Genome-wide analysis of RNA sites consistently edited in human blood reveals interactions with mRNA processing genes and suggests correlations with biological and drug-related variables.

Edoardo Giacopuzzi[1], Massimo Gennarelli[1,2], Chiara Sacco[1,2], Chiara Magri[2], Alessandro Barbon[2]

[1]Genetics Unit, IRCCS Centro S. Giovanni di Dio, Fatebenefratelli, 25123 Brescia, Italy

[2]Department of Molecular and Translational Medicine, Biology and Genetic Unit, University of Brescia, 25123 Brescia, ITALY.

edoardo.giacopuzzi@unibs.it

massimo.gennarelli@unibs.it

chiara.sacco@unibs.it

chiara.magri@unibs.it

alessandro.barbon@unibs.it

Corresponding Author:

Prof. Alessandro Barbon: Department of Molecular and Translational Medicine, Biology and Genetic Unit, University of Brescia, 25123 Brescia, ITALY.

alessandro.barbon@unibs.it

Abstract

Background

A-to-I RNA editing is a co-/post-transcriptional modification catalyzed by ADAR enzymes, that deaminate Adenosines (A) into Inosines (I). Most of known editing events are located within inverted ALU repeats, but they also occur in coding sequences and may alter the function of encoded proteins. RNA editing contributes to generate transcriptomic diversity and it is found altered in cancer, autoimmune and neurological disorders. However, little is known about how editing process could be influenced by genetic variations, biological and environmental variables.

Results

We analyzed RNA editing levels in human blood using RNA-seq data from 459 healthy individuals and identified 2,079 sites consistently edited in this tissue, that we considered the most biologically relevant editing sites. As expected, analysis of gene expression revealed that *ADAR* is the major contributor to editing on these sites, explaining ~13% of observed variability. After removing *ADAR* effect, we found significant associations for 1,122 genes, mainly involved in RNA processing. These genes were significantly enriched in genes encoding proteins interacting with ADARs, including 276 potential ADARs interactors and 9 ADARs direct partners. In addition, association analysis on 28 biological and drugs intake variables revealed several factors potentially influencing RNA editing in blood, including sex, age, BMI, drugs and medications. Finally, we identified genetic loci associated to editing levels, including known *ADAR* eQTLs and a small region on chromosome 7, containing *LOC730338* lincRNA gene.

Conclusions

Our data provides a detailed picture of the most relevant RNA editing events and their variability in human blood, giving interesting insights on the mechanisms behind this post-transcriptional modification and its regulation in this tissue.

Keywords

RNA editing, ADAR1, ADAR2, RNA-seq, Blood

Background

RNA editing is a co-/post-transcriptional process based on the enzymatic deamination of specific adenosines (A) into inosines (I). Since inosine has similar base-pairing properties to guanosine, it is read as guanosine by both splicing and translation machineries, thus generating different RNA molecules from those coded by DNA [1]. RNA editing contributes to the diversification of the information that is encoded in the genome of an organism, thereby providing a greater degree of complexity. Currently, the conversion of A to I is thought to be the most common RNA editing process in higher eukaryotic cells [2].

RNA editing is catalyzed by adenosine deaminase enzymes (ADARs) [3, 4]. In mammals, three members of the ADAR family have been characterized so far. ADAR1 (gene name: *ADAR*) and ADAR2 (gene name: *ADARB1*) are active enzymes expressed in many tissues, while ADAR3 (gene name: *ADARB2*) is expressed specifically in the Central Nervous System (CNS). To date, no functional RNA editing activity has been attributed to this enzyme. The critical role of ADAR enzymes is shown by phenotypes of knockout mice that resulted in embryonic lethality or death shortly after birth [5–7], clearly indicating that A-I RNA editing is essential for normal life and development. In addition, dysregulated RNA editing levels at specific re-coding sites have been linked with a variety of diseases, including neurological or psychiatric disorders and cancer [2, 8, 9]. Interestingly, ADARs mRNA and protein expression levels do not always reflect RNA editing levels [10]. It has been shown that the subcellular distribution of ADAR enzymes [11] and their interaction with inhibitors [12, 13] and activators [14, 15] influence ADARs activities.

Originally, A-to-I RNA editing in mammalian cells was described for a low number of mRNAs and it was responsible for deep changes of protein functions. These editing sites were discovered serendipitously comparing DNA and cDNA sequences by direct sequence analysis [16, 17]. The number of identified RNA editing sites has largely increased with the diffusion of RNA sequencing (RNA-Seq) studies based on next generation sequencing technologies (NGS), reaching over two millions of sites. The majority of RNA editing sites is located within intragenic non-coding sequences: 5'UTRs, 3'UTRs and intronic retrotransposon elements, such as ALU inverted repeats

[18, 19]. With lowering cost of NGS, many RNA-Seq datasets from human tissues, healthy and pathological conditions, have been deposited in sequence databases, available to the scientific community. In parallel, the development of computational pipelines to search for RNA editing sites on RNA-Seq data, allowed a global analysis of the editing reaction, shedding light on its evolutionary conservation [20], tissues specificity [21, 22], cellular specificity [23] and its role in diseases such cancer [24] or neurological disorders [9, 25].

About 2.5 million editing sites have been identified so far and are listed in RNA editing databases [26, 27], but only recently the dynamic and regulation of RNA editing has been systematically investigated in human tissues [22]. However, little is known about how editing process could be influenced by genetic variations [28, 29], biological and environmental variables [30]. Here, we want to go further in characterizing and understanding the complexity of RNA editing. Focusing on the most likely biologically relevant sites, we will unveil possible correlations with gene expression and genetic variations. To this aim, we investigated consistently edited sites from existing RNA seq dataset of whole blood from 459 healthy subjects [31], correlating editing levels with a collection of 28 biological and pharmacological variables, as well as with genes expression and genotyping data.

Results

RNA editing sites consistently edited in human blood samples

Of the ~ 2M editing sites reported in RADAR database, our dataset of 459 RNA-seq samples had adequate coverage for 709,184 sites, covering > 75% of the total sites reported for genes expressed in blood according to GTEx data. (Additional file 1: Figure S1). Most of these sites are edited only in a small fraction of samples and 691,304 (97.5%) have no detectable editing levels in our cohort (Additional file 1: Figure S2). To provide a picture of the most biologically relevant editing sites in human blood we focalized our attention on 2,079 consistently editing sites (CES), namely those with at least 5% of editing level in at least 20% of individuals. These sites are mainly distributed in ALU regions (1,805; 86.5%) and 3'UTR regions (1,234; 59.4%), distributed across 421 genes. Overall, we detected 1,266 sites in exons of protein coding genes, including 10 recoding sites (resulting in a missense substitution) and 12 synonymous sites. We also detected 53 sites annotated on ncRNAs (Figure 1a, b). Detailed statistics of the 2,079 trusted sites are reported in Additional file 2, while recoding sites in Table 1.

Considering mean values for each site, detected editing levels range from 0.05 to 1, with most sites showing moderate editing levels between 0.05 and 0.30 (Figure 1c). We also detected 33 sites highly edited (mean value ≥ 0.9), located mainly in intronic regions (Figure 1d). Highly edited sites are reported in Additional file 1: Table S1. To further assess reliability of detected sites, we compared the CES editing levels with those reported in the REDIportal database [26], a well-established resources containing multi-tissue estimations of RNA editing levels. The comparison revealed high concordance (concordance correlation coefficient 0.84, Additional file 1: Figure S3), for 2,035 and 2,003 sites when considering the whole REDIportal dataset or blood tissue data, respectively. For the sites included in this study, the editing levels from REDIportal are reported in Additional file 2.

We used Spearman correlation test to analyze correlation in editing level changes across the 2,079 CES to find sites with co-regulated RNA editing. We found 540 significant relationships (FDR < 0.05) involving 361 sites. Correlations were generally low with only 58 sites with relationships

above 0.5 rho value. Correlations become stronger for close sites, especially below 50 bp distance, with 60 out of 66 (91%) high rho (≥ 0.5) relationships located in this range. No strong negative correlations (rho < -0.5) were observed (Additional file 1: Figure S4).

Genes influencing the total editing rate of CES

We performed regression analysis to identify genes whose expression is associated to the CES total editing rate, calculated for each subject as the total sum of G-containing reads divided by the total number of reads observed at all the 2,079 CES. The analysis revealed 4,719 genes associated to the CES total editing rate (FDR < 0.05). Enrichment analysis on Gene Ontology biological processes (GO-BP) revealed a strong enrichment for genes involved in immune system and interferon signaling (FDR < 1e-6, Figure 2a). Among significant genes, *ADAR* emerged as the top influencing factor, explaining about 13% of the observed variability, while *ADARB1* showed no significant effect (Figure 2b). The influence of *ADAR* was similar on ALU (~10%) and non-ALU (~13%) sites, while *ADARB1* remains not associated also when considering groups separately (Additional file 1: Figure S5). *ADARB2* gene was not detectable in our gene expression data. When the same analysis was repeated removing *ADAR* effect, we obtained 1,122 genes associated to CES total editing rate (FDR < 0.05), including 376 with a strong association at FDR < 0.01 (Additional file 3). Enrichment analysis on GO-BP and REACTOME pathways revealed that these genes mainly impact ribonucleoprotein complex biogenesis and RNA metabolism / processing (Figure 2c).

To assess possible interactions between ADAR enzymes and genes whose expression is associated with CES total editing rate, we performed network analysis using data on protein-protein interactions from STRING v.10, BioPlex and BIOGRID databases. We created a 415 proteins network including genes significantly associated with CES total editing rate (FDR < 0.01) that interact with ADAR1 and ADAR2 proteins or one of their first neighbors (Figure 3a). Considering top associated genes with FDR < 0.01 we found 285 out of 376 (76%) interactors. The observed fraction of ADARs-connected genes represents a significant enrichment compared to random groups (empirical p-value < 1e-06, Figure 3b) and these genes are strongly enriched for RNA binding proteins (Figure 3c). Significant genes also includes 9 genes with direct interactions with

ADARs (Table 2). We estimated the role of each node in this network looking at degree and betweenness values. Degree value account for the number of interaction involving a single node in the network, while betweenness is a measure of centrality based on shortest paths. Node with higher betweenness centrality would have a major role in the network, because more information will pass through that node. Among ADARs proteins, ADAR1 is the main target of network interactions (0.077 betweenness centrality, 72 degree values) compared to ADAR2 (0.018 betweenness centrality, 29 degree). Among associated genes with a direct interaction with ADARs, *ELAVL1*, *RPA1* and *IFI16* act as relevant hub nodes, with betweenness centrality values of 0.137, 0.028, 0.020, respectively (Figure 3d). Detailed network-based statistics are reported in Additional file 4, together with adjusted P value for association with CES total editing rate.

Biological factors possibly influencing editing levels

In order to identify possible biological factors influencing editing levels, we studied the correlation of the 28 biological / pharmacological variables described in Additional file 1, table S2 with CES total editing rate and with the *ADAR* expression level. Overall, 5 variables (age, current and maximum BMI, sex and blood pressure medications) revealed a significant correlation (Figure 4, complete results in Additional File 1: Table S3). The same variables, with the exception of sex, resulted also significantly correlated with the expression level of *ADAR*. To better investigate the effect of biological / pharmacological variables and of *ADAR* expression and identify correlations between these variables and specific groups of CES, we performed principal component (PC) analysis of CES editing levels. Even if the variance explained by single component is generally low (PC1 ~ 0.025), our data revealed 24 factors with a significant correlation (p-value < 0.05) with one of the first 5 PCs (Figure 5 and Additional file 1: Table S4).

Sex, age and BMI, were confirmed as major contributors to editing variability being associated to PC1 and 2 and influencing also lower components. Substance intake variables, including alcohol, smoke and drugs assumption, were clearly associated to PC3, even if few of the observed associations, namely alcohol intake, thyroid and cholesterol lowering medications, may be influenced by sex biased distribution (Additional file 1: Table S5). As expected from the analysis of

genes influencing CES total editing rate, *ADAR* expression level resulted associated to the first 2 PCs, confirming its pivotal role in shaping editing levels variability. Correlation of editing levels for single sites with the first 5 PCs are reported in Additional file 5.

Identification of genetic variants influencing CES total editing rate

We performed genome wide association analysis between genotyping data of 573,801 SNPs and CES total editing rate to identify SNPs associated to editing levels in human blood (Figure 6a). Based on GRASP database of known SNP-phenotype associations, the 100 SNPs with the lowest p-values are involved in 44 human phenotypes with strongest impact on ADAR expression and AraC toxicity (Additional file 1: Table S6). Among these SNPs, 25 SNPs are known eQTLs regulating expression of 19 genes in blood tissue (Additional file 1: Table S7). These genes resulted to be nominally enriched for genes encoding for RNA-binding proteins involved in transcription regulation and response to cytokines (Additional file 1: Figure S6).

After variant clumping, our analysis identified a single significant locus on chromosome 7 (rs856554: p-value 1.86e-7, FDR 0.042), containing the lincRNA gene *LOC730338* (ENSG00000233539) (Figure 6b and Table 3). The SNP rs856554 showed a significant effect on CES total editing rate and seems to influence *ADAR* expression, despite this effect do not reach statistical significance (Figure 6c). Association results for single SNPs with nominal p-value < 0.05 and for loci after variants clumping are reported in Additional file 6.

Taken together, the 36 known *ADAR* eQTLs present in genotyping data explain 5.5% of CES total editing rate variability (p value 3.46e-4) and 5 of them resulted among the top 100 associated SNPs (Additional file 1: Table S8). The effect of the top associated *ADAR* eQTL (rs6699825) on *ADAR* expression and CES total editing rate is represented in Figure 6d.

Discussion

The process of A-to-I RNA editing has gained increasing attention in recent years, being implicated in multiple aspects of human physiology and, when dysregulated, in human diseases, such as neurological disorders and cancer [2, 9, 24]. Thanks to advances in next-generation sequencing technology, the prevalence and dynamic of "RNA editome" have been recently characterized across many tissues and developmental stages [18, 19, 21, 22].

Overall, more than 2 million editing sites have been described so far, but most of them occur at very low level in inverted repeat ALU sequences and likely represent random editing with low impact on biological functions [32]. To focus only on those sites that are most likely biologically relevant in human blood, we first selected consistently edited sites (CES) across our dataset of about 450 RNA-Seq samples, resulting in a group of 2,079 sites with at least 5 % editing in at least 100 individuals.

As expected, the majority of these sites are located in inverted repeat ALU sequences [18, 19] that facilitate the formation of a RNA double stranded secondary structure with high affinity for ADAR editing enzymes. Interestingly, near 60% of detectable editing sites are located in the 3'UTRs, that are known preferred binding sites for miRNAs. This suggests a potential extensive role of editing process in modulating the miRNA mediated regulation of gene expression in blood. Indeed, RNA editing in the 3'UTRs might introduce nucleotide changes to miRNA target sites or stabilize RNA secondary structure reducing the accessibility for AGO2-miRNAs complex [33–35].

We identified 22 editing sites located in coding sequences: 12 resulting in synonymous modifications and 10 inducing non-synonymous amino acid changes (re-coding sites). Among the latter, there were well studied re-coding sites, such as the S/G site of *AZIN1* [8], that mediates the binding to antizyme and cell cycle progression; the G/R site of *BLCAP* [36], that is involved in the regulation of STAT3 signaling pathway; and the L/R site of *NEIL1* [37], that might modulate the DNA repair capability of the enzyme. Their editing levels range from high (75% of *NEIL1*) to medium-low (14% and 16% for *BLCAP* and *AZIN1*, respectively), indicating that both edited and unedited isoforms are needed for the proper function of the tissues. Interestingly, among the re-coding sites, we also detected sites with an high editing level, such as two sites edited at 70% on the

small subunit processome component (*UTP14C*). It is worth to notice that in blood cells several editing sites in 3'UTR and intronic regions reach an editing level of more than 90%. The high efficiency of editing indicates a strong ADARs binding, affecting transcripts stability or structure [38], but the actual functional effect of these fully edited sites remain elusive. Overall, RNA editing process in human blood seems more pervasive than previously reported, prompting for further analyses to understand its biological effects also in healthy subjects.

Further, we investigated the association of genes expression with total editing rate of CES. *ADAR* (encoding ADAR1 enzyme) resulted as the top associated gene and its expression explained about 13% of observed variability, while *ADARB1* (encoding ADAR2 enzyme) was not associated to global editing level even when ALU and non-ALU sites were considered separately. *ADARB2* (encoding ADAR3 protein) is not expressed in blood cells, excluding that it could have a major negative effect on the editing levels in blood as observed for brain tissues [22]. Thus, ADAR1 emerges as the major contributor to editing process in blood, as already reported for human B cells and other tissues [22, 39], while other ADAR enzymes seems to have only a limited effect. Overall, association analysis revealed 4,719 genes that might have a potential effect on the editing process, strongly enriched for genes involved in the immune system and interferon signaling. This supports the association between inflammatory processes and A-to-I editing, that seems mediated by *ADAR* expression modulation. Indeed, ADAR1 is present in two main isoforms, a constitutive p110 and an interferon inducible p150 form that is active under an inflammatory response [40]. Moreover, RNA editing, especially ADAR1 activity, are important to modulate innate immunity [41, 42]. Modification in the global editing level has been reported after inflammation in mouse and in vitro studies using several inflammatory mediators [43].

When the effect of *ADAR* expression is removed from our analysis, new genes associated to global editing level emerged. These genes are mainly involved in RNA metabolism and ribonucleoprotein complex processing, confirming what recently found after a global analysis of GTEx data [22] and strengthening the role of RNA editing complex in the RNA processing [39].

Associated genes after *ADAR* correction are strongly enriched for potential ADARs interactors, as revealed by network analysis using data from protein-protein interaction databases. Moreover,

associated genes interacting with ADARs mainly encode for RNA binding proteins, as revealed by enrichment analysis, suggesting that they could be involved in RNA recognition or assembly of the editing complex. Network analysis showed that ADAR1 is the main target of these interactions, confirming its prevalent role in blood samples, compared to the other editing enzymes. We also identified 9 associated genes with direct interaction with ADARs. Among them, *ELAVL1*, *RPA1* and *IFI16* emerged as relevant hubs in the network, aggregating most of the interactions directed to ADARs proteins. The stabilizing RNA-binding protein human antigen R (HuR), encoded by *ELAVL1*, has been recently proposed as an ADAR1 interactor involved in the regulation of transcripts stability in human B cells [39]. It was unclear if this stabilizing effect is editing depended or not. However, ADAR1-mediated RNA editing of the 3'UTR of cathepsin S enables the recruitment of HuR to the 3' UTR, thereby controlling the cathepsin S mRNA stability and expression in endothelial cells and in human atherosclerotic vascular diseases [30]. The observed association between the global editing level and the *ELAVL1* expression strengthens a general role of RNA editing in RNA stability through the modulation of expression of genes involved in RNA metabolism. *RPA1* and *IFI16* have never been involved in ADARs activity and represent new interesting partners that could expand the understanding of ADAR1 function and regulation. *RPA1* gene encodes the largest subunit of the heterotrimeric Replication Protein A (RPA) complex, which binds to single-stranded DNA, forming a nucleoprotein complex that is involved in DNA replication, repair, recombination, telomere maintenance and response to DNA damage [44]. ADAR1 presents Z-DNA binding domains, that are not present in the other editing enzymes [45], helping to direct ADAR1 to active transcription sites and to interact with DNA. Thus, the interaction with RPA1 protein might broaden ADAR1 activity also in the field of DNA repair and maintenance. *IFI16*, interferon gamma inducible protein 16, encodes a member of the HIN-200 (hematopoietic interferon-inducible nuclear antigens with 200 amino acid repeats) cytokines family. This protein interacts with p53 and retinoblastoma-1 and localizes to the nucleoplasm and nucleoli [46], where also ADAR enzymes are present. Both IFI16 protein and ADAR1 were associated to response to viral DNA and regulation of immune and interferon signaling responses [46, 47]. Future studies will establish the actual functional meaning of these interactions and their role in RNA

editing.

Recently, global editing level have been investigated across tissues and in different species [21, 22] and also correlated to the genetic background of human population [30] and to common disease variants [29]. However, the published studies lack a detailed characterization of samples that allows assessing the role of biological and environmental factors.

Relying on our dataset containing several demographic, biological and pharmacological variables, we also investigated the potential impact of these external factors on RNA editing process genome-wide. Five variables showed significant correlations with CES total editing rate, namely blood pressure medications, sex, age and body mass index (BMI, current and max). Except for sex, their effect on editing levels seems mainly driven through modulation of *ADAR* expression. Correlation between age and editing was already reported during brain development both in rat [48] and in primates [49] and our data strengthens this correlation also outside the central nervous system. Moreover, it has been been previously suggested that in liver cancer, ALU editing is gender dependent, being lower in the tumor of female patients; however normal tissue do not showed this difference [50]. Here we showed that, at least in blood, gender is a main factor influencing RNA editing. Finally and for the first time, our study correlated CES editing levels with BMI and blood pressure medications, shedding light on new medical areas in which editing regulation may be involved.

A more detailed analysis using principal components of editing levels revealed twenty-four variables potentially influencing RNA editing for specific groups of sites in blood, even if the proportion of editing variability explained is low. Sex, age and BMI confirmed to have the strongest effect on RNA editing levels, being associated to the first principal component. PC1 and PC2 components are also strongly associated with *ADAR* expression, supporting that the observed effect of these biological factors could be due to modulation of *ADAR*. This data indicates that, when analyzing editing variations among different groups, such as in case / control studies, gender, age and other biological characteristics should be taken into account carefully to avoid biased results. Interestingly, the third principal component is associated with variables related to drugs and substances intake, but only weakly with *ADAR* expression, indicating that drugs might modulate

editing levels also through alternative mechanisms. A role of drugs, in particular antidepressants, have been reported for editing sites on specific neuronal transcripts [51–54]. Our analysis suggests a broader impact of drugs, with several substances able to influence RNA editing process in blood. Even if the contribution of single substances seems small, (PC3 explains ~0.7% of editing variability), substances intake overall may have a larger effect, as suggested by association of global intake variables ("all drugs" and "none treatments") with multiple principal components.

Finally, we analyzed genotyping data to identify SNPs associated to CES total editing rate. Known *ADAR* eQTLs resulted among the SNPs with the best p-values and, taken together, they explain about 5% of the observed variation in global editing. Our data confirmed that they actually influence expression of *ADAR* in blood and this explain also the observed effect on editing levels. We found a single locus significantly associated with global editing level in blood, localized on chromosome 7 and containing the lincRNA gene *LOC730338.* Long intergenic noncoding RNAs, or lincRNAs, are long RNA transcripts (longer than 200 nucleotides) that have been identified in mammalian genomes mainly by bioinformatic analysis of transcriptomic data. Despite thousands of lincRNAs are now validated, the exact functional role remains unknown for most of them. lincRNAs appear to contribute to the control of gene expression and have a role in cell differentiation and maintenance of cell identity [55]. It has been recently reported in *C. elegans* that lncRNAs are extensively down-regulated in the absence of ADARs as a result of siRNA generation [56]. The authors suggests that ADARs can interfere with the generation of siRNAs by endogenous RNAi and promote lncRNA expression. *LOC730339* expression cannot be measured in our dataset since it lack a poly-A tail; therefore, it is not possible to assess if the associated SNPs observed in the locus could act as eQTLs for this lincRNA. These SNPs seem to have only a marginal correlation with ADAR expression and the mechanism that link this locus and *LOC730339* to editing process remain to be investigated.

According to GRASP, HaploReg and GTEx databases, the 100 SNPs with the best p-values also contain several SNPs reported in previous GWAS studies, as well as known eQTLs of genes coding for RNA binding proteins involved in transcription, supporting a co-regulation of RNA editing and transcription and a possible role of editing in several human phenotypes. Overall, this data indicates

that genetic variations, especially those associated to *ADAR* expression, can influence observed editing levels. The analysis of these SNPs should be taken into account when investigating editing levels in different human populations both in physiological and pathological conditions.

Despite our RNA-seq dataset has only moderate coverage and thus may have limited power to investigate sites with very low editing levels, we assume that biologically relevant sites should be edited at detectable level consistently across samples [32] and thus our dataset is able to provide a detailed picture of the distribution and regulation of the most relevant editing events.

Conclusion

This study provides a detailed picture of the most consistent RNA editing sites and their variability in human blood. Our results confirm the pivotal role of ADAR1 in the regulation of RNA editing process in blood and suggest new genes, genetic variants, biological and environmental variables that are involved in the RNA editing process. Future studies will be required to confirm and clarify their role and their relationship with the ADAR family enzymes.

Methods

Description of data

RNA-Seq raw data (aligned reads) was obtained from NIMH repository, NIMH Study 88 / Site 621, dataset 7 (Levinson RNA Sequencing Data). The original data and samples details are described in [31]. This dataset includes poly(A)+ RNA sequencing and genotyping data from blood samples of 922 subjects, 463 MDD patients and 459 control subjects. The present study focuses only on the 459 controls. Data are provided as aligned reads on hg19 human genome assembly with transcript mapped to RefSeq canonical dataset. Samples are sequenced with a median of 65.6 M reads (31.6 - 258.3), resulting in a median of 14,289 (11,660 - 15,137) detectable genes addressed by at least 10 reads (Additional file 1: Figure S7). Only the 14,961 genes covered with at least 10 reads in at least 100 subjects were considered in the present study for association with editing levels.

A detailed phenotypic description including demographic, pharmacological and biological variables is also included for each subject. Among them, we considered only those relevant in at least 30 subjects and not related to MDD clinical evaluation or socio-economic variables. The 28 variables considered in this study are reported in Additional file 1: Table S2. Moreover, each experiment is annotated with a rich set of technical variables, representing quality metrics of RNA sequencing and characteristics of the blood sample. Normalized gene expression data is given as residuals of ridge regression of log-transformed read counts with 35 technical variables, to remove the effect of experimental biases (see the original paper [31]).

Assessment of editing levels and selection of consistently edited sites

The original aligned reads were de-duplicated using Picard and the editing levels were then determined genome-wide from BAM files using REDITools v.1.0.4 [57] with the following parameters: -t25 -m20 -c10 -q25 -O5 -l -V0.05 -n0.05. Only sites with a minimum coverage of 10 reads were considered, otherwise their editing level was considered as missing.

To reduce the chance of measuring false-positive editing sites, we selected only sites that met the following criteria: i) sites reported within RefSeq genes by RADAR database [27] and never seen as

Single Nucleotide Variants in the human population according to 1000G phase3 and ExAC v.0.3.1; ii) sites occurring in regions were incorrect alignments could have generated artifacts in editing detection were filtered out: known pseudogenes from GENCODE v25; segmental duplication with ≥ 99% identity; single exon genes, that are often retrotransposed genes with high similarity to the corresponding parent gene.

The filtered dataset resulted in 709,184 sites, representing > 75% RADAR editing sites occurring in blood expressed genes. Finally, to provide a picture of most biologically relevant editing events in blood, we decided to focus only on sites with detectable editing levels (at least 5%) in at least 100 subjects (~20% of total individuals) for subsequent quantitative analysis, resulting in a final dataset of 2,079 sites (consistently edited sites, CES).

Comparison with REDIPortal dataset

We compared editing levels detected in CES from blood samples with similar data obtained from REDI Portal [26]. Editing levels were retrieved directly from REDIPortal database, containing RNA editing values calculated from 55 body sites of 150 healthy individuals from GTEx project. Mean editing levels of our 2,079 CES were compared with corresponding data reported for blood tissue in REDI portal. To assess concordance between the two datasets, we calculated concordance correlation coefficient between mean editing values detected in our data and reported in REDIPortal blood tissue for overlapping sites.

Correlation between editing levels across sites

Using Spearman correlation test, we analyzed correlation of editing levels across the 2,079 CES. Each site was analyzed against all other sites for a total of 4,322,241 tests. FDR correction modified as in [58] was used to account for multiple tests with related variables. Corrplot R package v.0.84 was used to analyze correlation matrices and generate correlation plots.

Association between CES total editing rate and gene expression

To investigate which genes could influence the editing process, we used robust linear regression

(robust v.0.4 R package) to assess the association between gene expression levels and the CES total editing rate in each subject. CES total editing rate for each subject was calculated as in Equation 1.

$$\frac{\sum_{i=1}^{m} G_i}{\sum_{i=1}^{m} C_i}$$

The sum of number of G-containing reads (G_i) observed at all CES (m), divided by the sum of total reads observed (C_i) at all CES.

CES total editing rate was determined also for Alu sites and non-Alu sites, separately. To choose the set of phenotypic, biological and pharmacological variables to include as covariates in regression analyses, a stepwise model selection by AIC was performed (using stepAIC from MASS R package v.7.3-5). The 6 included variables are indicated in Additional file 1: Table S1. Moreover, since there was a correlation between the variance observed at each editing site and its sequencing coverage for sites with coverage below ~ 40 X (Additional file 1: Figure S8), the log2 of reads count was also included as covariate in the analysis. The strength of the association was determined by ANOVA test comparing the null ('background') model that includes only the set of covariates with the full model (covariates plus normalized expression levels). FDR was used to correct for multiple tests. Subsequently, association analyses were repeated including ADAR expression as additional covariate, to remove the effect of ADAR expression.

Gene set enrichment analysis and gene network analysis

The impact on biological functions and cellular pathways of genes found associated with CES total editing rate was investigated using hypergeometric test. We tested the over-representation of pathways among the subset of significant genes at 5% FDR level compared to all expressed genes. Enrichment analysis was performed separately for the following sets from MSigDB v.6.0: cellular pathways from REACTOME and the three main Gene Ontology categories (Cellular Components, GO:CC; Biological Process, GO:BP; Molecular Function, GO:MF). To verify if the proteins encoded by these genes could interact with ADAR proteins, the major enzymes involved in RNA-editing, we explored human protein-protein interaction (PPI) data. First, we created a

comprehensive human PPI network combining data from 3 different sources: BioPlex 2.0 [59], BioGRID 3.4.15 [60] and STRING 10.0 [61]. For the BioGRID dataset, only interactions marked as physical were taken in to account, whereas for the STRING dataset only interactions with a combined score above 400 and physical/biochemical evidences were considered. Proteins of the ubiquitin gene family were removed from the network, resulting in a final PPI dataset with 22,913 proteins (nodes) and 833,686 interactions, containing 108 direct interactors of ADARs (ADAR1, ADAR2 and ADAR3 proteins). Among the 376 genes strongly associated with global editing level (FDR < 0.01), we assessed the number of encoded proteins interacting with ADAR1, ADAR2 or one of their first neighbors. To test the significance of these overlap, we performed a random test on the overall set of 14,961 genes addressable in our RNA-Seq data (background genes). We randomly sampled among background genes 1 million groups of N genes (N = 376) and for each simulated group we counted how many elements interacts directly with ADARs or one of their neighbors. Empirical p-value was then calculated as the number of test resulting in an equal or higher number of interactors. Cytoscape v.3.4.0 [62] was used to visualize the PPI network and calculate network related statistics.

Identification of biological factors correlated with editing levels

To investigate which biological and pharmacological variables could influence editing levels in blood, we studied associations between the 28 biological / pharmacological variables described in Additional file 1: Table S2 and CES total editing rate across subjects. For the 5 variables resulting in significant associations, we also analyzed their correlation with *ADAR* gene expression levels. To further investigate the effect of biological and pharmacological variables on editing levels in blood, we studied their correlation with the Principal Components of editing levels (PCs). To compute PCs, the missing values of the sites were first imputed using a nonparametric imputation method based on random forest (missForest R package v.1.4 [63]). The PCs were then determined on the complete data using the prcomp R package. To identify the number of PCs to account for, we evaluated the percentage of explained variance by the top 30 PCs, and identified the 5th component as the point at which the explained variance plateaus.

In both analyses, Kruskal-Wallis test, Mann-Whitney-Wilcoxon test and Pearson's product-moment correlation test were used to assess association for categorical, binary and continuous variables, respectively. To identify which editing sites were most correlated with each PC, we analyzed the loadings, that could be interpreted as correlation coefficient between the original variables and components. Moreover, given a high number of sites and low loading values, to deepen the role of each site in the computation of the PCs, we performed the Pearson correlation test. We considered a "moderate" correlation when its absolute value was between 0.3 and 0.5 and the test passed the Bonferroni threshold, while a "weak" correlation was considered when the correlation absolute values ranged between 0 and 0.3 and the respective p-values were significant for Bonferroni correction.

Association study for SNPs and global editing levels

To identify SNPs associated to global editing level, we analyzed genotyping data and global editing levels in the 459 human blood samples. Starting from genotypes provided in the original dataset [31], we filtered raw data removing SNPs strongly deviating from Hardy-Weinberg equilibrium (fisher test p-value < 1e-4) and with a minor allele frequency below 0.10, to ensure that the least represented genotype accounts for at least 5 individuals. The final dataset contained 573,801 SNPs. We used plink v.1.9 linear association analysis with additive model, including the same 7 covariates used for analysis of gene expression (see above, Additional file 1: Table S1). FDR was used to correct for multiple tests. After association analysis, we used GCTA [64] to evaluate the impact of ADAR known eQTLs on observed global editing levels, using the same set of covariates included for the plink association analysis. This analysis was performed including 36 known ADAR eQTLs present in our genotyping data. The top 100 associated SNPs were overlapped with GRASP 2.0 [65] database, to assess their role in human phenotypes and diseases. We then evaluated genes potentially regulated by the top 100 SNPs based on known blood eQTLs from HaploReg 4.1 [66] and GTEx [67]. Enrichment analysis was conducted for potentially regulated genes on GO:BP, GO:MF, GO:CC and REACTOME pathways using hypergeometric test. Background gene group was obtained as all genes with a known eQTL among all the 573,801 tested SNPs. To identify

significant loci associated to global editing level, we performed variant clumping based on the association results, using plink with 1Mb window and 0.5 R2 thresholds. In this way all SNPs in 1Mb window and with R2 ≥ 0.5 are grouped together around the index SNP, that is the SNP with the lower association p-value.

List of abbreviations

CES: Consistently edited site(s); GO:MF: Gene Ontology Molecular Function; GO:BP: Gene Ontology Biological Processes; GO:CC: Gene Ontology Cellular Component: eQTL: expression Quantitative Trait Locus.

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Availability of data and material

RNA-Seq raw data and description of subjects are available from NIMH repository, Study 88, but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available.

The datasets supporting the conclusions of this article are included within the article (and its additional files). Editing levels calculated from RNA-Seq are available in the GitHub repository, https://github.com/gedoardo83/RNA_editing_blood.

RADAR database: http://rnaedit.com/

GTEx database: https://www.GTExportal.org/home/

Competing interests

The authors declare that they have no competing interests

Funding

This work was supported by research grants from the Italian Ministry of University (PRIN projects n. 2006058401 to AB), from Fondazione Cariplo (grant: 2017-0620) and from the University of Brescia (Project “Refract” to MG). EG has been supported by Fondazione Cariplo and Regione Lombardia (Grant Emblematici Maggiori 2015-1080).

Authors' contributions

EG, AB conceived and designed the analysis. EG, CS performed bioinformatic and statistical analysis. MG acquired the data. MG and CM provided intellectual input and conceptual advice and revised the paper. EG, AB, CS wrote the paper. All authors read and approved the final manuscript.

Acknowledgements

Analyzed data come from NIMH Study 88 – Data and biomaterials were provided by Dr. Douglas F. Levinson. This project was supported by National Institutes of Health/National Institute of Mental Health grants 5RC2MH089916 (PI: Douglas F. Levinson, M.D.; Co-investigators: Myrna M. Weissman, Ph.D., James B. Potash, M.D., MPH, Daphne Koller, Ph.D., and Alexander E. Urban, Ph.D.) and 3R01MH090941 (Co-investigator: Daphne Koller, Ph.D.).

References

1. Nishikura K. Functions and regulation of RNA editing by ADAR deaminases. Annu Rev Biochem. 2010;79:321–49.

2. Behm M, Öhman M. RNA Editing: A Contributor to Neuronal Dynamics in the Mammalian Brain. Trends Genet. 2016;32:165–75.

3. Bass BL. RNA editing by adenosine deaminases that act on RNA. Annu Rev Biochem. 2002;71:817–46.

4. Orlandi C, Barbon A, Barlati S. Activity regulation of adenosine deaminases acting on RNA (ADARs). Mol Neurobiol. 2012;45:61–75.

5. Hartner JC, Schmittwolf C, Kispert A, Müller AM, Higuchi M, Seeburg PH. Liver disintegration in the mouse embryo caused by deficiency in the RNA-editing enzyme ADAR1. J Biol Chem. 2004;279:4894–902.

6. Higuchi M, Maas S, Single FN, Hartner J, Rozov A, Burnashev N, et al. Point mutation in an AMPA receptor gene rescues lethality in mice deficient in the RNA-editing enzyme ADAR2. Nature. 2000;406:78–81.

7. Wang Q, Khillan J, Gadue P, Nishikura K. Requirement of the RNA editing deaminase ADAR1 gene for embryonic erythropoiesis. Science. 2000;290:1765–8.

8. 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.

9. Khermesh K, D'Erchia AM, Barak M, Annese A, Wachtel C, Levanon EY, et al. Reduced levels of protein recoding by A-to-I RNA editing in Alzheimer's disease. RNA. 2016;22:290–302.

10. Wahlstedt H, Daniel C, Ensterö M, Ohman M. Large-scale mRNA sequencing determines global regulation of RNA editing during brain development. Genome Res. 2009;19:978–86.

11. Sansam CL, Wells KS, Emeson RB. Modulation of RNA editing by functional nucleolar sequestration of ADAR2. Proc Natl Acad Sci U S A. 2003;100:14018–23.

12. Filippini A, Bonini D, Lacoux C, Pacini L, Zingariello M, Sancillo L, et al. Absence of the Fragile X Mental Retardation Protein results in defects of RNA editing of neuronal mRNAs in mouse. RNA Biol. 2017;14:1580–91.

13. Tariq A, Garncarz W, Handl C, Balik A, Pusch O, Jantsch MF. RNA-interacting proteins act as site-specific repressors of ADAR2-mediated RNA editing and fluctuate upon neuronal stimulation. Nucleic Acids Res. 2013;41:2581–93.

14. Garncarz W, Tariq A, Handl C, Pusch O, Jantsch MF. A high-throughput screen to identify enhancers of ADAR-mediated RNA-editing. RNA Biol. 2013;10:192–204.

15. Marcucci R, Brindle J, Paro S, Casadio A, Hempel S, Morrice N, et al. Pin1 and WWP2 regulate GluR2 Q/R site RNA editing by ADAR2 with opposing effects. EMBO J. 2011;30:4211–22.

16. Sommer B, Köhler M, Sprengel R, Seeburg PH. RNA editing in brain controls a determinant of ion flow in glutamate-gated channels. Cell. 1991;67:11–9.

17. Burns CM, Chu H, Rueter SM, Hutchinson LK, Canton H, Sanders-Bush E, et al. Regulation of serotonin-2C receptor G-protein coupling by RNA editing. Nature. 1997;387:303–8.

18. Levanon EY, Eisenberg E, Yelin R, Nemzer S, Hallegger M, Shemesh R, et al. Systematic identification of abundant A-to-I editing sites in the human transcriptome. Nat Biotechnol. 2004;22:1001–5.

19. Li JB, Levanon EY, Yoon J-K, Aach J, Xie B, Leproust E, et al. Genome-wide identification of human RNA editing sites by parallel DNA capturing and sequencing. Science. 2009;324:1210–3.

20. Bazak L, Haviv A, Barak M, Jacob-Hirsch J, Deng P, Zhang R, et al. A-to-I RNA editing occurs at over a hundred million genomic sites, located in a majority of human genes. Genome Res. 2014;24:365–76.

21. Picardi E, Manzari C, Mastropasqua F, Aiello I, D'Erchia AM, Pesole G. Profiling RNA editing in human tissues: towards the inosinome Atlas. Sci Rep. 2015;5:14941.

22. 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.

23. Picardi E, Horner DS, Pesole G. Single cell transcriptomics reveals specific RNA editing signatures in the human brain. Rna. 2017;:rna.058271.116.

24. Fritzell K, Xu L-D, Lagergren J, Öhman M. ADARs and editing: The role of A-to-I RNA modification in cancer progression. Semin Cell Dev Biol. 2017.

25. Filippini A, Bonini D, La Via L, Barbon A. The Good and the Bad of Glutamate Receptor RNA Editing. Mol Neurobiol. 2016.

26. Picardi E, D'Erchia AM, Lo Giudice C, Pesole G. REDIportal: a comprehensive database of A-to-I RNA editing events in humans. Nucleic Acids Res. 2017;45:D750–7.

27. Ramaswami G, Li JB. RADAR: a rigorously annotated database of A-to-I RNA editing. Nucleic Acids Res. 2014;42 Database issue:D109-13.

28. Gu T, Gatti DM, Srivastava A, Snyder EM, Raghupathy N, Simecek P, et al. Genetic Architectures of Quantitative Variation in RNA Editing Pathways. Genetics. 2016;202:787–98.

29. 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.

30. 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.

31. Mostafavi S, Battle A, Zhu X, Potash JB, Weissman MM, Shi J, et al. Type I interferon signaling genes in recurrent major depression: increased expression detected by whole-blood RNA sequencing. Mol Psychiatry. 2014;19:1267–74.

32. Ulbricht RJ, Emeson RB. One hundred million adenosine-to-inosine RNA editing sites: Hearing through the noise. BioEssays. 2014;36:730–5.

33. Brümmer A, Yang Y, Chan TW, Xiao X. Structure-mediated modulation of mRNA abundance by A-to-I editing. Nat Commun. 2017;8:1255.

34. Borchert GM, Gilmore BL, Spengler RM, Xing Y, Lanier W, Bhattacharya D, et al. Adenosine deamination in human transcripts generates novel microRNA binding sites. Hum Mol Genet. 2009;18:4801–7.

35. Soundararajan R, Stearns TM, Griswold AL, Mehta A, Czachor A, Fukumoto J, et al. Detection of canonical A-to-G editing events at 3' UTRs and microRNA target sites in human lungs using next-generation sequencing. Oncotarget. 2015;6:35726–36.

36. Levanon EY, Hallegger M, Kinar Y, Shemesh R, Djinovic-Carugo K, Rechavi G, et al. Evolutionarily conserved human targets of adenosine to inosine RNA editing. Nucleic Acids Res. 2005;33:1162–8.

37. Yeo J, Goodman RA, Schirle NT, David SS, Beal PA. RNA editing changes the lesion specificity for the DNA repair enzyme NEIL1. Proc Natl Acad Sci U S A. 2010;107:20715–9.

38. Daniel C, Widmark A, Rigardt D, Öhman M. Editing inducer elements increases A-to-I editing efficiency in the mammalian transcriptome. Genome Biol. 2017;18:195.

39. 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.

40. George CX, John L, Samuel CE. An RNA editor, adenosine deaminase acting on double-stranded RNA (ADAR1). J Interferon Cytokine Res. 2014;34:437–46.

41. Mannion NM, Greenwood SM, Young R, Cox S, Brindle J, Read D, et al. The RNA-editing enzyme ADAR1 controls innate immune responses to RNA. Cell Rep. 2014;9:1482–94.

42. Song C, Sakurai M, Shiromoto Y, Nishikura K. Functions of the RNA Editing Enzyme ADAR1 and Their Relevance to Human Diseases. Genes (Basel). 2016;7.

43. Yang J-H, Luo X, Nie Y, Su Y, Zhao Q, Kabir K, et al. Widespread inosine-containing mRNA in lymphocytes regulated by ADAR1 in response to inflammation. Immunology. 2003;109:15–23.

44. Liu T, Huang J. Replication protein A and more: single-stranded DNA-binding proteins in eukaryotic cells. Acta Biochim Biophys Sin (Shanghai). 2016;48:665–70.

45. Barraud P, Allain FH-T. ADAR proteins: double-stranded RNA and Z-DNA binding domains. Curr Top Microbiol Immunol. 2012;353:35–60.

46. Choubey D, Panchanathan R. IFI16, an amplifier of DNA-damage response: Role in cellular senescence and aging-associated inflammatory diseases. Ageing Res Rev. 2016;28:27–36.

47. Samuel CE. Adenosine deaminases acting on RNA (ADARs) are both antiviral and proviral. Virology. 2011;411:180–93.

48. Zaidan H, Ramaswami G, Golumbic YN, Sher N, Malik A, Barak M, et al. A-to-I RNA editing in the rat brain is age-dependent, region-specific and sensitive to environmental stress across generations. BMC Genomics. 2018;19:28.

49. Li Z, Bammann H, Li M, Liang H, Yan Z, Phoebe Chen Y-P, et al. Evolutionary and ontogenetic changes in RNA editing in human, chimpanzee, and macaque brains. RNA. 2013;19:1693–702.

50. Paz-Yaacov N, Bazak L, Buchumenski I, Porath HT, Danan-Gotthold M, Knisbacher BA, et al. Elevated RNA Editing Activity Is a Major Contributor to Transcriptomic Diversity in Tumors. Cell Rep. 2015;13:267–76.

51. Barbon A, Orlandi C, La Via L, Caracciolo L, Tardito D, Musazzi L, et al. Antidepressant treatments change 5-HT2C receptor mRNA expression in rat prefrontal/frontal cortex and hippocampus. Neuropsychobiology. 2011;63:160–8.

52. Barbon A, Popoli M, La Via L, Moraschi S, Vallini I, Tardito D, et al. Regulation of editing and expression of glutamate alpha-amino-propionic-acid (AMPA)/kainate receptors by antidepressant drugs. Biol Psychiatry. 2006;59:713–20.

53. Labasque M, Meffre J, Carrat G, Becamel C, Bockaert J, Marin P. Constitutive activity of serotonin 2C receptors at G protein-independent signaling: modulation by RNA editing and antidepressants. Mol Pharmacol. 2010;78:818–26.

54. Englander MT, Dulawa SC, Bhansali P, Schmauss C. How stress and fluoxetine modulate serotonin 2C receptor pre-mRNA editing. J Neurosci. 2005;25:648–51.

55. Cabili MN, Trapnell C, Goff L, Koziol M, Tazon-Vega B, Regev A, et al. Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. Genes Dev. 2011;25:1915–27.

56. Goldstein B, Agranat-Tamir L, Light D, Ben-Naim Zgayer O, Fishman A, Lamm AT. A-to-I RNA editing promotes developmental stage-specific gene and lncRNA expression. Genome Res. 2017;27:462–70.

57. Picardi E, Pesole G. REDItools: high-throughput RNA editing detection made easy. Bioinformatics. 2013;29:1813–4.

58. Benjamini Y, Yekutieli D. The Control of the False Discovery Rate in Multiple Testing under Dependency. The Annals of Statistics. 29:1165–88.

59. Huttlin EL, Bruckner RJ, Paulo JA, Cannon JR, Ting L, Baltier K, et al. Architecture of the human interactome defines protein communities and disease networks. Nature. 2017;545:505–9.

60. Chatr-Aryamontri A, Oughtred R, Boucher L, Rust J, Chang C, Kolas NK, et al. The BioGRID interaction database: 2017 update. Nucleic Acids Res. 2017;45:D369–79.

61. Szklarczyk D, Morris JH, Cook H, Kuhn M, Wyder S, Simonovic M, et al. The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible. Nucleic Acids Res. 2017;45:D362–8.

62. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 2003;13:2498–504.

63. Stekhoven DJ, Bühlmann P. MissForest--non-parametric missing value imputation for mixed-type data. Bioinformatics. 2012;28:112–8.

64. Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: a tool for genome-wide complex trait analysis. Am J Hum Genet. 2011;88:76–82.

65. Leslie R, O'Donnell CJ, Johnson AD. GRASP: analysis of genotype-phenotype results from

1390 genome-wide association studies and corresponding open access database. Bioinformatics. 2014;30:i185-94.

66. Ward LD, Kellis M. HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. Nucleic Acids Res. 2012;40 Database issue:D930-4.

67. GTEx Consortium, Laboratory DA &Coordinating C (LDACC)—Analysis WG, Statistical Methods groups—Analysis Working Group, Enhancing GTEx (eGTEx) groups, NIH Common Fund, NIH/NCI, et al. Genetic effects on gene expression across human tissues. Nature. 2017;550:204–13.

Tables

Table 1. Editing levels detected for the 10 recoding sites identified in human blood

Site id (hg19)	Gene	Strand	Aa change	Alu	Editing Minimum	Editing Mean	Editing Maximum
chr3_49398423	*RHOA*	-	Lys->Arg	yes	0.19	0.33	0.64
chr4_2835556	*SH3BP2*	+	Arg->Gly	no	0.05	0.08	0.16
chr4_2940026	*NOP14*	-	Asn->Ser	yes	0.1	0.23	0.56
chr4_77979680	*CCNI*	-	Arg->Gly	no	0.05	0.08	0.19
chr8_103841636	*AZIN1*	-	Ser->Gly	no	0.07	0.16	0.45
chr13_52604264	*UTP14C*	+	Ser->Gly	no	0.24	0.64	1
chr13_52604880	*UTP14C*	+	Gln->Arg	no	0.45	0.85	1
chr15_75646086	*NEIL1*	+	Lys->Arg	no	0.27	0.73	1
chr16_3292200	*MEFV*	-	Stp->Trp	yes	0.05	0.16	0.36
chr20_36147563	*BLCAP*	-	Gln->Arg	no	0.05	0.14	0.33

Table 2. Network based statistics for the 9 ADARs direct partners

Gene	Betweenness centrality	Degree	Association adjusted p	Associated genes rank
ELAVL1	0.137	168	0.0312	840
HDLBP	0.001	15	0.0400	986
HNRNPUL1	0.007	39	0.0004	10
IFI16	0.020	66	0.0055	246
RPA1	0.028	96	0.0140	496
SDAD1	0.002	19	0.0253	713
SUZ12	0.007	52	0.0022	110
THOC1	0.002	17	0.0184	587
USP39	0.002	20	0.0190	596

The table reports betweenness centrality and degree values for the 9 genes directly interacting with ADARs in PPI databases and significantly associated to global editing levels (adjusted $p < 0.05$). Adjusted p-values are calculated as FDR corrected p values from robust regression of global editing level and gene expression. The rank position among top associated genes is also reported.

Table 3. Top 4 SNPs associated to CES total editing rate define a locus on chromosome 7

SNP	Chr	Position	A1	Beta	p value (FDR clumped association)	R2	Gene (distance)
rs856554*	7	46760129	G	0.00352	1.87×10^{-07} (0.042)	-	*LOC730338*] (23.4kb)
rs856589	7	46734307	A	0.00326	2.44×10^{-07}	0.73	[*LOC730338*]
rs6463347	7	46780614	C	0.00328	5.14×10^{-07}	0.76	*LOC730338*] (43.9kb)
rs856565	7	46721854	A	0.00327	8.16×10^{-07}	0.88	[*LOC730338*]

For each SNP the table reports distance from *LOC730338* gene. Gene name within square brackets indicate SNPs located within the gene, while single bracket indicate 3' distance. Index SNP is marked with * and the FDR of association for LD-clumped association analysis is reported. R2 with the index SNP is reported for other SNPs in the locus. Genomic coordinates refer to hg19 genome assembly.

Figure legends

Figure 1. Distribution of 2,079 consistently edited sites (CES) analyzed in the study

(a) Distribution of the 2,079 CES within ALU regions and (b) based on functional classification. (c) Density plot representing overall distribution of editing levels. (d) Density plots of editing levels for different editing site categories and ALU/non-ALU sites.

Figure 2. Association between gene expression and CES total editing rate

We analyzed association between CES total editing rate and gene expression for 14,961 human genes. (a) Gene set enrichment analysis by hypergeometric test on GO-BP categories and REACTOME pathways revealed that associated genes are mainly involved in immune system response mediated by interferon I and alpha / beta. (b) When we analyze distribution of CES total editing rate and *ADAR* gene expression, *ADAR* expression levels explains ~ 13% of observed variability. No significant effect is observed for *ADARB1* expression. *ADAR* and *ADARB1* expression levels are reported as residuals of ridge regression with technical covariates (see description of data in methods section). The graphs report adjusted p-value and R2 value from robust regression analysis. (c) The 1,122 genes associated to CES total editing rate after removing *ADAR* expression effect were enriched for genes mainly involved in ribonucleoprotein and RNA processing.

Figure 3. Genes associated with CES total editing rate are enriched for ADAR interactors

(a) Reconstructed PPI network including ADARs and best genes significantly associated with global editing levels (FDR < 0.01). Among these genes, we observed 285 potential ADARs interactors, including 9 direct partners of ADARs proteins. (b) Boxplot of number of ADARs interacting genes observed in 1M random simulations. The observed number of interactions (285) resulted in empirical p-value < 1e-6. (c) ADARs interactors are strongly enriched for RNA binding proteins in GO-MF categories. (d) Distribution of degree and betweenness centrality values among network nodes are represented by violin plots. ADAR1 protein has a major role (higher values)

among ADAR proteins. Among ADARs direct partners, ELAVL1, RPA1 and IFI16 showed high values of degree and betweenness centrality, suggesting a central role in the network.

Figure 4. Impact of biological / pharmacological factors on CES total editing rate

Our analysis revealed significant associations with CES total editing rate for blood pressure medication (a), BMI current (b), Age (c) and Sex (d). The first three variable resulted associated to *ADAR* expression level, as well. Significance level (p) is reported in each plot based on Mann-Whitney-Wilcoxon or Pearson's product-moment correlation test for binary and continuous variables, respectively. For continuous variables the Pearson correlation coefficient (r) is also reported.

Figure 5. Impact of biological / pharmacological factors on PCs of editing levels

We analyzed correlation between 28 biological / pharmacological variables and principal components (PCs) calculated from editing levels of 2,079 CES. The heathmap represents strength of association, with significant p values < 0.05 colored in yellow-red scale. Sex, age and BMI are the strongest factors, correlated to PC1, while substance / drug intake variables were mostly associated with PC3. For each PC, variance explained is represented by the bar plot in the upper side, while association with *ADAR* expression with is represented in the lower panel.

Figure 6. Association study for SNPs and CES total editing rate

(a) Manhattan plot representing the association between 573,801 SNPs and CES total editing rate, where black line represents threshold for the top 100 SNPs (p value ~ 10e-4). (b) Detailed view of genotyped SNPs located in the region at chromosome 7 that showed significant association with CES total editing rate. Known GWAS associations for human phenotypes from GRASP database are reported in the lower panel. (c) The top associated SNP (rs856554) showed a significant effect on global editing level and seems to influence also *ADAR* expression, despite this effect is not significant. (d) The top associated *ADAR* eQTL (rs6699825) showed significant effect on both. P values reported in box-plots are based on Tukey post-hoc pairwise test in ANOVA. *ADAR* gene

expression level is reported as residual of ridge regression with technical covariates (see description of data in methods section).

Additional materials

Additional file 1 (pdf). Supplementary tables and figures

Supplementary tables S1-S8. Supplementary Figures S1-S8

Additional file 2 (xls). Detailed statistics for the 2,079 editing sites considered in the study

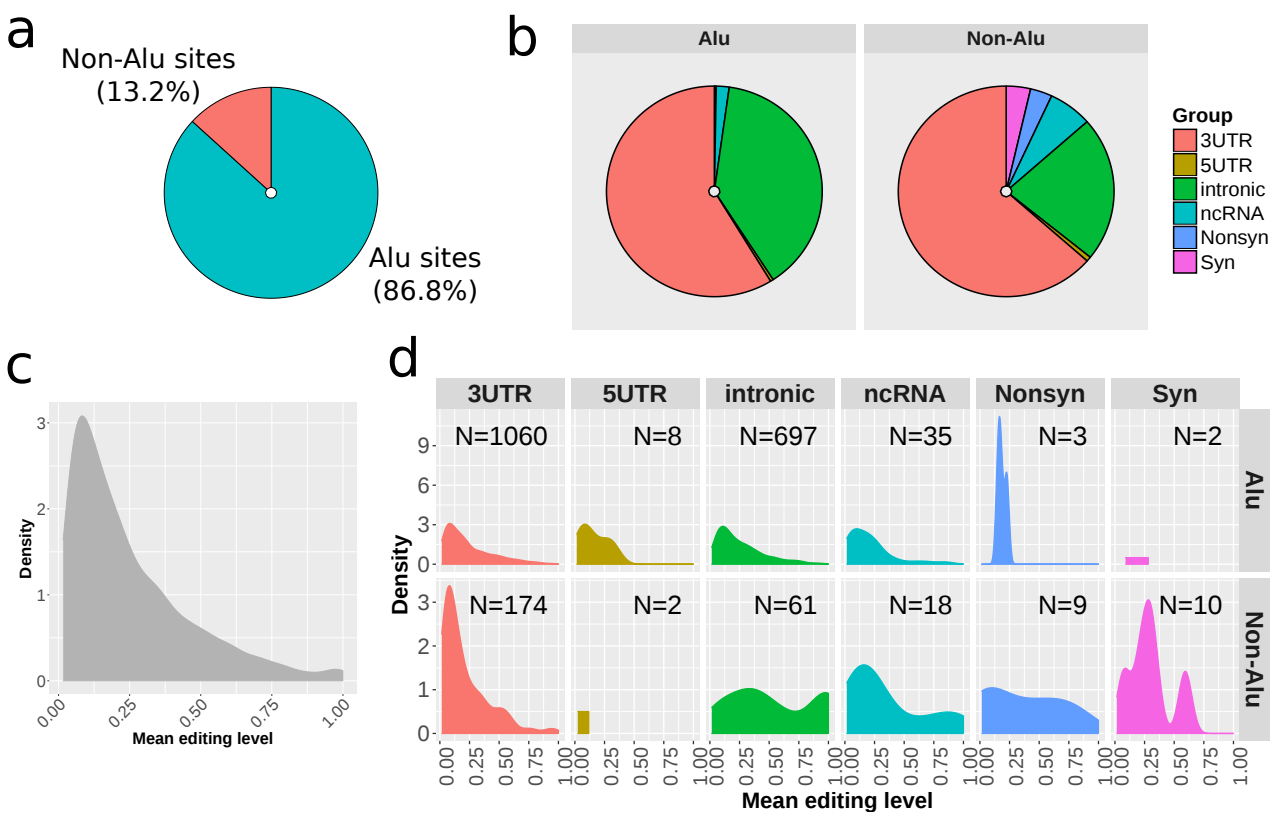
Additional file 3 (xls). Complete results of robust regression between CES total editing rate and gene expression levels where we removed the effect of ADAR expression.

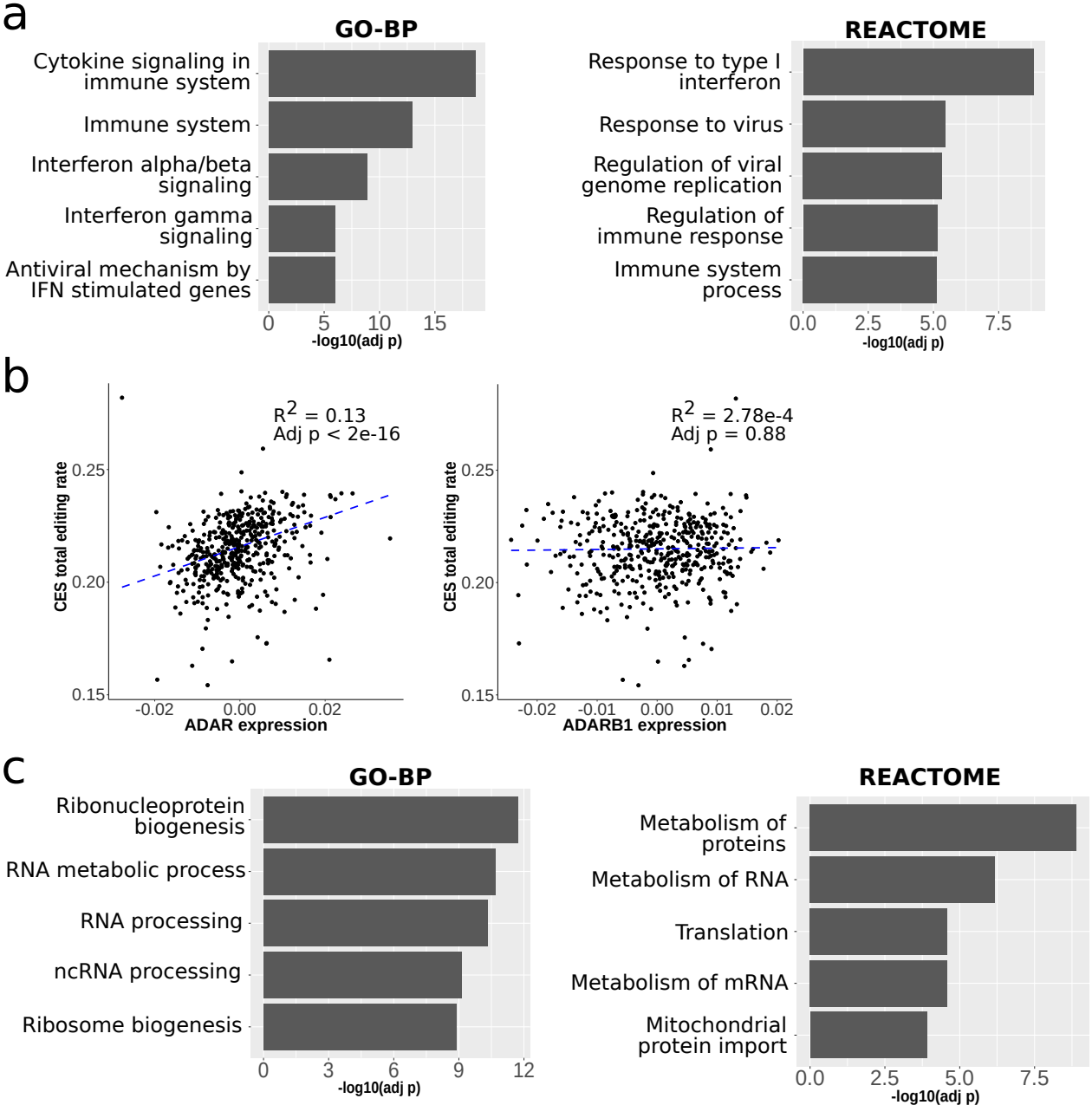
Additional file 4 (xls). Node properties in the protein-protein interaction network including genes associated to CES total editing rate (FDR<0.01) interacting with ADARs or one of their first neighbors

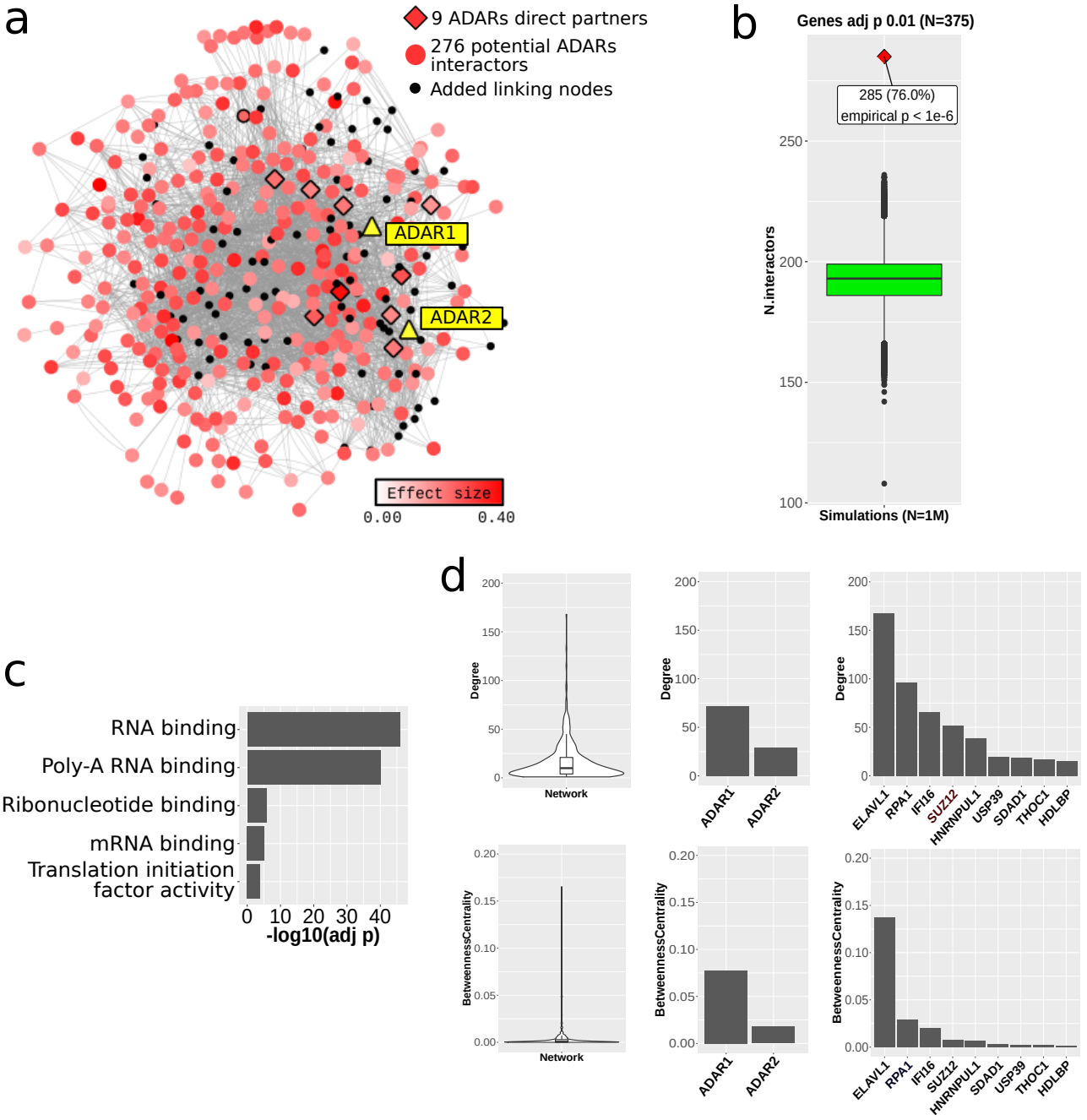
Additional file 5 (xls). Association of editing sites with principal components of editing.

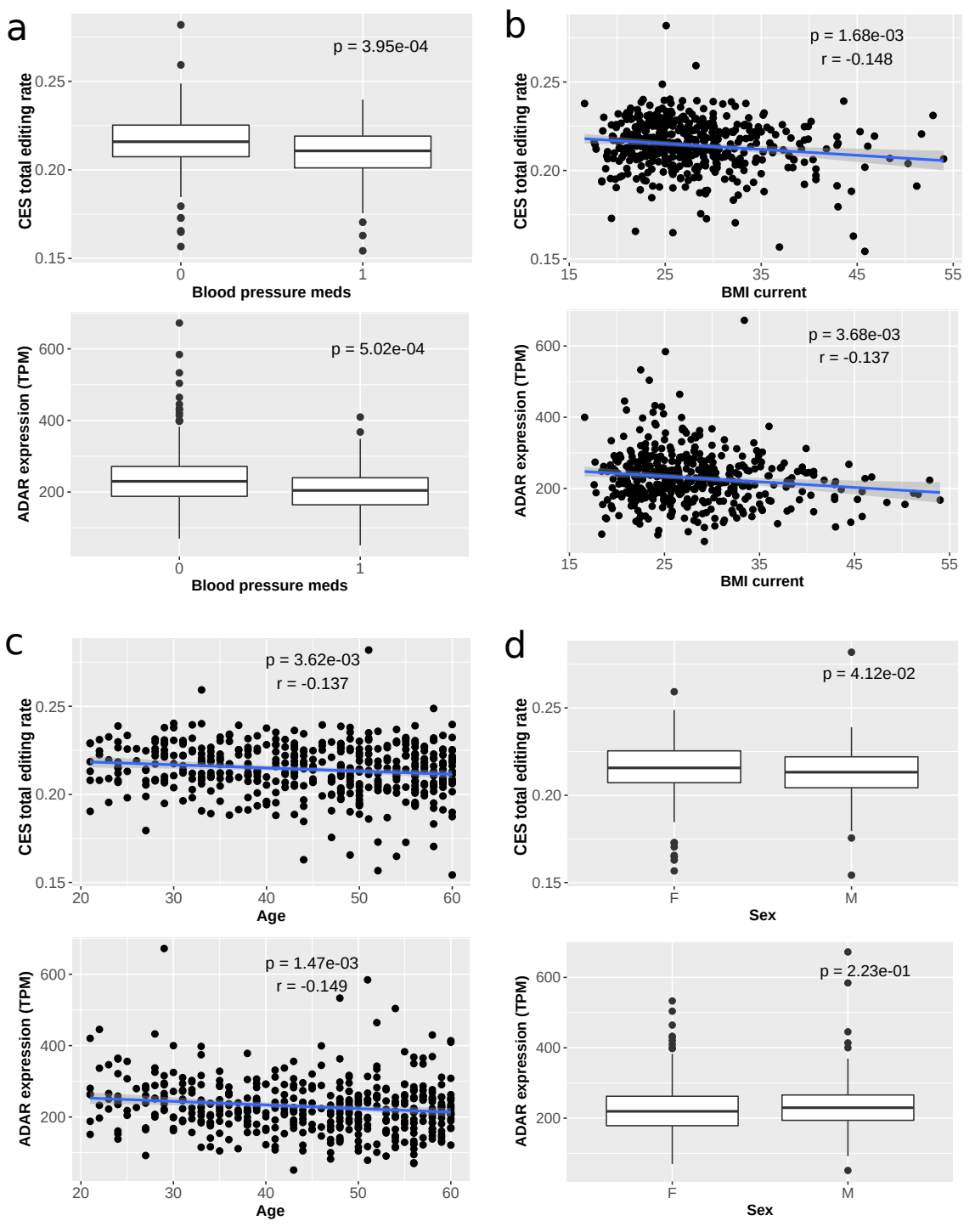
Additional file 6 (xls). Results of genome-wide association study for CES total editing rate

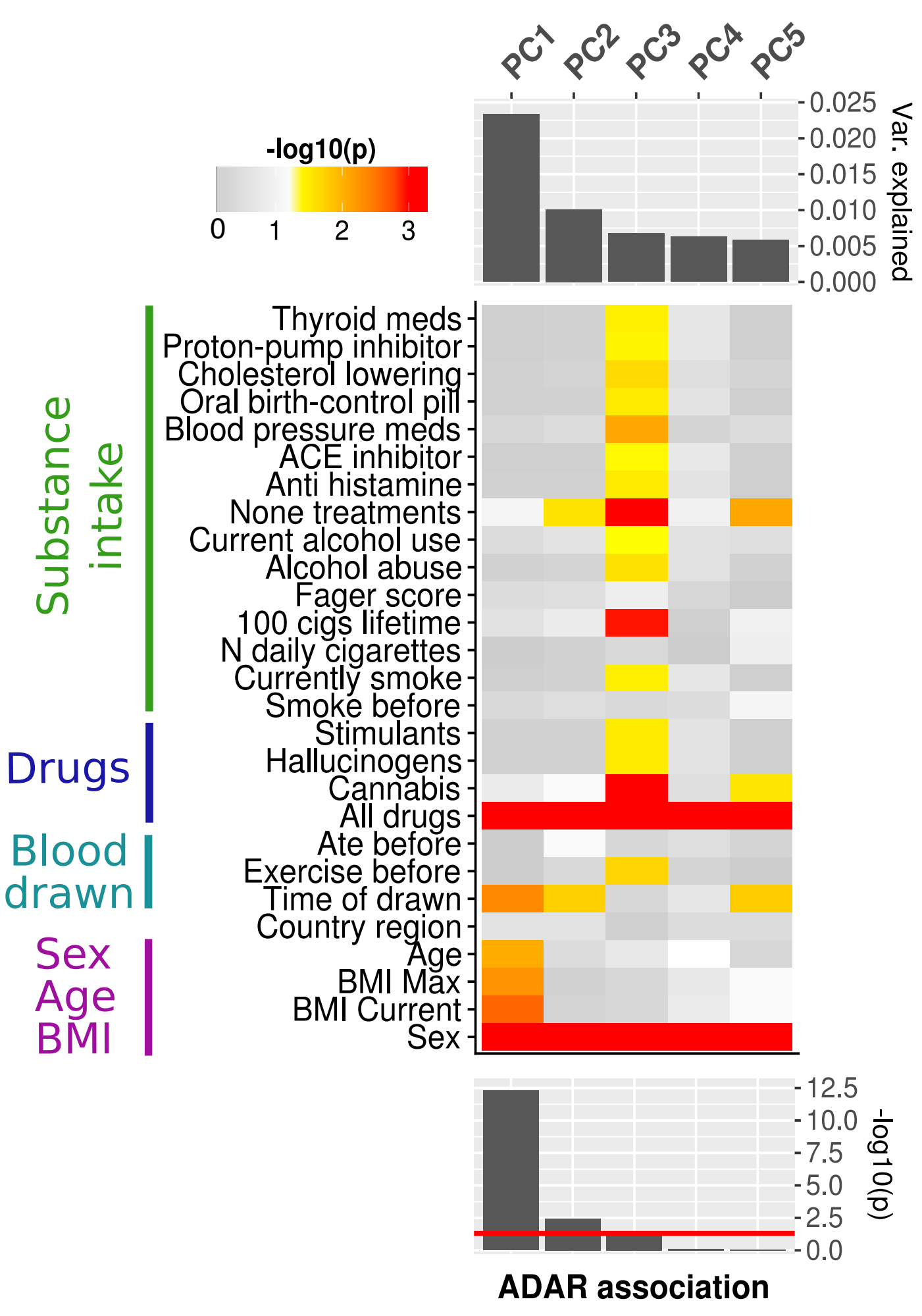
Association results for single SNPs and loci identified after variant clumping are reported.

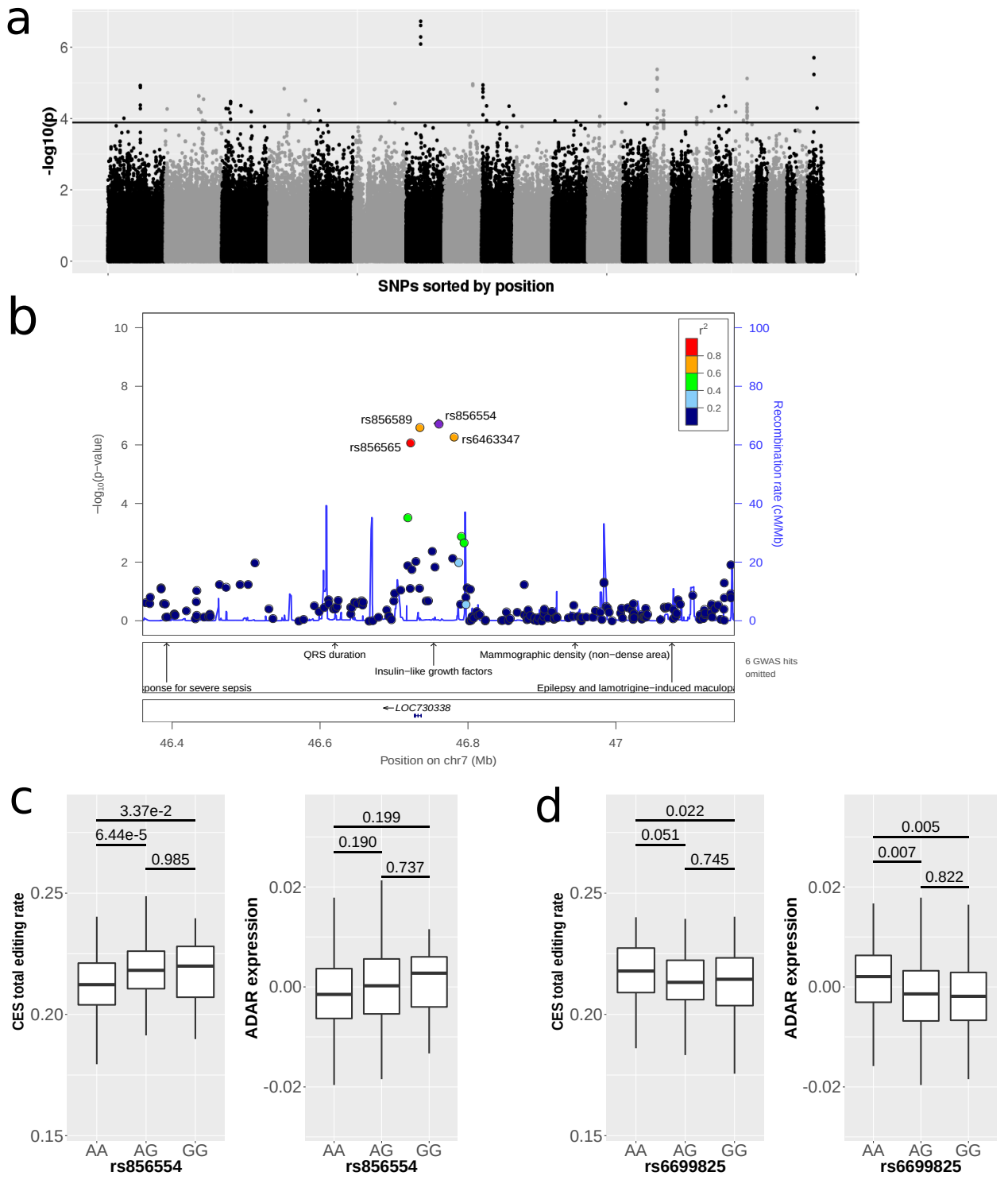

a
Non-Alu sites
(13.2%)
Alu sites
(86.8%)
b
Alu
Non-Alu
Group
3UTR
5UTR
intronic
ncRNA
Nonsyn
Syn
c
Density
Mean editing level
d
3UTR
5UTR
intronic
ncRNA
Nonsyn
Syn
N=1060
N=8
N=697
N=35
N=3
N=2
N=174
N=2
N=61
N=18
N=9
N=10
Alu
Non-Alu
Density
Mean editing level


a
GO-BP
Cytokine signaling in immune system
Immune system
Interferon alpha/beta signaling
Interferon gamma signaling
Antiviral mechanism by IFN stimulated genes
0
5
10
15
-log10(adj p)
REACTOME
Response to type I interferon
Response to virus
Regulation of viral genome replication
Regulation of immune response
Immune system process
0.0
2.5
5.0
7.5
-log10(adj p)
b
R2 = 0.13
Adj p < 2e-16
CES total editing rate
0.25
0.20
0.15
-0.02
0.00
0.02
ADAR expression
R2 = 2.78e-4
Adj p = 0.88
CES total editing rate
0.25
0.20
0.15
-0.02
-0.01
0.00
0.01
0.02
ADARB1 expression
c
GO-BP
Ribonucleoprotein biogenesis
RNA metabolic process
RNA processing
ncRNA processing
Ribosome biogenesis
0
3
6
9
12
-log10(adj p)
REACTOME
Metabolism of proteins
Metabolism of RNA
Translation
Metabolism of mRNA
Mitochondrial protein import
0.0
2.5
5.0
7.5
-log10(adj p)


a
9 ADARs direct partners
276 potential ADARs interactors
Added linking nodes
ADAR1
ADAR2
Effect size
0.00
0.40
b
Genes adj p 0.01 (N=375)
285 (76.0%)
empirical p < 1e-6
N.interactors
250
200
150
100
Simulations (N=1M)
c
RNA binding
Poly-A RNA binding
Ribonucleotide binding
mRNA binding
Translation initiation factor activity
0 10 20 30 40
-log10(adj p)
d
Degree
Network
ADAR1
ADAR2
ELAVL1
RPA1
IFI16
SUZ12
HNRNPUL1
USP39
SDAD1
THOC1
HDLBP
BetweennessCentrality


a
p = 3.95e-04
CES total editing rate
0.25
0.20
0.15
0
1
Blood pressure meds
p = 5.02e-04
ADAR expression (TPM)
600
400
200
0
1
Blood pressure meds
b
p = 1.68e-03
r = -0.148
CES total editing rate
0.25
0.20
0.15
15
25
35
45
55
BMI current
p = 3.68e-03
r = -0.137
ADAR expression (TPM)
600
400
200
15
25
35
45
55
BMI current
c
p = 3.62e-03
r = -0.137
CES total editing rate
0.25
0.20
0.15
20
30
40
50
60
Age
p = 1.47e-03
r = -0.149
ADAR expression (TPM)
600
400
200
20
30
40
50
60
Age
d
p = 4.12e-02
CES total editing rate
0.25
0.20
0.15
F
M
Sex
p = 2.23e-01
ADAR expression (TPM)
600
400
200
F
M
Sex


PC1
PC2
PC3
PC4
PC5
0.025
0.020
0.015
0.010
0.005
0.000
Var. explained
-log10(p)
0
1
2
3
Substance intake
Drugs
Blood drawn
Sex
Age
BMI
Thyroid meds
Proton-pump inhibitor
Cholesterol lowering
Oral birth-control pill
Blood pressure meds
ACE inhibitor
Anti histamine
None treatments
Current alcohol use
Alcohol abuse
Fager score
100 cigs lifetime
N daily cigarettes
Currently smoke
Smoke before
Stimulants
Hallucinogens
Cannabis
All drugs
Ate before
Exercise before
Time of drawn
Country region
Age
BMI Max
BMI Current
Sex
12.5
10.0
7.5
5.0
2.5
0.0
-log10(p)
ADAR association


a
-log10(p)
SNPs sorted by position
b
rs856589
rs856554
rs6463347
rs856565
r²
0.8
0.6
0.4
0.2
−log10(p−value)
Recombination rate (cM/Mb)
QRS duration
Insulin−like growth factors
Mammographic density (non−dense area)
Epilepsy and lamotrigine−induced maculop
6 GWAS hits omitted
LOC730338
Position on chr7 (Mb)
c
3.37e-2
6.44e-5
0.985
CES total editing rate
AA
AG
GG
rs856554
0.199
0.190
0.737
ADAR expression
rs856554
d
0.022
0.051
0.745
rs6699825
0.005
0.007
0.822
rs6699825