SUPPLEMENTAL FIGURES

Global landscape and genetic regulation of RNA editing in cortical samples from individuals with schizophrenia

Michael S. Breen[1,2,3], Amanda Dobbyn[2,4,5], Qin Li[6], Panos Roussos[1,2,7,8,9], Gabriel E. Hoffman[2,4,7], Eli Stahl[1,2,4,7,10], Andrew Chess[4,9,11], Pamela Sklar[4,9], Jin Billy Li[6], Bernie Devlin[12], Joseph D. Buxbaum[1,2,3,9,13] for the CommondMind Consortium (CMC)[14]

[1]Department of Psychiatry, [2]Department of Genetics and Genomic Sciences, [3]Seaver Autism Center for Research and Treatment, [4]Icahn Institute of Genomics and Multiscale Biology, [5]The Charles Bronfman Institute for Personalized Medicine, Icahn School of Medicine at Mount Sinai, New York, New York, 10029 USA; [6]Department of Genetics, Stanford University School of Medicine, Stanford, California, 94305, USA; [7]Pamela Sklar Division of Psychiatric Genomics, Icahn School of Medicine at Mount Sinai, New York, New York, 10029 USA; [8]Mental Illness Research, Education, and Clinical Center (VISN 2 South), James J. Peters VA Medical Center, Bronx, New York, 10468, USA; [9]Friedman Brain Institute, Icahn School of Medicine at Mount Sinai, New York, New York, 10029 USA; [10]Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, Massachusetts 02142 USA; [11]Department of Developmental and Regenerative Biology, Icahn School of Medicine at Mount Sinai, New York, New York, 10029 USA; [12]Department of Psychiatry, University of Pittsburgh School of Medicine, 3811 O'Hara Street, Pittsburgh, Pennsylvania 15213, USA; [13]Mindich Child Health and Development Institute, Icahn School of Medicine at Mount Sinai, New York, New York, 10029 USA; [14]A full list of authors can be found at the end of the article.

Correspondence to: michael.breen@mssm.edu and joseph.buxbaum@mssm.edu

Supplemental Figures 1-18, in brief:

Figure S1

CommonMind Consortium (CMC) Discovery cohort

CMC DLPFC (*N* tot=540)	Control (N=286)	SCZ (N=254)	*P*-value
Site: *n* (MSSM/Penn/Pitt)	166/38/82	141/58/55	NA
Gender, female: *n* (%)	124 (43%)	94 (33%)	NA
Ethnicity, Caucasian: *n* (%)	215 (75%)	211 (74%)	NA
Age (years):	65.86 ± 19.96	68.55 ± 16.97	NA
PMI (hours):	13.63 ± 8.16	20.5 ± 13.36	3.4e-08
RIN:	7.81 ± 0.87	7.40 ± 0.87	5.8e-06
Brain weight (*g*):	1244.99 ± 193.06	1231.41 ± 180.55	NA
pH:	6.57 ± 0.27	6.49 ± 0.27	0.02

CMC ACC (*N* tot=470)	Control (*N*=245)	SCZ (*N*=225)	*P*-value
Site: *n* (MSSM/Penn/Pitt)	143/24/78	125/41/59	NA
Gender, female: *n* (%)	100 (40%)	81 (36%)	NA
Ethnicity, Caucasian: *n* (%)	184 (75%)	184 (81%)	NA
Age (years):	64.94 ± 20.42	67.61 ± 17.24	NA
PMI (hours):	13.59 ± 7.87	19.85 ± 12.50	1.5e-06
RIN:	7.58 ± 0.85	7.34 ± 0.76	0.001
Brain weight (*g*):	1246.09 ± 193.72	1241.07 ± 174.17	NA
pH:	6.59 ± 0.26	6.51 ± 0.26	0.03

NIMH Human Brain Collection Core (HBCC) Replication cohort

HBCC DLPFC (*N* tot=317)	Control (*N*=217)	SCZ (*N*=100)	*P*-value
Gender, female: *n* (%)	60 (27%)	35 (35%)	NA
Ethnicity, Caucasian: *n* (%)	90 (41%)	38 (38%)	NA
Age (years):	30.05 ± 19.99	49.92 ± 13.34	9.2e0-7
PMI (hours):	29.04 ± 13.34	36.04 ± 22.00	0.005
RIN:	7.66 ± 0.89	7.32 ± 0.89	0.002
Brain weight (*g*):	1351.31± 219.53	1348.43 ± 162.12	NA
pH:	6.49± 0.29	6.37 ± 0.22	0.007

Figure S1. Summary of patient statistics. Patient level covariates were recorded and compared between SCZ cases and control samples, separately for each brain region (ACC and DLPFC) and cohort (discovery and validation). In our discovery cohort, a total of 358 unique schizophrenia cases and 380 unique controls were sampled in at least one brain region. A Shapiro-Wilk test was used to assess normality of covariates. If the resulting variable was normally distributed, a two-sided Student's *t*-test was applied, alternatively a two-sided Wilcoxon rank sum test with continuity correction was implemented.

Figure S2

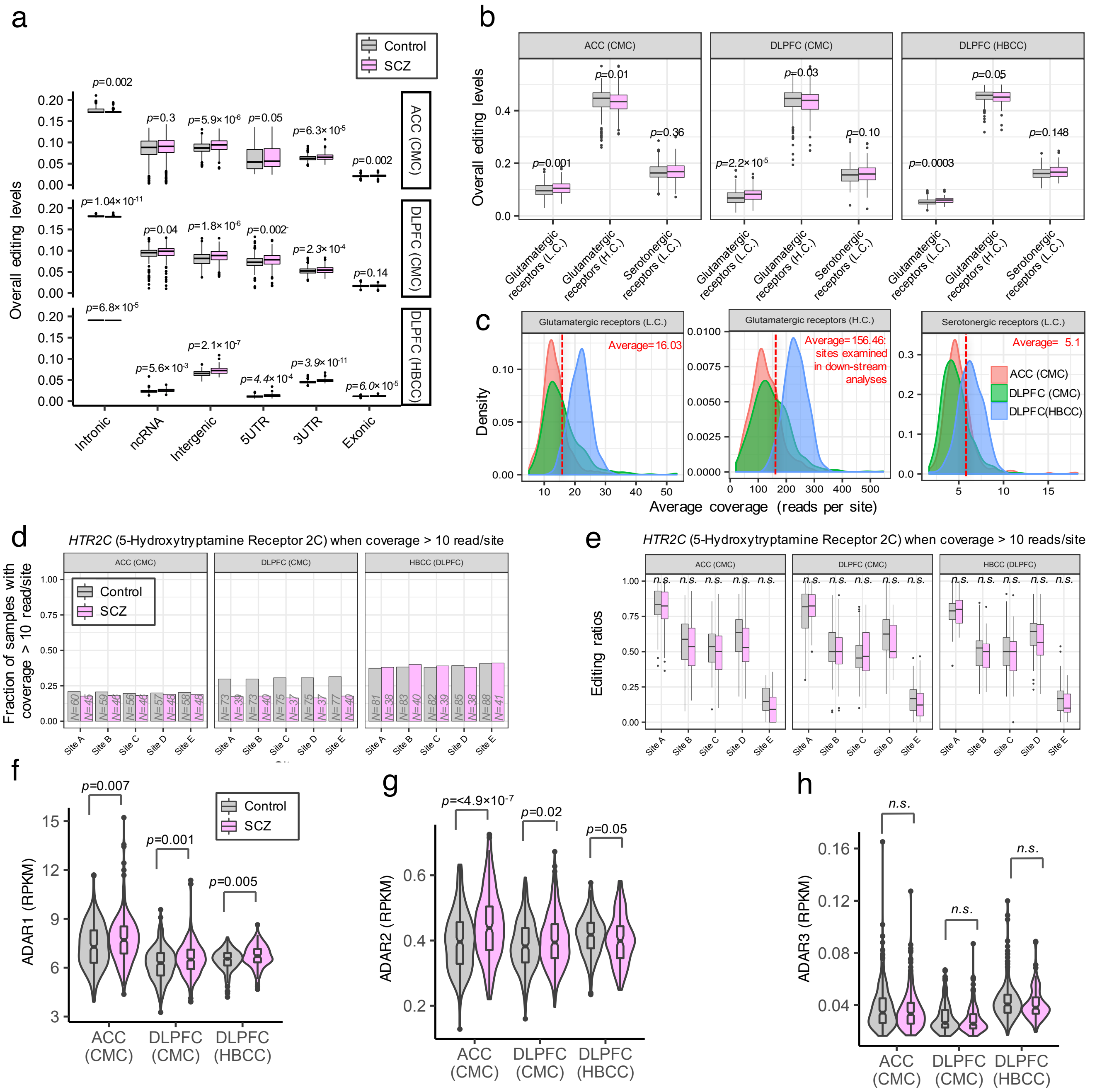

Figure S2. Overall RNA editing and ADAR expression. Unless otherwise specified, all comparisons made here include CMC ACC (*$n^{control}$=245, n^{SCZ}=225*) and DLPFC (*$n^{control}$=286, n^{SCZ}=254*) as well as the NIMG HBCC DLPFC (*$n^{control}$=217, n^{SCZ}=100*) samples. **(a)** Overall RNA editing levels are computed separately for each discrete genic region. **(b)** RNA editing levels were examined based on *a priori* defined glutamatergic and serotonergic receptor activity gene sets (GO:009589 and GO:0008066, respectively). For this analysis, RNA editing sites were parsed into two groups: 1) sites which were detected across all >80% of all samples with sufficient base coverage (>20) and examined in down-stream analyses, here termed high-confidence (H.C.) sites; and 2) all remaining sites here termed low-confidence (L.C.) sites. Note that serotonergic receptor activity genes were lowly expressed in these data sets and as a result could not be parsed into a separate H.C. group. **(c)** Average read coverage/site for each gene indicates high coverage for H.C. glutamatergic sites and low coverage for low-confidence glutamatergic sites;

averages are depicted with dashed red line. (**d**) A smaller fraction of samples with coverage two-fold higher than average (> 10 reads/site) for HTR2C A-E sites (sample numbers labeled in figure), **(e)** which were examined for differences in editing ratios between SCZ cases and controls. Reads per kilobase of transcript per million mapped reads (RPKM) expression levels for (**f**) *ADAR1*, (**g**) *ADAR2* and (**h**) *ADAR3*. Note that *ADAR2* expression for HBCC samples is trending opposite to that in CMC samples. For all comparisons, a two-sided Wilcoxon rank sum test with continuity correction was used to test significance between groups. Whisker box plots and violin plots used throughout this figure show median, lower and upper quartiles, and whiskers represent minimum and maximum of the data.

Figure S3

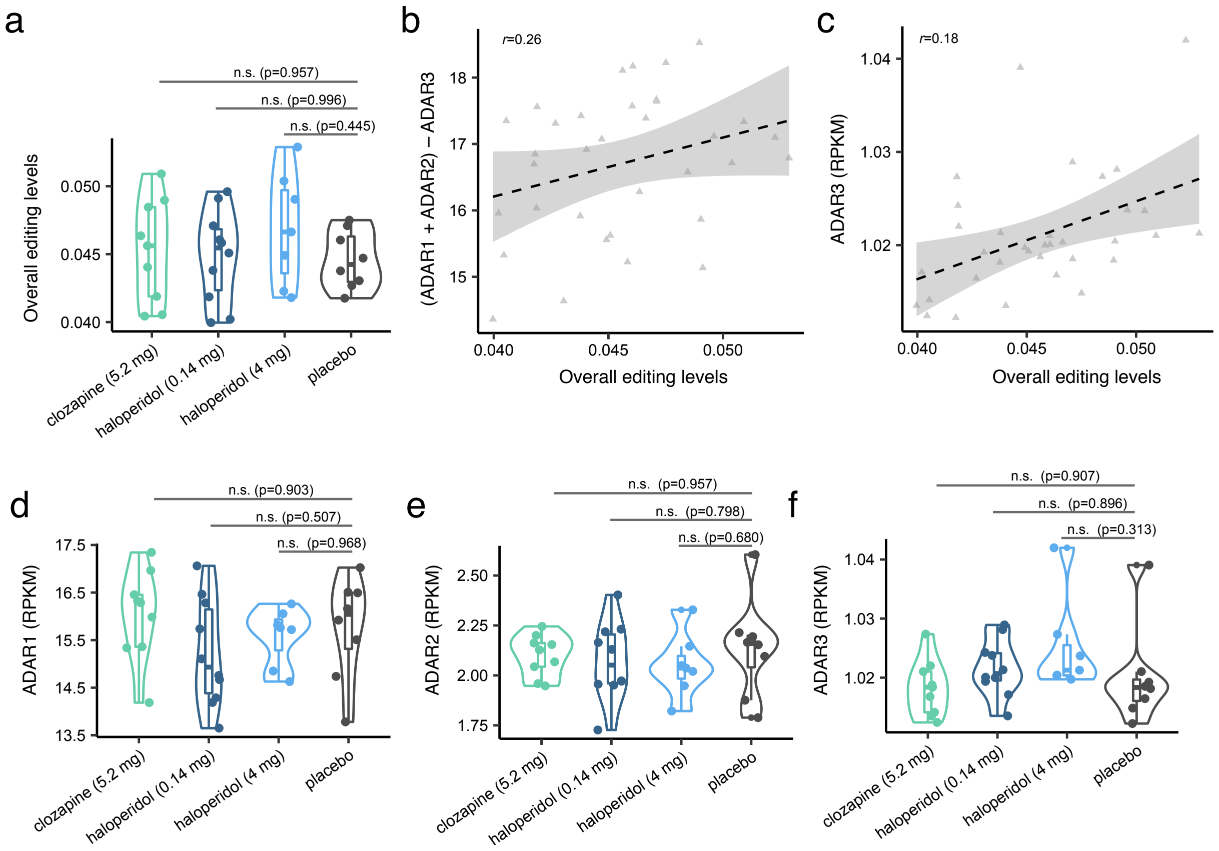

Figure S3. Overall RNA editing and ADAR expression in macaque samples. (a) Overall RNA editing levels are computed separately for DLPFC samples treated with different antipsychotic medications and dosages. Variance of overall RNA editing levels explained by (**b**) *ADAR1* and *ADAR2* and (**c**) *ADAR3* Reads per kilobase of transcript per million mapped reads (RPKM) expression levels. R^2 values were calculated by robust linear regressions on overall editing levels and logarithmic transformed RPKM values. RPKM expression levels for (**d**) *ADAR1*, (**e**) *ADAR2* and (**f**) *ADAR3*. For all comparisons, a Dunnett's multiple comparison of means test was used comparing each treatment group relative to placebo. A total of 34 rhesus macaque DLPFC samples were analyzed in all comparisons ($n^{\text{clozapine 5.2mg}}$=9, $n^{\text{haloperidol 0.14mg}}$=10, $n^{\text{haloperidol 4mg}}$=7, n^{placebo}=8). Whisker violin plots used throughout this figure show median, lower and upper quartiles, and whiskers represent minimum and maximum of the data.

Figure S4

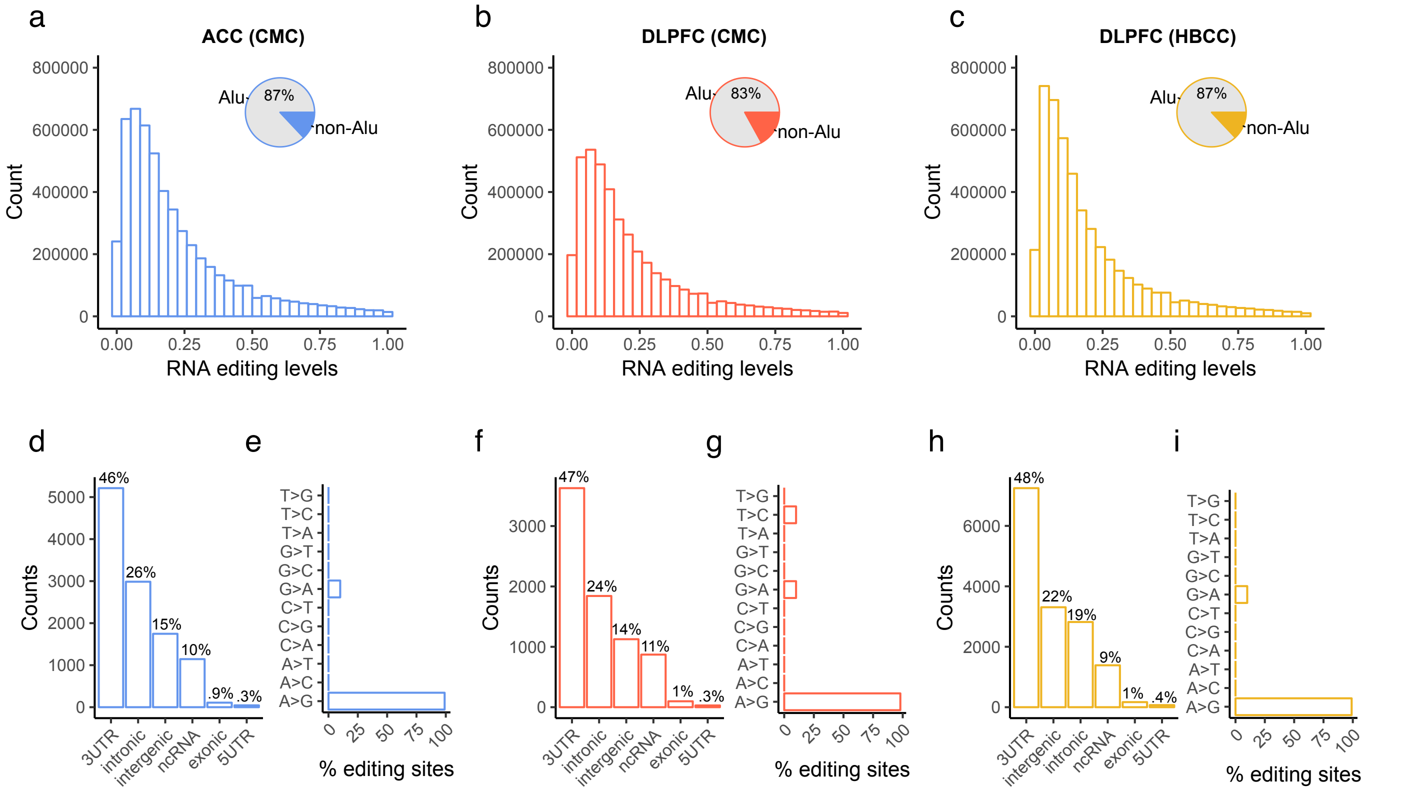

Figure S4. Characterization of detected RNA editing events. Frequency distributions of RNA editing levels for all detected sites in the (**a**) ACC and (**b**) DLPFC CMC samples and (**c**) DLPFC HBCC samples. Inset pie charts indicate the total fraction of all detected sites that map to Alu repeat elements. The total number of RNA editing events were summarized within each genic region and nucleotide conversion rates were assessed for the (**d-e**) ACC and (**f-g**) DLPFC CMC samples and (**h-i**) DLPFC HBCC samples.

Figure S5

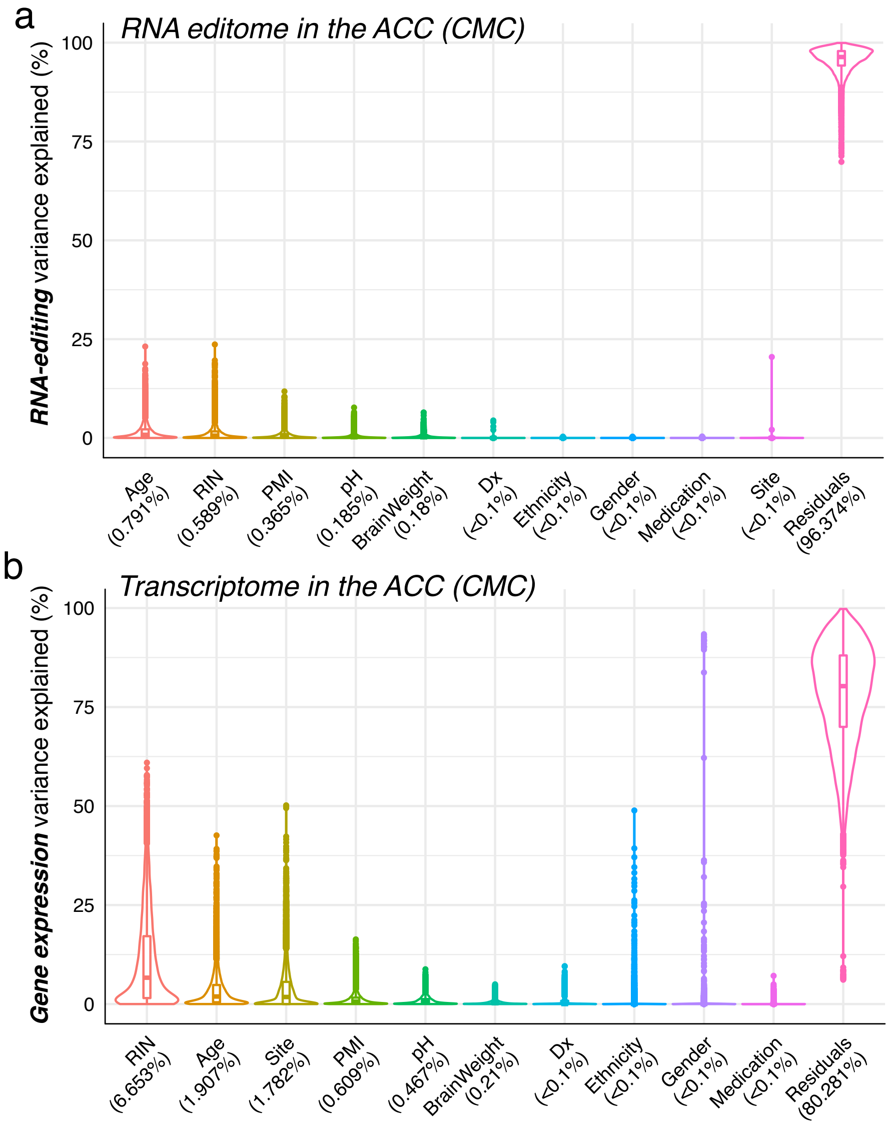

Figure S5. Computing variance explained. Variance explained according to nine covariates, which represent potential technical, biological and clinical sources of variability. The linear mixed model framework of the R package variancePartition was used to quantify variability explained in the (**a**) RNA editome and (**b**) transcriptome within all ACC samples (*$n^{control}$=245, n^{SCZ}=225*). The dynamic range of transcriptome data finds stronger relationships and is under greater influence with the recorded covariates than RNA editing measurements. Whisker violin plots used throughout this figure show median, lower and upper quartiles, and whiskers represent minimum and maximum of the data. Results from the CMC DLPFC and NIMH HBCC DLPFC samples were highly similar to those depicted here for the ACC and are not shown.

Figure S6

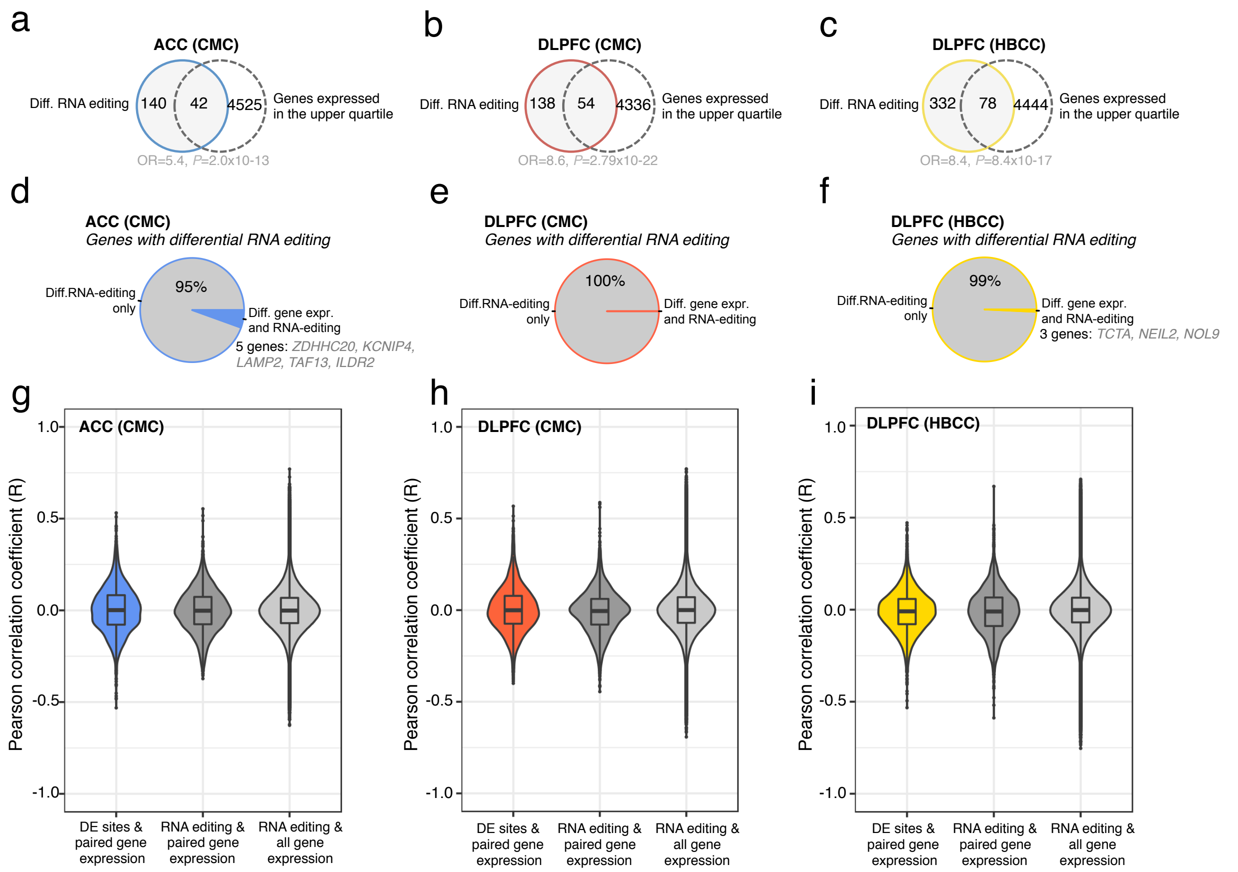

Figure S6. Relationship between RNA editing and gene expression. Enrichment analysis assessing whether genes harboring differentially edited sites overlap with genes that are highly expressed in the CMC (**a**) ACC (*$n^{control}$=245, n^{SCZ}=225*) and (**b**) DLPFC samples (*$n^{control}$=286, n^{SCZ}=254*) and (**c**) NIMH HBCC DLPFC samples *($n^{control}$=217, n^{SCZ}=100)*. Genes were labeled 'highly expressed' if they had an average gene expression value across all samples that was in the upper 3rd quartile. A one-sided Fisher's exact test was used to compute overlap significance and estimated odds-ratios. Overlap analysis also assessed whether genes with differential RNA editing sites displayed differential expression in the CMC (**d**) ACC and (**e**) DLPFC samples and (**f**) NIMH HBCC DLPFC samples. Inset gene symbols indicate genes that harbor differential editing sites and are dysregulated in SCZ. Correlation analysis of RNA editing levels were compared to gene expression levels in three instances: 1) for differentially edited sites relative to their respective gene expression levels; 2) all RNA editing sites relative to their respective gene expression levels; 3) all RNA editing sites relative to gene expression levels other than their respective gene. This analysis was carried out for the CMC (**g**) ACC and (**h**) DLPFC samples and (**i**) the NIMH HBCC DLPFC samples (two-sided Wilcoxon rank sum test with continuity correction). Whisker violin plots used throughout this figure show median, lower and upper quartiles, and whiskers represent minimum and maximum of the data.

Figure S7

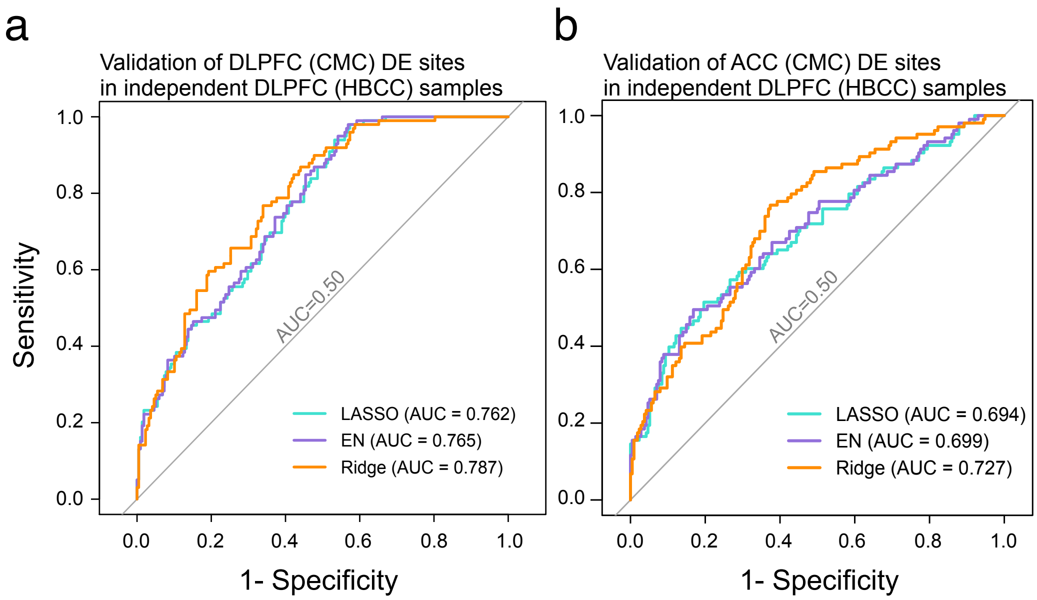

Figure S7. Multivariate supervised classification. Three regularized regression techniques, including ElasticNet (EN), Lasso and Ridge Regression were fit using the glmnet R package in order to assess cross-validation of the schizophrenia (SCZ)-related sites derived from CMC samples in withheld HBCC samples. Two prediction models were built using the differentially edited sites in the (**a**) DLPFC and (**b**) ACC derived from the CMC training sets to predict case/control status from withheld DLPFC data derived from the HBCC test set. Area under the receiver operator curve (AUC) values are used to assess the overall precision of these models. Ridge Regression achieved 78% and 72% prediction accuracy when using altered RNA editing and samples derived from DLPFC and ACC training data, respectively, to predict DLPFC HBCC test set samples.

Figure S8

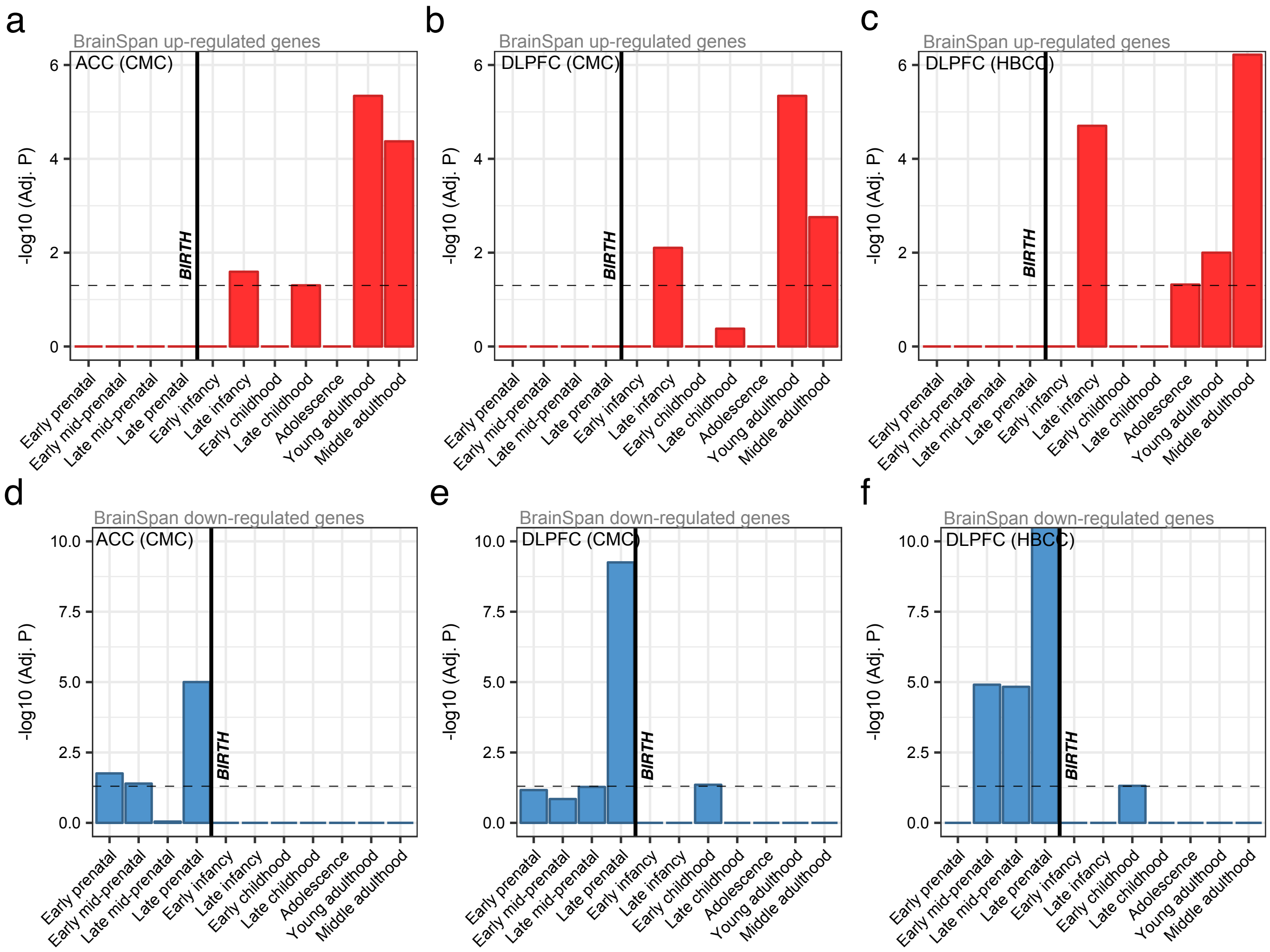

Figure S8. BrainSpan developmental gene expression profiles. Enrichment analysis examined whether differentially edited sites in SCZ mapped to genes with specific developmental trajectories using RNA-seq data from the BrainSpan Project. A total of 11 developmental stages (x-axis) were analyzed and gene sets, indicating whether genes are over-expressed or under-expressed at each stage relative to all other stages were used to compute overlap and enrichment (one-sided Fisher's exact test). A consistent enrichment of differentially edited sites mapping to genes, which are highly expressed during young and middle adulthood was observed for the (**a**) ACC and (**b**) DLPFC CMC samples and (**c**) DLPFC HBCC samples. Additionally, we observed that these genes were also predominately under-expressed during the fetal period for the (**d**) ACC and (**e**) DLPFC CMC samples and (**f**) DLPFC HBCC samples, indicating that the developmental expression properties of these genes gradually increase in expression from early prenatal periods and peak during adulthood. All comparisons made here were computed using differential editing results derived from CMC ACC ($n^{control}$=245, n^{SCZ}=225) and DLPFC ($n^{control}$=286, n^{SCZ}=254) and HBCC DLPFC ($n^{control}$=217, n^{SCZ}=100) samples.

Figure S9

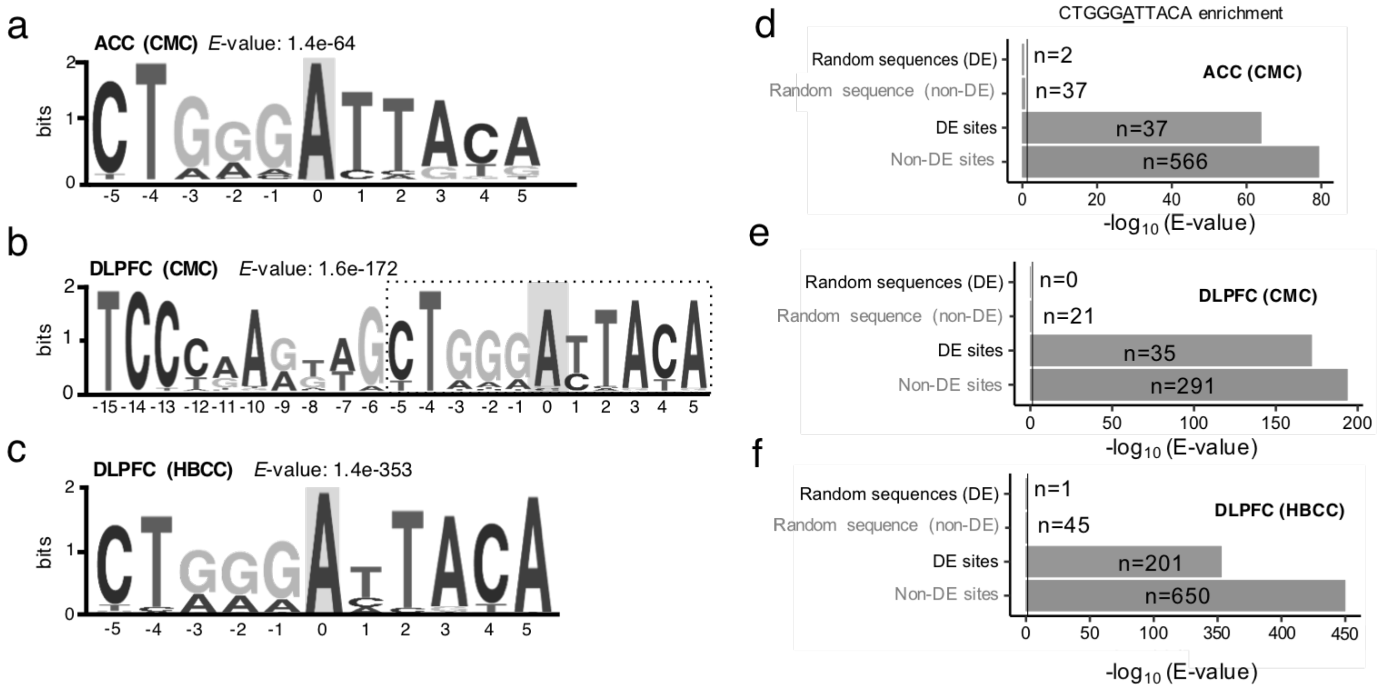

Figure S9. Motif enrichment analysis. All RNA editing sites were assessed for motif enrichment analysis ± 20bp from the editing site using MEME. Consistent and strong enrichment was observed for 10bp motif (± 5bp from the editing site) for differentially edited (DE) sites derived from the (**a**) ACC and (**b**) DLPFC CMC samples and (**c**) DLPFC HBCC samples. The overall height of each stack indicates the sequence conservation at that position (measured in bits), whereas the height of symbols within the stack reflects the relative frequency of the corresponding nucleic acid at that position. Subsequently, we tested whether non-differentially edited (non-DE) sites and randomly sampled sites with flanking sequences matched for GC contained enrichment for this same motif. Two sets of randomly sampled sites were chosen to match the exact number of DE and non-DE sites for each brain region. A general enrichment for DE and non-DE RNA editing sites were identified, for which no enrichment was identified for randomly selected sites in the (**d**) ACC and (**e**) DLPFC CMC samples and (**f**) DLPFC HBCC samples. MEME reports an E-value for each motif it finds, which is an estimate of the number of (equally or more interesting) motifs one would expect to find by chance if the letters in the input sequences were shuffled. Motifs with small E-values are very unlikely to be random sequence artifacts.

Figure S10

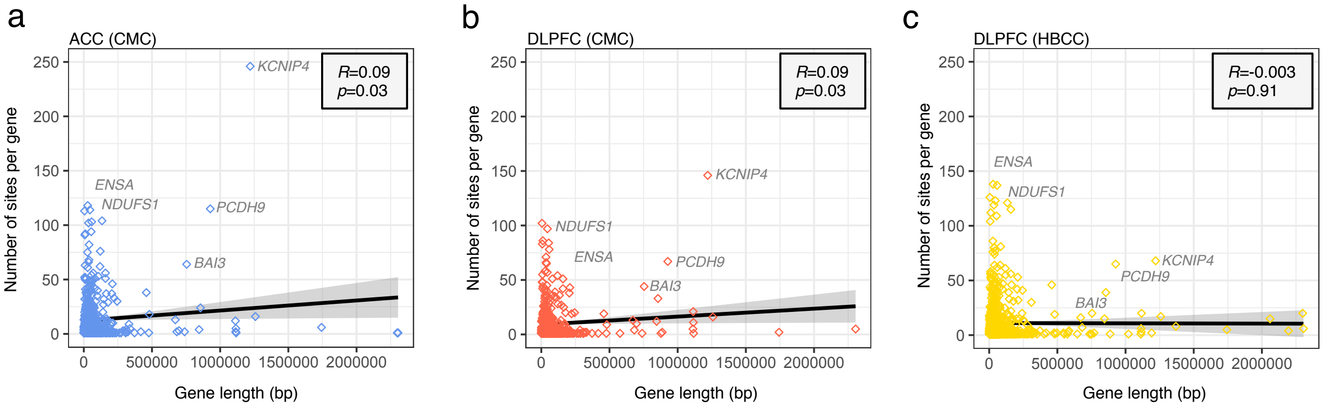

Figure S10. Gene length versus RNA editing sites. The total number of detected RNA editing events correlates with gene length for sites identified in the (**a**) ACC ($n^{control}$=245, n^{SCZ}=225) and (**b**) DLPFC CMC samples ($n^{control}$=286, n^{SCZ}=254) and (**c**) DLPFC HBCC samples ($n^{control}$=217, n^{SCZ}=100). Outlier genes are labeled in grey. R^2 values were calculated by robust linear regressions on overall editing levels and logarithmic transformed RPKM values.

Figure S11

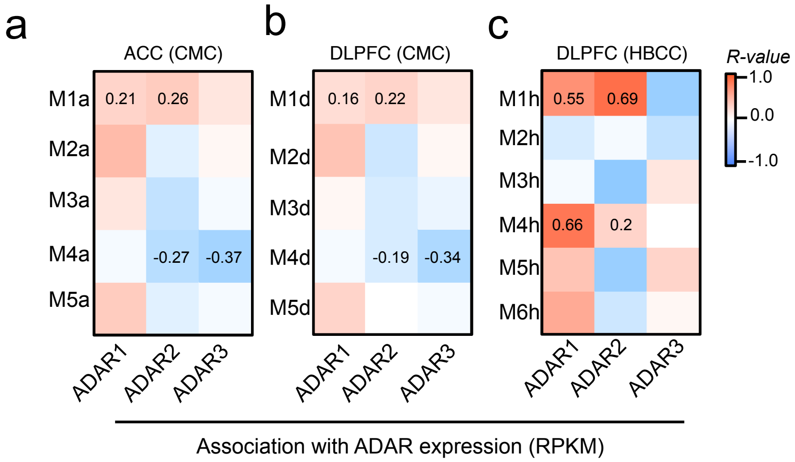

Figure S11. Module eigengene correlations with ADAR expression. Pearson correlation coefficients between module eigengene values and ADAR expression values for co-editing networks identified in the the (**a**) ACC and (**b**) DLPFC CMC samples and (**c**) DLPFC HBCC samples. Expression is quantified as the number of RNA-seq reads per kilobase of transcript per million mapped reads (RPKM). Correlation coefficients for modules of interest are presented in each corresponding cell. *R*-values were calculated using a Student's asymptotic *p*-value on overall editing levels and logarithmic transformed RPKM values for CMC ACC ($n^{control}$=245, n^{SCZ}=225) and DLPFC samples ($n^{control}$=286, n^{SCZ}=254) as well as NIMH HBCC DLPFC samples ($n^{control}$=217, n^{SCZ}=100).

Figure S12

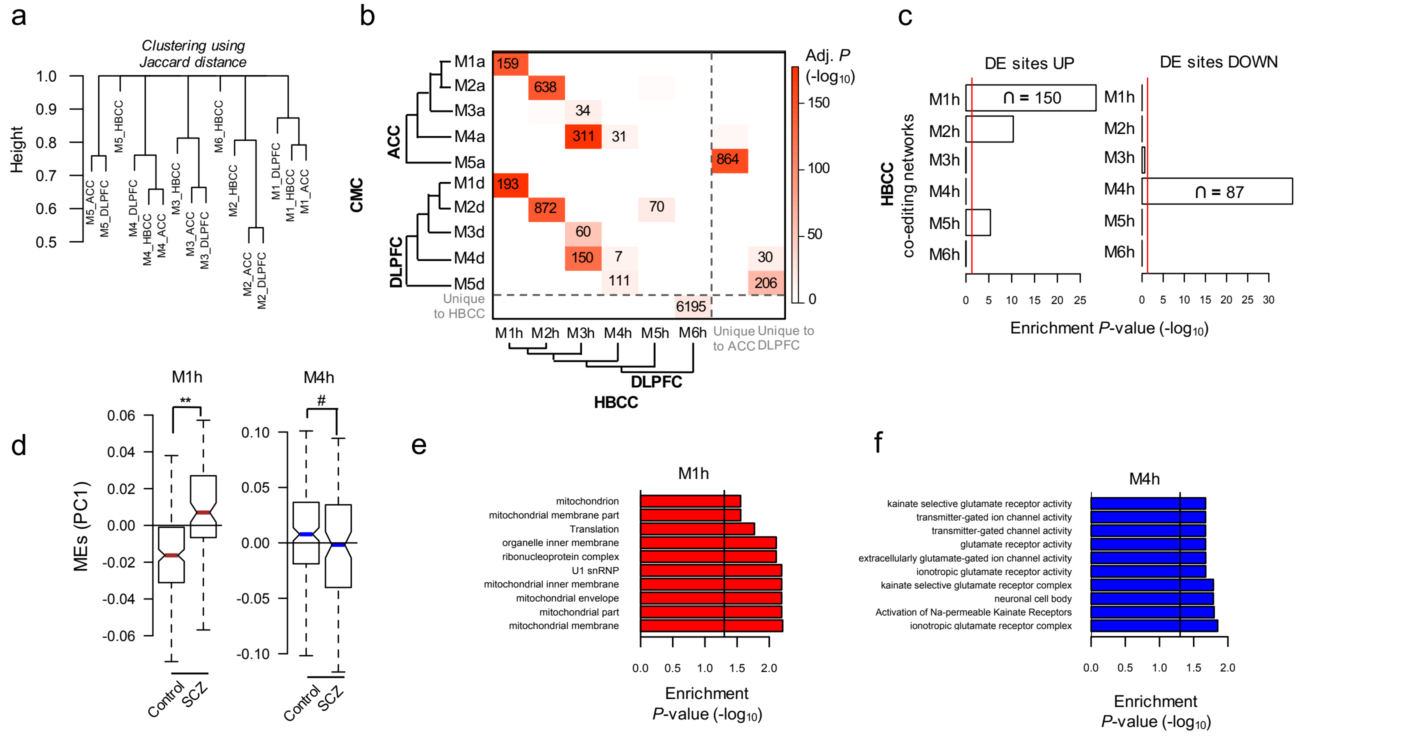

Figure S12. Validation of co-editing network analysis. Unsupervised co-editing network analysis was applied to NIMH HBCC DLPFC samples ($n^{control}$=217, n^{SCZ}=100). **(a)** Each site within a module (derived from discovery and validation samples) was converted into a binary matrix of site presence/absence calls and distance-based clustering with pairwise similarity was measured via Jaccard coefficient to confirm relationships between modules across brain regions/cohorts. **(b)** Overlap analysis of co-editing modules derived from these validation samples were compared to those previously identified within the ACC and DLPFC discovery. Unsupervised clustering was used to group modules by module eigengene (ME) values using Pearson's correlation coefficient and Ward's distance method. **(c)** Enrichment analysis of differentially edited sites within co-editing networks in the NIMH HBCC DLPFC samples ($n^{control}$=217, n^{SCZ}=100). Significance of overlap (for panel B and C) was computed using a one-sided Fisher's exact test and Bonferroni correction to adjust for multiple comparisons. **(d)** Differential ME analysis for modules M1h (over-edited, *p*=0.03) and M4h (under-edited, *p*=0.06) ($n^{control}$=217, n^{SCZ}=100) was conducted using a linear model and covarying for age, RIN, PMI, sample site and gender. Whisker box plots used throughout this figure show median, lower and upper quartiles, and whiskers represent minimum and maximum of the data. The top functional enrichment terms for the (**e**) over-edited module M1h and (**f**) under-edited module M4h using a one-sided hypergeometric test.

Figure S13

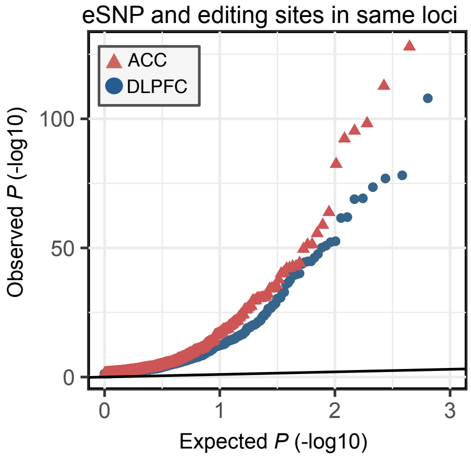

Figure S13. Quantile-Quantile plot. Quantile-quantile plot for association testing *P* values between RNA editing sites and genetic variants in the same gene as each editing site. Results are shown for the ACC (in red, $n^{control}$=180, n^{SCZ}=180) and DLPFC (in blue, $n^{control}$=210, n^{SCZ}=211). *P*-values were computed using a linear regression and FDR correction from the R package matrixEQTL, covarying for site, gender, age, PMI, DX and RIN.

Figure S14

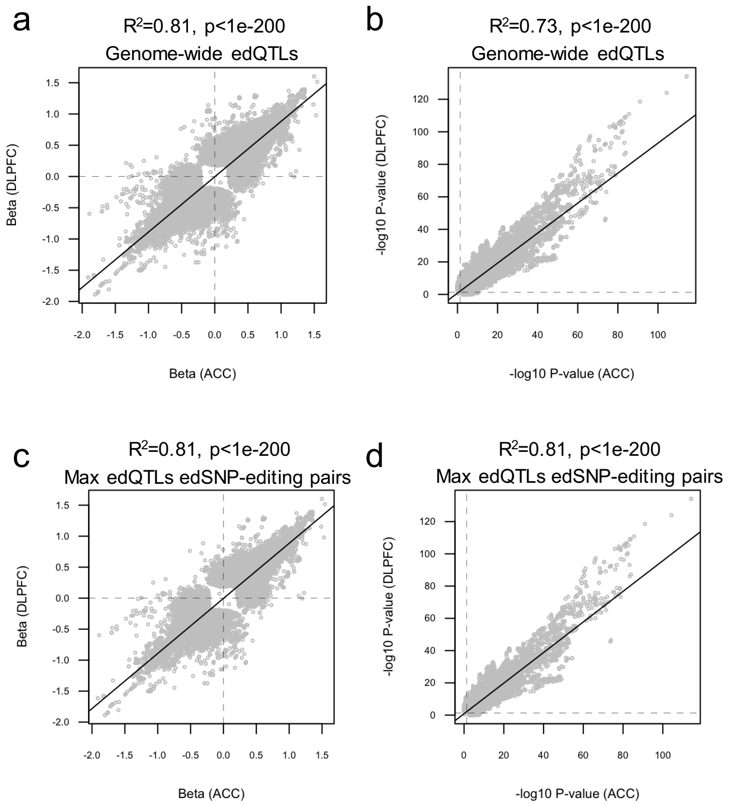

Figure S14. Concordance of edQTLs between brain regions. Genome-wide concordance of **(a)** effect sizes (beta-values) and **(b)** corresponding -$\log_{10}$ *p*-values for all edQTLs were compared for the ACC (x-axis; $n^{control}$=180, n^{SCZ}=180) and the DLPFC (y-axis; $n^{control}$=210, n^{SCZ}=211). Subsequently, a subset of the max edQTLs edSNP-editing pairs were similarly evaluated for concordance of **(c)** effect sizes (beta-values) and **(d)** corresponding -log10 p-values across these same samples. Beta-values and corresponding *p*-values were computed using a linear regression and FDR correction from the R package matrixEQTL, covarying for site, gender, age, PMI, DX and RIN. R^2 values were calculated by robust linear regressions.

Figure S15

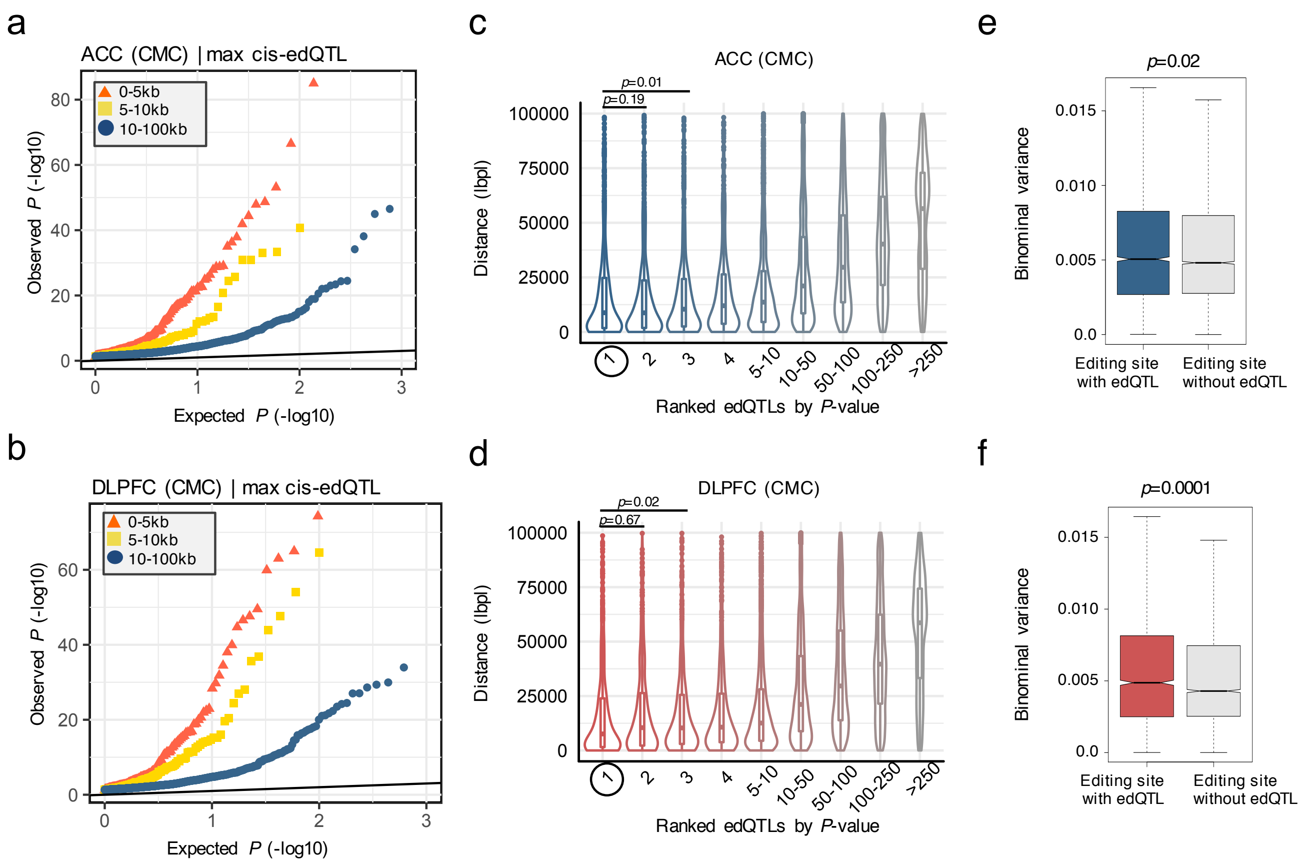

Figure S15. Distance plots for edQTL analysis. Quantile-quantile plot for edQTL associations. Corresponding p-values were computed using linear regression and FDR correction from the *R* package matrixEQTL, covarying for site, gender, age, PMI, DX and RIN. Editing sites fell within 0-5 kb (orange), between 5 kb-10 kb (gold), and between 10 kb-100 kb (blue) from the original best-associated editing site for (**a**) ACC ($n^{control}$=180, n^{SCZ}=180) and **(b)** DLPFC ($n^{control}$=210, n^{SCZ}=211). Violin plots quantify the strength of significance for all edQTLs (x-axis) as a function of distance (bp, y-axis) in the (**c**) ACC and (**d**) DLPFC. edQTLs are ranked by significance, with the max-edQTL prioritized as number one (circled) followed by the second most significant edQTL, and so on (two-sided Wilcoxon rank sum test with continuity correction). Whisker violin plots used throughout this figure show median, lower and upper quartiles, and whiskers represent minimum and maximum of the data. Furthermore, binomial variance analysis indicates RNA editing sites with edQTLs display more variances than those which do not have edQTLs in the (**c**) ACC and (**f**) DLPFC (two-sided Wilcoxon rank sum test with continuity correction).

Figure S16

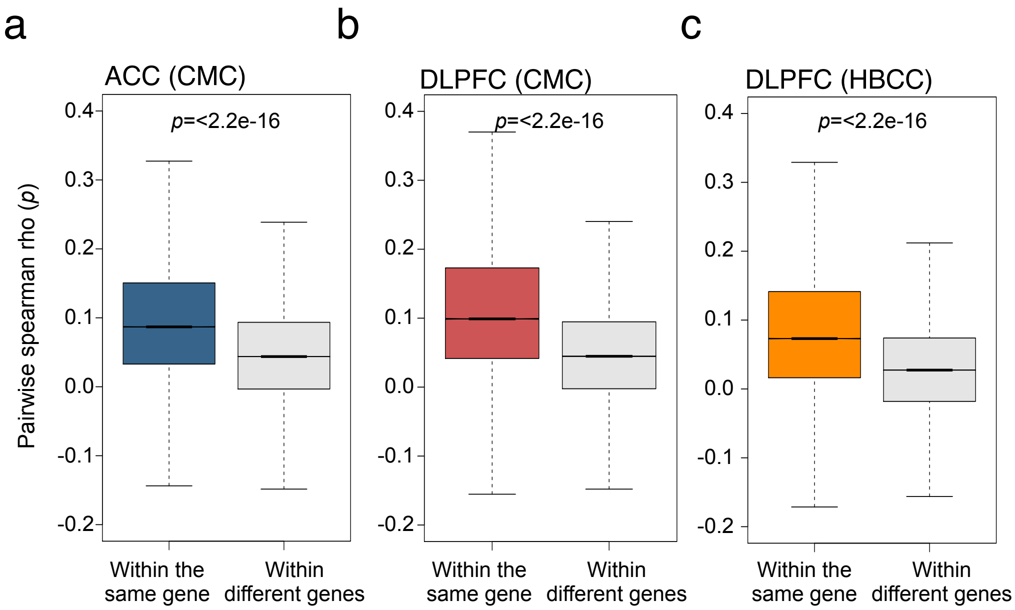

Figure S16. Correlations between RNA editing levels. Spearman correlation coefficient computed a series of pairwise associations between RNA editing levels for sites within the same gene as well as pairwise associations between RNA editing levels for sites in all other genes. Higher correlations were observed between sites in the same gene in the (**a**) ACC ($n^{control}$=245, n^{SCZ}=225) and (**b**) DLPFC samples ($n^{control}$=286, n^{SCZ}=254) as well as (**c**) NIMH HBCC DLPFC samples ($n^{control}$=217, n^{SCZ}=100) (two-tailed Student's *t*-test).

Figure S17

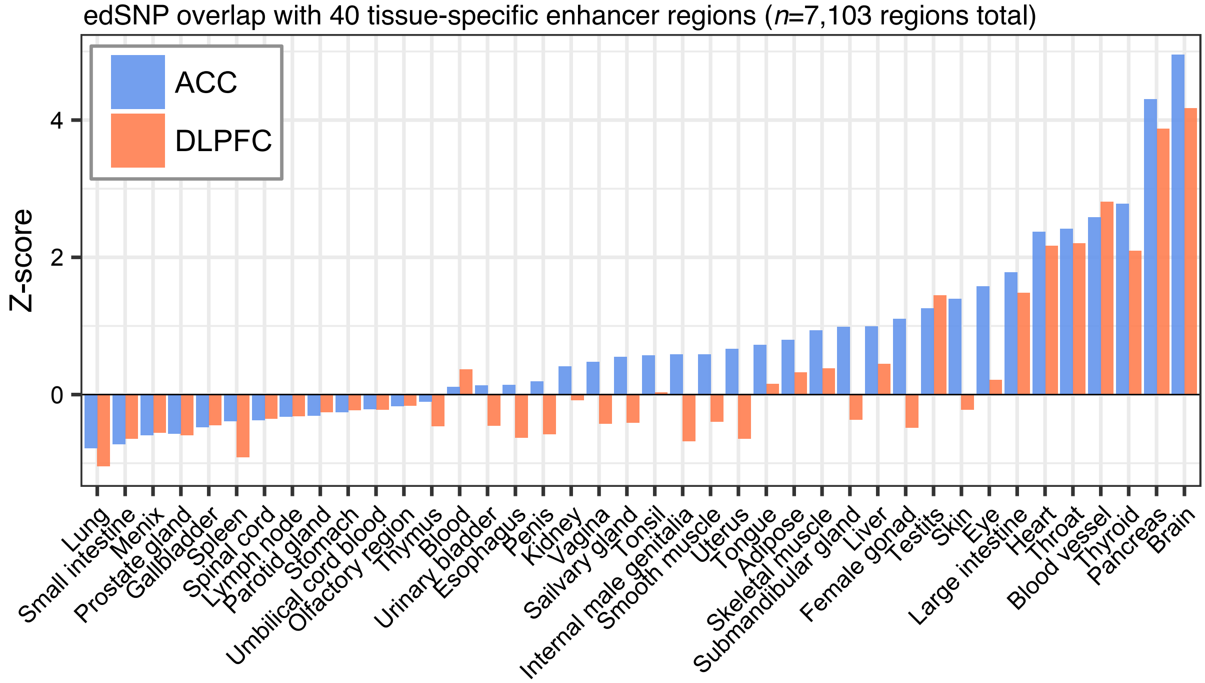

Figure S17. Tissue-specific enhancer enrichment analysis. Genomic coordinates for edSNPs were overlapped with tissue-specific enhancer regions derived from 40 different human tissues; data from the FANTOM project. The regioneR R package was used test overlaps of genomic regions based on permutation sampling. We repeatedly sampled random regions from the genome 1000 times, matching size and chromosomal distribution of the region set under study. By recomputing the overlap with the enhancer features in each permutation, statistical significance of the observed overlap was computed. We observed enrichment for many tissues, but the strongest enrichment was for brain tissue in the (**a**) ACC (blue) and (**b**) DLPFC (orange).

Figure S18

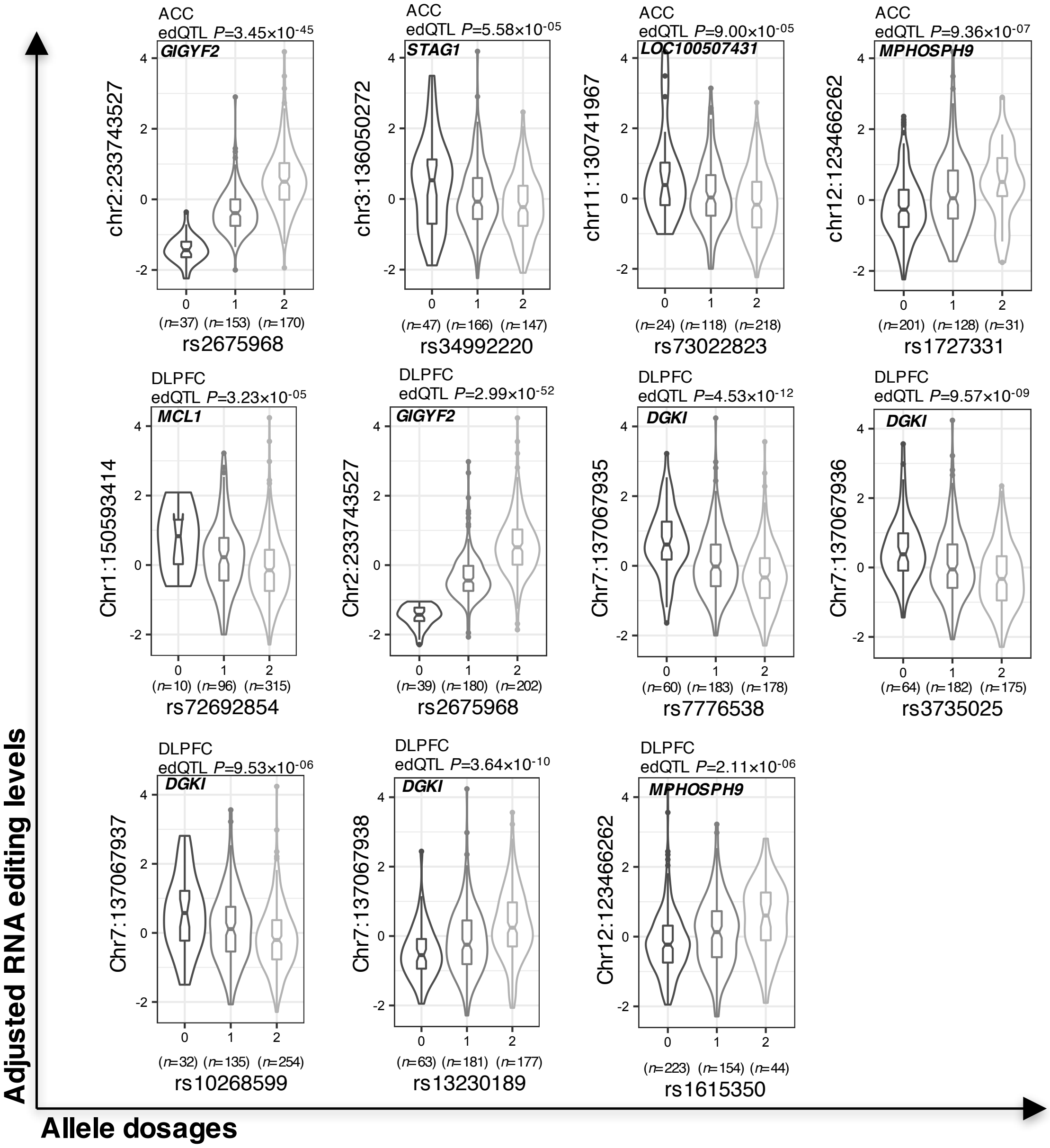

Figure S18. Cis-edQTLs that co-localize with GWAS loci using coloc2 software. Associations between adjusted RNA editing levels and imputed genotype dosages for edQTLs that co-localize with SCZ GWAS risk loci in the ACC and DLPFC (brain regions are labeled accordingly; *p*-values were computed using linear regression and FDR adjustment from the *R* package matrixEQTL). The allelic effect of the SNPs on editing levels are shown by boxplots within violin plots. Violin plot shows the density plot of the data on each side, the lower and upper border of the box correspond to the first and third quartiles, respectively, the central line depicts the median, and whiskers extends from the borders to ±1.5xinter-quantile range, the distance between the first and third quantiles.