Descriptions of Supplementary Tables

Miscellaneous notes

Tissue and cell abbreviations

Here and throughout the manuscript we use the following tissue abbreviations:

- HC: Hippocampus
- BA9: Dorsolateral prefrontal cortex
- BA24: Anterior cingulate cortex
- NAcc: Nucleus accumbens

For flow sorted data, we use the following suffixes:

- pos: $NeuN^+$ cells
- neg: $NeuN^-$ cells

For example, BA9_pos means $NeuN^+$ from the dorsolateral prefrontal cortex.

Genome build and coordinates

Here and throughout the manuscript we use:

- The GRCh37/hg19 build of the human reference genome is used in all analyses
    - We use 1-based genomic coordinates
    - We use closed genomic intervals. E.g., chr1:10-12 means positions 10, 11, 12 on chr1
- GENCODE v19 (protein-coding transcripts and long non-coding RNA transcripts) is used for all gene models (http://www.gencodegenes.org/releases/19.html)
    - The 'GENCODE v19 GTF' file was downloaded from ftp://ftp.sanger.ac.uk/pub/gencode/Gencode_human/release_19/gencode.v19.annotation.gtf.gz

Supplementary_Table_01.Demographics.xlsx

Demographic information for brain donor subjects.

- Individual: Donor ID
- PMI(h): Postmortem interval in hours
- Age(yr:day): Age in years and days at time of death
- Sex: M = Male or F = Female
- Ethnicity: Donor ethnicity
- Sorted Tissues: Semicolon delimited list of cell-sorted tissues available for whole-genome bisulfite-sequencing
- Bulk Tissues: Semicolon delimited list of bulk tissues available for whole-genome bisulfite-sequencing
- ATAC-seq Tissues: Semicolon delimited list of cell-sorted tissues available for ATAC-seq
- RNA-seq Tissues: Semicolon delimited list of cell-sorted tissues available for RNA-seq

Supplementary_Table_02.Summary_of_Sorted_Sequencing.csv

Summary of whole-genome bisulfite-sequencing data for FANS samples (n = 45).

- Sample ID: Sample ID in the form <Donor>_<Region>_<NeuNstatus>
- Number sequenced PE reads: Number of sequenced paired-end reads
- Number aligned PE reads: Number of aligned paired-end reads
- Alignment rate (%): Percentage of sequenced reads that were aligned
- Number covered CpGs: Number of autosomal CpGs in the reference genome covered by at least one read
- Covered CpGs (%): Percentage of autosomal CpGs in the reference genome covered by at least one read
- Mean depth: Average sequencing depth of autosomal CpGs in the reference genome covered by at least one read
- Bisulfite conversion rate (%): Estimated bisulfite-conversion rate from spike-in unmethylated lambda DNA

Supplementary_Table_03.Summary_of_Bulk_WGBS.csv

Summary of whole-genome bisulfite-sequencing data for bulk samples (n = 27).

- Sample ID: Sample ID in the form <Donor>_<Region>
- Number sequenced PE reads: Number of sequenced paired-end reads
- Number aligned PE reads: Number of aligned paired-end reads
- Alignment rate (%): Percentage of sequenced reads that were aligned
- Number covered CpGs: Number of autosomal CpGs in the reference genome covered by at least one read
- Covered CpGs (%): Percentage of autosomal CpGs in the reference genome covered by at least one read
- Mean depth: Average sequencing depth of autosomal CpGs in the reference genome covered by at least one read
- Bisulfite conversion rate (%): Estimated bisulfite-conversion rate from spike-in unmethylated lambda DNA

Supplementary_Table_04.Summary_of_CG-DMRs_and_blocks_from_F-stat_Analysis.xlsx

Summary of CG-BLOCKs and CG-DMRs from F-stat analysis.

- Reference autosomal genome: The autosomal reference genome (hg19)
- Analyzed autosomal genome (bulk): The set of autosomal CpGs used in the differential methylation analysis of the bulk tissue samples
- Analyzed autosomal genome (sorted): The set of autosomal CpGs used in the differential methylation analysis of the sorted tissue samples
- pos vs neg: Differential methylation analysis comparing $NeuN^+$ and $NeuN^-$ samples
- pos: Differential methylation analysis comparing $NeuN^+$ samples
- Non-NAcc pos: Differential methylation analysis comparing BA9_pos, BA24_pos, and HC_pos samples
- neg: Differential methylation comparing $NeuN^-$ samples
- Bulk tissue: Differential methylation analysis comparing bulk tissue samples

Columns:

- N: Number of features
- Number of CpGs: Number of CpGs in feature
- Genomic size (bp): Size of feature in base pairs
- Median size (bp): Median size of feature in base pairs

Supplementary_Table_05.CG-DMRs.csv

Coordinates and summary statistics of CG-DMRs identified by F-stat analysis of NeuN-positive and NeuN-negative samples (n = 45).

- chromosome: Chromosome of CG-DMR
- start: Start coordinate of CG-DMR
- end: End coordinates of CG-DMR
- n: Number of CpGs in CG-DMR
- areaStat: The ‘area’ of the F-statistic; equal to the sum of the F-statistics for the individual CpGs
- BA24_neg: Average methylation in CG-DMR of BA24_neg samples
- BA24_pos: Average methylation in CG-DMR of BA24_pos samples
- BA9_neg: Average methylation in CG-DMR of BA9_neg samples

- BA9_pos: Average methylation in CG-DMR of BA9_pos samples
- HC_neg: Average methylation in CG-DMR of HC_neg samples
- HC_pos: Average methylation in CG-DMR of HC_pos samples
- NAcc_neg: Average methylation in CG-DMR of NAcc_neg samples
- NAcc_pos: Average methylation in CG-DMR of NAcc_pos samples
- pos_vs_neg: Is this labeled a CG-DMR between pos and neg samples by our annotation pipeline (TRUE or FALSE)
- perm_P: P-value based on 1,000 permutations

Supplementary_Table_06.CG-blocks.csv

Coordinates and summary statistics of blocks identified by F-stat analysis of NeuN-positive and NeuN-negative samples (n = 45).

- chromosome: Chromosome of block
- start: Start coordinate of block
- end: End coordinates of block
- n: Number of CpGs in block
- areaStat: The 'area' of the F-statistic; equal to the sum of the F-statistics for the individual CpGs
- BA24_neg: Average methylation in block of BA24_neg samples
- BA24_pos: Average methylation in block of BA24_pos samples
- BA9_neg: Average methylation in block of BA9_neg samples
- BA9_pos: Average methylation in block of BA9_pos samples
- HC_neg: Average methylation in block of HC_neg samples
- HC_pos: Average methylation in block of HC_pos samples
- NAcc_neg: Average methylation in block of NAcc_neg samples
- NAcc_pos: Average methylation in block of NAcc_pos samples
- pos_vs_neg: Is this labeled a block between pos and neg samples by our annotation pipeline (TRUE or FALSE)
- perm_P: P-value based on 1,000 permutations

Supplementary_Table_07.Novel_CellType_CG-DMRs_Compared_to_Published_Data.xlsx

Summary of novel NeuN-positive vs. NeuN-negative CG-DMRs from F-stat analysis of NeuN-positive and NeuN-negative samples (n = 45). See Methods for how novelty was determined.

- Reference autosomal genome: The autosomal reference genome (hg19)
- Analyzed autosomal genome (sorted): The set of autosomal CpGs used in the differential methylation analysis of the sorted tissue samples
- Novel small CG-DMRs: Novel CG-DMRs identified in our analysis of NeuN-positive and NeuN-negative samples
- Novel non-NAcc pos: Novel CG-DMRs identiefied in our differential methylation analysis comparing BA9_pos, BA24_pos, and HC_pos samples

Columns:

- N: Number of features
- Number of CpGs: Number of CpGs in feature
- Genomic size (bp): Size of feature in base pairs
- Median size (bp): Median size of feature in base pairs

Supplementary_Table_08.Neuronal_CG-DMRs.csv

Coordinates and summary statistics of CG-DMRs identified by F-stat analysis of $NeuN^+$ samples (n = 23).

- chromosome: Chromosome of CG-DMR
- start: Start coordinate of CG-DMR
- end: End coordinates of CG-DMR
- n: Number of CpGs in CG-DMR
- areaStat: The 'area' of the F-statistic; equal to the sum of the F-statistics for the individual CpGs
- BA24_pos: Average methylation in CG-DMR of BA24_pos samples
- BA9_pos: Average methylation in CG-DMR of BA9_pos samples
- HC_pos: Average methylation in CG-DMR of HC_pos samples
- NAcc_pos: Average methylation in CG-DMR of NAcc_pos samples
- NAcc_pos_vs_BA9_pos: Is this labeled a CG-DMR between NAcc_pos and BA9_pos samples by our annotation pipeline (TRUE or FALSE)
- perm_P: P-value based on 1,000 permutations

Supplementary_Table_09.Non-Neuronal_CG-DMRs.csv

Coordinates and summary statistics of CG-DMRs identified by F-stat analysis of $NeuN^-$ samples (n = 22).

- chromosome: Chromosome of CG-DMR
- start: Start coordinate of CG-DMR
- end: End coordinates of CG-DMR
- n: Number of CpGs in CG-DMR
- areaStat: The 'area' of the F-statistic; equal to the sum of the F-statistics for the individual CpGs
- BA24_neg: Average methylation in CG-DMR of BA24_neg samples
- BA9_neg: Average methylation in CG-DMR of BA9_neg samples
- HC_neg: Average methylation in CG-DMR of HC_neg samples
- NAcc_neg: Average methylation in CG-DMR of NAcc_neg samples
- perm_P: P-value based on 1,000 permutations

Supplementary_Table_10.GREAT.xlsx

Results from GREAT analysis using the whole genome (hg19) as background. Settings used are: Basal+extension (constitutive 5.0 kb upstream and 1.0 kb downstream, up to 100.0 kb max extension) with curated regulatory domains included. Gene Ontology (GO) terms returned must be significant by both the binomial and hypergeometric tests using the multiple hypothesis correction false discovery rate (FDR) ≤ 0.05 whose binomial fold enrichment is at least 2.

Each sheet in this workbook is the GREAT results for a single set of regions:

- Hypo_DMRs_NAvALL: Hypomethylated NAcc neuronal CG-DMRs
- Hyper_DMRs_NAvALL: Hypermethylated NAcc neuronal CG-DMRs

The columns in each spreadsheet are:

- Ontology: Ontology source
- Term_Name: Term identifier from the ontology

- Binom_Rank: Ordinal rank of the binomial p-value compared to the p-values of other annotations
- Binom_Raw_P-Value: Uncorrected p-value from the binomial test over genomic regions
- Binom_FDR_Q-Val: False discovery rate q-value of the binomial test p-values
- Binom_Fold_Enrichment: (Observed / Expected)-fold enrichment of number of genomic regions in the test set with the annotation from the binomial test
- Binom_Observed_Region_Hits: Actual number of genomic regions in the binomial test set with the annotation
- Binom_Region_Set_Coverage: The proportion of all genomic regions in the binomial test set that lie in the regulatory domain of a gene with the annotation
- Hyper_Rank: Ordinal rank of the hypergeometric p-value compared to the p-values of other annotations
- Hyper_FDR_Q-Val: False discovery rate q-value of the hypergeometric test p-values
- Hyper_Fold_Enrichment: (Observed / Expected)-fold enrichment of number of genomic regions in the test set with the annotation from the hypergeometric test
- Hyper_Observed_Gene_Hits: Actual number of genomic regions in the hypergeometric test set with the annotation
- Hyper_Total_Genes: Number of genes in the genome with the annotation used in the hypergeometric test
- Hyper_Gene_Set_Coverage: Proportion of all genes with the annotation that are tagged by the test set

Supplementary_Table_11.Non-NAcc_CG-DMRs.csv

Coordinates and summary statistics of CG-DMRs identified by F-stat analysis of NeuN-positive, non-NAcc samples (n = 17).

- chromosome: Chromosome of CG-DMR
- start: Start coordinate of CG-DMR
- end: End coordinates of CG-DMR
- n: Number of CpGs in CG-DMR
- areaStat: The ‘area’ of the F-statistic; equal to the sum of the F-statistics for the individual CpGs
- BA24_pos: Average methylation in CG-DMR of BA24_pos samples
- BA9_pos: Average methylation in CG-DMR of BA9_pos samples
- HC_pos: Average methylation in CG-DMR of HC_pos samples
- perm_P: P-value based on 1,000 permutations

Supplementary_Table_12.Neuronal_CG-blocks.csv

Coordinates and summary statistics of blocks identified by F-stat analysis of NeuN positive samples (n = 23).

- chromosome: Chromosome of block
- start: Start coordinate of block
- end: End coordinates of block
- n: Number of CpGs in block
- areaStat: The ‘area’ of the F-statistic; equal to the sum of the F-statistics for the individual CpGs

- BA24_pos: Average methylation in block of BA24_pos samples
- BA9_pos: Average methylation in block of BA9_pos samples
- HC_pos: Average methylation in block of HC_pos samples
- NAcc_pos: Average methylation in block of NAcc_pos samples
- perm_P: P-value based on 1,000 permutations
- coversPCGene: Does the block cover the entirety of a protein-coding gene(s)

Supplementary_Table_13.CA_DMRs_and_CT-DMRs.xlsx

Coordinates and summary statistics of CA-DMRs and CT-DMRs identified by F-stat analysis of $NeuN^+$ samples (n = 23).
- chromosome: Chromosome of DMR
- strand: Strand of DMR
- start: Start coordinate of DMR
- end: End coordinates of DMR
- n: Number of Cs in DMR (CpAs for CA-DMRs, CpTs for CT-DMRs). Only Cs on the same strand as the DMR are counted
- areaStat: The 'area' of the F-statistic; equal to the sum of the F-statistics for the individual Cs
- BA24_pos: Average methylation in DMR of BA24_pos samples
- BA9_pos: Average methylation in DMR of BA9_pos samples
- HC_pos: Average methylation in DMR of HC_pos samples
- NAcc_pos: Average methylation in DMR of NAcc_pos samples
- perm_P: P-value based on 1,000 permutations

Supplementary_Table_14.RNA-seq.NAcc_posvsBA9_pos.csv

Results of RNA-seq differential gene analysis comparing NAcc_pos samples to BA9_pos samples (n = 11).
- gene_id: Ensembl gene ID in GENCODE v19 GTF file
- logFC: Estimate of the log2-fold-change corresponding to NAcc_pos vs. BA9_pos comparison
- BH_P: Benjamini-Hochberg adjusted P-value from the hypothesis test that logFC is zero

Supplementary_Table_15.RNA-seq.NAcc_negvsBA9_neg.csv

Results of RNA-seq differential gene analysis comparing NAcc_neg samples to BA9_neg samples (n = 9).
- gene_id: Ensembl gene ID in GENCODE v19 GTF file
- logFC: Estimate of the log2-fold-change corresponding to NAcc_neg vs. BA9_neg
- BH_P: Benjamini-Hochberg adjusted P-value for the hypothesis test that logFC is zero

Supplementary_Table_16.Summary_of_RNA-seq.csv

Summary of RNA-seq samples (n = 20).
- Sample ID: Sample ID in the form <Donor>_<Tissue>_<NeuNstatus>
- Number sequenced PE reads: Number of sequenced paired-end reads

- Number of quasi-mapped PE reads: Salmon reported number of quasi-mapped reads
- Alignment rate (%): Salmon reported quasi-mapping rate

Supplementary_Table_17.GO_Analysis.xlsx

Results from Metascape analysis of several genes sets against 'GO Biological Processes', 'Reactome Gene Sets', and 'KEGG Pathway' gene sets libraries. All genes in the genome were used as background. Only results with a P-value < 0.01, minimum count = 3, and an enrichment factor >1.5 are reported.

Each sheet in this workbook is the Metascape results for a single gene set:
- Genes_downregulated: Genes downregulated in NAcc vs. BA9 in neurons
- Genes_upregulated: Genes upregulated in NAcc vs. BA9 in neurons

The columns in each spreadsheet are:
- GroupID: 'Summary' is the representative enriched term for the cluster (selected based on significance) while 'Member' denotes membership in the cluster
- Category: The gene set library: 'GO Biological Processes', 'Reactome Gene Sets', or 'KEGG Pathway'
- Term: GO Biological Process, Reactome, or KEGG Pathway term
- Description: Term description
- Overlap: Input genes associated with term / all genes associated with term
- LogP: Log10(P-value), minimum value is -2 corresponding to a p-value of 0.01; p-values calculated based on accumulative hypergeometric distribution
- Log(q-value): Log of q-value calculated using the Benjamini-Hochberg method for correction for multiple hypotheses testing
- InTerm_InList: # input genes in term / total # of genes in term
- Genes: Genes (Entrez gene IDs) from the input that were found to be associated with that term
- Symbols: Gene symbols from the input that were found to be associated with that term

Supplementary_Table_18.ATAC-seq.pos_vs_neg.csv

Results of ATAC-seq differential accessibility analysis comparing $NeuN^+$ samples to $NeuN^-$ samples (n = 22).
- chromosome: Chromosome of peak
- start: Start coordinate of peak
- end: End coordinate of peak
- logFC: Estimate of the log2-fold-change corresponding to pos vs. neg
- BH_P: Benjamini-Hochberg adjusted P-value for the hypothesis test that logFC is zero

Supplementary_Table_19.ATAC-seq.NAcc_pos_vs_BA9_pos.csv

Results of ATAC-seq differential accessibility analysis comparing NAcc_pos samples to BA9_pos samples (n = 11).
- chromosome: Chromosome of peak

- start: Start coordinate of peak
- end: End coordinate of peak
- logFC: Estimate of the log2-fold-change corresponding to NAcc_pos vs. BA9_pos
- BH_P: Benjamini-Hochberg adjusted P-value for the hypothesis test that logFC is zero

Supplementary_Table_20.ATAC-seq.NAcc_neg_vs_BA9_neg.csv

Results of ATAC-seq differential accessibility analysis comparing NAcc_neg samples to BA9_neg samples (n = 11).
- chromosome: Chromosome of peak
- start: Start coordinate of peak
- end: End coordinate of peak
- logFC: Estimate of the log2-fold-change corresponding to NAcc_pos vs. BA9_pos
- BH_P: Benjamini-Hochberg adjusted P-value for the hypothesis test that logFC is zero

Supplementary_Table_21.Summary_of_ATAC-seq.csv

Summary of ATAC-seq samples (n = 22).
- Sample ID: Sample ID in the form <Donor>_<Tissue>_<NeuNstatus>
- Number sequenced PE reads: Number of sequenced paired-end reads
- Alignment rate (%): Bowtie2 reported alignment rate
- Duplicate rate (%): Percentage of reads marked as potential PCR duplicates by Picard's MarkDuplicates tool
- Mitochondrial contamination (%): Percentage of reads aligned to mitochondrial chromosome

Supplementary_Table_22.Haystack.xlsx

Results from Haystack analysis of neuronal DARs overlapping hyper- or hypomethylated neuronal CG-DMRs. A separate analysis of those neuronal DAR/CG-DMRs that also overlap promoters (defined as +/- 2 kb from transcription start site) is also included. Only results with a q-value < 0.05 and an enrichment ratio > 1 are reported. Additionally, only transcription factors that are expressed in our dataset are reported. Expressed genes had at least 1 cpm in at least 4 libraries (the size of the smallest group of samples) (24,161 / 33,351 genes).

Each sheet in this workbook is the output for a particular input, where the input is the set of genomic coordinates defining the neuronal DAR/CG-DMRs:
- Haystack_promoter: neuronal DARs that overlap hyper- or hypomethylated NAcc neuronal CG-DMRs that also overlap a gene promoter
- Haystack_all: neuronal DARs that overlap hyper- or hypomethylated NAcc neuronal CG-DMRs

The columns in each spreadsheet are:
- Motif ID: ID of motif in JASPAR database
- Motif Name: Name of motif in JASPAR database

- Presence in Target: Presence of motif in input
- Presence in BG: Presence of motif in background
- Ratio: Enrichment ratio of ‘Presence in Target’ to ‘Presence in BG’
- p-value: P-value from one-sided Fisher’s exact test of whether motif is enriched in target compared to background
- q-value: False discovery rate Q-value computed from ‘p-value’
- DEG: “YES” if gene was found to be differentially expressed between NAcc and BA9

Supplementary_Table_23.TEPIC_DYNAMITE_Analysis.csv

Results of the TEPIC/DYNAMITE pipeline to identify candidate transcription factors with a strong influence on differential gene expression observed between NeuN+ nuclei from NAcc and BA9 (n=2,952 differentially expressed genes).

- Transcription Factor: Gene symbol for transcription factors identified
- Normalized Correlation Coefficient: Resulting value from the logistic regression with elastic net regularization; all non-zero values are reported

Supplementary_Table_24.SLDSR_traits.xlsx

Description of the traits used in the SLDSR analysis.

- Trait: GWAS trait
- Type: Whether the Trait is broadly categorized as Psychiatric, Neurological, Behavioral-Cognitive, or an ‘Additional_Phenotype’ (BMI and the 3 negative controls, Height, Crohn’s Disease, and Coronary Artery Disease).
- N_cases: Number of cases in GWAS data (if applicable)
- N_controls: Number of controls in GWAS data (if applicable)
- N: Number of samples in GWAS data (if applicable)
- Link to GWAS data: URL from where GWAS data was downloaded
- Filename: Filename of GWAS data
- MD5 checksum: MD5 checksum of GWAS data (if available)
- Publication: Reference to original publication (if available)

Supplementary_Table_25.SLDSR_results.xlsx

Results of the stratified linkage disequilibrium score regression (SLDSR) analyses using the LDSC software for the CG-DMRs (NeuN+), CG-DMRs (NeuN+ vs. NeuN-), DARs (NeuN+), DARs (NeuN+ vs. NeuN-), Brain H3K27ac, CNS (LDSC), and chromHMM (union) features. GWAS sizes for each trait are reported in Supplementary Table 24.

Each sheet in this workbook is the SLDSR results for a set of analyses:

- Baseline adjustments: Running SLDSR on each of our 7 features separately combined with the 53 baseline features
- Stringent adjustments: Running SLDSR on each of our 4 differential features separately combined with the 3 non-differential brain-specific features and the 53 baseline features

- Adjusting for ndf excluding df: Running SLDSR on each of our 4 differential features having removed regions found by our differential approach from the non-differential features (but not from the baseline features)

The columns in each spreadsheet are:

- Trait: GWAS trait
- Stratum: Whether the Trait is believed to be linked to the brain (Brain-linked or Non-brain-linked) and whether we detect any evidence of the heritability of the Brain-linked trait being linked to a brain-derived Feature in the 'Baseline adjustments' analysis (Brain-linked (sig) vs. Brain-linked (non-sig)).
- Feature: Genomic feature
- Total width (bp): Total width of the Feature
- Proportion of SNPs: Proportion of SNPs tested in the Feature
- Coefficient: LDSC regression coefficient
- Coefficient_std_error: LDSC standard error of the regression coefficient
- Coefficient_z-score: LDSC regression coefficient Z-score
- Coefficient_p: 1-sided P-value for Coefficient_z-score, i.e. Pr(Z > z) where z ~ Normal(0, 1)
- Coefficient_holm: Coefficient_p adjusted for multiple testing using Holm's method. P-values are adjusted within each Trait.
- Proportion of h2: LDSC estimate of proportion of h^2 explained by the feature for the given trait
- Proportion of h2 standard error: LDSC estimate of the standard error of the proportion of h^2 explained by the feature for the given trait
- Enrichment: LDSC enrichment score
- Enrichment_std_error: LDSC standard error of the enrichment score

Supplementary_Table_26.Overlap_Gene_Symbols.xlsx

A list of 56 out of 237 genes (Supplementary Table 10 in Jaffe *et al.*, 2018[56]) recently shown to have differentially expressed features in schizophrenia vs control prefrontal cortex samples that contain at least one neuronal CG-DMR in our dataset.