nature genetics

Supplementary information

https://doi.org/10.1038/s41588-025-02249-4

Single-nucleus chromatin accessibility profiling identifies cell types and functional variants contributing to major depression

In the format provided by the authors and unedited

Demultiplexing of pooled subjects

We multiplexed nuclei extracted from a male and a female subject to reduce capture-cost and potential batch-effects between subjects and sexes (Supplementary Fig 1A-B). Genotypes estimated based on snATAC-seq read-counts at 1000 Genomes[1] based common variants were used to demultiplex pooled libraries at single-cell level. Additionally, principal component analysis (PCA) of top 1% most variable peaks clustered the demultiplexed subjects by sex (Supplementary Fig 1C; Supplementary method: Demultiplexing and assignment of sex). In addition, pseudo-bulk chromatin accessibility profiles at sex-specific gene, *XIST*, distinguished demultiplexed donors by sex (Supplementary Fig 1E). One subject was removed from downstream analysis as sex was not confirmed by the brain bank. Out of 83 remaining subjects used for differential analysis, one subject (reported as male) clustered with the female sex in PCA (Supplementary Fig 1C); however, both snRNA-seq and genotyping data consistently identified this subject as female. Furthermore, X chromosome inbreeding coefficient for this subject was very low (0.02383), which is in the range expected for a female, depicting robustness of snATAC-seq based demultiplexing approach.

We finally validated subject identities by assessing concordance between the estimated genotypes based on snATAC-seq read-counts and external genotypes obtained from the genotyping arrays. All snATAC-seq subjects showed the highest Pearson's correlation with their own genotypes (Supplementary Fig 1D), except one female subject which showed the highest correlation with the multiplexed male subject. Reassuringly, this female subject clustered with the female sex in PCA and showed female-specific accessibility at *XIST* (Supplementary Fig 1C, E). Notably, Pearson's correlation across differential peaks between original analysis and after removing this female subject was high across clusters (mean $r>0.95$). Taken together, we show that combining subjects by sex can allow robust demultiplexing of pooled snATAC-seq subjects based on sex-specific regions and/or common variants. Importantly, this method omits the need for obtaining external genotypes which can be costly and significantly reduces the cost of 10x single-cell capture and sequencing.

Robustness of differential accessibility results

In addition to the known covariates (age, sex, PMI, batch, no. of cells per subject) accounted for in the differential accessibility model, we evaluated the effects of other potential confounders on differential analysis by systematically adding them to the model: ethnicity (Caucasian 95.45% cases, 95% controls; African American: 0% cases, 5% controls; Hispanic: 2.27% cases; NA: 2.27% cases), presence of substance at death (Yes: 72.73% cases, 15% controls; No: 20.45% cases, 45% controls; NA: 6.82% cases, 40% controls) as well as pH of the brain tissue for each subject. We found high correlations between all differential peaks tested in the original and new analysis after adding pH (mean across all clusters: r=0.99) and modest to high correlations for other covariates (r>0.75, Supplementary Fig 7D). Additionally, to test whether a death by drug overdose (22% of cases, n=10/44) has an impact on differential analysis, we correlated differences in the mean values of normalized log2 counts per million for differential regions (DARs) in cases versus controls and intoxicated versus non-intoxicated cases and found no correlation between them (Pearson's r= -4.6e-05, p=1, Supplementary Fig 7C). This suggested that chromatin accessibility differences in MDD versus controls were not associated with a suicide by drug overdose in MDD individuals.

Furthermore, we assessed whether MDD-associated accessibility differences were likely to arise by chance by shuffling case/control labels 100 times within each batch in each cluster (nPerm=100 times, Supplementary Fig 7E). We found an overall low percentage overlap between actual and permuted DARs in almost all the clusters. In addition, and reassuringly, for clusters with highest MDD-associated differential accessibility, including Mic2 and ExN1, the real data consistently yielded higher no. of DARs than permuted data. Conversely, in clusters with low number of actual DARs, we found overlapping no. of permuted DARs with the original DARs. Since these clusters might inherently have lower variation, there are more chances to observe overlapping no. of permuted DARs. Together, these results confirmed that differential accessibility identified in cases vs controls is unlikely to be driven by chance.

In addition, we tested whether MDD-associated accessibility differences found in Mic2 cluster were driven by technical factors, including different sample sizes, cell type proportions, and

sequencing depth in cases vs controls. For this, firstly, we randomly downsampled case subjects (n=44) 100 times to match control sample size (n=39). Percentage common DARs and effect-size (logFC) Pearson's correlations between all differential peaks tested in original and downsampled datasets were consistently high (Supplementary Fig 8A). Additionally, we downsampled both MDD and control subjects to an equal size (n=39, n=35, n=30 each). Here again, effect size correlations between downsampled and original DARs were consistently high (r=0.99, 0.89, 0.84, Supplementary Fig 8C). We also compared known measures of sequencing depth between MDD versus control subjects in Mic2 by checking a) mean number of read pairs across subjects (non-duplicates, non-mitochondrial, and high mapq>30) b) mean fragment counts in peaks, and found no significant differences between groups (two-sided Wilcoxon test, Supplementary Fig 8B). Finally, we evaluated whether per-subject cell proportions for Mic2 were different between MDD (total no. of cells, n=899) versus controls (n=1365). We did not find significant differences in Mic2 cell proportions between the two groups (two-sided Wilcoxon test, FDR=0.08). However, three control subjects were identified to be potential outliers (>=1.5*IQR) with significantly higher proportions of Mic2 cells compared to remaining subjects. Therefore, we performed sensitivity analysis by removing these three subjects, which also resulted in overall high effect-size correlation (Pearson's r=0.94, p<2.2e-16) and majority overlap (62.8%) with original DARs. Notably, to mitigate the effect of the total number of cells sampled for each subject, we have added it as a covariate in our differential accessibility model.

Functional MDD sSNPs identified using cluster-specific gapped-kmer SVM

In total, 13,531 MDD-associated SNPs were derived from three GWAS[2–4] consisting of LD expanded (86.3%), fine-mapped (22.3%), and genome-wide significant SNPs (3.7%). As expected, these SNP sets overlapped with each other, but each of them had their exclusive contributions to MDD-associated SNPs (75.4%, 1.9% and 12%, respectively; Supplementary Fig 7E). Across 35 snATAC clusters (excluding ExN7, Mix1 and Mix2) and corresponding broad cell-types, we observed that 8.7% of MDD associated SNPs were overlapping with 1001-bp marker cCREs or candidate DARs (cdSNPs). In turn, 8.5% of cdSNPs were qualified as sSNP for at least one cluster by gkmSVM modelling. cdSNPs were relatively more abundant in broad cell-type and cluster marker peaks (43.7% and 83.4%) compared to DARs (2% and 12.5%). Further, 71.1% and 80% of sSNPs were observed in cell-type and cluster marker peaks, respectively. Overall, we

observed that majority of sSNPs were detected in neurons (78%), specifically excitatory neurons (75%), compared to glial cell types (22%). Moreover, 97.9% of the sSNPs were exclusively identified in clusters of a single broad cell-type as opposed to the few SNPs shared between multiple cell-types (2.1%; Ast-Oli and ExN-Mic). Furthermore, 53.1% of sSNPs were exclusively disrupting accessibility in a single cluster. In case of excitatory neuronal clusters, we found sSNPs mapping to specific cortical layers, where the highest percentage of cluster-exclusive sSNPs were observed in deep-layer neurons. Notably, our candidate cluster ExN1 had the highest percentage of cluster-exclusive SNPs (40%) among any other deep-layer clusters identified in our data. Only a handful sSNPs were observed to have an effect shared across all excitatory neuronal clusters (1.4%). On the other hand, in case of inhibitory neuronal clusters, the percentage of sSNPs (3%) were much lower than that of excitatory neuronal clusters.

In case of glial clusters, we identified a variety of sSNPs for astrocyte (6.1%), oligodendrocyte (7.2%), OPC (3.1%) and microglia (4.1%) clusters. Of note, rs9889058 in astrocytes had the highest observed consensus variant effect score across all clusters. This sSNP was located in astrocyte cell-type candidate DAR and showed the most significant variant-effect score in Ast3 among other astrocytic clusters. This MDD sSNP was found to disrupt a Zf family CTCF TF binding site and was linked to GNAO1 gene through peak-to-gene linkages. Interestingly, Ast3 cluster also showed significant accessibility disruption in MDD individuals (followed by ExN1 and Mic2). Likewise, rs174561 sSNP was located in cluster-specific marker peaks of an astrocytic Ast4 and oligodendrocyte Oli1 clusters in promoter, exon, intron and 5'UTR regions of various FADS1 transcripts, and was eQTL for FADS1, FEN1, TMEM258, NXF1 as well as sQTL for FADS3 gene.

Reliability assessment of cluster-specific SNP prioritization

Here, we interrogate individual components of gkmSVM based workflow to assess whether statistical significance analyses of cdSNPs is reliable.

The trained classifiers for all clusters except ExN7 attained between 80% and 90% median AUC-

ROC and AUC-PR computed on held out test folds of cluster-specific 5-fold cross validation schemes (Supplementary Fig 7A). Regardless of the performance of ExN7 models, this cluster was omitted from downstream analyses due to low signal-to-noise ratio. Additionally, for 28 clusters, we tested the trained models on datasets consisting of 1001bp sequences underlying cluster marker peak regions as well as GC content and chromosome matched non-peak sequences. We observed that the models had even better performance on marker peaks with respect to median AUC-ROC (ranging from 90.7% to 98.2%) and AUC-PR (ranging from 91% to 98.1%). Since most of the cdSNPs were in cluster marker peak regions, we concluded that predictive power of the trained models was suitable for conducting a reliable chromatin accessibility disruption analysis.

Next, we examined the agreement between consensus ISM, deltaSVM and gkmexplain variant effect scores. We observed that each pair of scores were significantly correlated with respect to cdSNP variant effect scores (Spearman's ρ, min: 0.977, max: 0.984). Since none of the score type pairs had perfect correlation (Supplementary Fig 7B-C), we further investigated the cdSNPs which were not eligible to be sSNPs but were deemed statistically significant with respect to at least one

score type. In line with the high observed correlations, number of such SNPs were much lower compared to sSNPs (Supplementary Fig 7D). Hence, incorporating all three score types in sSNP criteria makes our analysis more stringent and reduces false positive findings.

Finally, we investigated the percentage of high, medium and low confidence sets of sSNPs whose seqlets were successfully matched (q-value < 0.1) to TFBS via TOMTOM. We observed that a greater percentage of high and medium priority sSNPs matched to TFBS (47.2% and 45.2%; Supplementary Fig 7D), compared to low priority sSNPs (18.4%).

Cross disorder relevance of MDD SNPs

We sought to assess whether there is a specificity from MDD SNPs to brain-related traits (i.e., ADHD, bipolar disorder, schizophrenia) and relevant traits (i.e., insomnia, neuroticism and PTSD) or non-brain related traits (i.e., height and Crohn's disease). For this purpose, we checked if cdSNPs and sSNPs were enriched in genome-wide significant loci identified for these traits

through GWAS (Supplementary Figure 11D). First, we inquired cross-disorder relevance of MDD GWAS SNPs in snATAC-seq defined marker cCREs or MDD-related chromatin (cdSNPs) compared to all MDD GWAS SNPs. In context of brain-related traits, MDD-associated cdSNPs were observed to be significantly enriched in bipolar disorder (Fisher's exact test, OR: 2.09; p-value: 0.0009) and schizophrenia risk loci (OR: 1.19; p-value: 0.007), suggesting an enrichment of related psychiatric traits using our candidate chromatin regions. However, we also see an unexpected enrichment for non-brain related height loci (OR: 1.78; p-value: 7e-20). While polygenicity of height (and thus the high number of loci) potentially influences this result, reassuringly, MDD-associated cdSNPs showing significant allele-specific chromatin activity (i.e., sSNPs) were significantly depleted (OR: 0.64; p-value: 0.04) in height loci. Similarly to cdSNPs, we found an enrichment for bipolar disorder risk loci (OR: 2.49; p-value: 0.07) in sSNPs, but a modest depletion for schizophrenia risk loci in sSNPs (OR=0.67, p=0.1), suggesting association between functional sSNPs in MDD and Bipolar disorder. Interestingly, sSNPs were also enriched in Crohn's disease risk loci (OR: 2.21; p-value: 0.06). Although primarily a non-brain related trait, Crohn's disease has been significantly associated with depression risk[5,6], further suggesting associations between inflammatory changes and depression. While not statistically significant, odds ratio for insomnia was higher in sSNPs (OR=1.3) compared to cdSNPs (OR=1.1), while lower for neuroticism in sSNPs (OR=0.72) compared to cdSNPs (OR=1.03). (Supplementary Figure 11D)

While these results showed overall relevance of MDD-associated sSNPs and cdSNPs with other psychiatric disorders and traits, we further assessed cross-disorder relevance of our highlighted MDD sSNPs reported through multi-modal visualizations (Fig 6B-C, Extended Fig 7B-C, Supplementary Fig 12A-B) For this, we overlayed Manhattan plot tracks for MDD, PTSD, schizophrenia, bipolar disorder and ADHD for highlighted sSNPs. Individual dot colors in cross-trait GWAS loci tracks represent lead MDD SNP (blue), LD friends of lead MDD SNPs (yellow) and sSNP (pink) identified from MDD GWAS track in each plot (Supplementary Figure 11E).

Cross-trait plot for highlighted ExN1 sSNPs rs2276138, rs11601579 revealed that locus harboring rs2276138 was most significant for MDD followed by bipolar disorder. The locus harboring rs11601579 was implicated for MDD, schizophrenia and bipolar disorder. Moreover, for Mic2, we found that locus harboring rs10210757 was most significant in MDD followed by

schizophrenia and rs5995992 is located within a locus genome-wide significant for MDD and schizophrenia (Supplementary Fig 11E). Examples of ExN1 specific sSNPs (rs9299525, rs6063348, Supplementary figure 12A-B) were also tested for cross-disorder relevance (Supplementary Fig 11E), which included rs6063348 in a locus most significant for MDD but also modestly significant for other psychiatric disorders, PTSD, schizophrenia, bipolar disorder and ADHD.

Allelic imbalance

We generated single-cell allelic counts at germline variants using phased genotypes for each subject and snATAC-seq and snRNA-seq data using SALSA[7]. First, we examined allelic-imbalance in accessibility by measuring differences in allelic counts of individual sSNPs and cCREs-containing sSNPs by aggregating reference versus alternate allele counts across individuals. For measuring differences at the level of cCREs (1001bp from gkm-SVM pipeline), we summed snATAC-seq read-counts mapping to reference versus alternate alleles of all heterozygous variants falling within sSNP-containing cCREs, enabling the assignment of each cCRE to a haplotype, as described previously [7]. Nearly 50% of ExN1 sSNPs, including rs2276138 and rs11601579 (Fig 6C-D) showed allelic-imbalance at the level of sSNP or were within cCREs showing allelic differences in accessibility across individuals in snATAC-seq (11/22, exact binomial test; $p<0.1$, Supplementary Figure 14A). Next, we evaluated regulatory effects on gene expression by aggregating snRNA-seq read-counts across heterozygous variants in high LD ($r>0.8$) with each sSNP. This approach allowed us to assess the gene regulatory impact of non-coding sSNPs through their associated exonic variants, resulting in 45% of ExN1 sSNPs showing allelic imbalance across individuals in snRNA-seq (10/22, exact binomial test; $p<0.1$, Supplementary Figure 14A). Across both approaches using snATAC-seq and snRNA-seq data, 66% (44/67) of all MDD sSNPs and 73% (16/22) of ExN1 MDD sSNPs in subject genotypes demonstrated allele-specific effects on accessibility or expression.

Supplementary Methods

Nuclei Extraction

Nuclei were extracted from frozen DLPFC tissue sections as previously described[8] with some modifications. Briefly, frozen brain sections were dounced for 5mins in the lysis buffer. The lysis suspension was homogeneously mixed by pipetting up and down 10-15 times following which only 80% of the lysis solution was mixed with 5ml of wash buffer to make diluted nuclei suspensions. The samples were mixed and passed through a 30-um cell strainer to remove cellular debris and washed twice in 10ml wash buffer. Next, Optiprep cushion was created followed by gradient centrifugation for 30mins at 10,000g and resuspended in 1ml wash buffer. Nuclei were counted with Hoescht 33342 (1:2000) three times (average was used) with Cell countess (Olumpys). Nuclei were then spun down at 500g for 5mins and diluted in 1xNB (10x Genomics) to obtain a final concentration of 3080 nuclei/ul. We increased the loading concentration by 20% to compensate for potential sources of nuclei loss as done previously [5 6].

Multiplexing and library construction

Equal volume of nuclei suspensions from a male and female subject (either from case or control) were pooled together to obtain a final concentration of 3080 nuclei/ul. The combined nuclei were loaded in one of the lanes of 10x microfluidic chip. Each microfluidic chip was loaded with at least four multiplexed libraries representing both cases and controls. Single-cell capture and library preparations were done using 10x Genomics Chromium Single Cell v1.1 reagents following the protocol outlined by the user guide [https://www.10xgenomics.com/support/single-cell-atac/documentation/steps/library-prep/chromium-single-cell-atac-reagent-kits-user-guide-v-1-1-chemistry]. Cell Ranger atac v2.0 was used for alignment against the GRCh38 reference available on the 10x Genomics website (refdata-cellranger-arc-GRCh38-2020-A-2.0.0). The sequencing metrics are provided in Supplementary Table 1.

Demultiplexing and assignment of sex

10x snATAC-seq libraries with pooled male and female nuclei were demultiplexed at single-cell level into specified number of donors (n=2) using vireoSNP(v0.5.6, v0.5.7)[9] based on 1000 Genomes based common variants [1] piled-up at single-cell level using CellSNP[10]. The external genotyping data obtained from blood or brain tissues of these subjects was lifted over from hg19 to hg38 using LiftoverVcf from Picard(v2.26.3). cellsnp-lite(v1.2.2) was used to call variants from snATAC-seq demultiplexed subject bam files with the --minMAF flag set to 0.1 and the --minCOUNT flag set to 20. The --genotype flag was also used to obtain a vcf file with genotypes for each subject bam file. The Pearson correlation between the genotypes in the demultiplexed snATAC-seq subject bams and genotypes obtained from a genotyping array at a later date were calculated with Assign_Indiv_by_Geno.R[11] to verify the subject bam genotypes were most strongly correlated with the genotyping array genotypes for the correct individual. Finally, to determine sex of the demultiplexed donors, we used multiple approaches: a) PCA of top 1% most variable peaks using pseudo-bulked peak accessibility counts per subjects, b) psuedobulk accessibility profiles at sex-specific gene *XIST* (normalized by reads in TSS), c) finally, we trained a machine learning classifier on sex-specific chromatin accessibility data generated previously to confirm sex at single-cell level. Overall, these results agreed with pseuodbulk accessibility profiles identified by PCA and at sex-specific gene, *XIST*.

Evaluation of differential accessibility analysis

Principal component analysis (PCA) was performed on the normalized per-subject pseudobulk accessibility counts per million. Eigencorplot in PCAtools(v2.4.0) (https://github.com/kevinblighe/PCAtools) was used to assess Pearson's correlations between top 10 PCs and potential covariates. Variance partition analysis was performed on per cell-type and subject pseudobulk log2 normalized counts per million. % variance was sorted in order of median fraction of variance explained by each covariate before plotting. The effects of additional covariates (pH, ethnicity, and presence of substance at death) on differential accessibility results were assessed by adding each of them to the differential model and correlating effect size (logFC

values) of all the peaks between original and new analysis in each cluster using Pearson's method. Finally, differential accessibility analyses were performed by permuting case/control labels 100 times, within each batch, in each cluster, using limma-voom in muscat (with the same parameters). To examine robustness of differential results in Mic2, we took several steps. Firstly, case subjects were randomly down-sampled (n=44) 100 times to match control sample size (n=39). Percentage common DARs and effect-size (logFC) Pearson's correlations were computed for differential peaks tested in the original and downsampled datasets. Second, subjects in both case and control groups were downsampled once to equal sizes (n=39, n=35, n=30 each) and effect size correlations were computed. For differential accessibility in downsampling tests, muscat parameter was changed to filter="genes" (originally "both") to avoid filtering samples. We compared percentage of Mic2 cells in each subject and between cases and controls with Wilcoxon tests using rstatix [https://cran.r-project.org/package=rstatix (2021)] and bootstrapping p values 10000 times to mitigate the effect of outliers (boot: Bootstrap R (S-Plus) Functions. https://cran.r-project.org/package=boot (2021).

Genotypes

Genotypes were obtained using the Illumina Global Screening array including the Psych 30K panel. GenomeStudio(Software 2.0.5) was used to call genotypes from raw IDAT files, and the genotypes were exported from GenomeStudio according to Illumina's Plus/Minus convention. After filtering out SNPs with a call rate less than 98%, MAF < 0.01, or Hardy-Weinberg equilibrium p-value less than 0.000001, the VCF files per chromosome were uploaded to the TOPMed imputation server https://imputation.biodatacatalyst.nhlbi.nih.gov/#! . The selected reference panel was TopMed r3 (hg38), and the rsq filter was set to 0.1. The pipeline created chunks of 10 Mb, and a liftover step was performed since the input data was built using the hg19 reference, while the reference panel was hg38. The genotypes were phased with Eagle v2.4 (Loh et al., 2016) using an overlap of 5 Mb. Imputation was then performed with MINIMAC v4.1.6 using a window of 500kb. After imputation, SNPs with MAF < 0.05 or Hardy-Weinberg equilibrium p-value less than 0.000001 were removed.

Quality Control of Genotyping Array Samples

Sex mismatches were identified in the dataset using PLINK (v1.9, v2)[12,13] 1.9's --check-sex option to calculate the X chromosome inbreeding coefficient. Subjects annotated as male with X chromosome inbreeding coefficients less than 0.8, or annotated as female with inbreeding coefficients greater than 0.2 were deemed problematic. Potential duplications of samples as well as possible cross contamination of samples was examined using identity by descent calculated by PLINK 1.9's -- genome option, after LD pruning. Samples that were mismatched in both the snATAC-seq and snRNA-seq datasets were removed from the genotyping array dataset. Samples with call rates less than 98% were also removed.

PsychEncode data

PsychEncode DLPFC neuronal and non-neuronal (NeuN+/-) HiC loops[14] and NeuN+/- H3K27ac and H3K4me3 histone modification peaks were obtained[15]. For integrating NeuN+/- histone peaks in MDD sSNP accessibility tracks, we used per-subject DLPFC histone peaks (bed format) from the PsychEncode EpiMap database [https://psychencode.synapse.org/Explore/Studies]. and generated a merged non-overlapping peak-set across all subjects using GenomicRanges(v1.44.0, v1.48.0) reduce() function. Peaks were converted to hg38 using kentutils liftOver command line with hg19ToHg38.over.chain.gz UCSC liftover chain. For assessing overlaps with DARs, marker

cCREs, and peak-to-gene linkages, we used consolidated histone modification peaks from cortical cells [12].

MDD GWAS fine-mapping

Genome-wide fine-mapping of MDD GWAS[2] with SparsePro[16] was performed on summary statistics and reference LD panel based on White British population of UK Biobank[17]. Briefly, SparsePro utilizes variational inference to (a) aggregate correlated variants having similar association with a trait into *K* sparse effect sets; (b) infer set-level causal status, effect sizes and configuration. Moreover, effect sizes as well as causal status and configuration of effect sets were used to compute variant-level posterior inclusion probability (PIP) and effect size estimates. Fine-

mapping inputs were prepared as follows: GWAS variants were first binned into sliding windows (i.e., loci) of 3-Mbp with neighbouring windows having a 2-Mbp overlap. In turn, loci-level summary statistics were harmonized with LD matrix by inverting effect size and alleles of mismatching SNPs. SparsePro was run with K=5 sparse effect sets on each 3-Mbp loci. To mitigate boundary effects, estimates for the variants located within 1-Mbp windows centered at each loci were retained. Moreover, 95% credible set was computed on genome-wide variant-level PIP estimates. Next, sparse effect sets inferred at each loci were filtered to retain those including at least one variant qualified for the credible set. Across 22 autosomal chromosomes, 213 independent effect sets comprising of 3,013 variants in LD were identified to be likely causal for MDD across the genome.

Cluster-specific accessibility disruption analyses by MDD associated genetic variants.

We adapted gkmSVM variant-scoring workflow[18] (Supplementary Figure 9-10) in Python (v3.6, v3.8), biopython (1.79), gimmemotifs (0.15.0), h5py (3.6.0), matplotlib (3.7.0), numpy (1.22.2), pandas (1.4.1), pyranges (0.0.115), scanpy (1.9.5),scikit_learn (1.0.2), scipy (1.8.0), seaborn (0.11.2), logomaker (0.8). The workflow consists of the following main steps: (a) gkmSVM training and evaluation; (b) gkmSVM-based consensus variant effect scores of cdSNPs; (c) Statistical significance of cdSNPs; and (d) High, medium and low confidence subsets of sSNPs.

gkmSVM training and evaluation

For each of the 35 snATAC clusters (excluding Mix1, Mix2, and ExN7), we adopted a 5-fold cross validation setup and trained gkmSVM models to predict whether 1001bp genomic sequences are from cluster-specific cCREs (accessible, positive examples) or randomly sampled non-peak regions (inaccessible, negative examples) using default parameter setting defined as follows: gapped k-mer kernel with center weighting (wgkm, t = 4); a word length of 11 (l = 11); 7 informative columns in each word (k = 7); the maximum allowed number of mismatches to consider (d = 3); the initial value of the exponential decay function used for center weighting (M= 50), half-life parameter of 50 positions for exponential decay (H = 50); the regularization

penalty parameter of SVM (c = 1); and precision parameter (e = 0.001).

The datasets used for each cluster consisted of equally numerous positive and negative examples with matching chromosome and GC content distributions. To obtain positive example set per cluster, we applied iterative peak removal (based on MACS2 p-value) on 501bp peaks called for each cluster; extended resulting peaks by 250bp at both directions; extracted underlying sequences from the reference genome (hg38); and discarded the sequences containing an ambiguous base ("N" or "n"). In case of negative examples, we first generated a cluster-independent set of candidate negative examples by tiling hg38 reference genome 1001bp at a time with a stride of 50bp while discarding those having an ambiguous base. Moreover, individually for each cluster, we further removed the sequences intersecting with 1001-bp peak regions. Since the number of negative examples were much greater than those of positive examples in each cluster, we subsampled the former to match the size of the latter. For this goal, we first partitioned each positive example set into 20 equally populated bins according to the GC content distribution percentiles defined on the same set; and assigned the candidate negative examples according to the resulting cluster-specific bins. Next, we initialized an empty dataset per cluster and added pairs of positive and negative examples, one at a time, until the positive example set of the corresponding cluster is exhausted. At each step, we paired one randomly sampled positive example with one randomly sampled candidate negative example originating from the same chromosome and GC bin, where sampling was done without replacement. Resulting cluster-specific datasets were partitioned to cross-validation folds as follows: Fold 1 consisted of chromosomes 3, 4, 12 and 13;Fold 2 consisted of chromosomes 5, 6, 14 and 15; Fold 3 consisted of chromosomes 7, 8, 11, 16 and 17; Fold 4 consisted of chromosomes 2, 9, 10, 18 and 19; and Fold 5 consisted of chromosomes 1, 20, 21, 22, X and Y.

At each iteration of the cross-validation setup, gkmtrain function[19] was run with default parameter setting to train gkmSVM model on 4 folds; and gkmpredict[19] function was run on sequences in the remaining (test) fold to obtain unnormalized decision scores whose magnitude quantified model confidence and sign reflected whether the sequences are accessible (i.e., positive). To make model training computationally feasible, we sorted positive and negative example pairs in each training set (i.e., 4 folds) with respect to the ascending order of MACS2 p-value of positive examples and, trained the models on the top 60,000 pairs (120,000 examples). End2, InN3 and Mic2 clusters did not have 60,000 positive examples in their training sets and all available pairs of examples were used for model training. Finally, the trained models (175 in total) were evaluated with respect to AUC-ROC and AUC-PR metrics (scikit-learn package [17]) calculated according to test fold predictions (Supplementary Fig 7A).

gkmSVM-based consensus variant effect scores of cdSNPs

Here, we utilized the gkmSVM models to interrogate whether cdSNPs have an allelic impact on chromatin accessibility in the corresponding cluster. First, we extracted 201bp sequences centered at each cdSNP in each cluster from the reference genome (hg38); and generated two sequences per SNP by mutating the middle position (101st position) to the effect (A1) and non-effect (A2) alleles of the SNP, respectively. We employed three complementary approaches on 201-bp cdSNP sequences — in silico mutagenesis (ISM), deltaSVM [18] and gkmexplain [19] — to compute cluster-specific variant effect scores as follows:

To obtain *ISM* variant effect score per cdSNP in each cluster, we first executed gkmpredict function to predict whether 201bp allelic sequences of cdSNPs available for the corresponding cluster are accessible. Next, ISM variant effect score was computed by subtracting non-effect allele score from that of the effect allele resulting from the same model. Finally, consensus ISM variant effect scores for each cdSNP were calculated by averaging variant effect scores across the models trained for the corresponding cluster. In the case of *deltaSVM*, we first generated and predicted all possible non-redundant 11-mers with the trained models via gkmpredict. Resulting decision scores served as a model-specific basis for deltaSVM score calculation. Next, for each trained

model in each cluster, deltaSVM.pl script was run on 201-bp allelic sequences cdSNPs as well as 11-mer decision scores. Finally, we averaged resulting variant effect scores across models to get consensus deltaSVM variant effect scores per cdSNP. Unlike ISM and deltaSVM which report magnitude and direction of change in gkmSVM predictions between sequence pairs, *gkmexplain* measures contribution of the nucleotides in each input sequence on gkmSVM predictions in terms of base-resolution importance scores[20]. For each pair of cluster-specific cdSNPs and gkmSVM models, we first executed gkmexplain script on 201bp allelic sequences in importance score mode (-m 0) to obtain 201-by-4 gkmexplain allelic importance scores. Resulting matrices capture importance of base identities (i.e., A, G, C and T) available at each position (i.e., 201-bp) on model predictions and, thus only the base identities present at each position are non-zero. Next, we computed 201-by-4 consensus importance score matrices for each allele of cdSNPs as the mean of importance score matrices across models. Moreover, we extracted the consensus importance scores corresponding to the 51bp window centered at the cdSNP (i.e., from 75th to 126th position) and computed the respective sums of 51- by-4 consensus importance score matrices of both alleles. Finally, for each cdSNP in each cluster, consensus gkmexplain variant effect scores were computed by subtracting non-effect allele sum from that of effect allele.

Statistical significance of cdSNPs

To assess whether consensus variant effect scores of cdSNPs were statistically significant, we generated 10 randomly shuffled 201bp non-effect allele sequences (hereafter, null sequences) using fasta-dinucleotide-shuffle function of MEME Suite[21], followed by in silico mutation of the middle position of resulting sequences (i.e., 101st position) to both alleles of each cdSNP. Moreover, we computed null consensus variant effect scores for each score type (i.e., ISM, deltaSVM and gkmexplain) by repeating the previous step pairs of null allelic sequences resulting from the same random shuffle. Additionally, we fitted normal and t distributions using SciPy to null consensus variant effect scores available for each score type per cluster. In line with the previous work[18,22], t-distribution was found to be a better fit to empirical null consensus variant effect score distributions based on KS test. Finally, we denoted cdSNPs as sSNP within context of a cluster if and only if consensus variant effect scores for all score types reside outside of 95%

confidence interval (CI) of respective null variant effect score distributions of the corresponding cluster.

High, medium and low confidence sets of sSNPs

To further refine sSNPs, we identified high, medium and low confidence sets reflecting the quantitative support for SNP-driven accessibility disruption in its immediate local neighborhood (<201bp). First, we determined the *active allele* for each sSNP which had a larger sum of non-negative consensus gkmexplain importance scores within the 51bp region centered at the SNP compared to the other allele. Next, we constructed cluster-specific and position-independent null importance score distributions by pooling all positive scores in 201-by-4 null consensus gkmexplain score matrices computed for cdSNPs of the corresponding cluster. Moreover, we extracted the subsequences around sSNPs (hereafter, *seqlets*) whose base-resolution consensus importance scores are significantly elevated compared to the position-independent null consensus importance distribution as follows: Starting from the position of sSNP, we extended an empty sequence towards both directions by iteratively adding the flanking nucleotides in an alternating fashion. If two consecutive positions in a direction had consensus importance scores lower than 97.5th percentile of the position-independent null distribution of the corresponding cluster, we stopped extending towards that direction. Furthermore, if the resulting seqlet was smaller than 7bp upon termination of the procedure, it was alternatingly extended towards both directions to reach this minimum length. Next, we computed *prominence* and *magnitude scores* which depend on *seqlet score* and *seqlet signal-to-noise ratio (SNR)* scores based on seqlets for sSNPs in each cluster. Seqlet score for a given allele of an sSNP, is defined as the sum of non-negative consensus importance scores which overlap with seqlet of the SNP. Seqlet SNR for each allele of an sSNP is given by the ratio of seqlet score and the sum of non-negative values in 201-by-4 consensus importance score matrix. In turn, magnitude and prominence scores are given by the difference between non-effect and effect allele seqlet scores and seqlet SNRs, respectively. Magnitude score attempts to quantify the magnitude of chromatin accessibility disruption with respect to the change from non-effect to effect allele within seqlet region. On

the other hand, prominence score attempts to capture how prominent the chromatin accessibility disruption due to the SNP is within its 201bp neighborhood. Furthermore, we sought to assess the statistical significance of the prominence and magnitude scores. For this goal, we first computed null prominence and magnitude scores by repeating the above-mentioned procedure on null sequences of cdSNPs in all clusters. Next, we calculated prominence and magnitude score p-values for each sSNP using cluster-specific empirical null distributions constructed on null magnitude and prominence scores as well as their additive inverse. Finally, we partitioned sSNPs in each cluster to high, medium and low confidence sets according to the prominence and magnitude p-value based rule set defined here[18]. sSNPs having prominence score p-value less than 0.05 were assigned to high confidence set. These sSNPs disrupt chromatin accessibility of their seqlet regions which were substantially more accessible compared to other regions within 201bp window centered at the SNP. Next, the sSNPs which did not qualify for high confidence set were assigned medium confidence if their magnitude score p-value were less than 0.05 or their prominence score p-value were less than 0.1. These SNPs substantially disrupt chromatin accessibility of their seqlet region but the quantitative support for disruption is moderate. Remaining SNPs were grouped under low confidence set. Additionally, we ran TOMTOM[23] via , memes(v1.4.1) R package[24]) on seqlets of high and medium confidence sSNPs and identified significant matches (q-value < 0.1) to TFBS motifs (Homo Sapiens) available in CISBP-2.0[25] database.

Obtaining DLPFC eQTL and sQTLs

A total of 8 eQTL (RNA-seq) and sQTL (LeafCutter[26]) summary statistics corresponding to uniformly performed analysis on DLPFC tissue of individuals in BrainSeq[27], CommonMind[28] Gene expression elucidates functional impact of polygenic, ROSMAP[29] and GTEx[30] studies were downloaded from eQTL Catalogue[31] identified eQTLs and sQTLs with FDR<0.05 and used them for our downstream analyses.

Genomic location-based characterization of SNPs

MDD associated SNPs, cdSNPs and sSNPs were characterized with respect to their genomic

locations using several approaches. First, we downloaded human gene annotations (comrepensive set, hg38 build) from GENCODEv43[32] identified the genes nearest to each SNP. We used annotatr(v1.22.0)[33] in R(v4.2) package which outsources genic annotation data from org.Hs.eg.db and TxDb.Hsapiens.UCSC.hg38.knownGene, to obtain fine-grained functional information regarding the genomic regions containing the SNPs. Specifically, we focused on the following annotation categories provided in decreasing order of priority: coding sequence (CDS), 5'UTR, 3'UTR, Exon, Enhancer (FANTOM5[34]), Promoter (less than 1KBp distance to TSS), Intron, 1-to-5Kbp to TSS and Intergenic. Additionally, we determined DLPFC eQTL and sQTL hits (eQTL Catalogue, FDR < 0.05) within MDD associated SNP, cdSNP and sSNP sets.

Cross disorder relevance of MDD SNPs

First, we downloaded summary statistics for GWAS performed for the brain-related traits (i.e., ADHD, bipolar disorder, schizophrenia) and relevant traits (i.e., insomnia, neuroticism and PTSD) and non-brain related traits (i.e., height and Crohn's disease). Then, we retained genome-wide significant SNPs (P < 5e-8) in each summary statistics file. Next, for each trait, we identified lead SNPs by retaining genome-wide significant SNPs having lowest p-value within 1-MBp window centered at the SNPs. Finally, we defined genome-wide significant loci as 1-MBp windows centered at the lead SNPs. As the number of genome-wide significant loci for PTSD is very low (i.e., 2 loci), we omitted PTSD from enrichment analyses.

We performed Fisher's exact test for cdSNPs and sSNPs, to assess if they are enriched or depleted within trait loci compared to a background set of SNPs. As cdSNPs were derived from MDD-linked SNPs (i.e., 13,531 SNPs including genome-wide significant MDD SNPs, their LD friends as well as fine-mapped MDD GWAS SNPs), we used MDD-linked SNPs as background for analysis on cdSNPs. With the same reasoning, cdSNPs were used as background for enrichment analyses performed for sSNPs. We report results from these analyses below, as odds ratios (OR) and p-values for traits where modest or significant enrichment or depletion of genetic signals were observed.

Characterizing MDD sSNPs-associated risk genes

For visualizing all the nearest or linked risk-genes to MDD sSNPs, we computed "full String network" with organism="Homo Sapiens" and minimum required interaction score = 0.4 and maximum number of interactions (1st shell= 0, 2nd shell= 0). We then evaluated Monarch[35] based human phenotypes enriched in the network which resulted in the expected phenotypes (Supplementary Table 21).

Rank-rank hypergeometric overlap

We performed a threshold free, rank-rank hypergeometric overlap (RRHO) analysis with RRHO2[36]. For each cell-type, genes were scored using the product of the sign of logFC and -log10 uncorrected p-value from differential expression analysis in the snRNA-seq and differential accessibility analysis in snATAC-seq datasets. Each DAR was linked to the gene with highest correlation (peak-to-gene linkages) and p values of peaks linked to the same gene were combined using invchisq (k=2) function in https://rdrr.io/cran/metap/. Likewise, gene most correlated to DARs directly overlapping transcription start sites (TSS, Homo sapiens, hg38 genome assembly) were compared with snRNA-seq DEGs. Ranked gene lists were provided to RRHO2_initialize function (method "hyper" and log10.ind "TRUE") and the results were plotted using the RRHO2_heatmap function.

Allelic imbalance analysis

Single-cell allelic counts were generated with SALSA(v1.0)[7] after correcting for mapping biases with WASP[37]. For each MDD sSNP (67/90 sSNPs in donor genotypes), reference versus alternate allele counts from snATAC-seq were summed across subjects to examine potential allelic imbalance using binomial test with default parameters in R. In addition, allele-specific differences were compared at the level of cCREs by aggregating allelic counts across all heterozygous variants falling within 1001bp sSNP-containing cCREs (from gapped kmer SVM pipeline), enabling the assignment of each cCRE to a haplotype[7]. Finally, to measure potential regulatory effects of non-

coding sSNPs on gene-expression, allele-specific counts from snRNA-seq were aggregated across genetic variants in high LD (r>0.8) with each sSNP and summed over subjects to compute differences in reference versus alternate allele counts using binomial test. LD expansion was performed with respect to Phase 3 genotypes of 1000G EUR individuals (https://www.cog-genomics.org/plink/1.9/).

Luciferase Assay

HEK293 (CRL-1573) and BE(2)C (CRL-2268) cells were purchased from ATCC and were cultured in Eagle's Minimal Essential Medium (EMEM, ATCC) (HEK293) or 1:1 EMEM:F12 (BE(2)C), supplemented with 10% FBS (Invitrogen), in a 5% CO2 humidified incubator at 37°C. We examined 8 sequences (4 regions, each with two alleles). Regions were ordered as gBlock gene fragments (IDT), and cloned between the XhoI and HindIII sites of the pGL4.23 vector (Promega) using T4 DNA ligase (NEB). This vector contains a minimal promoter immediately following the multiple cloning site, and upstream of the transcription start site of the firefly luciferase gene. Constructs were verified by Sanger sequencing using RVprimer3.

rs2276138-T (clone 227T10)
CCGGCTCGAGCCTGCATCTCTTCCAAACACAAGCACCTCCAAAACTGGCAGGATCGGGGGTTGCAGGAAGATGGTGGGTCAG
AGTGGCCACTGGCAGGGCACAGTGGCCATGGCAACCTCAAGAACAGGGACACAGAGAAAGCCACTGCCCCAGAGCTGTGCCA
GGCAGCACGAAGATTGGGCATCGTGAGTTTCACACCCAGGGTTCCTAAAGCTTCCGG

rs2276138-C (clone 227C11)
CCGGCTCGAGCCTGCATCTCTTCCAAACACAAGCACCTCCAAAACTGGCAGGATCGGGGGTTGCAGGAAGATGGTGGGTCAG
AGTGGCCACTGGCAGGGCACAGTGGCCACGGCAACCTCAAGAACAGGGACACAGAGAAAGCCACTGCCCCAGAGCTGTGCCA
GGCAGCACGAAGATTGGGCATCGTGAGTTTCACACCCAGGGTTCCTAAAGCTTCCGG

rs11601579-C (clone 116C10)
CCGGCTCGAGCAGCACCGGGGTGTCCGTCACAGCGCCGGGTGTCCATTGCTGCCACCATGTGATCTTGATCTGTCCTTTTCCA
CCATCCGTCATGCCTTGCCGTGTGCCACCCACACCATTTAAGTGTAAAACACACTCTCTTCTGAATGCTCCCAGCAACACTA
CAAGAAAAGTGCTAGCAGCGTTGGCCTCTGCTGTACAGGTGAAGAAGAAGCTTCCGG

rs11601579-A (clone 116A9)
CCGGCTCGAGCAGCACCGGGGTGTCCGTCACAGCGCCGGGTGTCCATTGCTGCCACCATGTGATCTTGATCTGTCCTTTTCCA
CCATCCGTCATGCCTTGCCGTGTGCCACACACACCATTTAAGTGTAAAACACACTCTCTTCTGAATGCTCCCAGCAACACTA
CAAGAAAAGTGCTAGCAGCGTTGGCCTCTGCTGTACAGGTGAAGAAGAAGCTTCCGG

rs4509984-C (clone 450C9)

CCGGCTCGAGGTCCTTTCAGCAAGAACCCCTCTCTGCATAAAGGGTCACCTCACTTCACAAAGCCTCTGCTTCCTCTTCTAT
AAGAAGGGTGGTGGTAGCTGCCTCTCAGCTGCTGTGAAGATGGGAAGCACCCAGCCAGGCCTGGCACTTTCAGGGCACTCTG
CACACTGGTCAGCCCCGCAGGACCTGAGGCTGAGCCTCTCGTGGGACAAGCTTCCGG

rs4509984-G (clone 450G9)

CCGGCTCGAGGTCCTTTCAGCAAGAACCCCTCTCTGCATAAAGGGTCACCTCACTTCACAAAGCCTCTGCTTCCTCTTCTAT
AAGAAGGGTGGTGGTAGCTGCCTCTCAGGTGCTGTGAAGATGGGAAGCACCCAGCCAGGCCTGGCACTTTCAGGGCACTCTG
CACACTGGTCAGCCCCGCAGGACCTGAGGCTGAGCCTCTCGTGGGACAAGCTTCCGG

rs9299525-A (clone 929A10)

CCGGCTCGAGTTAATTCTTTTAAAGATGCTGCCCTAAATTCTACTGCTCAGAAGTGGGACATAACAAAACAGAAAATCACCA
CTAAAGAGTTTCAAAATGGATGGTAGGAATTAGTGCAAGGTCTTTGAAACTGTTTTGTAATTATTGCAGATTTTATCGATGT
TTTGGAAACAATACCCTGTGCTTGCTAAGAAAATGCAAAATAGGCTGAAGCTTCCGG

rs9299525-G (clone 929G9)

CCGGCTCGAGTTAATTCTTTTAAAGATGCTGCCCTAAATTCTACTGCTCAGAAGTGGGACATAACAAAACAGAAAATCACCA
CTAAAGAGTTTCAAAATGGATGGTAGGAGTTAGTGCAAGGTCTTTGAAACTGTTTTGTAATTATTGCAGATTTTATCGATGT
TTTGGAAACAATACCCTGTGCTTGCTAAGAAAATGCAAAATAGGCTGAAGCTTCCGG

rs10210757-C (clone 102C12)

CCGGCTCGAGTTTGCAGGCCCAGCAAGGACAGGGACCAAGGTTGAGGGCTTGGAGGAGTCTGACTAAAGTTTGGTCAAGGAG
AGAGTCTTTGTCAGGGGTTATGTGGGCTCTTCCTTCATGTTCCTTTAAAGAGCATGCCCATCTTCCGGGACTTCCTCTTTCT
GTTTAAGTTACCATTTACTGTGGCTTCTCACGCTCTTTACGCAGCTGAAGCTTCCGG

rs10210757-T (clone 102T11)

CCGGCTCGAGTTTGCAGGCCCAGCAAGGACAGGGACCAAGGTTGAGGGCTTGGAGGAGTCTGACTAAAGTTTGGTCAAGGAG
AGAGTCTTTGTCAGGGGTTATGTGGGCTTTTCCTTCATGTTCCTTTAAAGAGCATGCCCATCTTCCGGGACTTCCTCTTTCT
GTTTAAGTTACCATTTACTGTGGCTTCTCACGCTCTTTACGCAGCTGAAGCTTCCGG

Western Blots for TF expression

HEK293 and Be(2)C cells were washed 3x with ice-cold phosphate buffered saline (PBS), lysed in ice-cold RIPA lysis buffer (Millipore, MA, USA) enriched with cOmplete™ Protease Inhibitor Cocktail (Roche, Switzerland), and centrifuged at 12,000 x g for 10 min in a 4°C precooled centrifuge. Total protein concentration was measured by using Pierce™ BCA Protein Assay Kit

(Thermo Fisher, CA, USA) according to manufacturer's protocol. Protein samples were mixed with 4x Laemmli sample buffer (Bio-Rad, CA, USA) and boiled at 95 °C for 10 min. 10 ug of total protein per sample was loaded onto 4-20% Mini-PROTEAN® TGX™ Precast Gel (Bio-Rad, CA, USA), separated and transferred to nitrocellulose membranes by using semi-dry Trans-Blot Turbo Transfer System (Bio-Rad, CA, USA) according to manufacturer's protocol. The membranes were blocked for 1h at RT in Tris-buffered saline with 1% Tween20 (TBST) and 5% BSA, and incubated overnight at 4°C with the following primary antibodies:

ASCL1 rabbit monoclonal antibody (Cat# ab211327, Clone#: EPR19840, Abcam, UK. Western blot dilution: 1:1000), ZFP3 rabbit polyclonal antibody (Cat# PA5-62726, Thermo Fisher, CA, USA, Western blot dilution: 1:1000), FLI1 rabbit polyclonal antibody (Cat# SAB2100822, Millipore Sigma, MA, USA, Western blot dilution: 1:500), GLI1 rabbit polyclonal antibody (Cat# SAB4301901, Millipore Sigma, MA, USA, Western blot dilution: 1:1000), RFX3 (Cat# CF505916, Clone#: OTI3F7, OriGene, MD, USA, Western blot dilution: 1:2000), RFX1 (Cat# 68057-1-Ig, Clone#: 1G10H6, Proteintech, IL, USA, Western blot dilution: 1:2000). anti-rabbit IgG (H&L) goat polyclonal antibody HRP-conjugated (Cat# ab6721, Abcam, UK), anti-mouse IgG (H&L) sheep polyclonal antibody HRP-conjugated (Cat# NA931, Amersham, IL, USA).

Next, membranes were washed 3x in TBST and incubated with goat anti-rabbit IgG HRP (Cat# ab6721, Abcam, UK, Western blot dilution: 1:10000) and anti-mouse IgG (H&L) sheep polyclonal antibody HRP-conjugated (Cat# NA931, Amersham, IL, USA, Western blot dilution: 1:5000). Proteins were visualized using Clarity Western ECL Substrate (Bio-Rad, CA, USA). Protein signal was detected by using the ChemiDOC MP Imaging System (Bio-rad, CA, USA).

References

1. The 1000 Genomes Project Consortium *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
2. Howard, D. M. *et al.* Genome-wide meta-analysis of depression identifies 102 independent variants and highlights the importance of the prefrontal brain regions. *Nat. Neurosci.* **22**, 343–352 (2019).
3. Levey, D. F. *et al.* Bi-ancestral depression GWAS in the Million Veteran Program and meta-analysis in >1.2 million individuals highlight new therapeutic directions. *Nat. Neurosci.* **24**, 954–963 (2021).
4. Als, T. D. *et al. Identification of 64 New Risk Loci for Major Depression, Refinement of the Genetic Architecture and Risk Prediction of Recurrence and Comorbidities*. http://medrxiv.org/lookup/doi/10.1101/2022.08.24.22279149 (2022) doi:10.1101/2022.08.24.22279149.
5. Facanali, C. B. G. *et al.* The relationship of major depressive disorder with Crohn's disease activity. *Clinics* **78**, 100188 (2023).
6. Zhang, Y. *et al.* SUPERGNOVA: local genetic correlation analysis reveals heterogeneous etiologic sharing of complex traits. *Genome Biol.* **22**, 262 (2021).
7. Wilson, P. C. *et al.* Multimodal single cell sequencing implicates chromatin accessibility and genetic background in diabetic kidney disease progression. *Nat. Commun.* **13**, 5253 (2022).
8. Maitra, M. *et al.* Extraction of nuclei from archived postmortem tissues for single-nucleus sequencing applications. *Nat. Protoc.* **16**, 2788–2801 (2021).
9. Huang, Y., McCarthy, D. J. & Stegle, O. Vireo: Bayesian demultiplexing of pooled single-

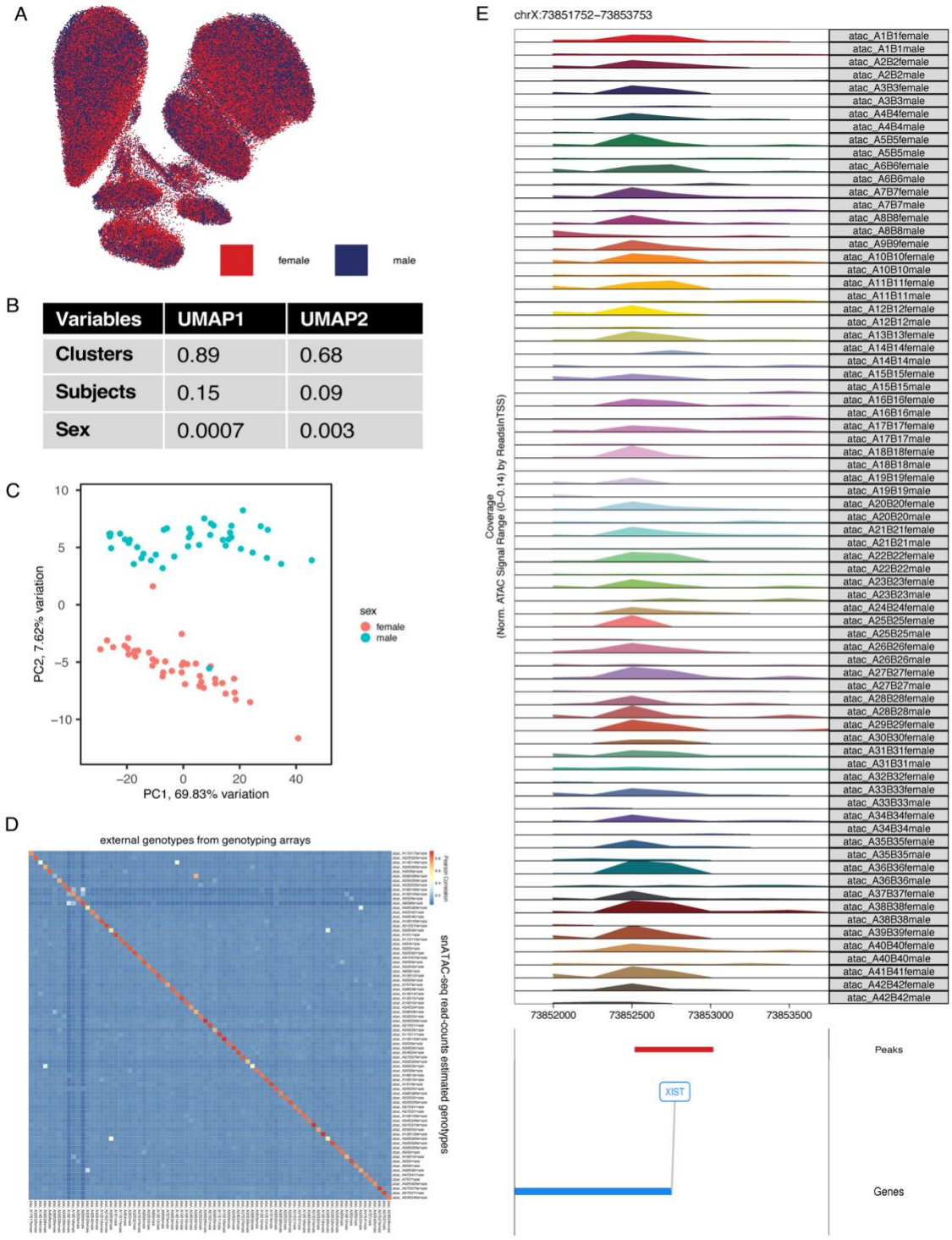
cell RNA-seq data without genotype reference. *Genome Biol.* **20**, 273 (2019).

10. Huang, X. & Huang, Y. Cellsnp-lite: an efficient tool for genotyping single cells. *Bioinformatics* **37**, 4569–4571 (2021).
11. Neavin, D. *et al.* Demuxafy *: Improvement in Droplet Assignment by Integrating Multiple Single-Cell Demultiplexing and Doublet Detection Methods*. http://biorxiv.org/lookup/doi/10.1101/2022.03.07.483367 (2022) doi:10.1101/2022.03.07.483367.
12. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience* **4**, 7 (2015).
13. Purcell, S. *et al.* PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
14. Hu, B. *et al.* Neuronal and glial 3D chromatin architecture informs the cellular etiology of brain disorders. *Nat. Commun.* **12**, 3968 (2021).
15. Girdhar, K. *et al.* Cell-specific histone modification maps in the human frontal lobe link schizophrenia risk to the neuronal epigenome. *Nat. Neurosci.* **21**, 1126–1136 (2018).
16. Zhang, W., Najafabadi, H. & Li, Y. *SparsePro: An Efficient Fine-Mapping Method Integrating Summary Statistics and Functional Annotations*. http://biorxiv.org/lookup/doi/10.1101/2021.10.04.463133 (2021) doi:10.1101/2021.10.04.463133.
17. Weissbrod, O. *et al.* Functionally informed fine-mapping and polygenic localization of complex trait heritability. *Nat. Genet.* **52**, 1355–1363 (2020).
18. Corces, M. R. *et al.* Single-cell epigenomic analyses implicate candidate causal variants at inherited risk loci for Alzheimer's and Parkinson's diseases. *Nat. Genet.* **52**, 1158–1168

(2020).

19. Lee, D. LS-GKM: a new gkm-SVM for large-scale datasets. *Bioinformatics* **32**, 2196–2198 (2016).
20. Shrikumar, A., Prakash, E. & Kundaje, A. GkmExplain: fast and accurate interpretation of nonlinear gapped *k* -mer SVMs. *Bioinformatics* **35**, i173–i182 (2019).
21. Bailey, T. L. *et al.* MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res.* **37**, W202-208 (2009).
22. Ober-Reynolds, B. *et al.* Integrated single-cell chromatin and transcriptomic analyses of human scalp identify gene-regulatory programs and critical cell types for hair and skin diseases. *Nat. Genet.* **55**, 1288–1300 (2023).
23. Gupta, S., Stamatoyannopoulos, J. A., Bailey, T. L. & Noble, W. Quantifying similarity between motifs. *Genome Biol.* **8**, R24 (2007).
24. Nystrom, S. L. & McKay, D. J. Memes: A motif analysis environment in R using tools from the MEME Suite. *PLoS Comput. Biol.* **17**, e1008991 (2021).
25. Weirauch, M. T. *et al.* Determination and inference of eukaryotic transcription factor sequence specificity. *Cell* **158**, 1431–1443 (2014).
26. Li, Y. I. *et al.* Annotation-free quantification of RNA splicing using LeafCutter. *Nat. Genet.* **50**, 151–158 (2018).
27. Collado-Torres, L. *et al.* Regional Heterogeneity in Gene Expression, Regulation, and Coherence in the Frontal Cortex and Hippocampus across Development and Schizophrenia. *Neuron* **103**, 203-216.e8 (2019).
28. Fromer, M. *et al.* Gene expression elucidates functional impact of polygenic risk for schizophrenia. *Nat. Neurosci.* **19**, 1442–1453 (2016).

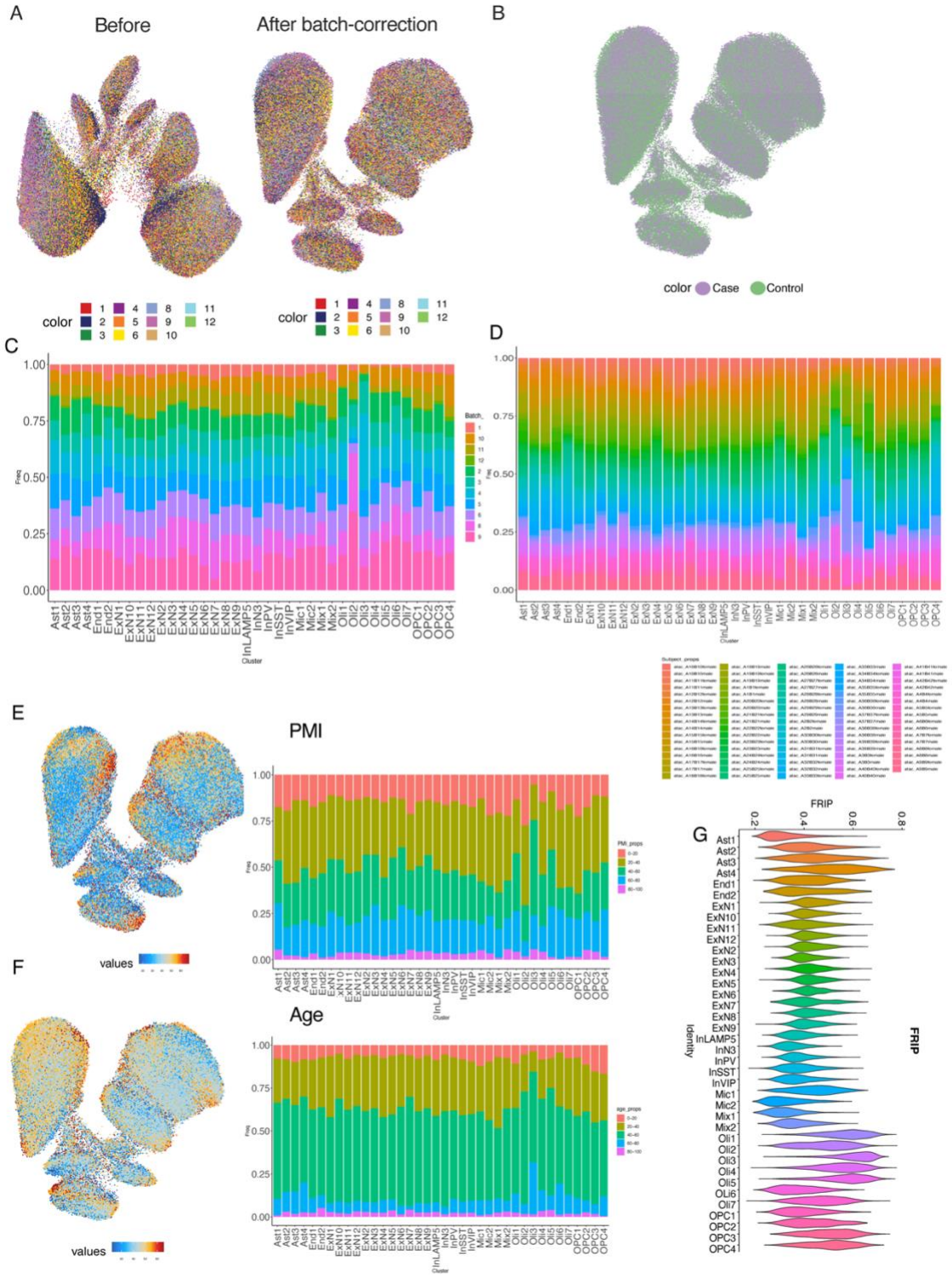
29. Ng, B. *et al.* An xQTL map integrates the genetic architecture of the human brain's transcriptome and epigenome. *Nat. Neurosci.* **20**, 1418–1426 (2017).
30. GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318–1330 (2020).
31. Kerimov, N. *et al.* A compendium of uniformly processed human gene expression and splicing quantitative trait loci. *Nat. Genet.* **53**, 1290–1299 (2021).
32. Harrow, J. *et al.* GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res.* **22**, 1760–1774 (2012).
33. Cavalcante, R. G. & Sartor, M. A. annotatr: genomic regions in context. *Bioinforma. Oxf. Engl.* **33**, 2381–2383 (2017).
34. The FANTOM Consortium *et al.* An atlas of active enhancers across human cell types and tissues. *Nature* **507**, 455–461 (2014).
35. Shefchek, K. A. *et al.* The Monarch Initiative in 2019: an integrative data and analytic platform connecting phenotypes to genotypes across species. *Nucleic Acids Res.* **48**, D704–D715 (2020).
36. Cahill, K. M., Huo, Z., Tseng, G. C., Logan, R. W. & Seney, M. L. Improved identification of concordant and discordant gene expression signatures using an updated rank-rank hypergeometric overlap approach. *Sci. Rep.* **8**, 9588 (2018).
37. van de Geijn, B., McVicker, G., Gilad, Y. & Pritchard, J. K. WASP: allele-specific software for robust molecular quantitative trait locus discovery. *Nat. Methods* **12**, 1061–1063 (2015).
38. Morabito, S. *et al.* Single-nucleus chromatin accessibility and transcriptomic characterization of Alzheimer's disease. *Nat. Genet.* **53**, 1143–1155 (2021).



Variables	UMAP1	UMAP2
Clusters	0.89	0.68
Subjects	0.15	0.09
Sex	0.0007	0.003

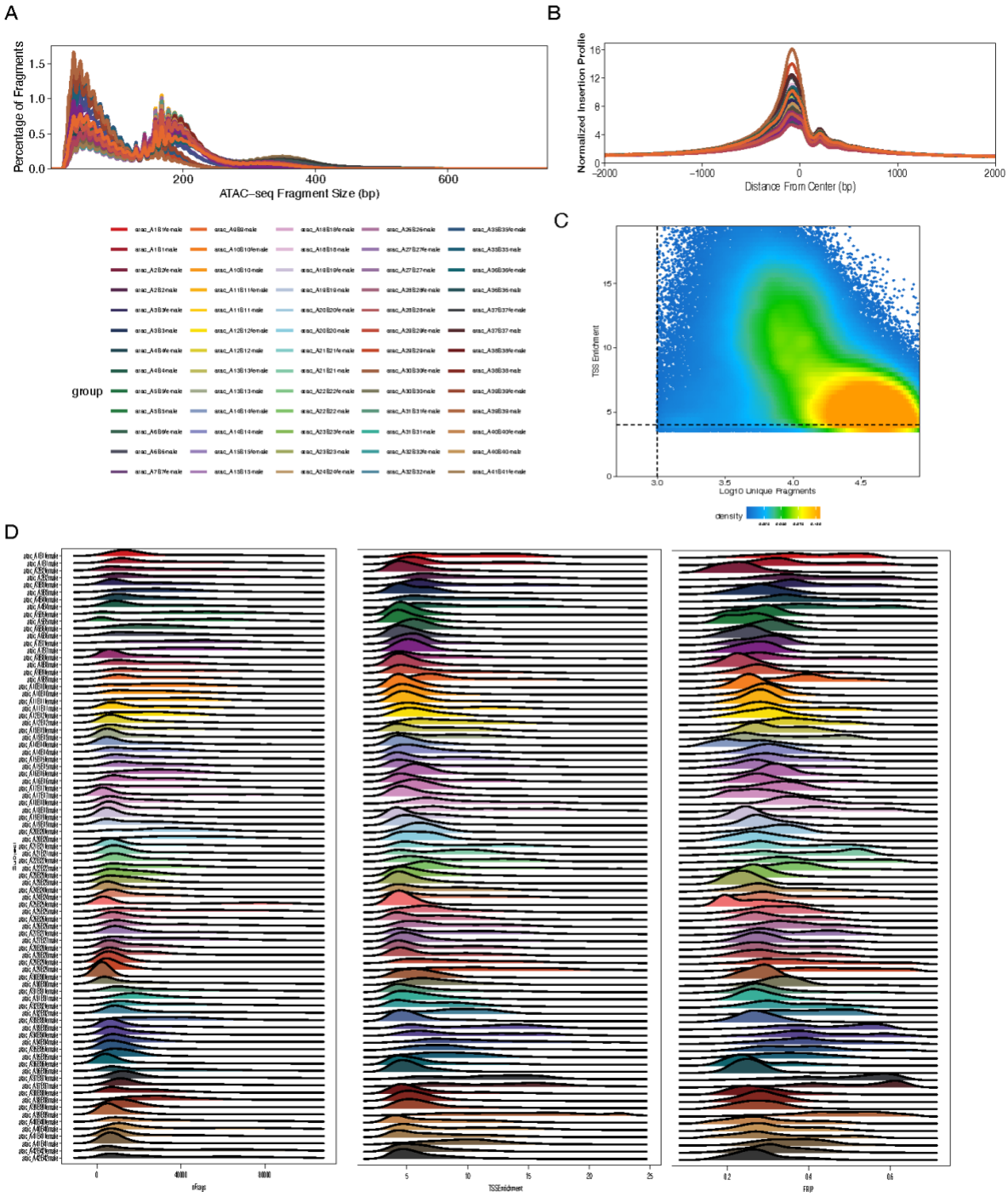
Supplementary Figure 1: Demultiplexing of pooled male and female subjects. A. UMAP plot of 201,456 cells clustered based on chromatin accessibility in 500bp genomic bins and colored by sex. B. The table shows R2 values reflecting cluster, individual and sex-specific variation

explained by UMAP1 and UMAP2 embeddings. C. PCA of top 1% most variable cCREs based on pseudobulk accessibility counts per subject and colored by the sex identified for each demultiplexed subject. D. Heatmap scaled by Pearson's correlations between genotypes estimated for each snATAC-seq subject based on chromatin accessibility and actual genotypes obtained from each individual. E. Pseudobulk chromatin accessibility (normalized by reads in TSS), at cCRE overlapping sex-specific gene, *XIST*, in each of the demultiplexed subject.



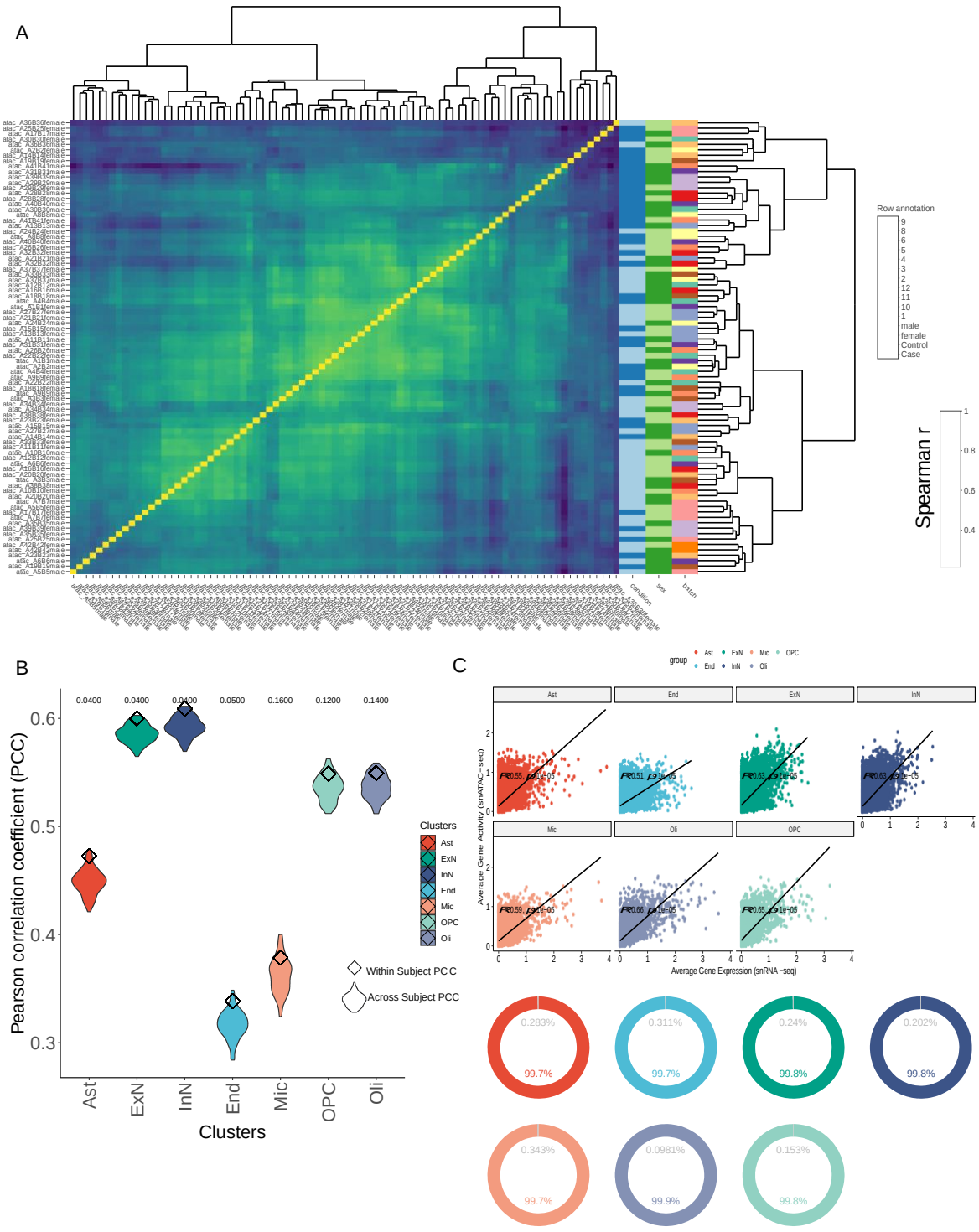
Supplementary Fig 2: **Quality-control of snATAC-seq clusters:** A. UMAP plots of 201,456 cells clustered based on chromatin accessibility in 500bp genomic bins before (left) and after

(right) batch correction with Harmony, and colored by each batch. B. UMAP plot colored by condition (MDD and control cells). C. Bar plots show proportions of cells contributed by each batch in each cluster. D. Bar plots show proportions of cells contributed by each subject in each cluster. E. UMAP plot (left) colored by PMI and corresponding bar plot (right) shows proportions of cells contributed by PMI groups in each cluster. F. UMAP plot (left) colored by age of individual subjects and bar plots (right) show proportions of cells contributed by each age group in each cluster. G. Violin plots depict distribution of fraction reads in peak (FRIP), a measure of snATAC-seq quality, across cells in each cluster.



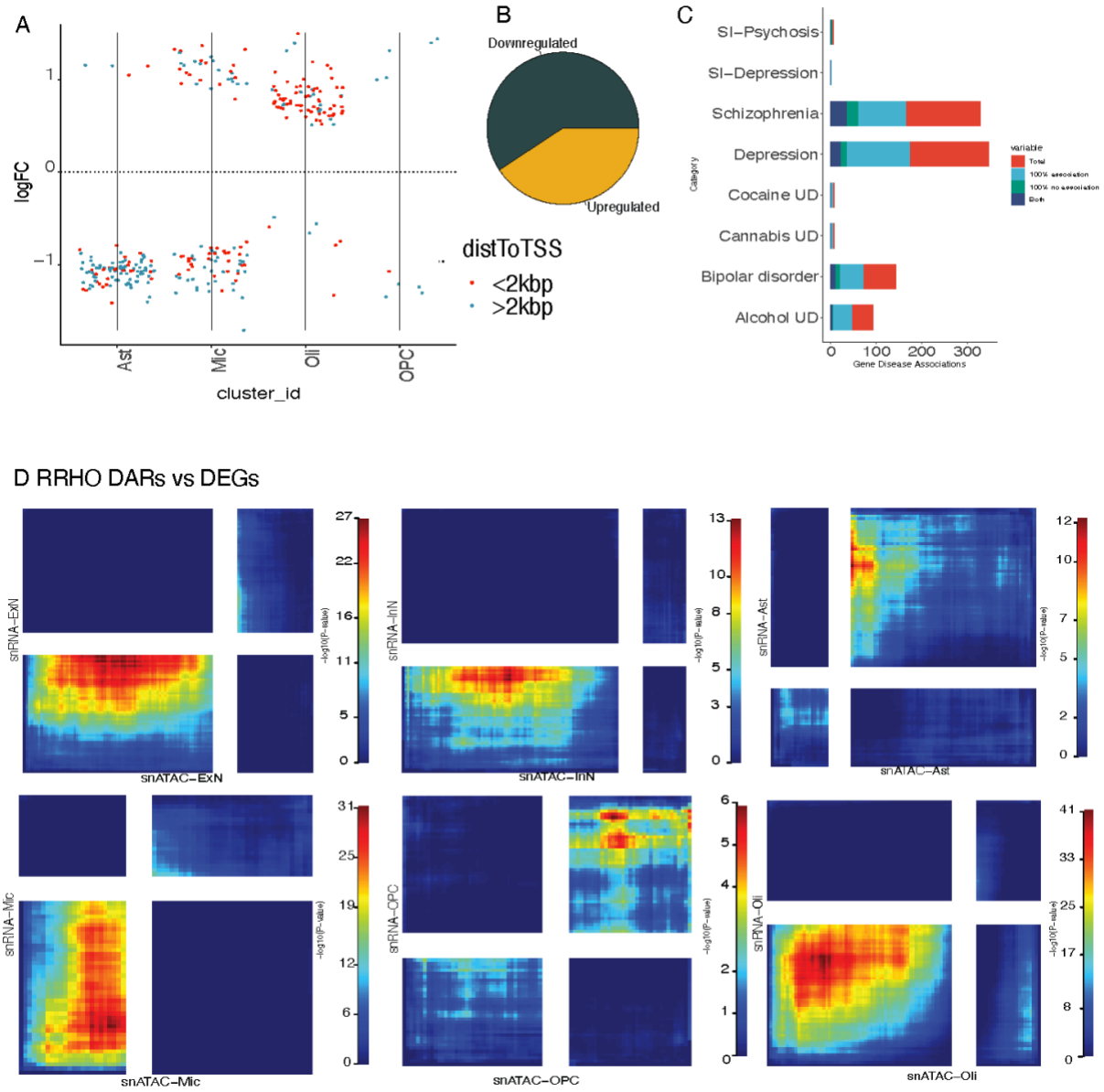
Supplementary Figure 3: **Quality-control of snATAC-seq subjects**. A. snATAC-seq fragment-size distribution in each subject. B. Normalized Tn5 insertions profiles aggregated across transcription start sites (TSS) in each subject computed using ArchR. C. Density scatter plot

depicting relationship between TSS Enrichment and log(10) unique fragments (nFrags) across 201,456 cells. D. Ridge plots showing distribution of the total number of fragments (nFrags), TSS enrichment scores, and fraction reads in peaks (FRIP) in each of the demultiplexed subjects.



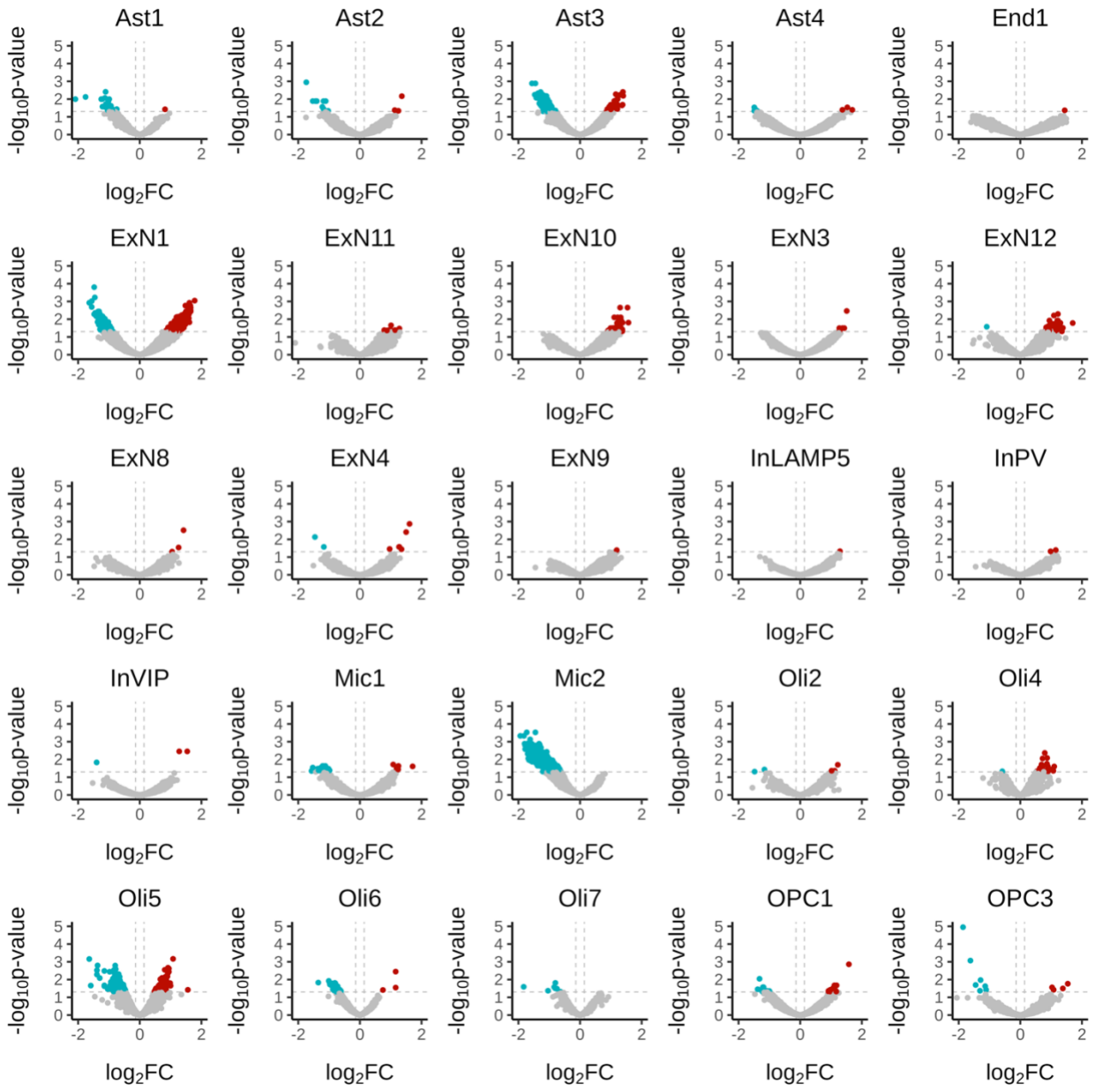
Supplementary Figure 4: snATAC-seq and snRNA-seq integration: A. Heatmap colored by Spearman's correlations (mean=0.49) based on psuedobulk accessibility log counts per million in

each cCRE for each subject and annotated by batch, sex, and condition. B. Comparison of Pearson's correlation between snATAC-seq promoter accessibility with snRNA-seq gene expression within and across subjects (one-sided permutation tests, p values are shown on top). Violin plots represent null distribution of a matched-number of across-subject correlations randomly sampled 100 times, without replacement, in each cluster. The diamond represents correlation between the same subject in snATAC-seq and snRNA-seq. C. Scatter plots show Pearson's correlations (R, p value are shown) between snATAC-seq promoter-accessibility and snRNA-seq gene-expression across all genes in each cell type. Donut plots show that less than 1% of genes (in grey) in each cell-type show high accessibility but low gene-expression (Genes in top 20% of highest promoter accessibility are categorized as "high accessibility" genes and "low gene-expression" are genes in bottom 20% of gene expression, as described previously[38]).

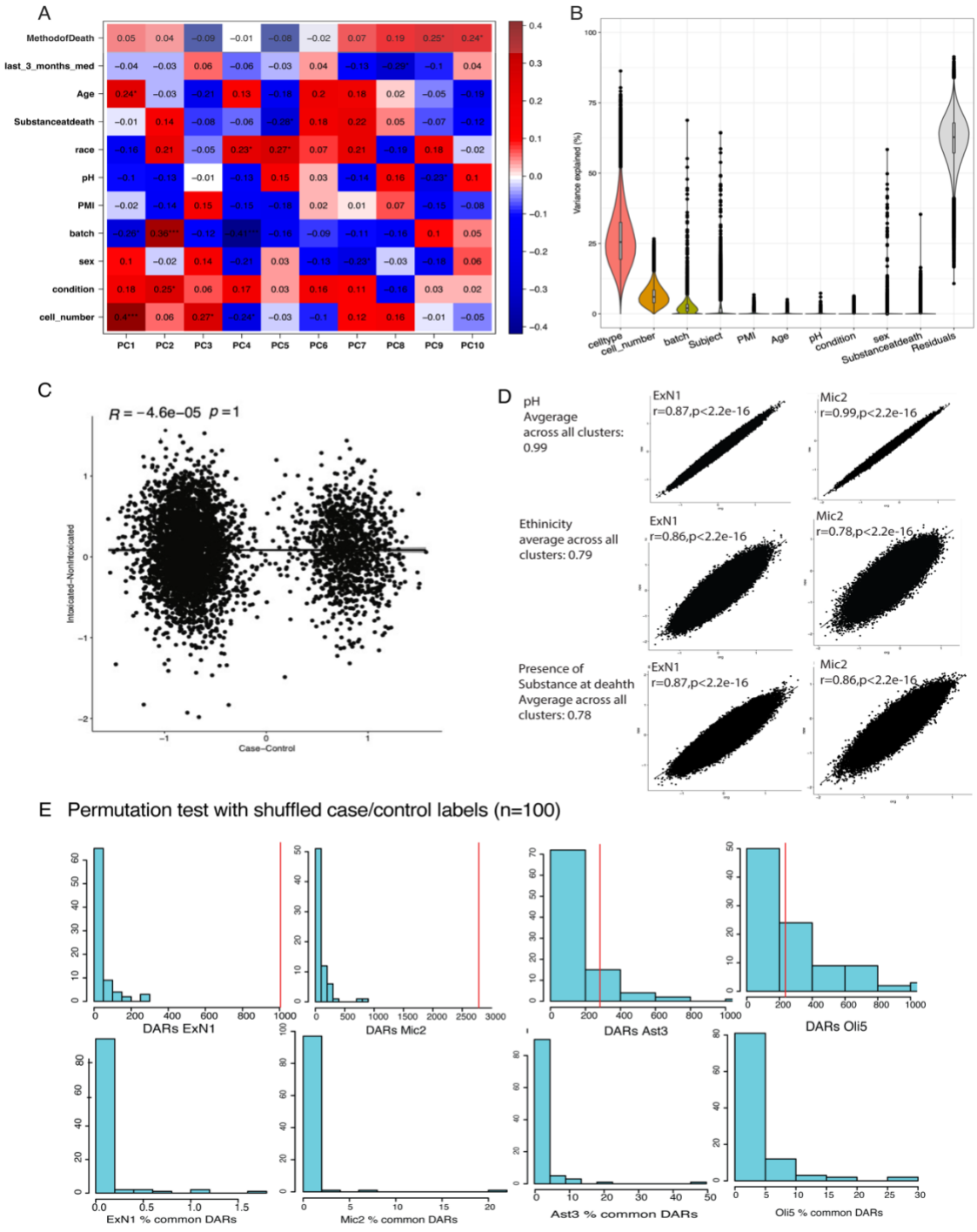


Supplementary Figure 5: Differential accessibility in MDD. A. Differentially accessible regions (DARs) in MDD (n=44) versus control (n=39) subjects in each cell type (Limma-voom; FDR < 0.05 & abs(LogFC) > log2(1.1)) split by the direction of effect in MDD versus controls and colored by the distance to the nearest TSS. B. Pie chart depicts the total number of DARs across cell types that were differentially more (upregulated, n=121) or less accessible (downregulated, n=176) in MDD. C. PsyGeNet analysis for psychiatric disorders represent the total number of gene-disease associations (GDA) for genes linked to cell-type DARs (peak-to-

gene linkages, r>0.45). For every GDA in PsyGeNET, an evidence index is provided, which keeps track of positive and negative findings. “100% association” indicates all evidence is in support, “100% no association” indicates the opposite, while “Both” indicates mixed support. SI: substance induced. D. Rank Rank Hypergeometric Overlap (RRHO2) based threshold-free comparisons between differentially expressed genes (DEGs) associated with MDD in snRNA-seq and genes whose expression showed the highest correlation (peak-to-gene linkages) with accessibility of a DAR peak in each cell-type colored by -log10(p-values) (one-sided hypergeometric tests). Color in the bottom left and top right quadrants reflect overlap in genes with concordant increased or decreased expression respectively, in MDD versus controls in snATAC-seq and snRNA-seq datasets. Colors in other quadrants reflect overlaps in genes with discordant effect-size directions between the two datasets. For each dataset, genes were ranked by the product of the sign of log fold change multiplied by the -log10 p-value from differential analyses.

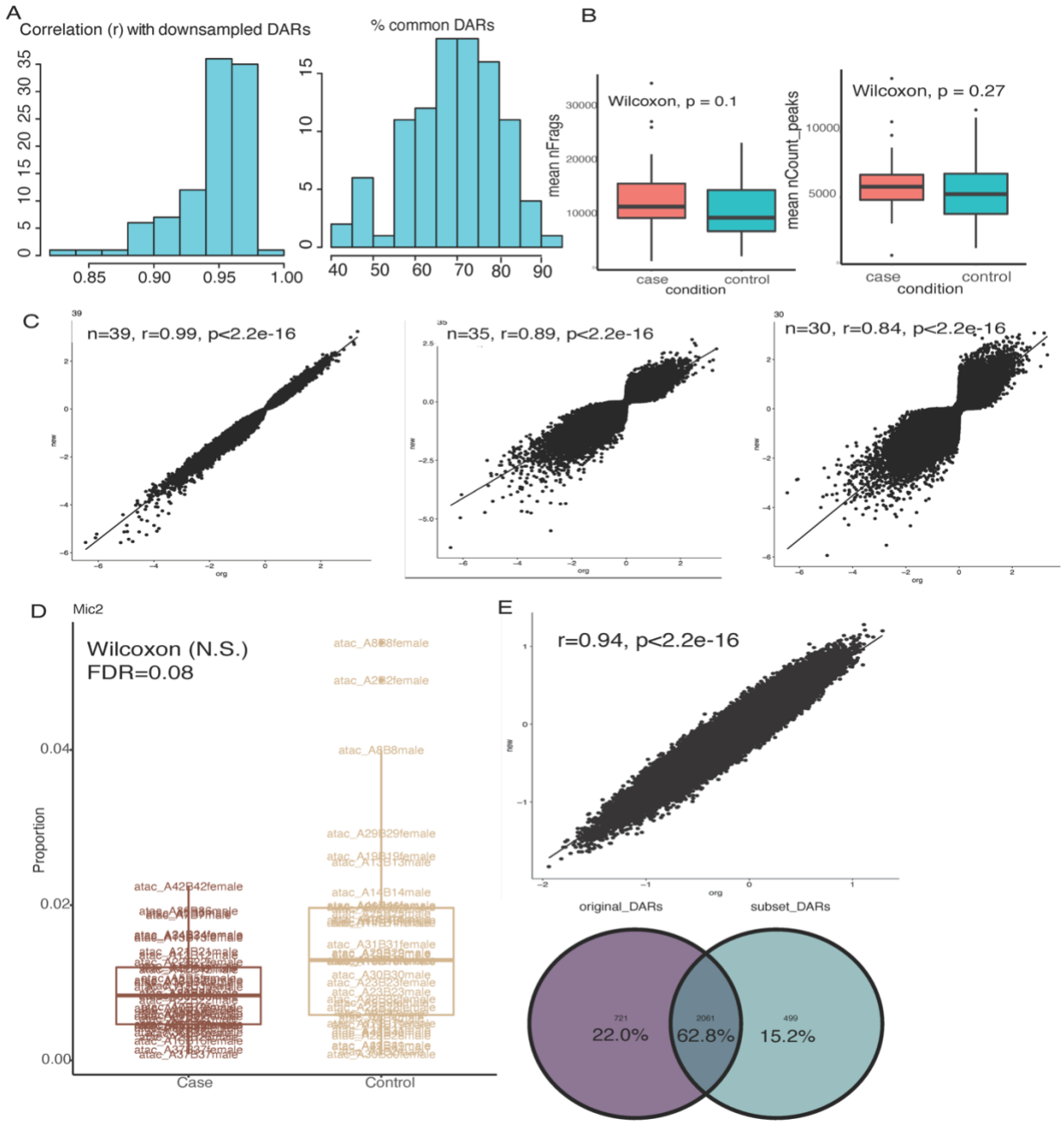


Supplementary Figure 6: Differential accessibility between MDD versus controls in each cluster: Volcano plots showing more (up-regulated) or less accessible (down-regulated) DARs in MDD versus controls in differential accessibility analysis performed in each cluster using limma-voom in muscat (DARs accessible in more than 3% of cells on an average in either cases or controls are shown). Non-significant cCREs were colored in grey.



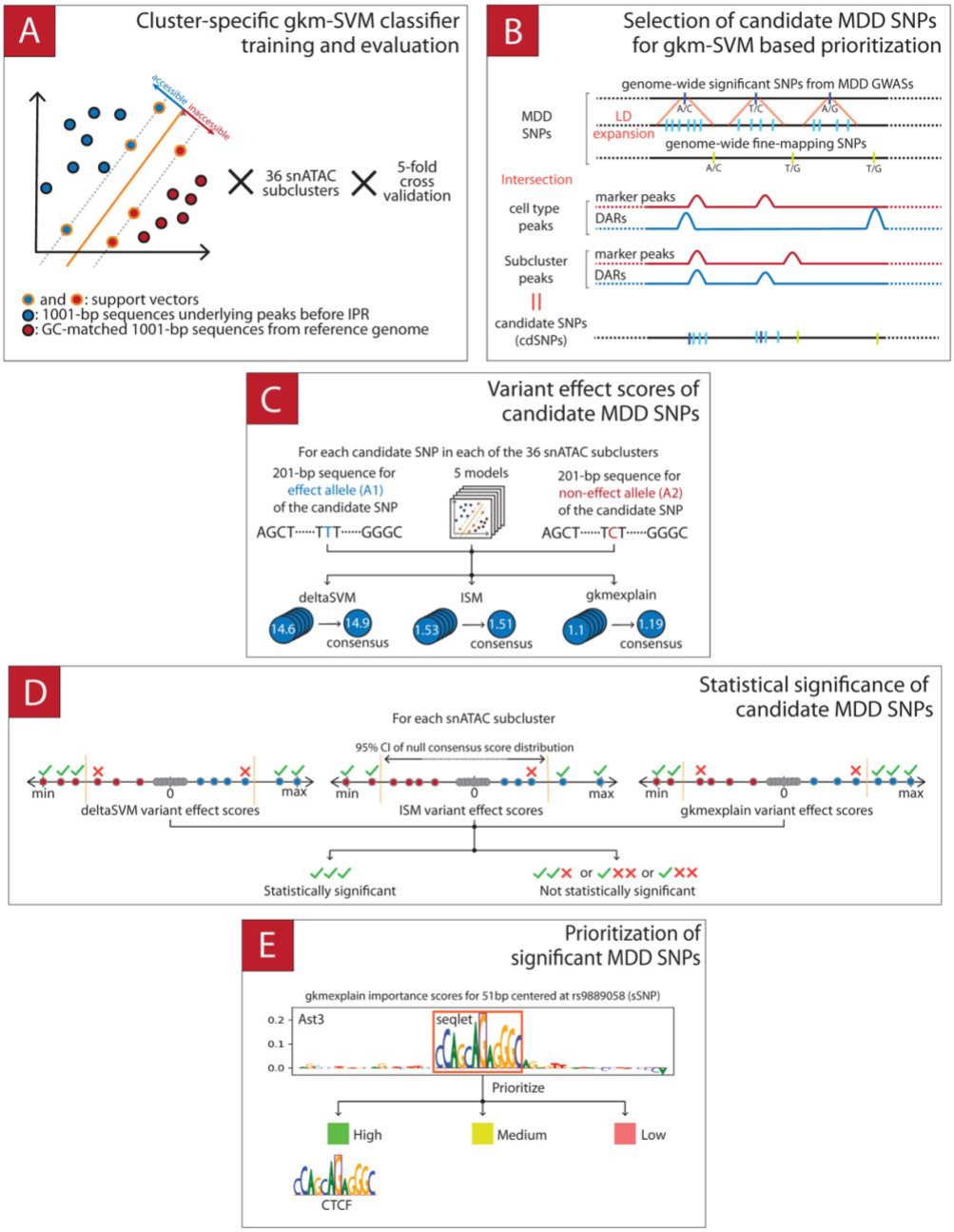
Supplementary Figure 7: Evaluation of the robustness of differential accessibility in MDD: A. Pearson's correlation between potential confounders and the top 10 principal components (PCs)

identified in PCA of per-subject pseudobulk log2 normalized counts per million. * Indicates significant correlation (FDR adjusted $p<0.05$). B. Variance partition analysis on pseudobulk log2 normalized counts per million for each cell-type in each subject (n=83). % variance explained by the potential covariates is shown (plotted by the median fraction of variance explained). C Correlation between differences in mean values of normalized log2 counts per million in cases vs. controls and intoxicated vs. non-intoxicated cases (22% cases died by a suicide with drug intoxication) (two-tailed Pearson's correlation test, R=-4.6e-05, p=1). D. Effect-size (logFC) correlation between peaks tested in the original and new differential analysis in each cluster after including potential covariates (two-tailed Pearson's correlation test, pH: mean r (across all clusters) =0.99, ethnicity: mean r =0.79, presence of substance at death: mean r=0.78). E. Differential accessibility analysis in each cluster (limma-voom in muscat) performed by permuting case/control labels 100 times within each batch. Histogram (above) shows the distribution of the number of DARs in each permutation test and red line depicts original no. of DARs in those clusters. Histogram (below) shows distribution of % common DARs (FDR adjusted $p<0.05$ & abs(logFC)>log2(1.1) in each of the permuted test with the original results. Permutation test results are shown for top 4 clusters (based on the no. of DARs). Boxplot represents median (50th percentile, line within the box), interquartile range (IQR; 25th to 75th percentiles, bounds of the box), and the whiskers (extend to 5th and 95^{th} percentile, data points within 1.5× IQR), Outliers beyond the whiskers are shown as dots.



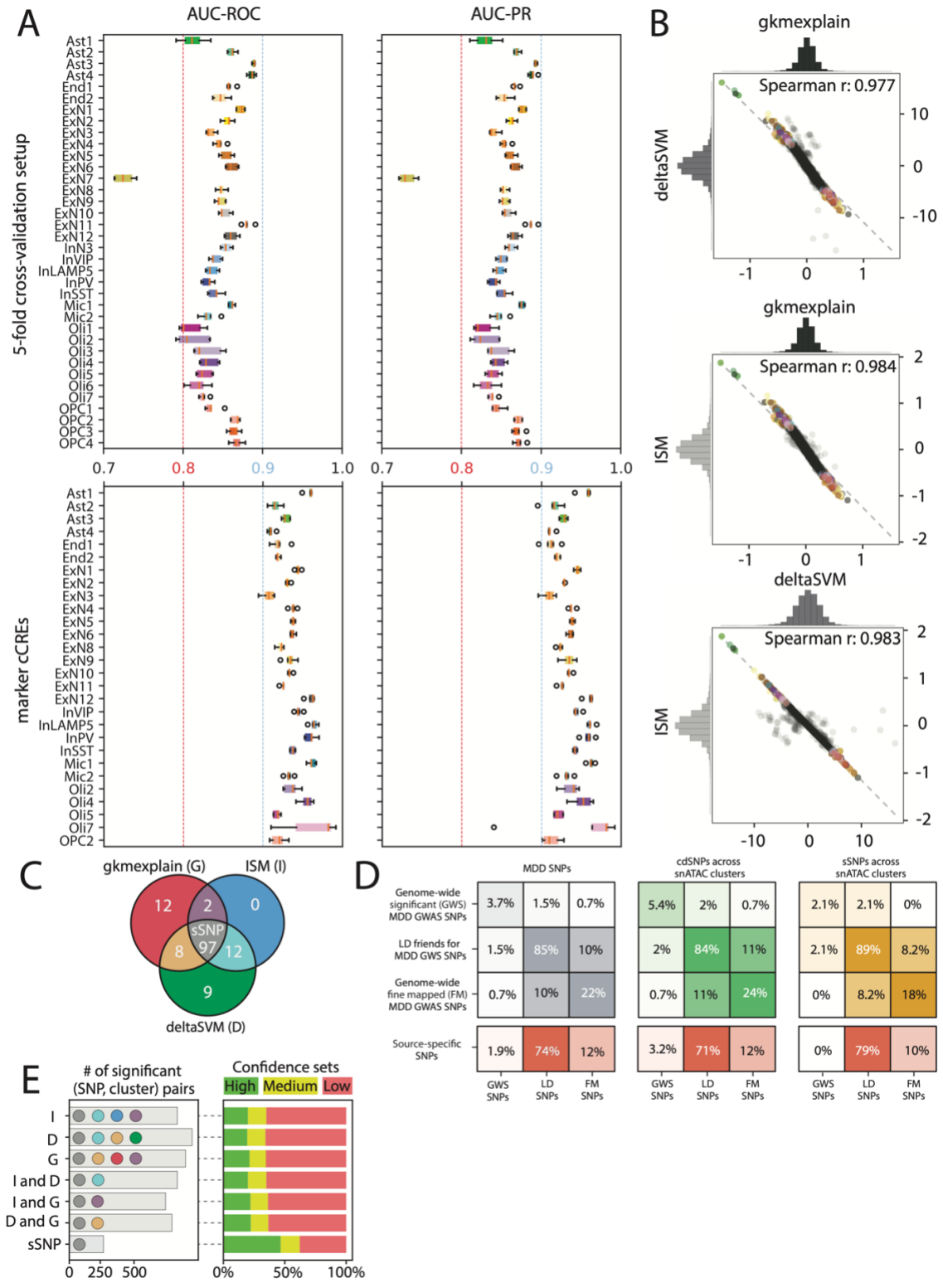
Supplementary Figure 8: Evaluation of the robustness of MDD-associated differential accessibility in Mic2 cluster: A. Sensitivity analysis performed by randomly downsampling MDD subjects (n=44), 100 times, to match the number of control subjects (n=39) in Mic2 cluster. Histogram (left) shows the distribution of effect-size (logFC) Pearson's correlation between differential peaks in original and downsampled data. Histogram (right) shows the % common DARs (FDR adjusted p<0.05 & abs(logFC)>log2(1.1) between the original and

downsampled data from each iteration (n=100). B. Boxplot shows comparison between MDD vs control subjects in Mic2 based on a) mean number of read-pairs across cells (non-duplicates, non-mitochondrial, and high mapq>30) (two-sided Wilcoxon test, p=0.1), and b) mean fragments in cCREs across cells in each subject (two-sided Wilcoxon test, p=0.27). C. Effect-size Pearson's correlation between original Mic2 DARs and DARs obtained from downsampling subjects in both cases and controls to an equal size (n=39, n=35, and n=30 subjects, respectively). D. Boxplot shows proportions of Mic2 cells in each subject compared between MDD versus control groups (two-sided Wilcoxon test, FDR adjusted p=0.08). E. Top: Effect-size Pearson's correlation between original Mic2 DARs and those obtained after removing control subjects (n=3) detected as potential outliers based on Mic2 cell proportions (>=1.5*IQR, shown in Supplementary Fig 8D). Bottom: Venn diagrams show % overlap between original Mic2 DARs (FDR adjusted p<0.05 & abs(logFC)>log2(1.1) and new analysis after removing 3 outlier control subjects. Boxplot represents median (50th percentile, line within the box), interquartile range (IQR; 25th to 75th percentiles, bounds of the box), and the whiskers (extend to 5th and 95^{th} percentile, data points within 1.5× IQR), Outliers beyond the whiskers are shown as dots.



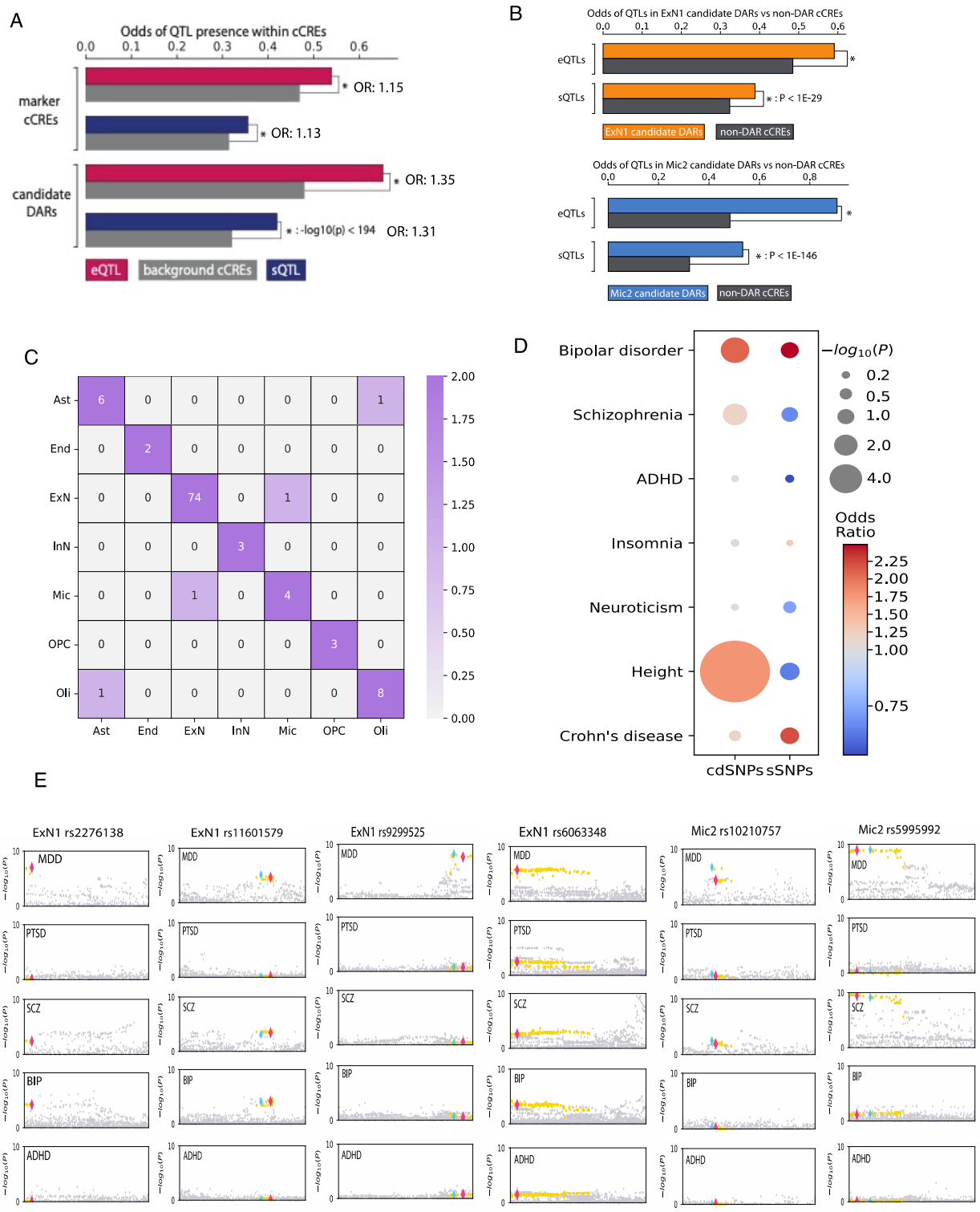
Supplementary Figure 9: Schematic describing gkm-SVM based variant-scoring workflow. A. Cluster-specific gkm-SVM classifiers were trained to predict cCRE status of input 1001-bp genome sequences via 5-fold cross validation. B. MDD SNPs located within 1001-bp marker

cCREs and candidate DARs are pooled to obtain candidate SNP set per cluster (cdSNPs). C. gkm-SVM-based cdSNP variant effect scores were computed by running ISM, deltaSVM and gkmexplain on 201-bp allelic sequences of cdSNPs. D. Statistical significance of cdSNPs were assessed with respect to score-specific null t-distributions fitted to null variant effect scores computed on shuffled (di-nucleotide preserved) 201-bp allelic sequences. For each score type and cluster, cdSNPs whose variant effect scores unanimously lie outside of 95% CI of respective null distributions were coined as statistically significant (sSNP). E. sSNPs were further partitioned into three groups according to the prominence and magnitude of their allelic impact to their immediate (seqlet) as well as 51-bp neighborhood. Additionally, seqlets of sSNPs having high or medium importance were interrogated for potential TFBS matches in CIS-BP 2.0 (TOMTOM, FDR adjusted p value (q)< 0.1). Please see supplementary methods for implementation specific details regarding the workflow, and supplementary results for reliability analysis of workflow components.



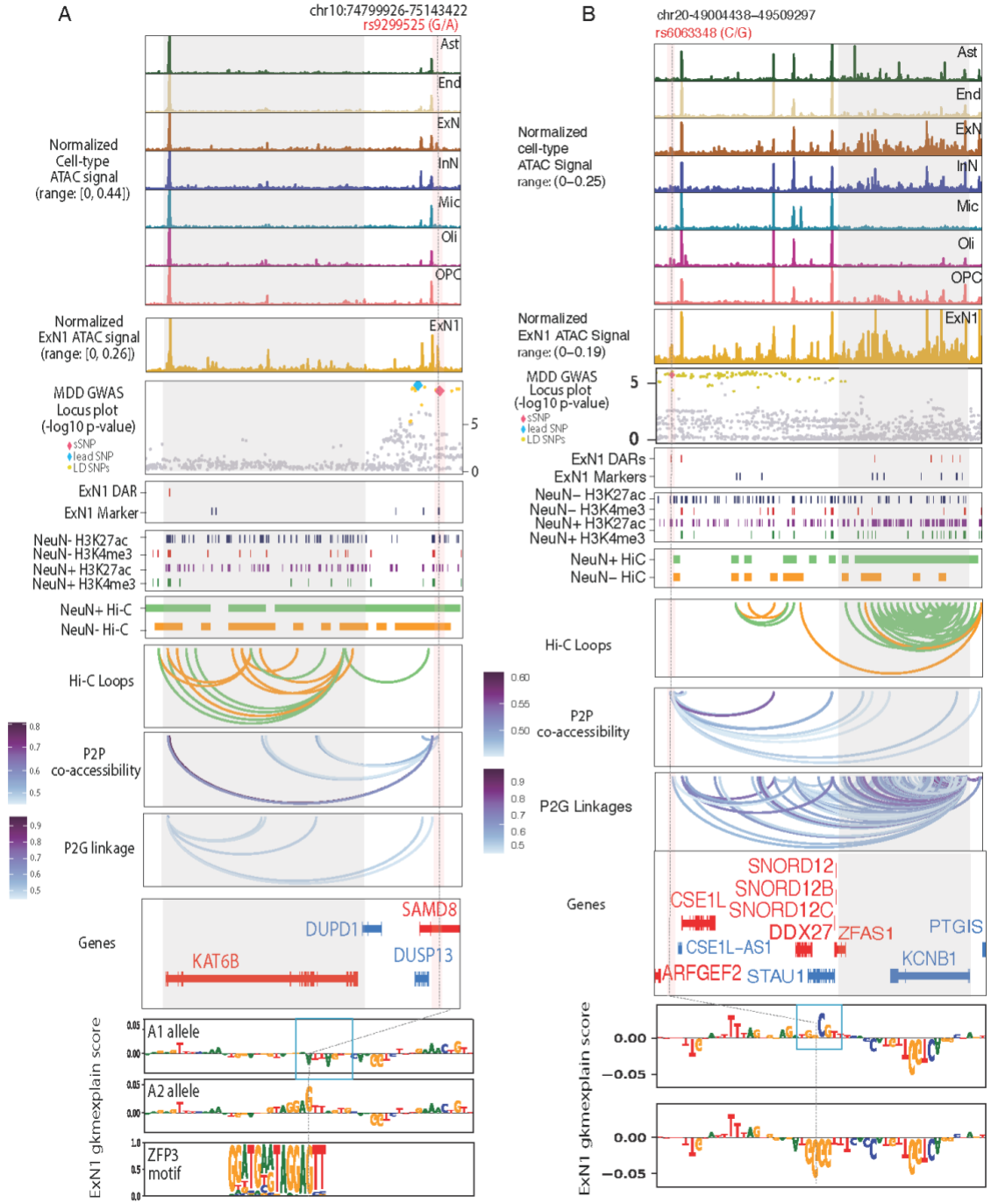
Supplementary Figure 10: **Cluster-specific gkm-SVM models for identification of MDD sSNPs.** A. Performance comparison of cluster-specific gkm-SVM classifiers across five-fold

cross-validation test splits (n=5, top) and marker cCREs (n=5, bottom) with respect to AUC-ROC (left) and AUC-PR (right). B. Pairwise correlation of cdSNP variant effect scores across snATAC clusters. C. Venn diagram showing the overlap between unique cdSNPs statistically significant according to ISM, deltaSVM and/or gkmexplain scores (97 sSNPs were defined as those significant across the three methods). D. Composition of MDD SNP (left), cdSNP (middle) and sSNP (right). E. The distribution of cluster-specific confidence set membership (right) for cdSNPs having marginal, pairwise or consensus statistical significance in at least one snATAC cluster (left). The circles inside bars in left panel are colored according to venn diagram (panel C). Boxplot represents median (50th percentile, line within the box), interquartile range (IQR; 25th to 75th percentiles, bounds of the box), and the whiskers (extend to 5th and 95^{th} percentile, data points within 1.5× IQR), Outliers beyond the whiskers are shown as dots.



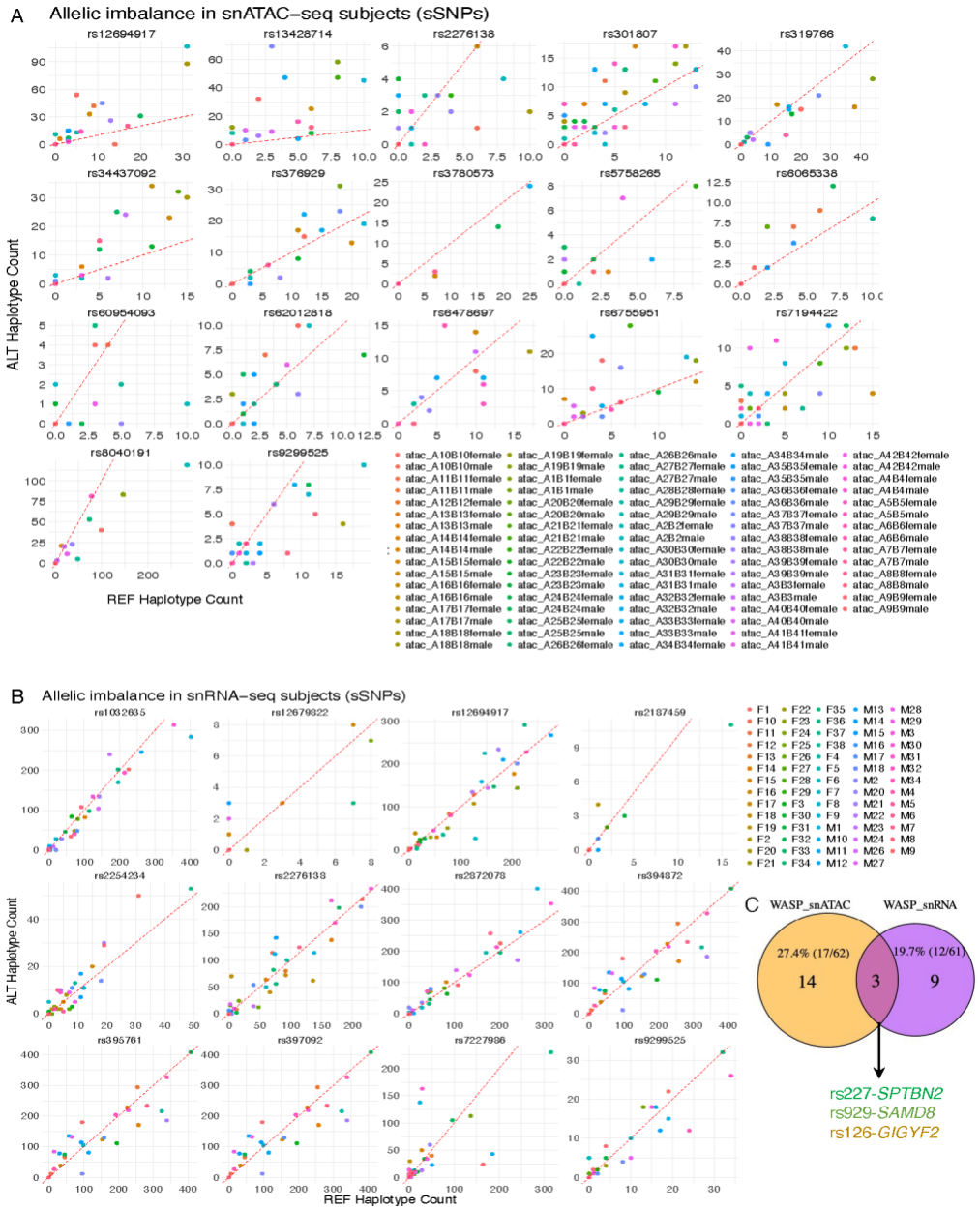
Supplementary Figure 11. Evaluation of MDD sSNPs. A. Overlap enrichment (one-sided Fisher's exact test, p values are shown) of DLPFC eQTLs and sQTLs with marker cCREs and

candidate DARs compared to non-marker or non-DAR cCREs, respectively. B. Overlap enrichment (one-sided Fisher's exact test, p values are shown) of DLPFC eQTLs and sQTLs with ExN1 and Mic2 cluster-specific candidate DARs compared to non-DAR cCREs across clusters. C. Heatmap showing sSNP overlap across cell-types. Each cell in the plot is scaled with respect to the total number of sSNPs available for the cell type in the column. D. Dot plot showing enrichment or depletion (two-sided Fisher's exact test) of cdSNPs and sSNPs in psychiatric disorder and control trait (height) associated significant loci (1-MBp windows centered at the lead SNPs) compared to a background set of SNPs. Eact dot size is scaled by the odds ratio and colored by -log10(unadjusted p value). Blue dots show depletion while red dots show enrichment. E. Manhattan plots show highlighted ExN1 and Mic2 sSNPs compared between MDD-associated GWAS loci (-log10(unadjusted p value) and other psychiatric disorders.

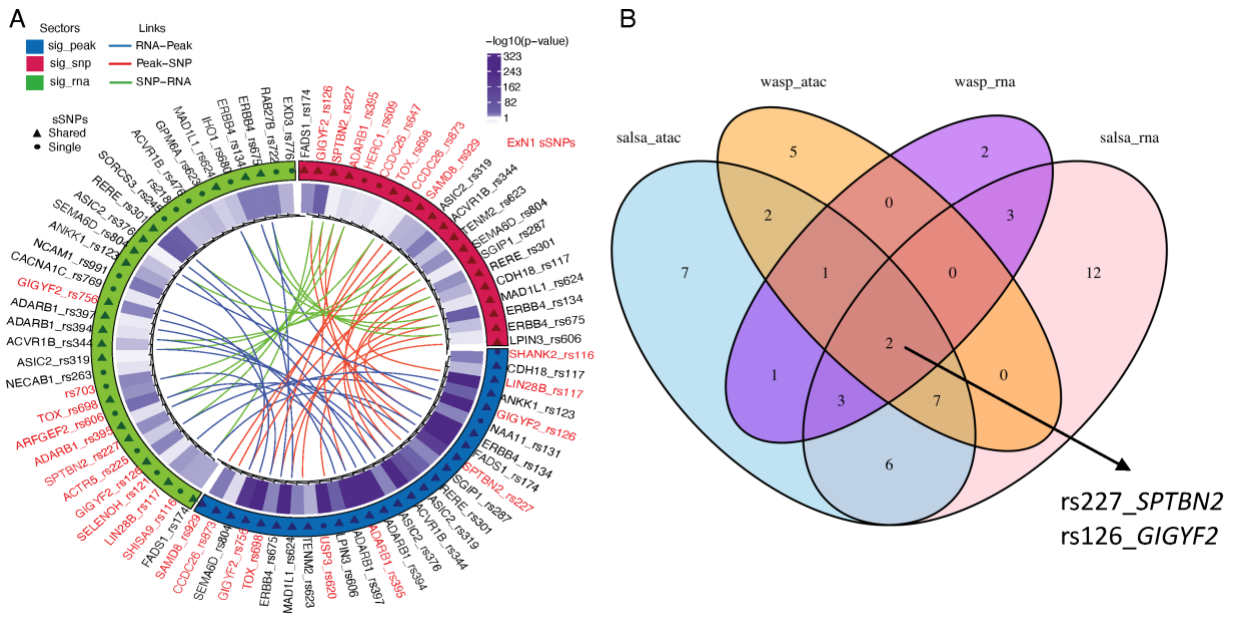


Supplementary Figure 12. MDD sSNPs with allelic-effects exclusively in ExN1 cluster. A. Multi-modal visualization plot for ExN1 sSNP (rs9299525). Top to bottom: snATAC-seq-derived pseudo-bulk tracks for each cell-type and ExN1 cluster (normalized by reads in TSS), Manhattan

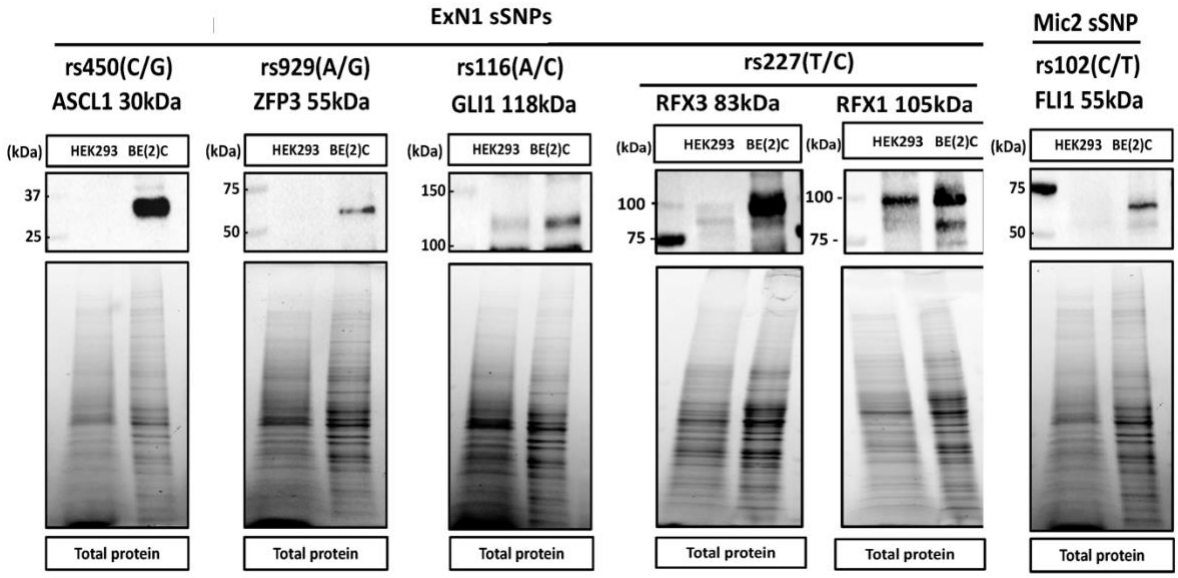
plot showing lead SNP and SNPs in LD (r>0.8) with MDD sSNP (red, chr10: G>A) overlapping ExN marker cCRE (r=0.4, linked gene: *KAT6B*), ExN1 candidate DARs and marker cCREs within the plotted region, DLPFC neuronal and non-neuronal (NeuN+/-) H3K27ac and H3K4me3 ChIP-seq peaks, DLPFC NeuN+/- promoter-anchored Hi-C fragments and 3D loops, snATAC-seq peak-to-peak co-accessible loops (r>0.45) present within 10kbps of sSNP associated cCRE, *KAT6B* gene locus (chr10:74824927-75032624), gkmExplain importance scores for each base in 50-bp region surrounding sSNP for A1 and A2 alleles from ExN1-specific gkm-SVM model corresponding, TF binding site (ZFP3) disrupted by sSNP. B. Multi-modal visualization plot for ExN1 sSNP (rs6063348). Top to bottom: snATAC-seq-derived pseudo-bulk tracks for each cell-type and ExN1 cluster (normalized by reads in TSS), Manhattan plot showing lead SNP and SNPs in LD (r>0.8) with MDD sSNP (red, chr20: C>G) overlapping ExN1 candidate DAR (FDR adjusted p=0.15) (r=04, linked: *KCNB1*), ExN1 candidate DARs and marker cCREs within the plotted region, DLPFC neuronal and non-neuronal (NeuN+/-) H3K27ac and H3K4me3 ChIP-seq peaks, DLPFC NeuN+/- promoter-anchored Hi-C fragments and 3D loops, snATAC-seq peak-to-peak co-accessible loops (r>0.45) present within 10kbps of sSNP associated cCRE, *KCNB1* gene locus (chr20-49293394-49484297), no known TF motif was significantly disrupted by sSNP.



Supplementary Figure 13: **Allelic imbalance by WASP.** A. Dot plots show WASP-corrected allelic read-counts for each individual (adjusted for GC content and total read depth) in snATAC-seq data comparing reference versus alternate alleles for each significant MDD sSNP detect using WASP (two-sided CHT, unadjusted p<0.1, in color). B. Dot plots shows WASP-corrected allelic read-counts for each individual (adjusted for GC content and total read depth) in snRNA-seq data comparing reference versus alternate alleles for each significant MDD sSNP (two-sided CHT, unadjusted p<0.1, in color). C. Venn diagram shows overlap between snATAC-seq and snRNA-seq based MDD sSNPs with significant allelic imbalance (WASP) in donor genotypes.



Supplementary Figure 14: **Allelic imbalance by SALSA**. A. Circos plot shows MDD sSNPs with significant allelic-imbalance across donor genotypes (two-sided exact binomal test, unadjusted p<0.1, Supplementary Table 2). Here, allelic accessibility counts were computed at single cell level using SALSA. The significance was measured at three levels: "sig_snp" represent sSNPs with significant differences between reference versus alternate allele counts across subjects, "sig_peak" represent sSNPs with significant allelic read-count differences across all the heterozygous variants in sSNP-containing cCREs, and "sig_rna" represent sSNPs with significant allelic read-count differences across exonic variants in high LD (r>0.8) with non-coding MDD sSNP. Circos plot is divided into three sectors based on each category (sig_snp, sig_peak, sig_rna). sSNPs are shaped based on the shared effects (triangle sSNPs are significant in atleast two categories and dots represent sSNPs that are significant only in one category). Heatmap shows -log10(unadjusted p value) for each sSNP in each category. Links connect the sSNPs shared across the categories. Each sSNP is associated with the nearest gene and colored (in red) if an ExN1-associated MDD sSNP with significant allele-imbalance (two-sided exact binomal test, unadjusted p<0.1). B. Venn diagram shows overlap between MDD sSNPs with significant allelic imbalance in WASP (two-sided CHT, unadjusted p<0.1) and SALSA (two-sided binomial test, unadjusted p<0.01) in snATAC-seq and snRNA-seq donor genotypes.

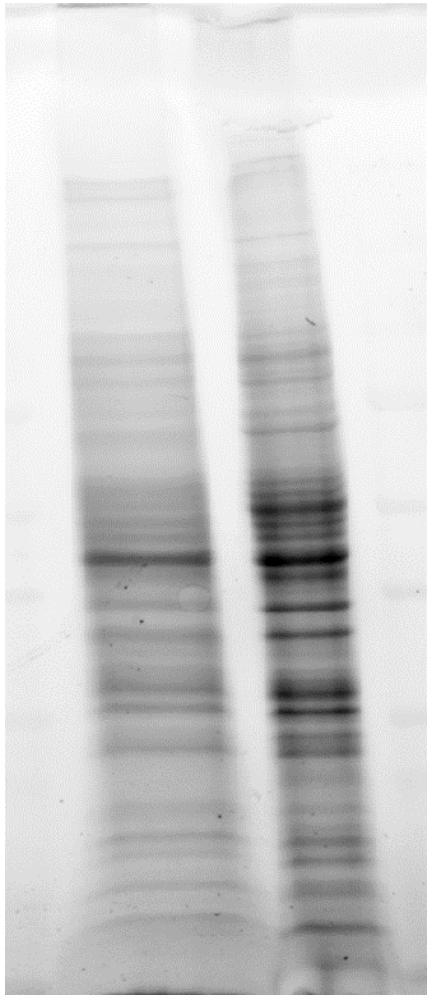


Supplementary Figure 15: **TF expression to validate MDD risk variants in luciferase assay**. Western blots show endogenous protein expression in HEK293 and BE(2)C (neuronal) cell lines for each of the TFs identified to be disrupted by MDD sSNPs in ExN1 (n=4) and Mic2 clusters (n=1).

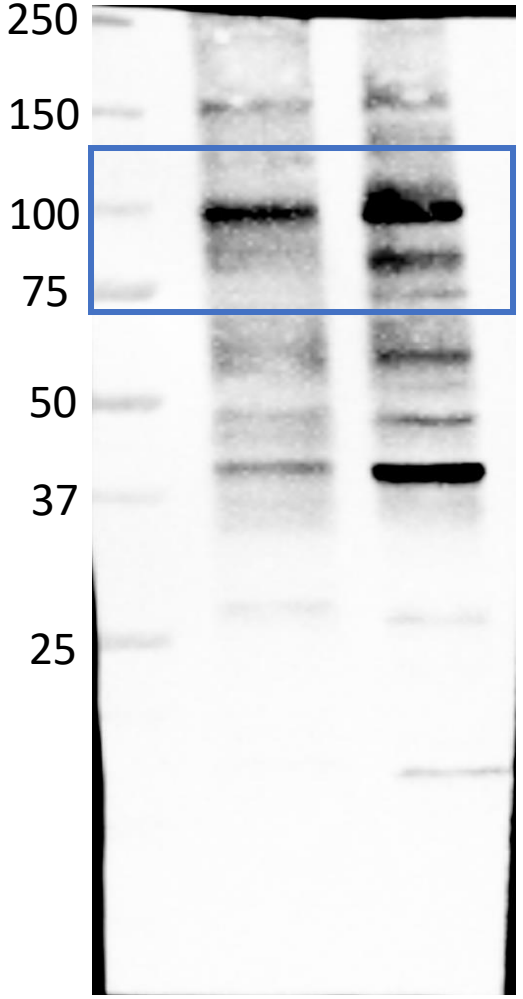
RFX1 105kDa

Total protein

HEK293 BE(2)C



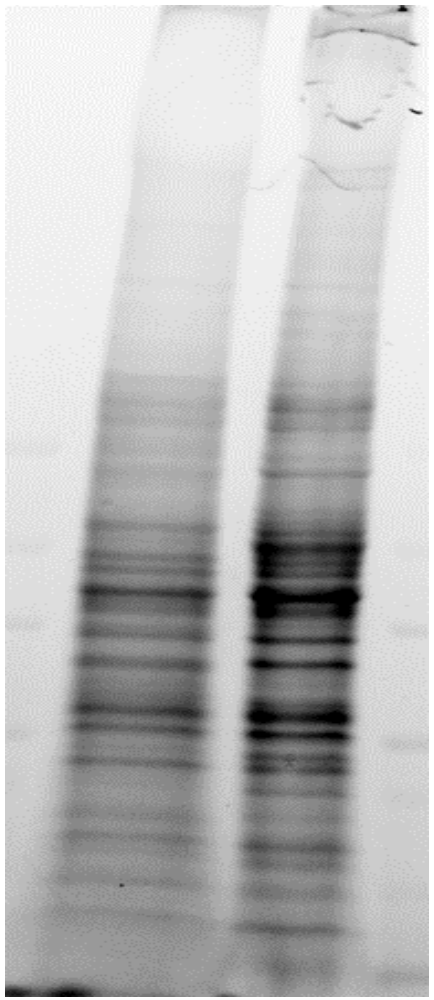
HEK293 BE(2)C



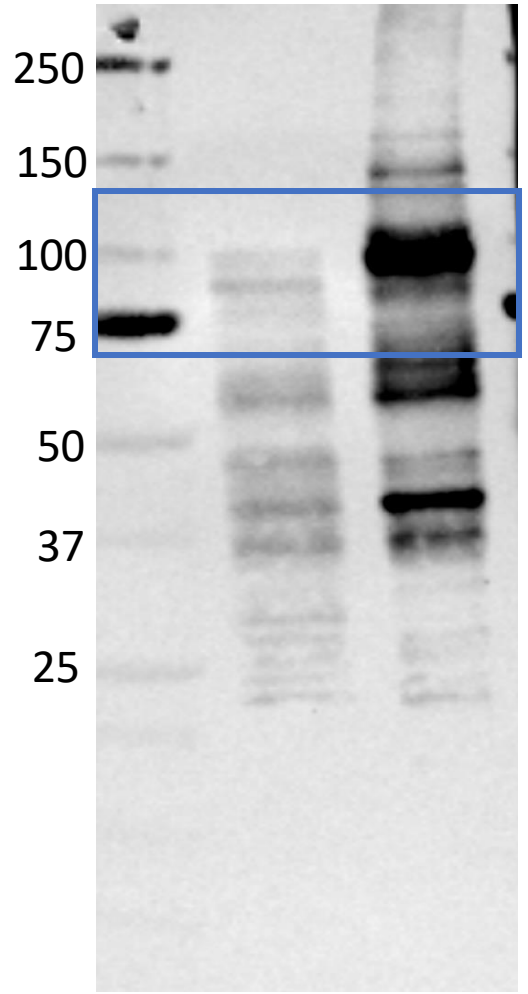
RFX3 83kDa

Total protein

HEK293 BE(2)C



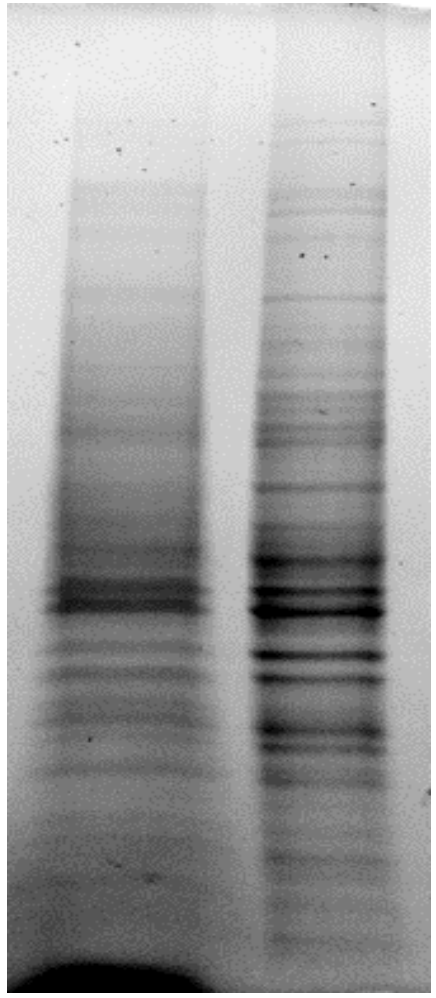
HEK293



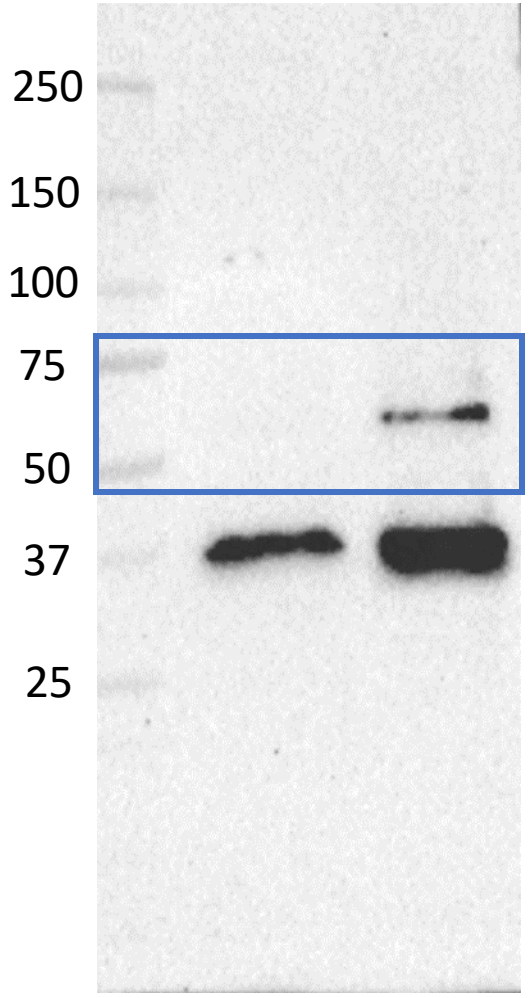
ZFP3 55 kDa

Total protein

HEK293 BE(2)C



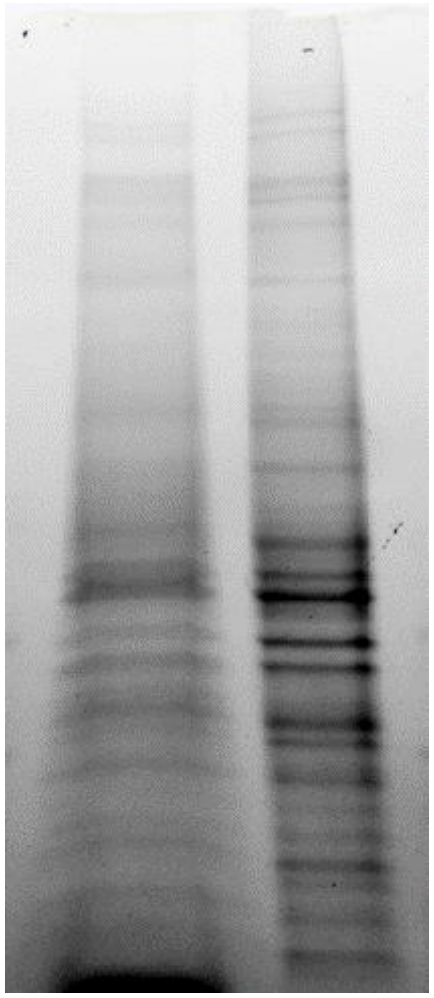
HEK293 BE(2)C



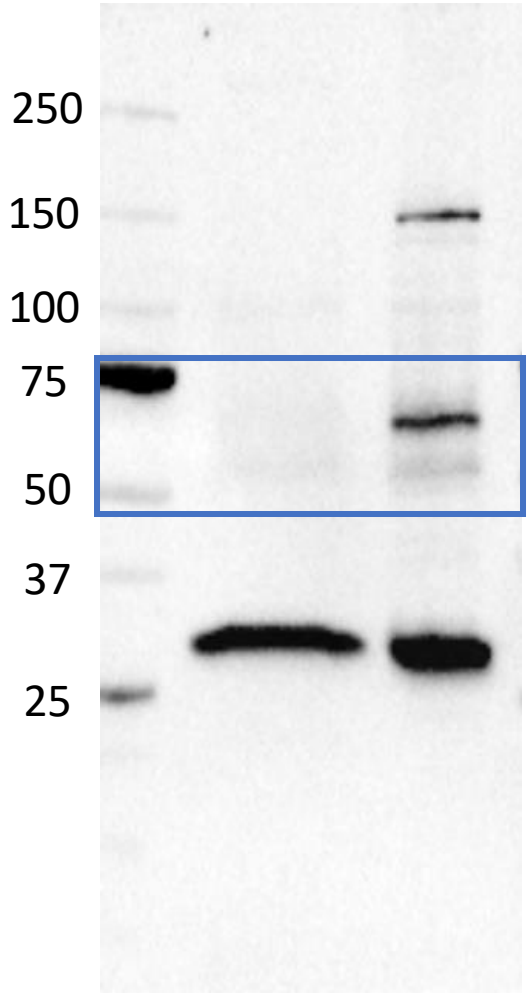
FLI1 55kDa

Total protein

HEK293 BE(2)C



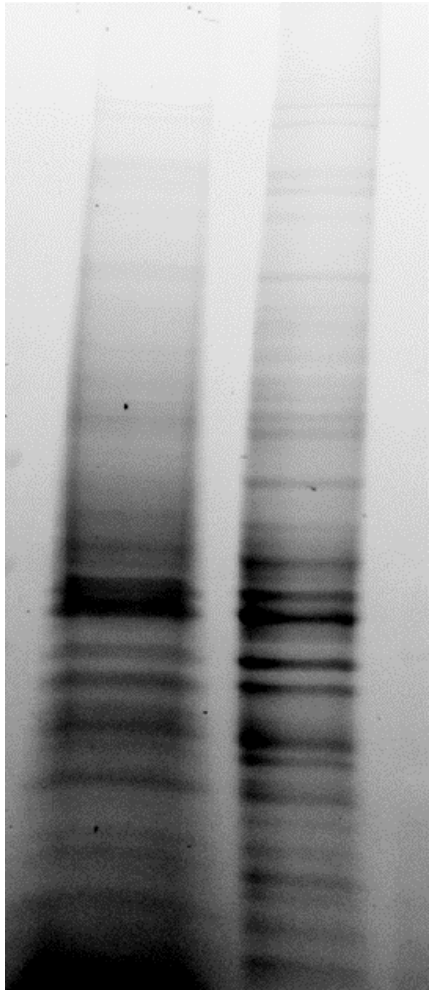
HEK293 BE(2)C



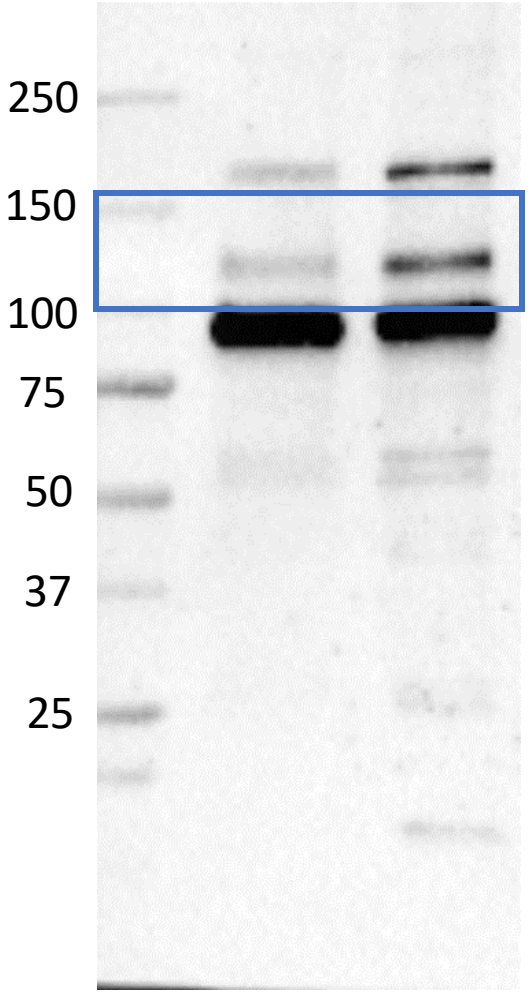
GLI1 118 kDa

Total protein

HEK293 BE(2)C



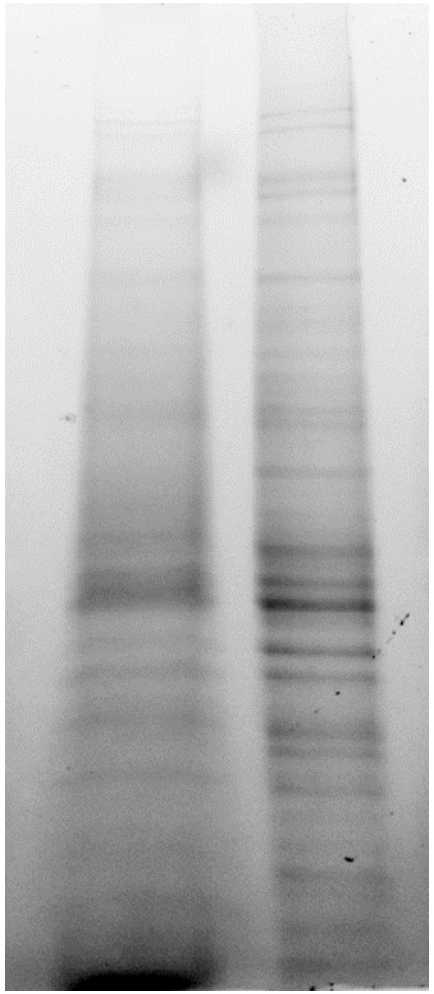
HEK293 BE(2)C



ASCL1 30 kDa

Total protein

HEK293 BE(2)C



HEK293 BE(2)C