant annotation and prioritization with ANNOVAR and wANNOVAR. Nat Protoc. 2015;10:1556–66.

28. R: a language and environment for statistical computing. https://www.gbif.org/tool/81287/r-a-language-and-environment-for-statistical-computing. Accessed 8 Nov 2018.
29. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. GigaScience. 2015;4:7.
30. Delaneau O, Zagury J-F, Marchini J. Improved whole-chromosome phasing for disease and population genetic studies. Nat Methods. 2013;10:5–6.
31. Howie BN, Donnelly P, Marchini J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genet. 2009;5:e1000529.
32. Walkley CR, Li JB. Rewriting the transcriptome: adenosine-to-inosine RNA editing by ADARs. Genome Biol 2017;18:205.
33. Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 1994;22:4673–80.
34. Lorenz R, Bernhart SH, Höner Zu Siederdissen C, Tafer H, Flamm C, Stadler PF, et al. ViennaRNA Package 2.0. Algorithms Mol Biol Amb. 2011;6:26.
35. Roth SH, Levanon EY, Eisenberg E. Genome-wide quantification of ADAR adenosine-to-inosine RNA editing activity. Nat Methods. 2019;16:1131–8.
36. Robinson JT, Thorvaldsdóttir H, Wenger AM, Zehir A, Mesirov JP. Variant review with the integrative genomics viewer. Cancer Res 2017;77:e31–4.
37. Kleinberger Y, Eisenberg E. Large-scale analysis of structural, sequence and thermodynamic characteristics of A-to-I RNA editing sites in human Alu repeats. BMC Genomics. 2010;11:453.
38. Li M, Wang IX, Li Y, Bruzel A, Richards AL, Toung JM, et al. Widespread RNA and DNA sequence differences in the human transcriptome. Science. 2011;333:53–8.
39. Kleinman CL, Majewski J. Comment on "Widespread RNA and DNA sequence differences in the human transcriptome.". Science. 2012;335:1302. author reply 1302.
40. Harjanto D, Papamarkou T, Oates CJ, Rayon-Estrada V, Papavasiliou FN, Papavasiliou A. RNA editing generates cellular subsets with diverse sequence within populations. Nat Commun. 2016;7:12145.
41. Wu Y, Shi B, Ding X, Liu T, Hu X, Yip KY, et al. Improved prediction of RNA secondary structure by integrating the free energy model with restraints derived from experimental probing data. Nucleic Acids Res. 2015;43:7247–59.
42. Zhou Z-Y, Hu Y, Li A, Li Y-J, Zhao H, Wang S-Q, et al. Genome wide analyses uncover allele-specific RNA editing in human and mouse. Nucleic Acids Res. 2018;46:8888–97.