Published in final edited form as:

Nat Neurosci. 2021 July ; 24(7): 941–953. doi:10.1038/s41593-021-00858-w.

Cell-type specific effects of genetic variation on chromatin accessibility during human neuronal differentiation

Dan Liang[1,2], **Angela L. Elwell**[1,2], **Nil Aygün**[1,2], **Oleh Krupa**[1,2], **Justin M. Wolter**[1,2], **Felix A. Kyere**[1,2], **Michael J. Lafferty**[1,2], **Kerry E. Cheek**[1,2], **Kenan P. Courtney**[1,2], **Marianna Yusupova**[3,4,5], **Melanie E. Garrett**[6], **Allison Ashley-Koch**[6,7], **Gregory E. Crawford**[8,9], **Michael I. Love**[1,10], **Luis de la Torre-Ubieta**[3,4,5,11], **Daniel H. Geschwind**[3,4,5,11], **Jason L. Stein**[1,2,12]

[1)]Department of Genetics, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA

[2)]UNC Neuroscience Center, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA

[3)]Neurogenetics Program, Department of Neurology, David Geffen School of Medicine University of California, Los Angeles, Los Angeles, CA 90095, USA

[4)]Center for Autism Research and Treatment, Semel Institute, David Geffen School of Medicine University of California, Los Angeles, Los Angeles, CA 90095, USA

[5)]Department of Human Genetics, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA 90095, USA

[6)]Duke Molecular Physiology Institute, Duke University, Durham, NC 27701, USA

[7)]Department of Medicine, Duke University, Durham, NC 27708, USA

[8)]Center for Genomic and Computational Biology, Duke University, Durham, NC 27708, USA

[9)]Department of Pediatrics, Division of Medical Genetics, Duke University, Durham, NC 27708, USA

[10)]Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA

[11)]Department of Psychiatry and Biobehavioral Sciences, Semel Institute, David Geffen School of Medicine University of California, Los Angeles, Los Angeles, CA 90095, USA

[12)]Lead Contact: jason_stein@med.unc.edu.

Author Contributions

JLS, DHG, and LTU conceived the study. JLS directed and supervised the study. JLS along with DHG provided funding. ALE, KEC, KPC, MY, LTU, and JLS cultured human neural progenitor cells. ALE performed library preparation. MJL pre-processed the RNA-seq data for eQTL. NA performed eQTL analysis. OK performed immunocytochemistry. OK, JMW, FAK and DL performed the functional validation assays. MEG, AAK, GEC provided access to adult dlPFC caQTL data. MIL aided in ASCA methodology. DL performed pre-processing, differential accessibility, caQTL, ASCA, co-localization, and motif analyses. JLS and DL wrote the manuscript. All authors commented on and approved the final version of the manuscript.

Declaration of Interests

The authors declare no competing interests.

Abstract

Common genetic risk for neuropsychiatric disorders is enriched in regulatory elements active during cortical neurogenesis. However, it remains poorly understood how these variants influence gene regulation. To model the functional impact of common genetic variation on the non-coding genome during human cortical development, we performed ATAC-seq and analyzed chromatin accessibility quantitative trait loci in cultured human neural progenitor cells and their differentiated neuronal progeny from 92 donors. We identified significant genetic effects on 988/1,839 neuron/progenitor regulatory elements, with highly cell-type and temporally specific effects. A subset (~30%) of caQTLs were also associated with changes in gene expression. Motif-disrupting alleles of transcriptional activators generally led to decreases in chromatin accessibility, whereas motif-disrupting alleles of repressors led to increases in chromatin accessibility. By integrating cell-type specific caQTLs and brain-relevant genome-wide association data, we were able to fine-map and identify regulatory mechanisms underlying non-coding neuropsychiatric disorder risk loci.

Genome-wide association studies (GWAS) have revealed hundreds of common single nucleotide polymorphisms (SNPs) that are associated with risk for neuropsychiatric disorders and interindividual differences in brain structure[1,2]. A crucial next step is to understand the molecular mechanisms underlying the effect of these variants[2]. This is complicated by many factors, including unknown causal variant(s) at an associated locus due linkage disequilibrium (LD)[3], unknown cell-type(s), tissue-type(s), or developmental time period(s) in which a genetic risk variant exerts its effects[4], and unknown regulatory function of non-coding risk variants[5]. Nevertheless, a commonly assumed model to explain molecular mechanisms underlying risk loci is that non-coding risk alleles disrupt transcription factor (TF) binding within cell-type specific regulatory elements (REs) leading to alterations in gene expression and downstream impacts on risk[6,7]. Thus, understanding genetic effects on cell-type specific regulatory activity is an essential aspect of moving from genetic association to a meaningful biological understanding of disorder risk.

With this in mind, several consortia including ENCODE, GTEx and PsychENCODE have taken major steps to build maps of non-coding genome function across the body[8–12]. These and other efforts have connected non-coding genetic variation to genes in developing and adult brain tissue by profiling 3-dimensional chromatin interactions and by measuring the genetic effects on gene expression, called expression quantitative trait loci (eQTLs)[13,14]. Although these studies are an important first step in connecting non-coding risk loci to genes, they do not elucidate how the gene is regulated via genetic variation.

Risk variants for multiple neuropsychiatric disorders are enriched in REs active at mid-gestation in humans, during cortical neurogenesis[15]. Histone acetylation QTLs (haQTLs) and chromatin accessibility QTLs (caQTLs) are powerful tools to identify the effect of genetic variation on non-coding REs and provide further understanding of regulatory mechanisms at tissue or cell-type specific levels[11,16,17]. However, the ability to connect human genetic variation to longitudinal changes in regulatory architecture during brain development is limited by the inaccessibility of brain tissue from the same individual over multiple time points. Here, we leveraged a well validated model of human brain development based on *in vitro* culture of primary human neural progenitors[18] to study the functional effects of genetic variation on chromatin architecture during neurogenesis. We measured chromatin accessibility[19] in cell culture in a cell-type specific manner at two key stages of neural development, during progenitor proliferation (N_{donor}=76) and after differentiation using their sorted neuronal progeny (N_{donor}=61). We identified thousands of caQTLs and allele specific chromatin accessibility (ASCA) sites, the majority of which were highly cell-type specific. We use the effects of these genetic variations to understand how disrupting TF binding motifs impact chromatin accessibility and gene expression, as well as to understand the cell-type specific regulatory mechanisms underlying genetic risk for neuropsychiatric disorders.

Results

Genome-wide chromatin accessibility profiling

We generated primary human neural progenitor cell (phNPC) lines from 14–21 gestation weeks genotyped fetal brains (N=92) using a neurosphere isolation method that results in cultures with high fidelity to the *in vivo* developing brain[18] (Figure 1a; Methods). phNPCs were cultured and isolated at two stages: progenitor cells and 8-week differentiated and sorted neurons (Extended Data Figure 1a–1b). Using immunofluorescence of neural cell markers, we found over 90% of the progenitor cells were positive for SOX2 and PAX6, indicating a highly homogenous population of radial glia cells[20] (Figure 1a; Extended Data Figure 1c). After 8-weeks of differentiation, we FACS sorted neurons labeled using a viral construct (AAV2-hSyn1-EGFP) which showed typical neuronal morphology (Figure 1a; Extended Data Figure 1b and 1d; Methods). We performed ATAC-seq on intact nuclei and found that libraries were high quality based on a comparison of quality metrics relative to previous *in vivo* developing brain data, as well as a sensitivity analysis and nucleosome periodicity (Extended Data Figure 1e and 2a–2b; Methods)[15,19]. We quantified accessibility as batch-effect-corrected reads within accessible peaks normalized for GC content, peak length and sequencing depth (Extended Data Figure 2c, 2e and 2f). We found higher correlations of chromatin accessibility for libraries from the same donors cultured at different times as compared to correlations across donors (Extended Data Figure 2d). To ensure independence for subsequent analyses, we randomly selected one library from each donor for each cell-type (N_progenitor=76 and N_neuron=61) to identify accessible peaks (N=90,227; average peak length of 409 bp; Methods).

To determine the *in vivo* relevance of these accessible peaks, we performed an overlap analysis utilizing previously classified chromatin states from 93 *in vivo* human tissues and cell types (Figure 1b; Methods). The accessible peaks from progenitors and neurons most strongly overlap with enhancers and promoters in brain germinal matrix and fetal brain tissue, followed by other brain regions, indicating that these peaks were highly representative of the *in vivo* fetal brain. Principal component analysis of chromatin accessibility revealed that progenitors and neurons clearly separate along the first principal component (Figure 1c), indicating that cell-type was associated with the largest variability in chromatin accessibility profiles (64.91% of variance explained). These results demonstrate that chromatin accessibility measured from phNPC cultures are representative of REs present in the developing human brain and that chromatin accessibility patterns are different between progenitors and neurons, consistent with previous data from fetal brain tissues[15].

Identifying cell-type specific regulatory elements

To reveal cell-type specific REs involved in neuronal differentiation, we performed an analysis to determine which peaks had significantly different chromatin accessibility between progenitors and neurons (Figure 1d; Methods). We identified 35,379 peaks with greater accessibility in progenitors than neurons (progenitor peaks) and 44,729 peaks with greater accessibility in neurons than progenitors (neuron peaks; FDR < 0.05; Supplementary Table 1). At the promoter of *SYN1*, which was used to label neurons for sorting after differentiation, we observed considerably higher accessibility in neurons, as expected (LFC= −2.88, P-value=2.83e-55; Figure 1d). Among significant differentially accessible peaks, we found greater accessibility in progenitors at the promoters of genes highly or uniquely expressed in progenitors, such as the dorsal telencephalic marker *EMX2* (Figure 1e)[21]. Moreover, promoters of genes highly expressed in neurons, such as *DCX*, *BDNF*, *CAMK2B* and *SYT13*[22], showed greater chromatin accessibility in neurons (Figure 1d–1e).

We found an expected enrichment of Gene Ontology terms related to neurogenesis in genes with differentially accessible promoters (Extended Data Figure 3a; Methods). We also found that differentially accessible peaks were significantly enriched in ATAC-seq peaks from the relevant *in vivo* fetal brain laminae (Figure 1f)[15]. Specifically, progenitor peaks were more enriched in peaks with higher accessibility in the progenitor-enriched germinal zone. Conversely, neuron peaks were more enriched in peaks with higher accessibility in the neuron-enriched cortical plate. These results showed differentially accessible peaks represent cell-type specific active REs and were in strong agreement with biological processes and gene regulatory behavior present in *in vivo* fetal brain tissues.

To detect TFs involved in neuronal differentiation, we conducted a differential motif enrichment analysis to predict TF binding sites more active in either progenitors or neurons. We detected 62 TFs (FDR < 0.05) with binding sites present more often in progenitor peaks than neuron peaks (here, called progenitorTFs), and 208 TF motifs presents more often in neuron peaks than progenitor peaks (neuronTFs) (Methods; Extended Data Figure 3b; Supplementary Table 2). Within progenitorTFs and neuronTFs, we found TFs previously characterized with key roles in neurogenesis, which provides further support that TF binding within accessible peaks from this *in vitro* system reflect the expected *in vivo* developmental processes (Extended Data Figure 3c; Supplementary Table 2)[15]. We also identified several TFs that have not been previously associated with neuronal differentiation, such as *MEF2A*, *MIX-A* and *HOXB5*, which may be useful for directed differentiation of human neural progenitors.

Chromatin accessibility quantitative trait loci (caQTLs)

To identify genetic variants that influence chromatin accessibility within cell types representing longitudinal changes during cortical development, we performed caQTL analyses separately for progenitors and neurons using in total 90,227 peaks and 10M genetic variants (Figure 2a; Extended Data Figure 4a–4b). We stringently controlled for population stratification (Extended Data Figure 4c) in association tests using a mixed linear model including a kinship matrix as a random effect, and 10 genotype MDS components as fixed effects[23,24]. In addition, we included principal components (PCs) across chromatin accessibility profiles and sorter in neurons as fixed effect covariates to reduce the impact of unmeasured technical variation[25].

After the stringent multiple testing correction (Methods), we identified 1,839 progenitor caPeaks and 988 neuron caPeaks at FDR < 5%. caPeaks were significantly enriched in active REs defined in the fetal brain (Extended Data Figure 4d), consistent with their expected regulatory function. The most significant caSNPs of each caPeak are most often found near the peaks they are associated with (Figure 2b). We found that the most significant caSNP or an LD-proxy was located in annotated functional regions for over 80% of caPeaks (Figure 2c; Supplementary Table 3). These results imply that most genetic variants affect chromatin accessibility by altering the sequence (and presumably transcription factor binding sites) at the caPeak or disrupt chromatin accessibility at distal peaks which have secondary effects on the caPeak[7].

To identify if genetic influences on chromatin accessibility also affected gene expression, we compared progenitor/neuron caQTLs with eQTLs derived from the same cell lines and eQTLs from the mid-gestation bulk cortical wall[14,26]. For the most significant caSNPs for each caPeak, we estimated the posterior probability that the effect is shared with cell-type specific or cortical wall eQTLs (m-value > 0.9; Supplementary Table 4). Thirty percent of progenitor caQTLs and 34.9% of neuron caQTLs have shared effects with eQTLs in the same cell types, but a smaller proportion are shared with bulk cortical wall eQTLs (Figure 2d). Those SNPs with shared effects between caQTLs and eQTLs showed strongly positive correlations in effect sizes (r=0.85 in neurons and r=0.84 in progenitors; Figure 2e), indicating that alleles associated with increased chromatin accessibility tend to be associated with increased gene expression.

We then compared the number and effect size differences between caQTLs and eQTLs. First, we subsampled the eQTL dataset to ensure that caQTLs and eQTLs have the same sample sizes in order to avoid winner's curse[27] (Methods). The proportions of peaks (2.62%/5.81% in neuron/progenitor) or genes (1.85%/5.70%) influenced by genetic variation were comparable in caQTL or eQTLs. However, we observed that caQTLs generally explain more variance than eQTLs, implying that caQTL studies have higher power than eQTL studies[28] (Figure 2f; Extended Data Figure 4e).

Genetic effects on cell-type specific regulatory elements

We next aimed to identify the cognate genes of cell-type specific REs, fine map causal variants at eQTL loci, and predict regulatory mechanisms underlying eQTLs by overlapping cell-type specific caQTLs with cell-type specific or bulk fetal cortical eQTLs (r^2>0.8 between index ca/eSNPs; Supplementary Table 5). Using cell-type specific eQTLs, we identified ca/eQTL overlaps in 152/373 RE-Gene pairs in neurons/progenitors. Using the larger sample size of fetal cortical eQTLs, we identified 303/282 RE-Gene pairs using neuron/progenitor caQTLs. Within these RE-Gene pairs, we found many genes involved in neuronal differentiation such as *FABP7, VAT1* and *FGF1*[29–31]. We also identified RE-gene pairs where the caSNP disrupted TF motifs that have known function in neuronal differentiation. For example, the G allele of rs185220 was associated with increased chromatin accessibility of a caPeak (chr5:56,909,141–56,910,860) near the *SETD9* TSS and was associated with increased expression of *SETD9* in neurons and progenitors (Figure 3a–3c). Several TF motifs were disrupted by this caSNP, but we prioritized *REST* based on its expression in progenitors[22,26] (Figure 3d; logFC=−1.85, FDR = 2.75e-17) and evidence of binding at this site in ES cells and to a lesser degree in neurons differentiated from ES cells[8,9] (Figure 3a). The G allele of rs185220 led to disruption of the REST motif and increased chromatin accessibility, consistent with the function of REST as a repressor[32]. Through integration of ca/eQTL data, we hypothesize a regulatory mechanism where the G allele of rs185220 disrupts REST binding, resulting in increased chromatin accessibility and increased expression of *SETD9* (Figure 3e). As experimental validation of this caQTL, we found that the G allele of rs185220 increased the activity of this enhancer in progenitors relative to the A allele using a luciferase assay (Figure 3f).

In contrast with the previous example, we found the C allele of the caSNP, rs11544037, matched the motif of RAD21 and was associated with increased chromatin accessibility of the progenitor-specific enhancer (chr4:158,667,771–158,667,860 located ~5 kb upstream from the *ETFDH* TSS) and increased expression of *EFTDH* in fetal brain (Extended Data Figure 5a–5e). Experimental validation via a luciferase reporter assay showed a consistent result with the caQTL for this enhancer (Extended Data Figure 5f). As a final example of regulatory mechanisms underlying a cell-type specific eQTL, we found the C allele of rs11960262 associated with increased chromatin accessibility of a caPeak (chr5:142,684,441–142,686,700) located in the intron of the gene *FGF1* and also associated with increased *FGF1* expression specifically in progenitors (Extended Data Figure 5g–5i). The C allele of rs11960262 matched the motif of EGR1 (Extended Data Figure 5j–5k), which suggests that EGR1 binding at this caPeak was associated with increased chromatin accessibility and increased expression of *FGF1* in progenitors.

Allele Specific Chromatin Accessibility (ASCA)

We next tested for ASCA at heterozygous SNPs within accessible peaks. ASCA contrasts accessibility between two alleles within an individual at a given heterozygous SNP, so it is not susceptible to cross-individual confounding factors, such as population structure[33]. In total, we identified 1,602 significant progenitor ASCA and 3,288 significant neuron ASCA (FDR < 0.05; Supplementary Table 6). To determine if caQTLs also show ASCA, we retained significant caQTLs (non-clumped, FDR < 0.05) using the same heterozygous donors and read level criteria for ASCA, observing that 90.1% of filtered neuron caQTLs were shared with neuron ASCA (Fisher's test: OR=51.48, p-value=1.32e-228) and 86.9% of filtered progenitor caQTLs were shared with progenitor ASCA (Fisher's test: OR=45.54, p-value=1.37e-239). This demonstrates extremely high overlap between caQTLs and ASCA (Figure 4a), which indicates minimal influence of cross-individual confounding effects on the caQTL results. Similarly, for all filtered caQTLs and significant ASCA in Figure 4a, we found high correlations of effect sizes between caQTLs and ASCA (r=0.61 for neurons; r=0.69 for progenitors), indicating a shared direction and degree of effect (Figure 4b). The alternative allele showed a slight bias, even after controlling reference mapping bias, for a higher correlation with increasing chromatin accessibility from the reference allele (neuron log2(ALT/REF) are 53.4% positive, two-sided sign test p=0.034; progenitor log2(ALT/REF) are 52.5% positive, two-sided sign test p=0.121).

However, we also detected significant ASCAs that were not significant caQTLs (Figure 4b). These variants were found in larger peaks than those detected in both caQTL and ASCA (Extended Data Figure 6a). These ASCA-but-not-caQTL variants likely have an effect on the accessibility of a sub-region of the larger active RE. We posit that they are more detectable using ASCA because only reads containing the variant in the accessible region are tested, whereas they are not detectable in caQTLs, which integrate reads across the entire region (Extended Data Figure 6b). Other ASCA-but-not-caQTL sites were presumably due to lower power for caQTL detection (Extended Data Figure 6c).

We identified several loci that shared caQTLs, ASCA, and eQTLs. For example, previously described *SETD9* locus also demonstrated strong ASCA at rs185220 in neurons and progenitors (Extended Data Figure 6d). We were also interested in *FABP7* (also known as *BLBP*), which is a marker for radial glia that plays an important role in establishment of radial glial fibers spanning the cortical anlage during cortical development[30] (Figure 4c). The C allele of rs144376334 was associated with increased chromatin accessibility of the caPeak (chr6:122,832,401–122,834,160) in both progenitors and neurons and increased expression of *FABP7* (Figure 4d). The C allele of rs144376334 also showed increased ASCA in both progenitors and neurons (Figure 4e). rs144376334 disrupted several TF motifs that may drive the effect in both cell types, and we highlight JUN due to its higher expression in progenitors (Figure 4e; logFC = −1.22, FDR = 1.816794e-11)[22,26]. The motif disrupting allele was associated with decreased chromatin accessibility, consistent with activating REs (Figure 4g). These results suggest the potential regulatory mechanism underlying this locus in progenitors is that the genetic variation disrupts JUN binding to a distal RE leading to decreased expression of *FABP7*.

Cell-type specificity of caQTLs

To determine the cell-type specificity of ca/eQTLs, we estimated the posterior probability that the allelic effect is shared between the two cell types (m-value > 0.9). caQTLs showed a lower proportion of effect sharing between neurons and progenitors (45.6% and 41.0%) than eQTLs (78.0% and 56.7%) (Figure 5a). We found the estimated proportion of true alternative hypotheses that the variant is associated with the trait (π_1) of the most significant neuron/progenitor caSNP-caPeak pairs in progenitors/neurons is 0.73/0.70; however, the π_1 of the most significant neuron/progenitor eSNP-eGene pairs in progenitors/neurons is 0.92/0.77, providing additional support that caQTLs have higher cell-type specificity than eQTLs. We found 19%/35% of progenitor/neuron caPeaks overlapped with neuron/progenitor caPeaks (Figure 5b). For ASCA, we found 24%/12% of progenitor/neuron ASCA are shared between cell types, which was in agreement with the cell-type specificity observed in caQTLs. These results suggest that genetic variants often impact chromatin accessibility only within specific cell-types.

We further characterized the cell-type specificity of caQTLs by assessing differential accessibility of caPeaks (Figure 5c). We found 71.0% of progenitor caPeaks were more accessible in progenitors (LFC > 0). Similarly, 69.8% neuron caPeaks were more accessible in neurons (LFC < 0). This implies that a RE must be accessible to or bound by DNA-binding proteins within a specific cell-type in order to observe genetic effects on that RE. We next characterized the location of caPeaks relative to promoters (Figure 5d). We found there was a higher percentage of cell-type specific caPeaks that were distal to promoters than shared caPeaks. This result indicates cell-type specific caQTLs are more likely to affect the chromatin accessibility of distal REs, which is consistent with observations that distal REs have higher cell-type specificity than promoters[34].

To determine the direction and magnitude of the effect in caQTLs, we related the effect sizes of caQTLs or eQTLs between neurons and progenitors. We found that caQTLs showed a lower correlation between neurons and progenitors (r=0.73 and r=0.70) as compared to eQTLs (r=0.81 and r=0.81) (Figure 5e–5f; p value < 2.2e-16 in neurons and progenitors[35]), which is consistent with the observation that caQTLs showed a lower proportion of shared effects between neurons and progenitors than eQTLs (Figure 5a). Together, these results suggest that caQTLs have higher cell-type specificity than eQTLs, within the two cell types tested here.

Comparison to adult dorsolateral prefrontal cortex caQTLs

Previous work identified genetic variants associated with chromatin accessibility in adult post-mortem dorsolateral prefrontal cortex (DLPFC) using a sample of 272 individuals[16]. We tested whether the caQTLs identified in our work, modeling a prenatal time period, were also present in the adult cortex. We found 56% of adult peaks are shared with neurons and progenitors (Extended Data Figure 7a–b). We re-mapped caQTLs in neurons and progenitors using shared peaks and genetic variants with the adult data. We did not find any significant neuron/progenitor caQTLs shared with significant caQTLs in adult cortex. For the 27 significant neuron caQTLs, we found the correlation (r=0.61) of effect sizes with adult caQTLs are higher than the correlation (r=0.34) in 35 significant progenitor caQTLs (Extended Data Figure 7c), which may be expected given that progenitors are not present in the adult cortex. Together, these results indicate that caQTLs have high temporal specificity, as well as cell-type specificity.

Prediction of disrupted transcription factor binding

One favored model of how genetic variation influences chromatin accessibility is that SNPs disrupt TF motifs, decreasing the probability of TF binding to REs, altering chromatin accessibility[36]. To identify which TF motifs are disrupted by cell-type specific caSNPs, we mapped TF motifs to the sequence surrounding the neuron-specific/progenitor-specific caSNPs and determined if an allele at the caSNP sufficiently decreases the relative entropy of TF motifs (Methods; Supplementary Table 7). We then performed an enrichment analysis to infer which TF's binding is often affected by caQTLs within each cell type. In progenitors, we found an enrichment of caSNP disrupted *REST* and *SOX11* motifs, which are known to contribute to neurogenesis[32,37]. In neurons, we found an enrichment of caSNP disrupted RARb motifs, which is involved in prefrontal synaptogenesis and axon development[38] (Figure 6a). These results suggest that the motif disrupted TFs are involved in neuronal differentiation, indicating that the genetic variants that impact the activity of REs by disrupting the binding of TFs play functional roles during neurogenesis.

We next tested the impact of the TF motif-disrupting alleles on chromatin accessibility. Among 532 tested motifs in progenitors and 514 tested motifs in neurons, we found that motif-disrupting alleles led to decreased accessibility for 40 (72.7% of significant TFs) TFs in progenitors and 44 (97.8% of significant TFs) TFs in neurons (Figure 6b–6c), such as the motif of *POU3F2* (also known as BRN2) in progenitors and *ASCL2* in neurons, which are both involved in neurogenesis[39,40] (Figure 6d). Conversely, we found the motif-disrupting allele was associated with increased chromatin accessibility at the motif of *ZEB1*, a known transcriptional repressor[41] (Figure 6e). These results suggest that binding of transcriptional activators is associated with increased chromatin accessibility. However, binding of transcriptional repressors is associated with decreased chromatin accessibility.

Regulatory mechanisms underlying GWAS loci

To investigate if genetic variants associated with brain related traits are enriched in differentially accessible peaks during neurogenesis, we calculated partitioned heritability enrichment (Figure 7a–7b). We found cell-type specific enrichments for neuropsychiatric disorders and associated behaviors in accessible regions. Genetic variants associated with several childhood or adult onset neuropsychiatric disorders or traits, including ASD, schizophrenia, major depressive disorder (MDD), neuroticism and depressive symptoms, showed significant partitioned heritability enrichment in progenitor peaks. With the exception of schizophrenia, these disorders did not show significant enrichment in neuron peaks. We observed partitioned heritability enrichment for both intelligence and educational attainment within neuron peaks and progenitor peaks. As a negative control, we did not observe enrichment of inflammatory bowel disease (IBD) heritability in differentially accessible peaks. These results are consistent with the model that genetic variants alter the function of REs during cortical neurogenesis, which then leads to risk for neuropsychiatric disorders or related traits in childhood or adulthood[14,15].

We found genetic variants associated with the cortical global surface area showed significant partitioned heritability enrichment in progenitor peaks, as well as the surface area of multiple cortical subregions including caudal anterior cingulate, entorhinal, lateral occipital, lingual, and pericalcarine[1]. (Figure 7b). Genetic variants associated with the thickness of the entorhinal cortex, but not average thickness across the entire cortex, also showed significant partitioned heritability enrichment in progenitor peaks. These results are consistent with the radial unit hypothesis, which posits that expansion of the neural progenitor pool in prenatal development leads to alterations in adult cortical surface area[42].

To study the cell-type specific gene-regulatory impact of genetic variants associated with neuropsychiatric disorders and brain structure traits, we performed a co-localization analysis of progenitor/neuron caQTLs with existing GWAS data (Supplementary Table 8). We identified overlapped signals (pairwise LD $r^2 > 0.8$ between the GWAS index and caQTL index) and then performed a conditional analysis to verify that the two variants mark the same locus (Methods). We found co-localized loci in neuropsychiatric disorders, including schizophrenia, MDD, neuroticism and bipolar disorder, as well as IQ and educational attainment (Figure 7c). We also found co-localizations with global surface area and other brain structure associated loci[1]. We found additional ASCA sites located within GWAS loci (Extended Data Figure 8a; Supplementary Table 9). These results suggest that SNPs impact risk for brain-relevant traits and disorders by regulating the activity of REs in these two cell types during mid-fetal brain development, and provide a framework for exploring the mechanistic bases for these specific loci.

We next investigated regulatory mechanisms underlying co-localized loci using cell-type specific caQTLs. Combining fetal cortical/cell-type specific eQTL data, we found a co-localized locus across progenitor specific caQTLs, fetal cortical eQTLs and MDD GWAS (Figure 7d). We found more than 30 SNPs in high LD with an MDD GWAS index SNP (rs1950826). Eight of these variants were located in a caPeak (chr14:41,604,471–41,610,540). We prioritized one putatively causal SNP by testing for ASCA, finding that the A allele of the caSNP rs1950834 (protective allele for MDD), was associated with decreased accessibility of this caPeak in progenitors (Figure 7e–7f and 7h), which is consistent with luciferase reporter assay in previous work[43]. We also found the A allele of rs1950834 was associated with decreased expression of lncRNA AL121821.1 (ENSG00000258636) in fetal cortex and progenitors (Figure 7g). After conditioning on the MDD index SNP, rs1950826, the caQTL was no longer significant, indicative of co-localization (Figure 7d). We found evidence to support that this SNP disrupts the binding of ETV1 (Figure 7i; Methods)[22]. This suggests that the protective mechanism of this locus for MDD is via the protective allele at the caSNP disrupting ETV1 binding at an RE in progenitors, decreasing chromatin accessibility of this caPeak, and resulting in decreased expression of lncRNA AL121821.1.

As an additional example, we detected a co-localized locus between a neuron specific caQTL and schizophrenia GWAS (Extended Data Figure 8b). We found the C allele (schizophrenia protective allele) of rs9930307 was associated with decreased chromatin accessibility of a neuron caPeak (chr16:9,805,221–9,805,420) within an intron of *GRIN2A* (Extended Data Figure 8c). This caSNP was also a neuron-specific ASCA site, providing further evidence of this allele’s impact on chromatin accessibility (Extended Data Figure 8c). After conditioning on the schizophrenia index SNP (rs7191183) in the caQTL analysis, the caSNP was no longer significant (Extended Data Figure 8b). The motif of TP53 was disrupted by this caSNP (Extended Data Figure 8d–8e). Using a luciferase reporter assay, we found the C allele decreased the activity of this enhancer which is consistent with the caQTL result (Extended Data Figure 8f; Supplementary Table 10).

Discussion

Our caQTL analysis identified regulatory mechanisms underlying risk variants for neuropsychiatric disorders and brain-relevant traits. Currently, the function of individual non-coding brain-relevant risk loci is commonly understood through co-localization with eQTLs in adult post-mortem brain tissue or chromatin interaction[11,13]. Our work is able to complement previous studies in several ways: (1) caQTLs allow fine mapping of causal variants within LD-blocks by identifying putatively causal variants within peaks; (2) cell-type specific caQTLs can prioritize cell-types mediating the risk for neuropsychiatric illness because genetic effects on REs are highly cell-type specific; (3) most previous eQTL studies have been performed in post-mortem adult brain cortex[10,44], but cell-types contributing to the heritability for multiple disorders and traits are not present at this time period suggesting that temporal specificity matters for understanding risk for these disorders[15]; and (4) integration of caQTL, eQTL, and brain-trait GWAS allows a more complete understanding of regulatory mechanisms leading to risk for neuropsychiatric disorders, where non-coding genetic variants disrupt TF binding to REs, affecting chromatin accessibility, influencing expression of genes, leading to downstream risk for neuropsychiatric disorders. While we prioritized TFs driving the regulatory effects based on motif disruption and expression in the cell-types of interest, further experimental validation using complementary techniques (e.g. ChIP-seq) is necessary to determine if the caSNP does disrupt binding of the prioritized TF. Our study provides a resource to understand the impact of genetic variation on gene regulation during human cortical neurogenesis and provides an additional layer of information to explain the function of common variants associated with risk for neuropsychiatric illness and brain-related traits.

We found schizophrenia risk variants are enriched in progenitor REs (Figure 7a), consistent with previous post-mortem human studies[45], but contrary to a mouse gene expression study that found enrichments in neuronal but not progenitor cell types[46]. Given that the mouse study may miss human specific REs and assigns variants to genes by proximity, we believe that our findings in combination with previous literature suggest that genetic alterations in both human progenitor and adult neuronal REs contribute to risk for schizophrenia[12,15,45].

We provide evidence to support that caQTLs have higher effect sizes and more cell-type specificity than eQTLs in the two cell types we measured. This suggests that there are a limited number of mechanisms whereby genetic variation impacts chromatin accessibility, including TF binding to DNA, whereas there are considerably more mechanisms by which variation can impact transcript levels, such as altering TF binding, impacting methylation, or altering miRNA binding sites[47–49]. This also suggests that caQTL analyses will identify more genetic variants involved in gene regulation than eQTLs given a limited sample size. However, because our comparison was conducted between only two cell types, other cell-type specific caQTL and eQTL studies will be necessary to confirm the higher cell-type specificity of caQTLs more broadly.

caQTL analysis is able to prioritize causal variants associated with REs, but cannot be used directly to predict the genes regulated by these elements. Most caQTLs did not result in changes in gene expression in either cell types or in bulk fetal cortical tissue. Previous work has suggested that multiple transcription factors, including those that translocate to the nucleus after response to an external stimulus, are needed to change gene expression levels at certain loci[50]. caQTLs may therefore be more likely to co-localize with risk loci even in the absence of external stimuli (context-independence), whereas eQTLs would require additional stimuli (context-dependence). This also suggests that future work identifying ca/eQTLs in response to environmental stimuli relevant to neural proliferation, differentiation, or function will be especially useful to interpret GWAS risk loci.

Methods

Data availability

Data generated in this manuscript (including metadata) can be accessed via dbGaP (phs001958.v1.p1; https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs001958.v1.p1).

The RNaseq and genotype dataset used for fetal cortical eQTL analysis are available at dbGaP with accession number: phs001900. REST ChIP-seq data in H1 embryonic stem cells and neurons differentiated from H1 cells are available via the ENCODE portal (https://www.encodeproject.org/) with the following identifiers: ENCSR000BTV and ENCSR000BHM.

Code availability

All code used in this manuscript is deposited on bitbucket at https://bitbucket.org/steinlabunc/celltypespecificcaqtls_wasp/src/master/.

Tissue acquisition and cell culture of phNPCs:

Human fetal brain tissue was obtained from the UCLA Gene and Cell Therapy Core following IRB regulations. The tissue is often fragmented during acquisition from the surgical procedure. In the lab of Daniel Geschwind, flat, thin pieces of tissue that have the morphology of developing cortex were selected, and in some cases the tissue was sufficiently intact to be certain of cortical identity. Presumed cortical tissue from 14–21 gestation weeks was dissociated into a single cell suspension, cultured as neurospheres, plated for a low number of passages (2.5 ± 1.8 s.d.) on laminin/fibronectin and polyornithine coated plates, and then cryopreserved as human neural progenitors (HNPs) following our previous work[18].

Cryopreserved HNPs were shipped to UNC Chapel Hill after a signed material transfer agreement by both institutions. All proliferation, differentiation, sorting, library preparation, and analysis were performed at UNC Chapel Hill (following IRB regulations under the Office of Human Research Ethics). In total, HNPs from 92 donors were cultured (34% are female and 66% are male).

Donors were thawed in “rounds” of approximately 10 donors, so as to create a manageable workload of cell-culture (Extended Data Figure 1a). Donors were randomly assigned into groups and thawed 3 weeks apart. We performed specific experimental events on the same day of the week and had the same interval of time between events for each round. Experimental events included thawing cells, feeding cells, splitting cells, counting and plating cells, washing cells prior to differentiation, coating plates with attachment factors, adding virus, lifting cells for sorting, sorting, and ATAC-seq library preparation. As much as possible, the same researcher performed the same experimental events. We documented if a different researcher performed an experimental event in the database described below. To determine the impact of different rounds, we cultured cells from the same donors in different rounds as technical replicates, for progenitors (N_donor=11, N_replicate=13, in round 1–7,12 and 13) and neurons (N_donor=4, N_replicate=4, in round 2,6,12 and 13).

We used PBS with 5μg/ml Fibronectin (Sigma-Aldrich; SLBN9832V) and 10μg/ml Poly-L-Ornithine(Sigma-Aldrich;P3655–500MG) to make the coating stock for proliferation and PBS with 5μg/ml Laminin (Life Technologies; 23017015) and 10μg/ml Poly-L-Ornithine(Sigma-Aldrich;P3655–500MG) for differentiation.

We prepared 1x proliferation media using Neurobasal A (Life Technologies; 10888–022) with 100μg/ml Primocin (Invivogen; ant-pm-2), 10% BIT 9500 (Stemcell Technologies; 09500), 1% Glutamax (100x) (Life Technologies; 35050061), 1μg/mL Heparin (Sigma-Aldrich; H3393–10KU), 20μg/ml EGF/FGF (Life Technologies;PHG0313/PHG0023), 2ng/mL LIF (Life Technologies; PHC9481) and 20 ng/mL PDGF (Life Technologies; PHG1034). And we prepared 2x proliferation media using Neurobasal A with 100μg/ml Primocin, 10% BIT 9500, 1% Glutamax (100x), 1μg/mL Heparin, 40μg/ml EGF/FGF, 4ng/mL LIF and 40 ng/mL PDGF (all items have the same lot number with 1x proliferation media). Then, we prepared 1x differentiation media using Neurobasal A (Life Technologies; 10888–022) with 100μg/ml Primocin (Invivogen; ant-pm-2), 2% B27 (Life Technologies; 17504–044), 1% Glutamax (100x)(Life Technologies; 35050061), 10ng/mL NT-3 (Life Technologies; PHC7036) and 10ng/mL BDNF (Life Technologies; PHC7074). The 2x differentiation media used Neurobasal A with 100μg/ml Primocin, 2% B27, 1% Glutamax (100x), 20ng/mL NT-3 and 20ng/mL BDNF (all items have the same lot number with 1x differentiation media).

We thawed cells on a Monday (Extended Data Figure 1a). HNPs were cultured for 8 days using full feeds of proliferation media (1x proliferation media). On day 9, HNPs were split 1:2 and proliferated with half feeds of proliferation media with twice the concentration of growth factors (2x proliferation media) from day 10 to day 14. On day 15, HNPs were split 1:3 and proliferated with half feeds of 2x proliferation media from day 16 to day 21. On day 22, cells were plated for differentiation onto 8 × 6-well plate wells per donor at a concentration of 42,000 cells/cm^2 (differentiation library preparation wells). Two × 6-well plate wells were also plated for ATAC-seq preparation of progenitor cells (progenitor library preparation wells) in 1x proliferation media. On day 23, all differentiation wells were washed three times with Neurobasal A and then fed with 1x differentiation media (see media descriptions below). On day 24, the progenitor cells in proliferation media were lifted with trypsin and ATAC-seq libraries were prepared for progenitors. From day 24 through day 84 cells were half fed every Monday, Wednesday and Friday with 2x differentiation media. Virus for labeling neurons (AAV2-hsyn1-eGFP; https://www.addgene.org/50465/; acquired from the UNC Vector Core;[64]) was added at 20,000 MOI for library preparation wells on day 64. On day 84, cells were lifted using Papain (Worthington) with DNase (Worthington) and sent to cell sorter (BD FACS Aria II or Sony SH800S) to sort for live neurons labeled with GFP (data were analysed using the native softwares from these two sorters). Labeled GFP neurons were collected and aliquoted for immediate ATAC-seq library preparation of the neuron cell-type.

All cultures were visually evaluated and ranked with a subjective measure of cell health before ATAC-seq library preparation. Cell health was based on morphology and growth with the highest rank of 2 (mostly healthy cells by brightfield microscopy) and the lowest ranking (many dead cells) of 0.

Immunocytochemistry

HNPs were plated onto polyornithine/laminin coated German glass coverslips at ~80,000 cells/cm^2. Proliferation and 8 week differentiated cultures were fixed with 4% paraformaldehyde in parallel with samples being processed for ATAC-seq. Fixed cells were permeabilized with 0.1% Triton X-100/PBS, blocked with 10% normal goat serum in 0.1% Triton X-100/PBS, and subjected to immunocytochemistry. Coverslips were treated with the following primary antibodies in 0.02% Tween-20/PBS overnight at 4°C: Pax6 (1:300, BioLegend, Catalog#:901301, Lot#:B277104), Sox2 (1:300, Cell Signaling Technology, Catalog#:3579, Lot#:5), Sox2 (1:500, EMD Millipore, Catalog# AB5603, Lot#: 3187396), GFP (1:500, Millipore, Catalog#: AB16901, Lot#:2712295), HOPX (1:1000, Sigma-Aldrich, Catalog#:HPA030180, Lot#: C105752), Nkx2.1 (1:500, Millipore, Catalog#: MAB5460, Lot#:3074948), Tbr2 (1:300, eBioscience, Catalog#: 14-4877-82, Lot#: 2042087), Gad67 (1:500, EMD Millipore, Catalog#: MAB5406, Lot#: 3015328), TUBB3 (1:1000, BioLegend, Catalog#: 801202, Lot#: B249869), Satb2 (1:200, Santa-Cruz, Catalog#: sc-81376, Lot#: 132317), Ctip2 (1:500, Abcam, Catalog#: ab18465, Lot#: GR3242845–3). Coverslips were then treated with the following secondary antibodies for 1 hour at RT at 1:1000 dilution: Goat anti-RB AF488 (Thermo Fisher, Catalog#: A-11034, Lot#: 1812166), Goat anti-RB AF568 (Thermo Fisher, Catalog#: A-11011, Lot#: 1832035), Goat anti-CH AF488 (Thermo Fisher, Catalog#: A-11039, Lot#: 1759025), Goat anti-RT AF647 (Thermo Fisher, Catalog#: A-21247, Lot#: 2119156), Goat anti-MS AF488 (Thermo Fisher, Catalog#: A-11001, Lot#: 1939600), Goat anti-MS AF568 (Thermo Fisher, Catalog#: A-11031, Lot#: 2026148).

Library Preparation for human neural progenitors and neurons

Library preparation was conducted using the published ATAC-seq protocol[65]. ATAC-seq libraries were prepared immediately following cellular dissociation. Progenitor nuclei were counted using a hemocytometer while neuron nuclei were counted during sorting. 50,000 nuclei were aliquoted into the first step of the ATAC-seq published protocol. Libraries were prepared following the published instructions except that the last clean up step was modified to use KAPA pure beads (AmpureXP beads at a 1:1 ratio to remove dNTPs, salts, primers or primer dimers) instead of Qiagen Minelute clean-up kit. All libraries were sequenced to a minimum depth of 13.6M and an average depth of 25.5M using 50 bp PE sequencing on an Illumina HiSeq2500 or MiSeq machine (Extended Data Figure 1b). In total, we acquired 98 ATAC-seq libraries from progenitors (N_{donor}=85, $N_{libraries\ replicated}$=13) and 70 ATAC-seq libraries from neurons (N_{donor}=66, $N_{libarary\ replicated}$=4). All libraries were sequenced to an average depth of 25.5 (± 7.21 s.d.) million read pairs (Extended Data Figure 1b), which resulted in an average depth of 14 (± 4.8 s.d.) million reads pairs per sample after filtering for mitochondrial contamination and duplicates. We performed a sensitivity analysis for read depth vs peak calling that showed greater than 15 million filtered read pairs per library led to a fewer number of new peaks called, indicating a reasonable balance between read depth and peaks called on the libraries generated here (Extended Data Figure 2a).

Recording technical variables and randomization

To reduce the impact of batch effects on interpretation of our results, we attempted to either have no batches when possible (e.g., perform all experiments using the same lot of a reagent) or when this was not possible, randomize technical variables (round a donor was thawed in, sequencing pool) such that they had minimal correlation with variables of interest. In order to extensively document the impact of technical variables on outcome measures, we maintained a relational MySQL database which allowed us to keep track of many technical and biological variables throughout each experimental event. Each downstream ATAC-seq library preparation therefore is able to be tracked back to all technical and biological variables associated with its cell culture. The variables recorded were:

Media: Basal media lots, growth factor lots, supplement lots, antibiotic lots; *Virus*: Lot number; *Donor*: sex inferred from genotype; gestation week; *Culture*: passage, round, thaw date, each split date, split ratio, trypsin lot, PBS lot, polyornithine lot, fibronectin lot, plate, and well position, cells per well, date of virus addition, differentiation time, date of differentiation media addition, person plating for differentiation, virus used, person performing splits, person performing virus addition, virus lot, virus multiplicity of infection (MOI), laminin lot, dissociation lot, person performing dissociation of neuronal cultures; *Cell sorting*: Sort date and time, number of live cells, number of GFP+ cells, total number of cells, FACS machine; *ATAC-seq library preparation*: number of cells input to the library preparation, person performing the cell lifting, lysis date, PBS lot, lysis buffer batch, Illumina Kit lot, PCR master mix lot, PCR clean up kit lot, number of time pipetting up and down during lysis, number of times pipetting up and down during transposase reaction, transposase reaction volume, barcode indices used for multiplexing of each sample, number of PCR cycles added in the ATAC-seq protocol, final DNA concentration after library preparation complete; *Sequencing*: sequencing date, sequencing company, type of sequencer, read length.

Randomization was performed multiple times. First, randomization was performed to assign each donor to a thawing “round”. Randomization was performed at this stage by randomly ordering all donors and selecting those to go in each round (generally about 10 donors per round). After culture and library preparation were complete, randomization was performed to assign each library preparation to a pool for sequencing. Randomization was performed using custom R code to minimize the correlation of sequencing pool with concentration of the library, barcode index (assuring that no barcodes were represented in more than one pool), cell type (either neuron or progenitor), round cells were cultured in, and donor.

ATAC-seq data pre-processing

Sequencing reads were first quality controlled via fastqc (v0.11.7) to check for sequence quality in each library. We observed high quality sequencing for all libraries (PHRED > 20, average duplication rate = 43.07% which is almost entirely mtDNA contamination

(Extended Data Figure 1e) which is in agreement with previous studies using the same ATAC-seq method[15], and average GC content = 45%) (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Sequencing adapters were removed using BBMAP/BBDUK (https://jgi.doe.gov/data-and-tools/bbtools/bb-tools-user-guide/bbduk-guide/). We also calculated the number of total reads, the number of unique non-mitochondrial reads, duplicate rate, mitochondrial duplicate rate, TSS enrichment score, and FRiP score (The fraction of reads in called peak regions) for neuron samples and progenitor samples to check the library quality using atacqv[66].

Then, sequencing reads were mapped to the human genome including decoy sequences (GRCh38/hg38) using bwa mem[67] (v0.7.17) and WASP[68] to remove mapping bias at any bi-allelic SNP using imputed genotype data from each sample. Duplicate reads were then removed using WASP. Only uniquely mapped reads mapping to chr 1–22, X and Y were kept (mitochondrial genome and unmapped contigs were removed) using samtools[69] (v1.9). Sequencing reads mapped to ENCODE blacklist regions (http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeMapability/wgEncodeDacMapabilityConsensusExcludable.bed.gz, converted to hg38 using UCSCtools/liftOver (v320), were then removed by bedtools[70] (v2.26).

We did a sensitivity analysis for peak calling using pre-processed bam files. It showed acquiring 9x higher read depth resulted in only 70,000 additional peaks by MACS2[71] https://github.com/taoliu/MACS (Extended Data Figure 2a). So we reasoned that our acquired sequence depth obtained a reasonable balance between additional read depth and number of peaks called. We calculated the insert size of pre-processed bam files using Picard tools (v2.18.22). The insert size histogram shows clear periodicity representative of preferential Tn5 binding around nucleosomes (Extended Data Figure 2b).

To ensure independence, we randomly selected one library from each donor for each cell-type (technical replicates where one donor was cultured multiple times for a given cell-type) were excluded. Peaks were called for these selected samples (N_Neuron=61, N_Progenitor=76) using CSAW[72] (v1.16.1). For CSAW, peaks were identified in 10 bp bins with average read number greater than 5 across all samples (both neurons and progenitors). Bins directly next to each other (100 bp minimum distance) were merged to call a peak. In total, CSAW identified 90,227 peaks.

R v3.4.1 was used for all subsequent analyses. The number of reads within each CSAW-called peak were counted and then normalization factors for each peak across samples were calculated accounting for GC content, peak width, and total number of unique non-mitochondrial fragments sequenced using conditional quantile normalization[73] from the cqn package (v1.28.1). Variance stabilizing transformed (VST) counts were calculated using DESeq2[60] (v1.22.2) for batch effect correction and differential accessibility analysis by limma[61] (v3.38.3).

As two different sorters were used to sort GFP+ neurons (63 neuron cell lines (N_{donor}=61) for one sorter and 7 cell lines (N_{donor}=5) for another sorter), and we detected that sort location had a strong effect on PCA from neuron samples, sorter locations in neuron VST counts were first corrected (limma batcheffectremove()) (Extended Data Figure 2c). Corrected neuron VST counts and progenitor VST counts were combined so that the potential batch effect from cell culture rounds were corrected (limma removeBatchEffect).

Mycoplasma contamination checks

To check if there was any contamination from mycoplasma while in culture, we downloaded 98 mycoplasma genomes (from NCBI) and then mapped all ATAC-seq data to every mycoplasma genome. Less than 0.01% of each ATAC-seq sample mapped to any mycoplasma genome, which demonstrated that our cultures were not contaminated with mycoplasma.

Replicate correlations, principal component analysis, and correlation with technical factors

To determine the reliability of our experiment, we cultured the same donor multiple times. We correlated the batch corrected VST counts in CSAW peaks for neuron and progenitor replicates either within donors or calculated correlations across donors (Extended Data Figure 2d). The correlations of replicates within donors are higher than that of samples across donors in both neuron and progenitor samples.

The principal component analysis for batch corrected VST counts for all samples was done using the prcomp() R function. Then, the correlations for the first 10 PCs with technical and biological variables that we recorded were calculated using R. The technical variables include round of cell culture and sorter locations (only in neurons). Significant correlations with technical variables were removed after batch correction (Extended Data Figures 2e–2f).

ATAC-seq differential accessibility analysis

In order to find differentially accessible peaks across cell type controlling for technical factors, the dependent variable was the batch corrected number of reads within CSAW peaks and the linear regression model independent variables included a regressor for cell type (neuron or progenitor) and a factor regressor for donor IDs included in the analysis.

Enrichment of peaks within annotated regions of the genome

Enrichment of differentially accessible peaks within annotated genetic regions of the genome or epigenetically annotated regions of the genome was calculated using the ratio between the (#bases in state AND overlap feature)/(#bases in genome) and the [(#bases overlap feature)/(#bases in genome) × (#bases in state)/(#bases in genome)] as described previously by the Roadmap Epigenomics Consortium[74]. The significance of this enrichment was calculated using a binomial test as in the GREAT algorithm[75].

Chromatin state definitions from an imputed 25-state model were derived from fetal brain tissue (E081) and other *in vivo* tissues/cell types by the Roadmap Epigenomics project[74,76] and acquired from (http://www.broadinstitute.org/~jernst/MODEL_IMPUTED12MARKS/) after liftOver to hg38 (0.001% of peaks could not be lifted over). We generated the following combined states by merging states of similar genomic context:

Promoter(2_PromU, 3_PromD1, 4_PromD2, 22_PromP, 23_PromBiv), Enhancer (13_EnhA1, 14_EnhA2, 15_EnhAF, 16_EnhW1, 17_EnhW2, 18_EnhAc), Heterochromatin (21_Het), Quiescent (25_Quies), Transcribed (1_TssA, 5_Tx5', 6_Tx, 7_Tx3', 8_TxWk, 9_TxReg, 10_TxEnh5', 11_TxEnh3', 12_TxEnhW), Polycomb (24_ReprPC) and ZNF_Rpts (20_ZNF/Rpts).

Locations of ATAC-seq peaks in fetal brain tissue were acquired from previously published work[15]. After liftOver to hg38, 0.001% of peaks could not be lifted over.

Gene ontology analysis at the TSS for differentially accessible peaks (Extended Data Figure 3) was completed by first overlapping differentially accessible peaks with a region 2kb upstream and 1kb downstream of the TSS of genes defined by *Homo sapiens* gene ensembl version 78 GRCh38.p12. Protein-coding genes with promoter overlapped with selected differentially accessible peaks (|LFC|>0.5) were input into the TopGO[77] package (v2.34.0), with all protein-coding genes as background.

Gene based annotations of the genome were derived from *Homo sapiens* gene ensembl version 78 (GRCh38.p12) for plotting loci.

Differential transcription factor binding analysis

We performed an analysis to identify motifs with differential prevalence in differentially accessible peaks (Supplementary Table 2, Figure 3b). To avoid the bias caused by different numbers of progenitor peaks and neuron peaks, here we only used top 2000 progenitor peaks with highest LFC, and top 2000 neuron peaks with lowest LFC.

Potential transcription factor binding sites were called in the human genome using TFBSTools (v1.4.0) with a minimum score threshold of 80% based on position weight matrices from the JASPAR2016[78] core database, selecting vertebrates as the taxonomic group. Only the most recent version of the PWM for a given TF was used. To select regions of the genome that are highly conserved among vertebrates, and likely functional, 100-way phastCons[79] scores > 0.4 in regions ≥ 20 bp were saved (downloaded from UCSC genome browser). Only TFBS sites within conserved regions were retained for further analyses. Differential motif enrichment analysis was performed using a logistic regression model to identify motifs present more often in progenitor peaks as compared to neuron peaks, or vice versa. Logistic regression explicitly controlled for differences in peak width and peak conservation between progenitor and neuron differentially accessible peaks. The analysis was implemented in R as: glm(TFBS ~ ProgenitorNeuron + peakwidth + conservedbppercent, family = "binomial"). The dependent variable (TFBS) was a binary representation of whether each differentially accessible peak contained a motif of a TF or not. The independent variable of interest marked whether a peak was progenitor (ProgenitorNeuron=1) or neuron (ProgenitorNeuron=0). Other covariates included peak width (peakwidth) and the percentage of the peak with conservation (conservedbppercent) as defined above. Significant differential motif enrichment was determined by FDR adjusted P-value < 0.05 threshold of the ProgenitorNeuron covariate. progenitorTFs were defined as significantly differentially enriched motifs present more often in progenitor peaks as compared to neuron peaks, whereas neuronTFs were defined as significantly differentially enriched motifs present more often in neuron peaks as compared to progenitor peaks.

Genotype pre-processing

Genotyping was performed using Illumina HumanOmni2.5 or HumanOmni2.5Exome platform. SNP genotypes were exported into PLINK format. SNP marker names were converted from Illumina KGP IDs to rsIDs using the conversion file provided by Illumina. Quality control was performed in PLINK v1.90b3[80] (Extended Data Figure 4a). SNPs were filtered based on Hardy-Weinberg equilibrium (--hwe 1e-6), minor allele frequency (--maf 0.01), individual missing genotype rate (--mind 0.10), variant missing genotype rate (--geno 0.05) resulting in 1,760,704 directly genotyped variants. Multidimensional scaling (MDS) analysis of genotypes from all individuals used in the study was completed in PLINK v1.90b3. We did not see a strong effect of genotyping batch on genotype data based on MDS1 and MDS2 from different genotyping batches. We used PLINK v1.90b3 to call sex from genotype data. For the samples with unknown sex from genotype data, we ploted PCA (PC1 vs PC2) of ATAC-seq reads on sex chromosomes (chromsome X and Y) to identify sex (Extended Data Figure 4b).

Sample Swap and contamination Identification

Quality controlled genotype data and BAM files were used to identify any sample swaps between the ATAC-seq and genotyping data using VerifyBamID v1.1.3[81]. We removed any BAM file with [FREEMIX] > 0.02 or [CHIPMIX] > 0.02 (N_donor=5), and corrected sample swaps (N_donor=7). After this filtering step, our sample size was 76 unique donors for progenitor samples and 61 unique donors for neuron samples for the caQTL studies.

Imputation

Filtered genotype data were pre-phased by SHAPEIT[82] v2.837. Minimac4[83] (v1.0.0) was used to impute the filtered genotyped markers using reference haplotype panels from the 1000 Genomes Project (The 1000 Genomes Project Consortium Phase 3) that contain a total of 37.9 million SNPs in 2,504 individuals from any ancestry, including those from West Africa, East Asia and Europe. We separated chrX into pseudoautosomal regions and non-pseudoautosomal regions, then pre-phased and imputed them separately.

After genotype imputation, we extracted the genotypes for all individuals assayed for chromatin accessibility. Imputed genotype data were filtered for variant missing genotype rate < 0.05, Hardy-Weinberg equilibrium p-value $< 1 \times 10^{-6}$ and minor allele frequency (MAF) 1%. We retained variants with imputation quality Rsquared > 0.3 by Minimac4, resulting in ~13.6 million SNPs.

caQTL mapping

We calculated multidimensional scaling (MDS) for genotype data of our samples and genotype data from HapMap3 (https://www.sanger.ac.uk/resources/downloads/human/hapmap3.html) following the protocol from ENIGMA consortium (http://enigma.ini.usc.edu/wp-content/uploads/2012/07/ENIGMA2_1KGP_cookbook_v3.pdf). We identified multiple ancestries of donors of our samples in the MDS plot (MDS1 vs. MDS2) (Extended Data Figure 4c).

To control for population stratification and cryptic relatedness of our samples when mapping caQTLs, we ran caQTL analysis with EMMAX[23], which accounts for population structure using a genetic relatedness or kinship matrix. We used the emmax-kin function (-v -h -s -d 10) to create the IBS kinship matrix for each tested genetic variant from non-imputed genotype data excluding all genetic variants on the same chromosome with the tested genetic variant[84].

We performed proximal caQTL mapping using a window of 100 kb up- and down-stream of the center of 90,227 csaw peaks using VST normalized read counts of each peak for each donor (Extended Data Figure 4a). We performed caQTL analysis separately in neurons and progenitors using imputed genotype data. To prevent results driven by only one minor allele homozygous donor, we retained the variants where the number of minor allele homozygous donors is not 1 and at least 2 heterozygous donors. In addition to the kinship matrix[84], for the progenitor caQTLs, we include sorter locations, the first 10 genotype MDSs and 8 PCs across VST counts of the chromatin accessibility data. For neurons, we include the first 10 genotype MDSs and 7 PCs of VST counts of the chromatin accessibility data. These PC numbers were chosen to maximize the number of caQTLs for each cell-type. Nominal EMMAX p-values were corrected for multiple testing using the software eigenMT[85] and Benjamini–Hochberg FDR correction[86] within neuron caQTLs and within progenitor caQTLs separately (eigenMT-FDR < 0.05). We retained the most significant caSNP for each significant caPeak that survives the eigenMT-FDR threshold, and defined the caSNP-caPeak pair as the most significant caQTL for a given caPeak. The percent variance explained was calculated using the method from a previous study[87].

Identify correlated caPeaks

To identify correlated caPeaks, we defined primary caPeaks as the caPeaks harboring caSNP(s). We then defined secondary caPeaks as peaks which are associated with the caSNP of a primary peak. We calculated Perason's correlation between the primary caPeak and all caPeaks within +/− 2Mbp from the center of its secondary caPeak (including the secondary caPeak), then corrected the Pearson's correlation p-value using the Benjamini–Hochberg FDR correction[86]. If the secondary caPeak was significantly (FDR < 0.05) correlated with the primary caPeak, this caSNP-caPeak pair was classified as "caSNP in correlated caPeak".

Allele specific chromatin accessibility

We used GATK tools to extract allele specific read counts for every SNP. We first filtered for SNPs within each donor that had sufficient read depth by retaining SNPs with total counts greater than or equal to 10 for neuron and progenitor samples, separately. Then to calculate allele specific chromatin accessibility, we retained those SNPs with average read counts for all heterozygous donors greater than or equal to 15. Finally, we retained only those SNPs that meet these previous thresholds for at least 5 heterozygous donors. DESeq2 was used to calculate the LFC (Alternative read counts/Reference read counts) for filtered SNPs across all heterozygous donors. The non-heterozygous donors were excluded from the differential analysis for each SNP using sample-specific weights, and maximum likelihood estimation was used for dispersion estimation followed by Wald tests of the estimated LFC. FDR < 0.05 was used as the threshold for significance.

Bulk fetal brain eQTL mapping

Bulk fetal cortical wall eQTL data described in a previous publication[14], was re-analyzed in this study with the following modifications: (1) we used a linear mixed model implemented in EMMAX to more stringently control for population stratification, and (2) we add 23 more donors to the analysis because these donors were genotyped after the publication of the previous manuscript. rRNA-depleted RNA-seq data from flash frozen human fetal brain cortical wall tissues derived from 235 donors at 14–21 gestation weeks were used for eQTL analysis. 41% samples in the progenitor eQTL analysis were overlapped with the samples in fetal brain eQTL analysis, and 36% samples in the neuron eQTL analysis were overlapped with the samples in fetal brain eQTL analysis. Gene based annotations of the genome were derived from *Homo sapiens* gene ensembl version 92 (GRCh38) for eQTLs. Only genes which are expressed in more than 5% of donors with at least 10 counts were included in the analysis. VST normalized expression values were used as phenotypes for eQTL analysis. Genomic DNA from human fetal brain cortical wall tissues derived from 235 donors at 14–21 PCW was extracted. Each donor tissue was genotyped on a dense array (Illumina Omni 2.5+Exome) and imputed to a common reference panel (1000 Genomes; described above). Variants were retained in the analysis if there were at least 2 heterozygous donors and no homozygous minor allele donors, or if there were at least 2 minor allele homozygous donors. For the effect size comparison analysis fetal brain eQTL vs caQTLs (Figure 2e–2f), we subsampled fetal brain eQTL donors to the same sample size as the caQTL while maintaining the population composition similar to the larger donor list.

Cis-eQTL analysis was performed by evaluating association between each gene’s expression and variants within ±1 Mb window of transcription start site of each gene by implementing linear mixed model association software, EMMAX[88]. All markers on the chromosome of this candidate marker were excluded from the IBS kinship matrix generated with emmax-kin function (-v -h -s -d 10), and added as a random variable into linear mixed model for association test. In addition to kinship matrix, 10 MDS components of genotype and first 10 PCs of gene expression were included into the covariate matrix. After association, nominal p-values were corrected for multiple testing using the eigenMT and Benjamini Hochberg FDR correction, and associations with lower than 5% eigenMT-FDR threshold value were accepted as significant. We retained only the most significant eSNP for each significant eGene in this study.

M-value calculation

Using Metasoft (v2.0.1)[62], we calculated m-values between caQTLs and eQTLs. First, we selected the most significant caSNP for a given caPeak in either neurons or progenitors. Then, we found the SNP-Gene pair corresponding to that caSNP in bulk fetal brain or in the cell type specific eQTL (for sharing in Figure 2d). The caSNP may or may not be an eSNP and the eSNP may be associated with multiple genes. Then, we selected the most significant eSNP for an eGene and found the corresponding SNP-Peak pairs in neuron/progenitor caQTL analysis (for sharing in Figure 2f, Extended Data Figure 4e). The eSNP may not be a caSNP and the eSNP may be associated with multiple peaks. We calculated the m-values for SNPs in all the SNP-Peak-Gene combinations we found above. The SNP-Peak-Gene combinations with m-value greater than 0.9 in both caQTL and eQTL analysis are identified as shared SNP-Peak-Gene combinations.

Using the same approach, we identified shared neuron/progenitor SNP-Peak-Gene combinations using the most significant neuron/progenitor caSNPs and the most significant neuron/progenitor eSNPs.

Overlap of caQTLs with eQTLs

To identify RE-gene pairs neurons/progenitors, we listed SNPs with pairwise LD $r^2 > 0.8$ with the caSNPs in the caPeak using genotype data from neuron samples/progenitor samples, separately, then we listed SNPs with pairwise LD $r^2 > 0.8$ with index eSNP using the LD matrix from neuron samples/progenitor samples for cell type specific eQTL analysis. We labelled the caPeak and the eGene as an RE-Gene pair if any SNP from the above two categories is overlapped in neurons/progenitors (Supplementary Table 5).

To identify RE-gene pairs in fetal bulk cortical tissue eQTLs we listed SNPs with pairwise LD $r^2 > 0.8$ with the caSNPs in the caPeak using genotype data from neuron samples/progenitor samples, separately, then we listed SNPs with pairwise LD $r^2 > 0.8$ with index eSNP using the LD matrix from fetal brain samples for fetal bulk cortical eQTLs. We labelled the caPeak and the eGene as a RE-Gene pair if any SNP from the above two categories is overlapped in fetal brain tissues (Supplementary Table 5).

Estimation of sharing via π_1

The R package 'qvalue' (v2.20.0)[89] was used to estimate the π_0 of the input nominal p values of the cell type specific eQTL and caQTL data, then we used 1 minus the estimated π_0 to get π_1. We found all the neuron SNP-Peak pairs using the most significant progenitor caSNP-caPeak pairs, then used the nominal p values of the neuron SNP-Peak pairs to estimate the proportion of true neuron caQTLs in the SNP-Peak pairs (π_1). In the same way, we estimated the proportion of true progenitor caQTLs using the most significant neuron caSNP-caPeak pairs. For neuron eQTLs, we listed all the neuron SNP-Gene pairs using the most significant progenitor eSNP-eGene pairs, then used the nominal p values of the neuron SNP-Gene pairs to estimate the proportion of true neuron eQTLs in the SNP-Gene pairs (π_1). Similarly, we estimated the proportion of true progenitor eQTLs using the most significant neuron eSNP-eGene pairs.

Comparison to adult dorsolateral prefrontal cortex caQTLs

We acquired adult DLPFC ATAC-seq data from Sage Bionetworks-Synapse website via the psychENCODE Knowledge Portal under the accession number [syn5321694] https://www.synapse.org/#!Synapse:syn5321694[16]. To calculate the overlap of caQTLs between cultured neural cells and adult DLPFC, we first extracted read counts within adult DLPFC peaks in ATAC-seq data from neurons and progenitors. We found 65,573 DLPFC peaks have an average read counts greater than 5 across all neuron and progenitor samples, and these peaks demonstrate cell-type/tissue type specificity in chromatin accessibility as visualized in a PCA plot (Extended Data Figure 7a–b). Then using the shared peaks and the same SNPs with DLPFC caQTL, we re-mapped caQTLs in neurons and progenitors using the same models as previously described. We found 27 significant neuron caQTLs and 35 significant progenitor caQTLs using the same eigenMT-FDR threshold as previously used in caQTL mapping in neurons and progenitors. Using the same SNP-Peak pairs from DLPFC caQTLs, we found the π_1 is 0.001 in neuron caQTLs and 0.04 in progenitor caQTLs, which indicates a highly temporal specificity in caQTLs. We also found low correlations of effect sizes in significant neuron/progenitor caQTLs and DLPFC caQTLs (Extended Data Figure 7c).

Determining the impact of caSNPs on motifs

In order to determine if genetic variation within peaks impacts transcription factor (TF) binding motifs, we used motifBreakR (v1.14.0)[90] to map known TF motifs to the sequence surrounding the neuron-specific/progenitor-specific significant caSNPs located in an ATAC-peak have significant association with the caPeak (parameter setting: threshold = 1e-4). All annotated motifs (in total 630 TF motifs) are from JASPAR2016 vertebrate in MotifDb (1.26.0)[91]. We calculated the relative entropy (parameter setting: method=”ic”) for reference allele and alternative allele, then only kept the TFBSs which are strongly affected by the SNPs (motifbreakR parameter setting: effect=”strong”). We calculated the enrichment of neuron and progenitor caSNP-disrupted motifs in accessible peaks using the binomial test[75]. First, we found all TFBSs for a given TF in accessible regions (n) and calculated the fraction of base pairs of the motif compared to the overall base pairs of accessible peaks. Then, we counted the number of SNP-disrupted motifs for this TF (k). The final step was to calculate $P=Pr_{binom}(x>=k|n,p)$ using the binomial test to get the significance of the enrichment. We further filtered the enrichment results by differential expression from the same set of cells, and only kept the TFs with cell-type specific caSNP-disrupted motifs significantly enriched in accessible regions and significantly differentially expressed in the cell type[26].

To determine if the motif disrupting allele is associated with increased/decreased chromatin accessibility, we first identified the motif-disrupting allele. The motif-disrupting allele decreases the relative entropy of the position possibility matrix of a TFBS. Then, we aligned the motif-disrupting allele with the effect allele for caQTLs. Finally, we used linear regression to determine the relationship between decreased relative entropy and effect size for all motif-disrupting alleles for this TFBS (lm(effect size ~ decreased relative entropy+0). We fit the line through zero because we assume that if a motif is not disrupted by an allele, it will also have no effect on chromatin accessibility. The significance of the coefficient for effect size on decreased relative entropy was tested and the p-values adjusted to control FDR[86] (Figures 6c–6f).

Partitioned Heritability

Partitioned heritability was measured using LD Score Regression v1.0.0[92] to identify enrichment of GWAS summary statistics among differentially accessible peaks. First, an annotation file was created which marked all HapMap3 SNPs that fell within neuron peaks or progenitor peaks. To avoid bias caused by different numbers of progenitor peaks and neuron peaks, we randomly selected the same number of neuron peaks as progenitor peaks.

LD-scores were calculated for these SNPs within 1 cM windows using the 1000 Genomes European data. These LD-scores were included simultaneously with the baseline distributed annotation file from[92]. Subsequently, the heritability explained by these annotated regions of the genome was assessed from these genome-wide association studies: Attention-Deficit/hyperactivity disorder[93], autism spectrum disorder[94], IQ[95], major depressive disorder[96], Bipolar disorder[97], schizophrenia[98], insomnia[99], educational attainment[100], subjective well-being[101], depressive symptoms[101], neuroticism[102], anorexia nervosa[103], anxiety[104], Alzheimer's disease[105], epilepsy[106], Parkinson's disease[107], brain structures[1].

The enrichment was calculated as the heritability explained for each phenotype within a given annotation divided by the proportion of SNPs in the genome corresponding to the annotation and Benjamini–Hochberg FDR correction[86] was used to correct for multiple comparisons.

Co-localization with GWAS data

We used conditional caQTLs to detect the co-localization of caQTLs with multiple GWAS datasets, which are previously listed above in the “Partitioned Heritability” section. First, to identify co-localized loci: 1) we listed SNPs with pairwise LD $r^2 > 0.8$ with the caSNPs in the caPeak using genotype data from neuron samples and progenitor samples, separately; 2) we listed SNPs with pairwise LD $r^2 > 0.8$ with index GWAS SNP ($p<5e-8$ and exhibited the strongest association in upstream/downstream 100kb from the center of this caPeak) using the LD matrix from European genotype data from 1000 Genome project phase 3 with population code EUR. Second, we labelled the caPeak as a potentially co-localized locus if any SNP was shared between the above two categories. Third, we performed a conditional caQTL analysis for significant (eigenMT-BH FDR < 0.05) caSNPs conditioning on the index GWAS SNP[108]. If the caQTL is no longer significant (eigenMT-BH FDR > 0.05), then we called the caQTL as a co-localized locus with this GWAS trait.

Luciferase reporter assay

DNA fragments of differentially accessible chromatin peaks containing SNPs for functional validation were synthesized using Thermo Fisher Scientific's Gene String service. Fragments were amplified by PCR with primers containing KpnI and HindIII restriction sites. Digested fragments were then cloned into the multiple cloning region of the pGL4.23 vector (Promega), containing a minimal promoter upstream of the *luc2* luciferase reporter gene. To generate corresponding alternate alleles, we performed site-directed mutagenesis on the cloned, insert-containing luciferase plasmids using the Q5 Site-directed mutagenesis kit (NEB). All cloned sequences were verified by Sanger sequencing for the correct mutations and analyzed by NanoDrop to ensure high concentration and transfection-grade quality. Oligonucleotide sequences used for cloning are listed in Supplementary Table 10.

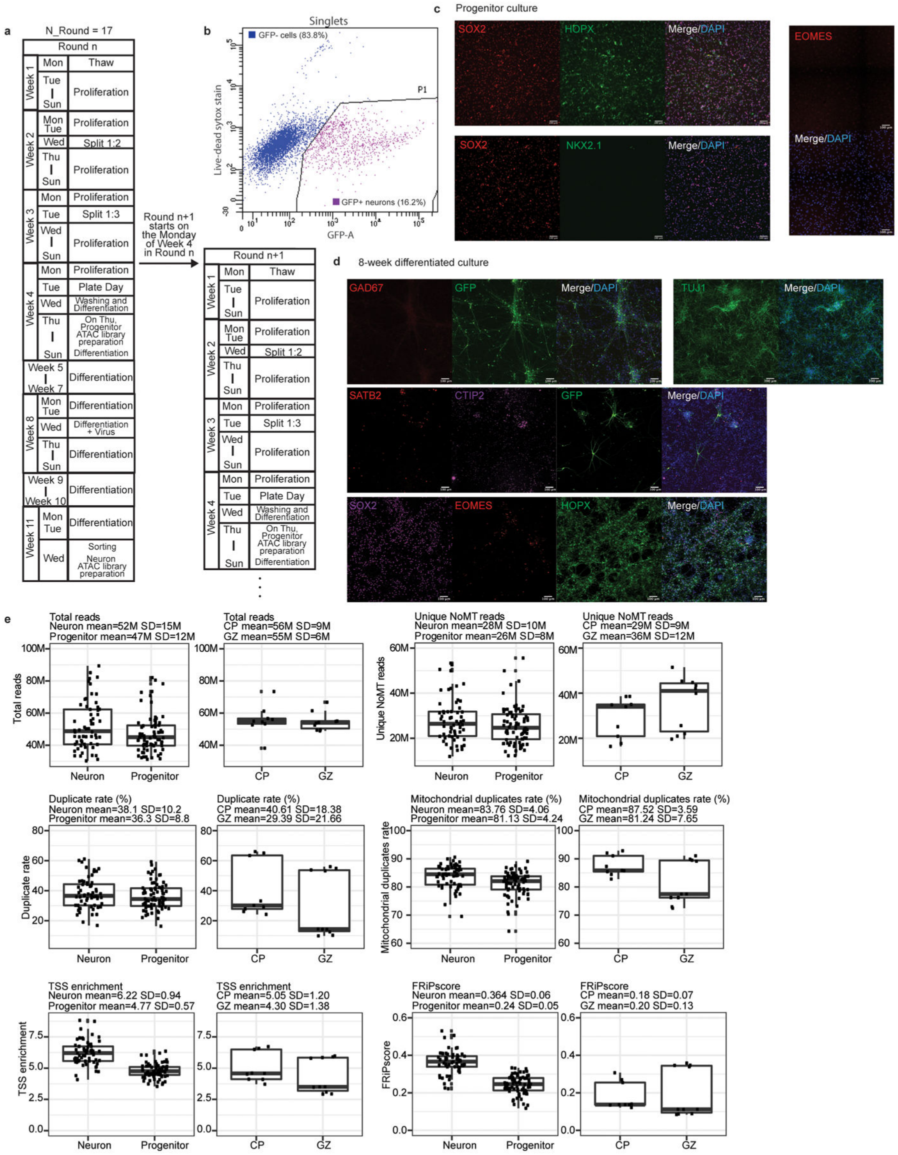
Human NPCs grown in 96 well plates were co-transfected with 120 ng/well luciferase reporter plasmid and 30 ng/well renilla control plasmid (pRL-SV40; Promega) using Lipofectamine STEM Transfection Reagent (Thermo Fisher). NPCs were then cultured for 72 hours prior to processing with the Dual-Glo Luciferase Assay System (Promega). Luciferase and renilla expression was measured using a CLAIROStar Plus Plate Reader

(BMG Labtech). Each luciferase reading was then normalized by its corresponding renilla reading to control for transfection efficiency and to calculate RLU. A total of 8 unique donors with at least 3 well replicates per plasmid per donor were used for analysis.

REST ChIP-seq data in H1 cells and neurons differentiated from H1 cells

We acquired the alignments of REST ChIP-seq data in H1 embryonic stem cells and neurons differentiated from H1 cells from the ENCODE portal[8,9] (https://www.encodeproject.org/) with the following identifiers: ENCSR000BTV and ENCSR000BHM. We normalized the read counts by library sizes then plotted the coverage using Gviz[109].

Extended Data



a N_Round = 17

Round n		
Week 1	Mon	Thaw
	Tue I Sun	Proliferation
Week 2	Mon Tue	Proliferation
	Wed	Split 1:2
	Thu I Sun	Proliferation
Week 3	Mon	Proliferation
	Tue	Split 1:3
	Wed I Sun	Proliferation
Week 4	Mon	Proliferation
	Tue	Plate Day
	Wed	Washing and Differentiation
	Thu I Sun	On Thu, Progenitor ATAC library preparation Differentiation
Week 5 I Week 7		Differentiation
Week 8	Mon Tue	Differentiation
	Wed	Differentiation + Virus
	Thu I Sun	Differentiation
Week 9 I Week 10		Differentiation
Week 11	Mon Tue	Differentiation
	Wed	Sorting Neuron ATAC library preparation

Round n+1 starts on the Monday of Week 4 in Round n

Round n+1		
Week 1	Mon	Thaw
	Tue I Sun	Proliferation
Week 2	Mon Tue	Proliferation
	Wed	Split 1:2
	Thu I Sun	Proliferation
Week 3	Mon	Proliferation
	Tue	Split 1:3
	Wed I Sun	Proliferation
Week 4	Mon	Proliferation
	Tue	Plate Day
	Wed	Washing and Differentiation
	Thu I Sun	On Thu, Progenitor ATAC library preparation Differentiation

Extended Data Fig. 1.

Flowchart for cell culture and pre-processing of ATAC-seq data.

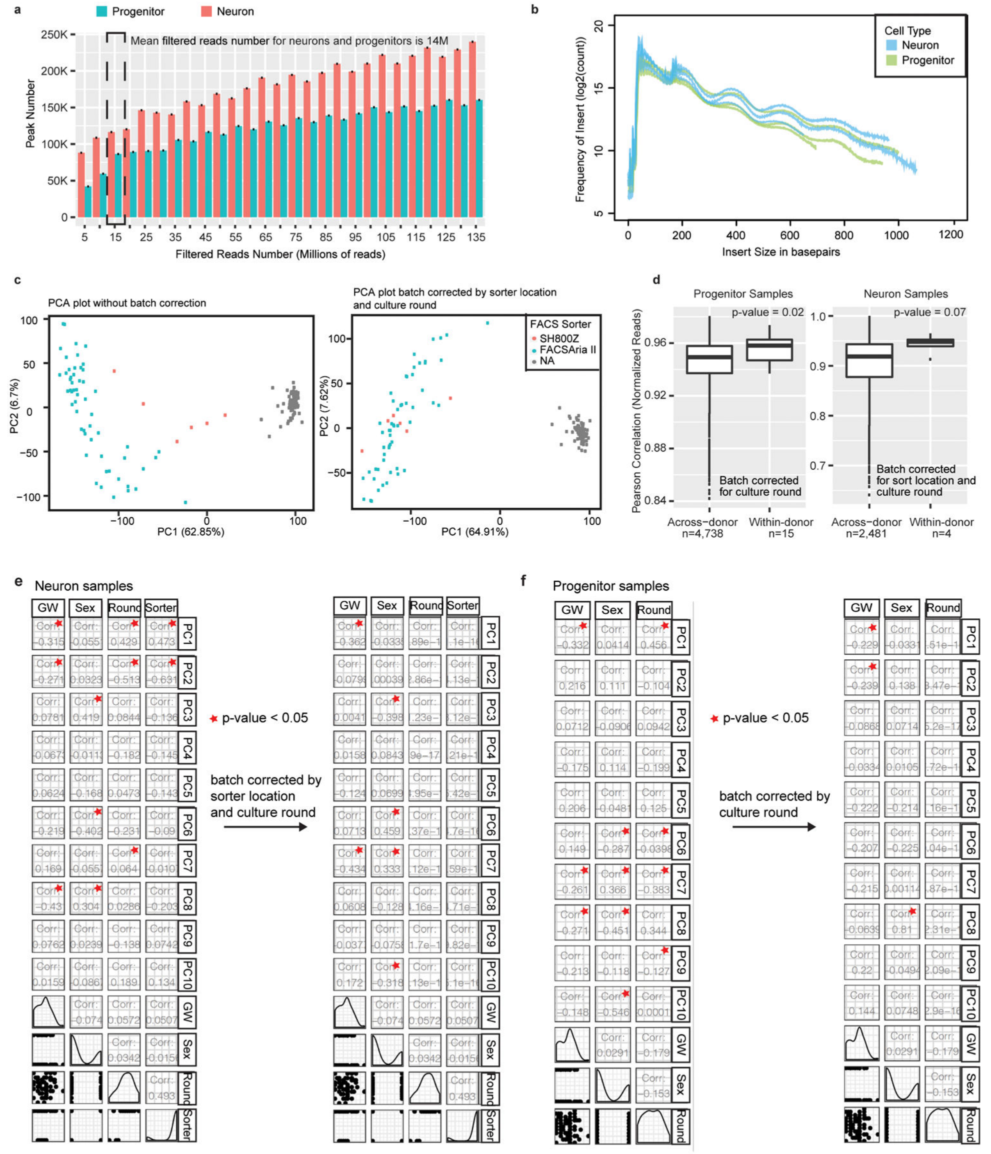
(a) Flowchart of cell culture for 17 rounds.

(b) The FACS gates for sorting EGFP+ neurons.

(c) Images of immunofluorescence for cell markers in progenitor cultures. Immunolabeling experiments were repeated in at least 10 unique donor cell lines with similar results. The scale bar presents 100 μm.

(d) Images of immunofluorescence for cell markers in 8-week differentiated cultures. Immunolabeling experiments were repeated in at least 10 unique donor cell lines with similar results. The scale bar presents 100 μm.

(e) Box plot for total sequence depth (forward reads and reverse reads), unique read number (forward reads and reverse reads), duplicate rate, mitochondrial duplicate rate, TSS enrichment and the fraction of reads in called peak regions (FRiP score) in neurons (N=61) and progenitors (N=76) compared to previously published data (N_GZ=3 biologically independent samples with 3–4 replicates for each sample, N_CP=3 biologically independent samples with 3 replicates for each sample)[15]. The center of the box is median of the data, the bounds of the box are 25th percentile and 75th percentile of the data, and the whisker boundary is 1.5 times the IQR. Maximum and minimum are the maximum and minimum of the data.



Extended Data Fig. 2.

ATAC-seq data QC.

(a) Peak calling versus library sequencing depth. We observed a slower rise in the number of new peaks called after 15 millions filtered read pairs. This indicates a reasonable balance between read depth and number of peaks called using an average of 14 million read pairs after filtering in our samples.

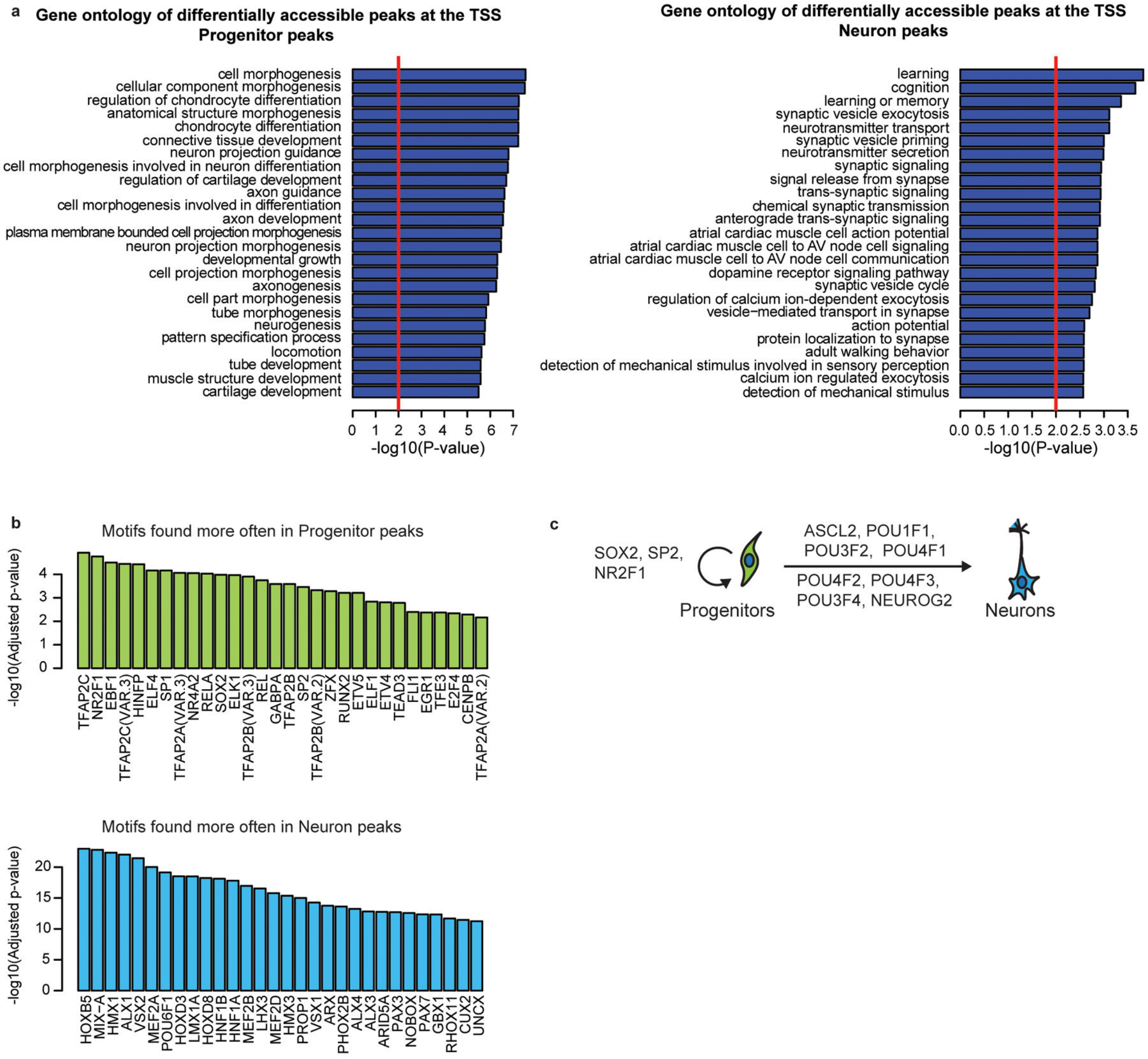
(b) Insert size histograms for 3 randomly selected neuron and progenitor samples.

(c) PCA plot for ATAC-seq data (N=137) before batch correction (*left*) and after batch correction (*right*), colored by sorter. We corrected normalized reads within ATAC-seq peaks in neurons by sorter locations. Then, we corrected normalized reads within ATAC-seq peaks in neurons and progenitors by cell culture round.

(d) Correlations of batch corrected normalized reads across donors and within donors. Correlations within donors was significantly higher than correlations across donors in progenitor (n=15). Correlations within donors were higher than correlations across donors in neurons (n=4), but not significant (p=0.07). P values are estimated by two-sided wilcoxon tests. The center of the box is median of the data, the bounds of the box are 25th percentile and 75th percentile of the data, and the whisker boundary is 1.5 times the IQR. Maximum and minimum are the maximum and minimum of the data.

(e) Correlations between PC1 to PC10 from normalized reads in neurons with known technical and biological factors.

(f) Correlations between PC1 to PC10 from batch correction normalized reads in progenitors with known technical and biological factors.



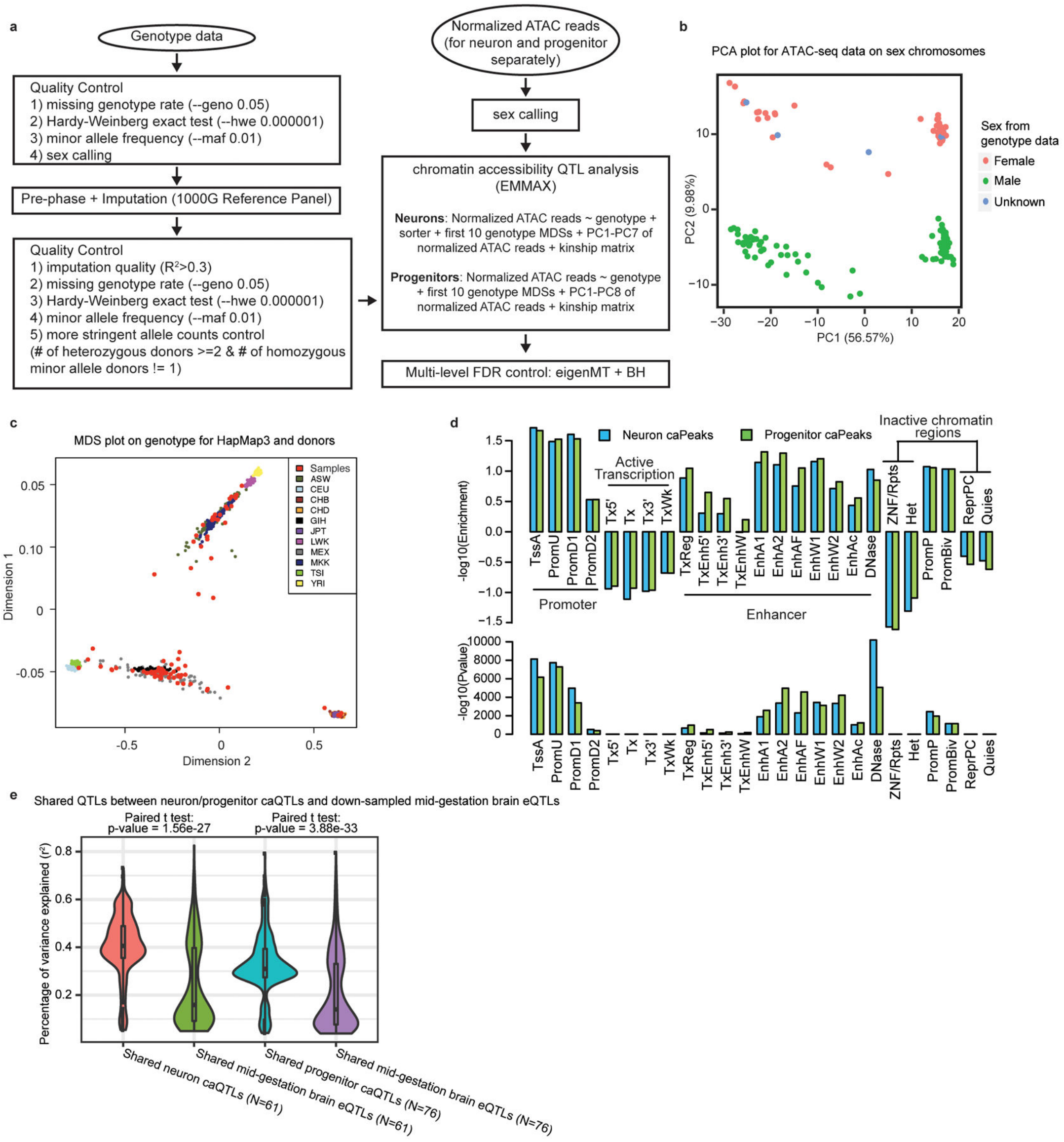
Extended Data Fig. 3.

Annotating differentially accessible peaks during neuronal differentiation.

(a) Gene ontology (GO) enrichment of differentially accessible peaks at the TSS. Progenitor peaks (left) and neuron peaks (right) showed enrichment for GO terms related to proliferation and differentiation, as expected.

(b) TFs with significantly differentially enriched conserved binding sites in differentially accessible peaks. The statistical test identifies TFs likely involved in neural progenitor proliferation and maintenance (progenitorTFs; top) or neurogenesis and maturation (neuronTFs; bottom). The top 30 significantly enriched TFs were shown in this figure, and the full list can be found in Supplementary Table 2. Within progenitorTFs, we found TFs previously characterized to have key roles for neural stem cell renewal and reprogramming, such as *SOX2*[52,53], and those known to be required for the maintenance of stem cells in cortex, such as *NR2F1*, *ETV5*, and *SP2*[54–56]. Within neuronTFs, *NEUROG2* and *LMX1A* were identified, which are known to drive neuronal differentiation[57,58], as well as TFs shown to induce neuronal identity from fibroblasts, including *ASCL2* and the *POU* family[39]. NeuronTFs also included *CUX1/2*, a marker for layer II-III neurons[59,60] and other laminar markers such as *TBR1* and *FOXP1*.

(c) Schematic of known functions for selected progenitorTFs and neuronTFs.



Extended Data Fig. 4.

Features of caQTLs.

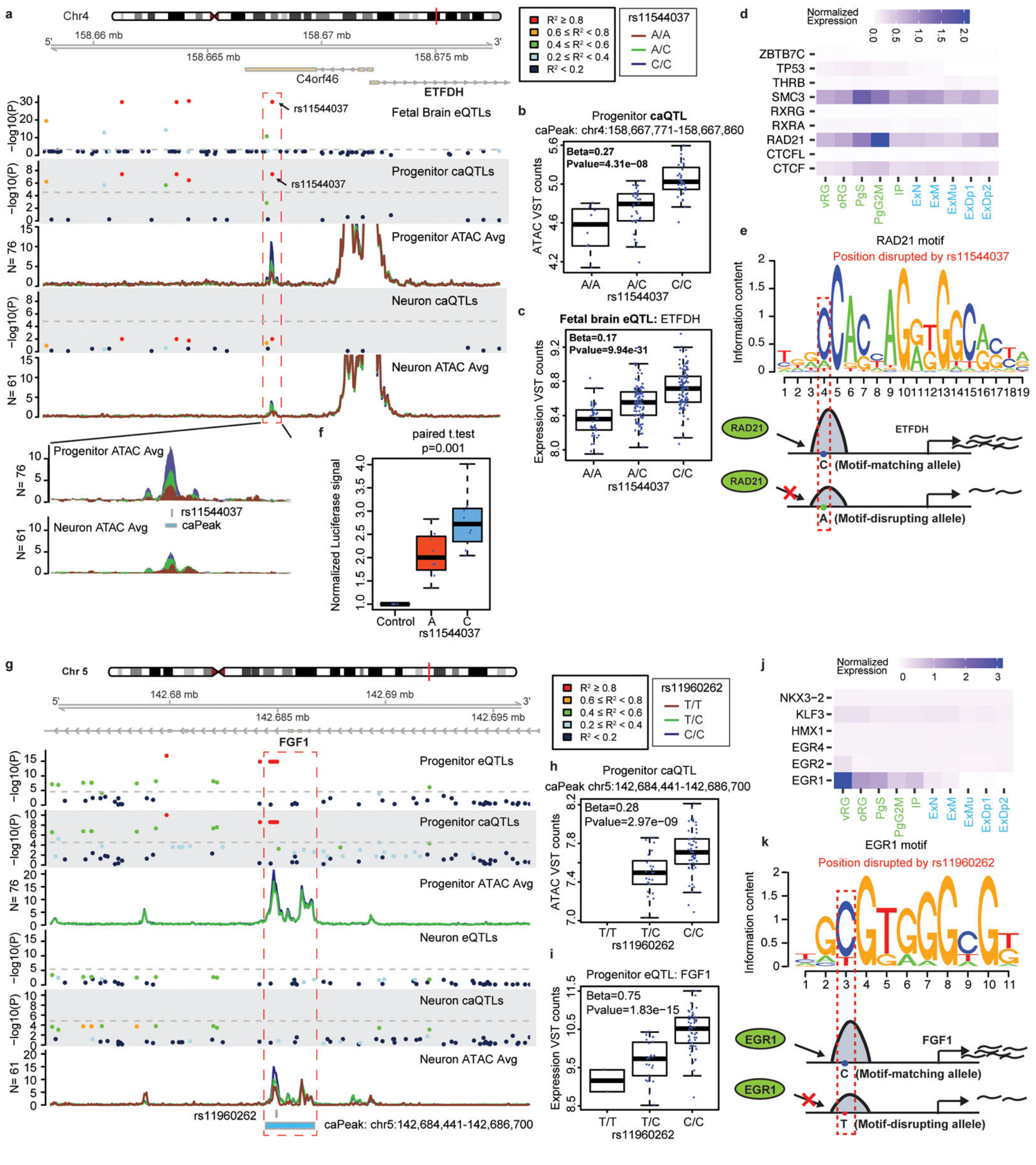
(a) Flowchart for caQTL data analysis.

(b) PCA plot for ATAC-seq data on sex chromosomes (chrX and chrY), colored by sex from genotype data, showing sex could be called using ATAC-seq data.

(c) MDS plot for genotype data of HapMap3 and donors in this study, colored by populations from HapMap3 data. ASW: African ancestry in Southwest USA; CEU: Utah residents with Northern and Western European ancestry from the CEPH collection; CHB: Han Chinese in Beijing, China; CHD: Chinese in Metropolitan Denver, Colorado; GIH: Gujarati Indians in Houston, Texas; JPT: Japanese in Tokyo, Japan; LWK: Luhya in Webuye, Kenya; MEX: Mexican ancestry in Los Angeles, California; MKK: Maasai in Kinyawa, Kenya; TSI: Toscans in Italy; YRI: Yoruba in Ibadan, Nigeria.

(d) Neuron and progenitor caPeaks enrichment at epigenetically annotated regulatory elements from fetal brain (Epigenetics Roadmap ID = E081).

(e) Comparison of percent variance explained (r^2) for shared neuron/progenitor caQTLs and fetal brain eQTLs (subset to the same sample size). P values are estimated by two-sided paired student-t tests. The center of the box is median of the data, the bounds of the box are 25th percentile and 75th percentile of the data, and the whisker boundary is 1.5 times the IQR. Maximum and minimum are the maximum and minimum of the data.



Extended Data Fig. 5.

Examples of fine-mapping and regulatory mechanisms underlying eQTLs.

(a) Co-localization of a progenitor-specific caQTL and fetal cortical eQTL for ETFDH.

(b) caQTL for rs11544037 and the labeled peak in progenitor (N=76). P-values are estimated by a mixed linear effects model using a two-sided test (Methods).

(c) eQTL of ETFDH in bulk fetal cortex (N=235). P-values are estimated by a mixed linear effects model using a two-sided test (Methods).

(d) The expression of TFs whose motifs are disrupted by rs11544037[22] (logFC=−0.32, FDR=7.55e-18)[26].

(e) The motif Logo of RAD21, where the red box shows the position disrupted by rs11544037. Schematic cartoon of mechanisms for rs11544037 regulating chromatin accessibility and gene expression.

(f) Luciferase signals for alleles of rs11544037 in progenitors (N=8). P-value is from two-sided paired t-tests.

(g) Co-localization of a progenitor-specific caQTL and eQTL for FGF1.

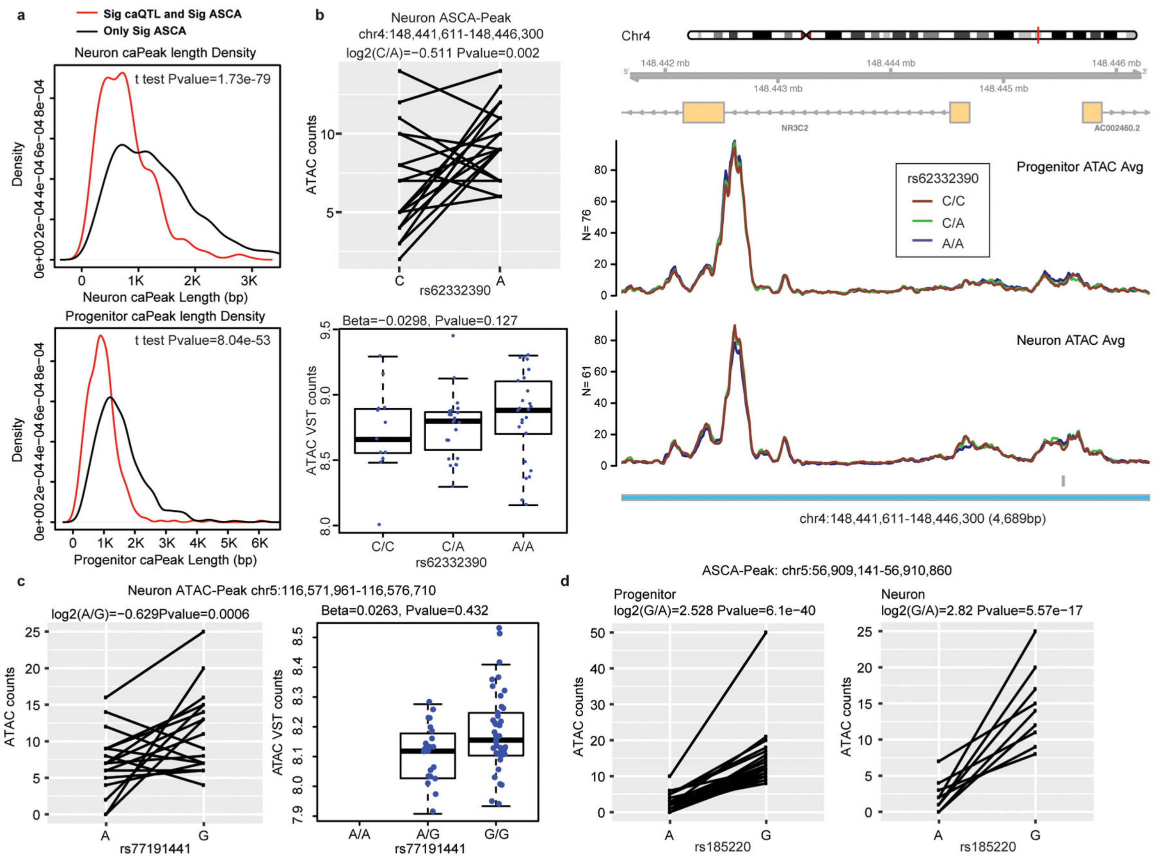
(h) CaQTL for rs11960262 and the labeled peak in progenitor (N=76). P-values are estimated by a mixed linear effects model using a two-sided test (Methods).

(i) eQTL of ETFDH in progenitors (N=85). P-values are estimated by a mixed linear effects model using a two-sided test (Methods).

(j) The expression of TFs in which motifs are disrupted by rs11960262.

(k) The motif Logo of EGR1, where the red box shows the position disrupted by rs11960262. Schematic cartoon of mechanisms for rs11960262 regulating chromatin accessibility and gene expression.

(For box plots in (b-c), (f) and (h-i), the center of the box is the median, the bounds of the box are 25th percentile and 75th percentile of the data, and the whisker boundary is 1.5 times the IQR. Maximum and minimum are the maximum and minimum of the data.)



Extended Data Fig. 6.

Features of ASCA.

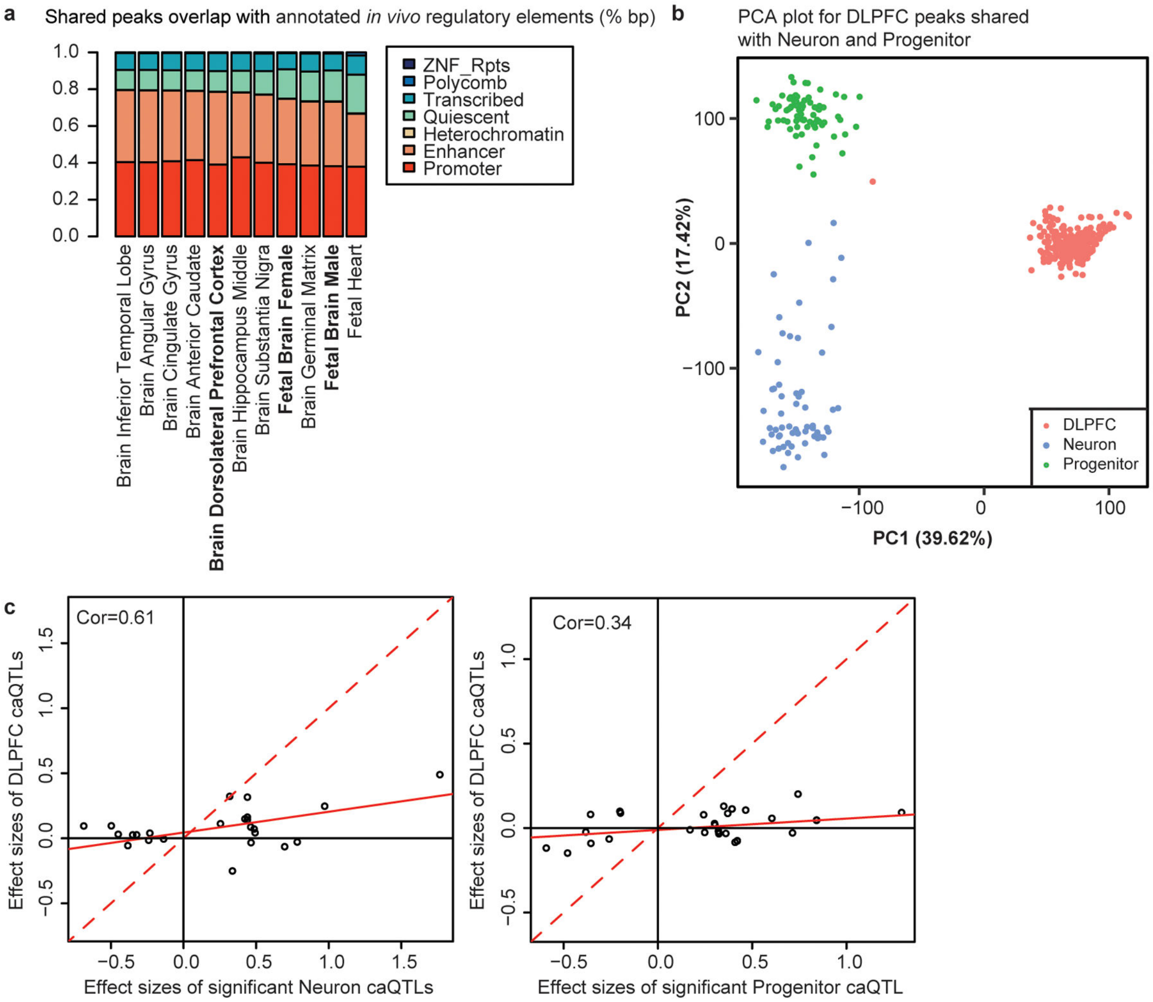
(a) Density plot for caPeak length from shared caQTLs and ASCA, and from peaks only significant in ASCA in neurons (*top*) and progenitors (*bottom*). P values are estimated by two-sided Student’s t-tests.

(b) The neuron ASCA (caSNP: rs62332390; caPeak: chr4:148,441,611–148,46,300; P values are estimated by the negative binomial generalized linear models from DESeq2 using a two-sided test[61]) is not a significant caQTL (N=61; P values are estimated by the mixed linear model using a two sided test) in neurons because the caPeak was very wide (4,689bp) and only the region near the ASCA SNP shows an association with genotype.

(c) The neuron ASCA (caSNP:rs77191441; caPeak:chr5:116,571,961–116,576,710; P values are estimated by the negative binomial generalized linear models from DESeq2 using a two-sided test[61]) is not a significant caQTL (N=61; P values are estimated by the mixed linear effects model with a two-sided test) in neurons due to low minor allele frequency leading to less power to detect a caQTL.

(d) ASCA between rs185220 (see Figure 3) and chromatin accessibility in progenitors (left) and neurons (right). P-values are estimated by the negative binomial generalized linear models from DESeq2 using a two-sided test[61].

(For box plots in (b) and (c), the center of the box is the median, the bounds of the box are 25th percentile and 75th percentile of the data, and the whisker boundary is 1.5 times the IQR. Maximum and minimum are the maximum and minimum of the data.)



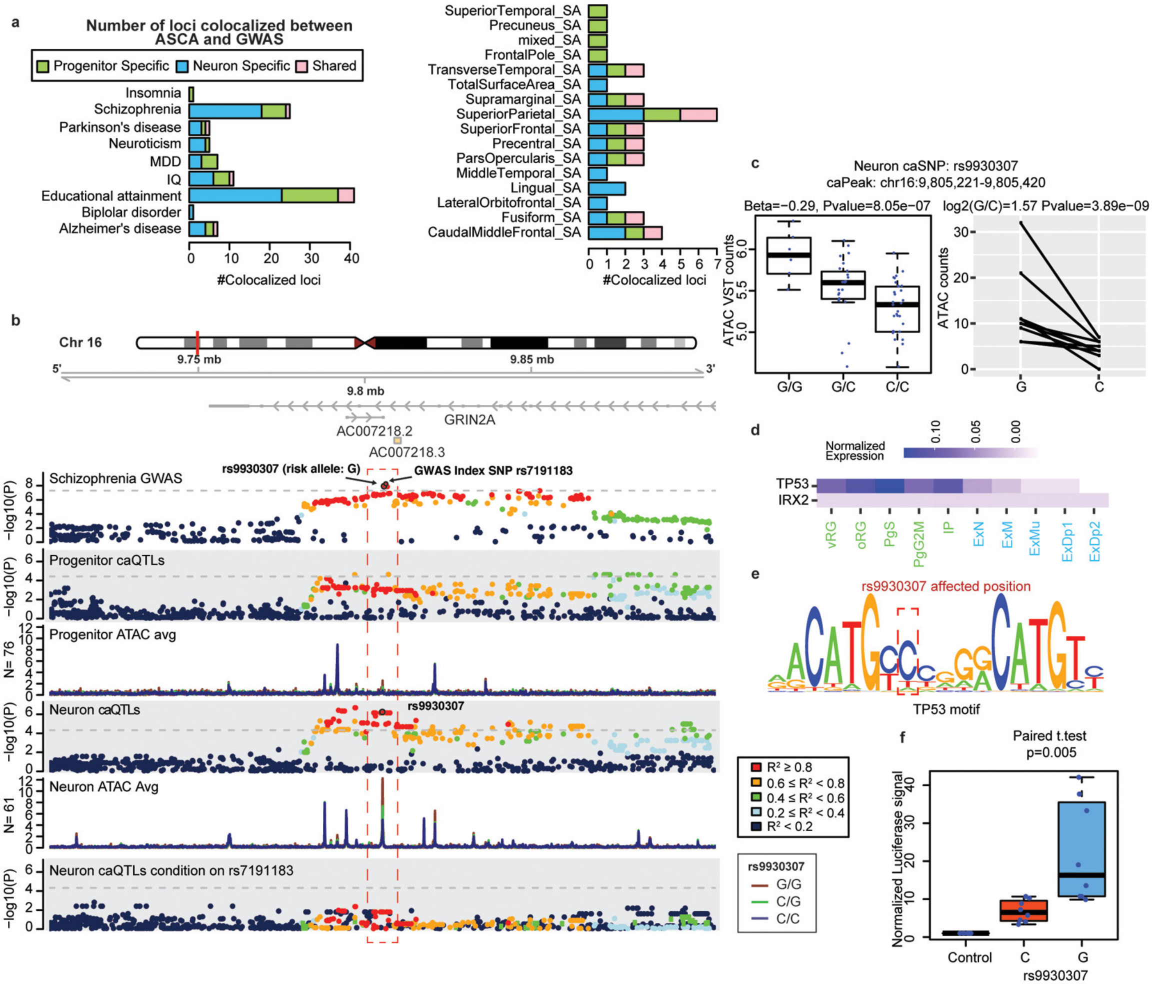
Extended Data Fig. 7.

Comparison to adult dorsolateral prefrontal cortex (DLPFC) caQTLs.

(a) Shared accessible peaks overlap at epigenetically annotated regulatory elements from different tissues. Accessible peak bp percentage overlapped with epigenetically annotated regulatory elements. From left to right, tissues ordered by bp percentage overlap with enhancers and promoters. Shared peaks overlap with both adult and fetal regulatory elements.

(b) PCA plot for read counts from shared peaks in adult DLPFC, neurons and progenitors.

(c) Correlations of effect sizes for significant neuron caQTLs and the same SNP-Peak pairs in adult DLPFC (left). Correlations of effect sizes for significant progenitor caQTLs and the same SNP-Peak pairs in adult DLPFC (right).



Extended Data Fig. 8.

An example of a neuron-specific caQTL leading to regulatory mechanisms underlying GWAS loci.

(a) Numbers of colocalizations between ASCA and GWAS loci.

(b) The neuron-specific significant caQTL (caSNP: rs9930307; caPeak: chr16: 9,805,221–9,805,420) co-localized with schizophrenia GWAS locus (index SNP: rs7191183).

(c) Box plot for the caQTL (left, N=61; P values are estimated by the mixed linear effects model using a two-sided test) and ASCA (right) (caSNP: rs9930307; caPeak: chr16: 9,805,221–9,805,420; P values are estimated by the negative binomial generalized linear models from DESeq2 using a two-sided test[61]).

(d) he expression of TFs in which motifs are disrupted by rs9930307.

(e) The motif logo of TP53 and the position disrupted by rs9930307.

(f) The box plot for luciferase signal for alleles of rs9930307 in progenitors (N=8). P value is from two-sided paired student-t tests.

(For box plots in (c) and (f), the center of the box is median of the data, the bounds of the box are 25th percentile and 75th percentile of the data, and the whisker boundary is 1.5 times the IQR. Maximum and minimum are the maximum and minimum of the data.)

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported by NIH (R00MH102357, U54EB020403, R01MH118349, R01MH120125), Brain Research Foundation, and NC TraCS Pilot funding to JLS. DHG was supported by NIH (R37 MH060233, R01 MH094714, UO1MH116489 and R01 MH110927). The following core facilities were utilized for this project: UNC Neuroscience Center Microscopy Core (P30NS045892), UNC Mammalian Genotyping Core, CGIBD Advanced Analytics Core (NIH grant P30 DK034987), UNC Flow Cytometry Core Facility, UNC Vector Core, UNC Research Computing. Additional core facilities utilized for this project were: UCLA CFAR (5P30 AI028697), and the UCLA Neuroscience Genomics Core. We thank Dr. Karen L. Mohlke for helpful comments.

Adult caQTLs were supported by the PsychENCODE Consortium: U01MH103392, U01MH103365, U01MH103346, U01MH103340, U01MH103339, R21MH109956, R21MH105881, R21MH105853, R21MH103877, R21MH102791, R01MH111721, R01MH110928, R01MH110927, R01MH110926, R01MH110921, R01MH110920, R01MH110905, R01MH109715, R01MH109677, R01MH105898, R01MH105898, R01MH094714, P50MH106934, U01MH116488, U01MH116487, U01MH116492, U01MH116489, U01MH116438, U01MH116441, U01MH116442, R01MH114911, R01MH114899, R01MH114901, R01MH117293, R01MH117291, R01MH117292 awarded to: Schahram Akbarian (Icahn School of Medicine at Mount Sinai), Gregory Crawford (Duke University), Stella Dracheva (Icahn School of Medicine at Mount Sinai), Peggy Farnham (University of Southern California), Mark Gerstein (Yale University), Daniel Geschwind (University of California, Los Angeles), Fernando Goes (Johns Hopkins University), Thomas M. Hyde (Lieber Institute for Brain Development), Andrew Jaffe (Lieber Institute for Brain Development), James A. Knowles (University of Southern California), Chunyu Liu (SUNY Upstate Medical University), Dalila Pinto (Icahn School of Medicine at Mount Sinai), Panos Roussos (Icahn School of Medicine at Mount Sinai), Stephan Sanders (University of California, San Francisco), Nenad Sestan (Yale University), Pamela Sklar (Icahn School of Medicine at Mount Sinai), Matthew State (University of California, San Francisco), Patrick Sullivan (University of North Carolina), Flora Vaccarino (Yale University), Daniel Weinberger (Lieber Institute for Brain Development), Sherman Weissman (Yale University), Kevin White (University of Chicago), Jeremy Willsey (University of California, San Francisco), and Peter Zandi (Johns Hopkins University. Acknowledge the ENCODE Consortium and the ENCODE production laboratory(s) generating the particular dataset(s).

References

1. Grasby KL et al. The genetic architecture of the human cerebral cortex. bioRxiv 399402 (2018) doi:10.1101/399402.
2. Sullivan PF & Geschwind DH Defining the Genetic, Genomic, Cellular, and Diagnostic Architectures of Psychiatric Disorders. Cell 177, 162–183 (2019). [PubMed: 30901538]
3. Barešić A, Nash AJ, Dahoun T, Howes O & Lenhard B Understanding the genetics of neuropsychiatric disorders: the potential role of genomic regulatory blocks. Mol. Psychiatry (2019) doi:10.1038/s41380-019-0518-x.
4. Gamazon ER et al. Using an atlas of gene regulation across 44 human tissues to inform complex disease- and trait-associated variation. Nat. Genet 50, 956–967 (2018). [PubMed: 29955180]
5. Lee PH et al. Principles and methods of in-silico prioritization of non-coding regulatory variants. Hum. Genet 137, 15–30 (2018). [PubMed: 29288389]
6. Albert FW & Kruglyak L The role of regulatory variation in complex traits and disease. Nat. Rev. Genet 16, 197–212 (2015). [PubMed: 25707927]
7. Kumasaka N, Knights AJ & Gaffney DJ High-resolution genetic mapping of putative causal interactions between regions of open chromatin. Nat. Genet 51, 128–137 (2019). [PubMed: 30478436]
8. Davis CA et al. The Encyclopedia of DNA elements (ENCODE): data portal update. Nucleic Acids Res 46, D794–D801 (2018). [PubMed: 29126249]
9. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature 489, 57–74 (2012). [PubMed: 22955616]
10. Consortium GTEx et al. Genetic effects on gene expression across human tissues. Nature 550, 204–213 (2017). [PubMed: 29022597]

11. Wang D et al. Comprehensive functional genomic resource and integrative model for the human brain. Science 362, (2018).
12. PsychENCODE Consortium et al. The PsychENCODE project. Nat. Neurosci 18, 1707–1712 (2015). [PubMed: 26605881]
13. Won H et al. Chromosome conformation elucidates regulatory relationships in developing human brain. Nature 538, 523–527 (2016). [PubMed: 27760116]
14. Walker RL et al. Genetic Control of Expression and Splicing in Developing Human Brain Informs Disease Mechanisms. Cell 179, 750–771.e22 (2019). [PubMed: 31626773]
15. de la Torre-Ubieta L et al. The Dynamic Landscape of Open Chromatin during Human Cortical Neurogenesis. Cell 172, 289–304.e18 (2018). [PubMed: 29307494]
16. Bryois J et al. Evaluation of chromatin accessibility in prefrontal cortex of individuals with schizophrenia. Nat. Commun 9, 3121 (2018). [PubMed: 30087329]
17. Schwartzentruber J et al. Molecular and functional variation in iPSC-derived sensory neurons. Nat. Genet 50, 54–61 (2018). [PubMed: 29229984]
18. Stein JL et al. A quantitative framework to evaluate modeling of cortical development by neural stem cells. Neuron 83, 69–86 (2014). [PubMed: 24991955]
19. Buenrostro JD, Giresi PG, Zaba LC, Chang HY & Greenleaf WJ Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nat. Methods 10, 1213–1218 (2013). [PubMed: 24097267]
20. Hansen DV, Lui JH, Parker PRL & Kriegstein AR Neurogenic radial glia in the outer subventricular zone of human neocortex. Nature 464, 554–561 (2010). [PubMed: 20154730]
21. Pollen AA et al. Molecular identity of human outer radial glia during cortical development. Cell 163, 55–67 (2015). [PubMed: 26406371]
22. Polioudakis D et al. A Single-Cell Transcriptomic Atlas of Human Neocortical Development during Mid-gestation. Neuron 103, 785–801.e8 (2019). [PubMed: 31303374]
23. Kang HM et al. Variance component model to account for sample structure in genome-wide association studies. Nat. Genet 42, 348–354 (2010). [PubMed: 20208533]
24. Yang J, Zaitlen NA, Goddard ME, Visscher PM & Price AL Advantages and pitfalls in the application of mixed-model association methods. Nat. Genet 46, 100–106 (2014). [PubMed: 24473328]
25. Pickrell JK et al. Understanding mechanisms underlying human gene expression variation with RNA sequencing. Nature 464, 768–772 (2010). [PubMed: 20220758]
26. Aygün N et al. Genetic influences on cell type specific gene expression and splicing during neurogenesis elucidate regulatory mechanisms of brain traits. bioRxiv 2020.10.21.349019 (2020) doi:10.1101/2020.10.21.349019.
27. Huang QQ, Ritchie SC, Brozynska M & Inouye M Power, false discovery rate and Winner's Curse in eQTL studies. Nucleic Acids Res 46, e133 (2018). [PubMed: 30189032]
28. Gate RE et al. Genetic determinants of co-accessible chromatin regions in activated T cells across humans. Nat. Genet 50, 1140–1150 (2018). [PubMed: 29988122]
29. Loeb-Hennard C, Cousin X, Prengel I & Kremmer E Cloning and expression pattern of vat-1 homolog gene in zebrafish. Gene Expr. Patterns 5, 91–96 (2004). [PubMed: 15533823]
30. Feng L, Hatten ME & Heintz N Brain lipid-binding protein (BLBP): a novel signaling system in the developing mammalian CNS. Neuron 12, 895–908 (1994). [PubMed: 8161459]
31. Hsu Y-C et al. Brain-specific 1B promoter of FGF1 gene facilitates the isolation of neural stem/progenitor cells with self-renewal and multipotent capacities. Dev. Dyn 238, 302–314 (2009). [PubMed: 18855895]
32. Ballas N, Grunseich C, Lu DD, Speh JC & Mandel G REST and its corepressors mediate plasticity of neuronal gene chromatin throughout neurogenesis. Cell 121, 645–657 (2005). [PubMed: 15907476]
33. Pastinen T Genome-wide allele-specific analysis: insights into regulatory variation. Nat. Rev. Genet 11, 533–538 (2010). [PubMed: 20567245]
34. Heinz S, Romanoski CE, Benner C & Glass CK The selection and function of cell type-specific enhancers. Nat. Rev. Mol. Cell Biol 16, 144–154 (2015). [PubMed: 25650801]

35. Diedenhofen B & Musch J cocor: a comprehensive solution for the statistical comparison of correlations. PLoS One 10, e0121945 (2015). [PubMed: 25835001]

36. Behera V et al. Exploiting genetic variation to uncover rules of transcription factor binding and chromatin accessibility. Nat. Commun 9, 782 (2018). [PubMed: 29472540]

37. Bergsland M, Werme M, Malewicz M, Perlmann T & Muhr J The establishment of neuronal properties is controlled by Sox4 and Sox11. Genes Dev. 20, 3475–3486 (2006). [PubMed: 17182872]

38. Shibata M et al. Regulation of Prefrontal Patterning, Connectivity and Synaptogenesis by Retinoic Acid. doi:10.1101/2019.12.31.891036.

39. Tsunemoto R et al. Diverse reprogramming codes for neuronal identity. Nature 557, 375–380 (2018). [PubMed: 29743677]

40. He X et al. Expression of a large family of POU-domain regulatory genes in mammalian brain development. Nature 340, 35–41 (1989). [PubMed: 2739723]

41. Wang H et al. ZEB1 Represses Neural Differentiation and Cooperates with CTBP2 to Dynamically Regulate Cell Migration during Neocortex Development. Cell Rep. 27, 2335–2353.e6 (2019). [PubMed: 31116980]

42. Rakic P Specification of cerebral cortical areas. Science 241, 170–176 (1988). [PubMed: 3291116]

43. Li S et al. Regulatory mechanisms of major depressive disorder risk variants. Mol. Psychiatry (2020) doi:10.1038/s41380-020-0715-7.

44. Dobbyn A et al. Landscape of Conditional eQTL in Dorsolateral Prefrontal Cortex and Co-localization with Schizophrenia GWAS. Am. J. Hum. Genet 102, 1169–1184 (2018). [PubMed: 29805045]

45. Li M et al. Integrative functional genomic analysis of human brain development and neuropsychiatric risks. Science 362, (2018).

46. Skene NG et al. Genetic identification of brain cell types underlying schizophrenia. Nat. Genet 50, 825–833 (2018). [PubMed: 29785013]

47. Gaffney DJ et al. Dissecting the regulatory architecture of gene expression QTLs. Genome Biol. 13, R7 (2012). [PubMed: 22293038]

48. Bell JT et al. DNA methylation patterns associate with genetic and gene expression variation in HapMap cell lines. Genome Biol. 12, R10 (2011). [PubMed: 21251332]

49. Chen K & Rajewsky N Natural selection on human microRNA binding sites inferred from SNP data. Nat. Genet 38, 1452–1456 (2006). [PubMed: 17072316]

50. Alasoo K et al. Shared genetic effects on chromatin and gene expression indicate a role for enhancer priming in immune response. Nat. Genet 50, 424–431 (2018). [PubMed: 29379200]

51. Ellis P et al. SOX2, a persistent marker for multipotential neural stem cells derived from embryonic stem cells, the embryo or the adult. Dev. Neurosci 26, 148–165 (2004). [PubMed: 15711057]

52. Han DW et al. Direct reprogramming of fibroblasts into neural stem cells by defined factors. Cell Stem Cell 10, 465–472 (2012). [PubMed: 22445517]

53. Liang H et al. Neural development is dependent on the function of specificity protein 2 in cell cycle progression. Development 140, 552–561 (2013). [PubMed: 23293287]

54. Naka H, Nakamura S, Shimazaki T & Okano H Requirement for COUP-TFI and II in the temporal specification of neural stem cells in CNS development. Nat. Neurosci 11, 1014–1023 (2008). [PubMed: 19160499]

55. Liu Y & Zhang Y ETV5 is Essential for Neuronal Differentiation of Human Neural Progenitor Cells by Repressing NEUROG2 Expression. Stem Cell Rev Rep 15, 703–716 (2019). [PubMed: 31273540]

56. Araújo JA de M et al. Direct Reprogramming of Adult Human Somatic Stem Cells Into Functional Neurons Using Sox2, Ascl1, and Neurog2. Front. Cell. Neurosci 12, 155 (2018). [PubMed: 29937717]

57. Fathi A, Rasouli H, Yeganeh M, Salekdeh GH & Baharvand H Efficient differentiation of human embryonic stem cells toward dopaminergic neurons using recombinant LMX1A factor. Mol. Biotechnol 57, 184–194 (2015). [PubMed: 25380985]

58. Cubelos B, Briz CG, Esteban-Ortega GM & Nieto M Cux1 and Cux2 selectively target basal and apical dendritic compartments of layer II-III cortical neurons. Dev. Neurobiol 75, 163–172 (2015). [PubMed: 25059644]
59. Zimmer C, Tiveron M-C, Bodmer R & Cremer H Dynamics of Cux2 expression suggests that an early pool of SVZ precursors is fated to become upper cortical layer neurons. Cereb. Cortex 14, 1408–1420 (2004). [PubMed: 15238450]
60. Love MI, Huber W & Anders S Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 15, 550 (2014). [PubMed: 25516281]
61. Ritchie ME et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 43, e47 (2015). [PubMed: 25605792]
62. Han B & Eskin E Interpreting meta-analyses of genome-wide association studies. PLoS Genet. 8, e1002555 (2012). [PubMed: 22396665]
63. Touzet H & Varré J-S Efficient and accurate P-value computation for Position Weight Matrices. Algorithms Mol. Biol 2, 15 (2007). [PubMed: 18072973]
64. Thiel G, Greengard P & Südhof TC Characterization of tissue-specific transcription by the human synapsin I gene promoter. Proc. Natl. Acad. Sci. U. S. A 88, 3431–3435 (1991). [PubMed: 1849657]
65. Buenrostro JD, Wu B, Chang HY & Greenleaf WJ ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide. Curr. Protoc. Mol. Biol 109, 21.29.1–9 (2015).
66. Orchard P, Kyono Y, Hensley J, Kitzman JO & Parker SCJ Quantification, Dynamic Visualization, and Validation of Bias in ATAC-Seq Data with ataqv. Cell Syst 10, 298–306.e4 (2020). [PubMed: 32213349]
67. Li H Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv [q-bio.GN] (2013).
68. van de Geijn B, McVicker G, Gilad Y & Pritchard JK WASP: allele-specific software for robust molecular quantitative trait locus discovery. Nat. Methods 12, 1061–1063 (2015). [PubMed: 26366987]
69. Li H et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25, 2078–2079 (2009). [PubMed: 19505943]
70. Quinlan AR & Hall IM BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26, 841–842 (2010). [PubMed: 20110278]
71. Feng J, Liu T & Zhang Y Using MACS to identify peaks from ChIP-Seq data. Curr. Protoc. Bioinformatics Chapter 2, Unit 2.14 (2011).
72. Lun ATL & Smyth GK csaw: a Bioconductor package for differential binding analysis of ChIP-seq data using sliding windows. Nucleic Acids Res 44, e45 (2016). [PubMed: 26578583]
73. Hansen KD, Irizarry RA & Wu Z Removing technical variability in RNA-seq data using conditional quantile normalization. Biostatistics 13, 204–216 (2012). [PubMed: 22285995]
74. Roadmap Epigenomics Consortium et al. Integrative analysis of 111 reference human epigenomes. Nature 518, 317–330 (2015). [PubMed: 25693563]
75. McLean CY et al. GREAT improves functional interpretation of cis-regulatory regions. Nat. Biotechnol 28, 495–501 (2010). [PubMed: 20436461]
76. Ernst J & Kellis M Large-scale imputation of epigenomic datasets for systematic annotation of diverse human tissues. Nat. Biotechnol 33, 364–376 (2015). [PubMed: 25690853]
77. Alexa A & Rahnenfuhrer J topGO: enrichment analysis for gene ontology. R package version 2, (2010).
78. Mathelier A et al. JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles. Nucleic Acids Res. 44, D110–5 (2016). [PubMed: 26531826]
79. Pollard KS, Hubisz MJ, Rosenbloom KR & Siepel A Detection of nonneutral substitution rates on mammalian phylogenies. Genome Res. 20, 110–121 (2010). [PubMed: 19858363]
80. Chang CC et al. Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience 4, 7 (2015). [PubMed: 25722852]
81. Jun G et al. Detecting and estimating contamination of human DNA samples in sequencing and array-based genotype data. Am. J. Hum. Genet 91, 839–848 (2012). [PubMed: 23103226]

82. Delaneau O, Marchini J & Zagury J-F A linear complexity phasing method for thousands of genomes. Nat. Methods 9, 179–181 (2011). [PubMed: 22138821]

83. Das S et al. Next-generation genotype imputation service and methods. Nat. Genet 48, 1284–1287 (2016). [PubMed: 27571263]

84. Price AL, Zaitlen NA, Reich D & Patterson N New approaches to population stratification in genome-wide association studies. Nat. Rev. Genet 11, 459–463 (2010). [PubMed: 20548291]

85. Davis JR et al. An Efficient Multiple-Testing Adjustment for eQTL Studies that Accounts for Linkage Disequilibrium between Variants. Am. J. Hum. Genet 98, 216–224 (2016). [PubMed: 26749306]

86. Benjamini Y & Hochberg Y Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. J. R. Stat. Soc. Series B Stat. Methodol 57, 289–300 (1995).

87. Shim H et al. A multivariate genome-wide association analysis of 10 LDL subfractions, and their response to statin treatment, in 1868 Caucasians. PLoS One 10, e0120758 (2015). [PubMed: 25898129]

88. Kang HM et al. Efficient control of population structure in model organism association mapping. Genetics 178, 1709–1723 (2008). [PubMed: 18385116]

89. Dabney A, Storey JD & Warnes GR qvalue: Q-value estimation for false discovery rate control. R package version 1, (2010).

90. Coetzee SG, Coetzee GA & Hazelett DJ motifbreakR: an R/Bioconductor package for predicting variant effects at transcription factor binding sites. Bioinformatics 31, 3847–3849 (2015). [PubMed: 26272984]

91. Shannon P & Richards M MotifDb: An annotated collection of protein-DNA binding sequence motifs. R package version 1, (2014).

92. Finucane HK et al. Partitioning heritability by functional category using GWAS summary statistics. bioRxiv 014241 (2015) doi:10.1101/014241.

93. Demontis D et al. Discovery of the first genome-wide significant risk loci for attention deficit/hyperactivity disorder. Nat. Genet 51, 63–75 (2019). [PubMed: 30478444]

94. Grove J et al. Identification of common genetic risk variants for autism spectrum disorder. Nat. Genet 51, 431–444 (2019). [PubMed: 30804558]

95. Savage JE et al. Genome-wide association meta-analysis in 269,867 individuals identifies new genetic and functional links to intelligence. Nat. Genet 50, 912–919 (2018). [PubMed: 29942086]

96. Wray NR et al. Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression. Nat. Genet 50, 668–681 (2018). [PubMed: 29700475]

97. Stahl EA et al. Genome-wide association study identifies 30 loci associated with bipolar disorder. Nat. Genet 51, 793–803 (2019). [PubMed: 31043756]

98. Pardiñas AF et al. Common schizophrenia alleles are enriched in mutation-intolerant genes and in regions under strong background selection. Nat. Genet 50, 381–389 (2018). [PubMed: 29483656]

99. Jansen PR et al. Genome-wide analysis of insomnia in 1,331,010 individuals identifies new risk loci and functional pathways. Nat. Genet 51, 394–403 (2019). [PubMed: 30804565]

100. Lee JJ et al. Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals. Nat. Genet 50, 1112–1121 (2018). [PubMed: 30038396]

101. Okbay A et al. Genetic variants associated with subjective well-being, depressive symptoms, and neuroticism identified through genome-wide analyses. Nat. Genet 48, 624–633 (2016). [PubMed: 27089181]

102. Nagel M et al. Meta-analysis of genome-wide association studies for neuroticism in 449,484 individuals identifies novel genetic loci and pathways. Nat. Genet 50, 920–927 (2018). [PubMed: 29942085]

103. Duncan L et al. Significant Locus and Metabolic Genetic Correlations Revealed in Genome-Wide Association Study of Anorexia Nervosa. Am. J. Psychiatry 174, 850–858 (2017). [PubMed: 28494655]

104. Otowa T et al. Meta-analysis of genome-wide association studies of anxiety disorders. Mol. Psychiatry 21, 1391–1399 (2016). [PubMed: 26754954]

105. Jansen IE et al. Genome-wide meta-analysis identifies new loci and functional pathways influencing Alzheimer's disease risk. Nat. Genet 51, 404–413 (2019). [PubMed: 30617256]

106. International League Against Epilepsy Consortium on Complex Epilepsies. Genome-wide mega-analysis identifies 16 loci and highlights diverse biological mechanisms in the common epilepsies. Nat. Commun 9, 5269 (2018). [PubMed: 30531953]

107. Nalls MA et al. Parkinson's disease genetics: identifying novel risk loci, providing causal insights and improving estimates of heritable risk. bioRxiv 388165 (2018) doi:10.1101/388165.

108. Civelek M et al. Genetic Regulation of Adipose Gene Expression and Cardio-Metabolic Traits. Am. J. Hum. Genet 100, 428–443 (2017). [PubMed: 28257690]

109. Hahne F & Ivanek R Visualizing Genomic Data Using Gviz and Bioconductor. Methods Mol. Biol 1418, 335–351 (2016). [PubMed: 27008022]

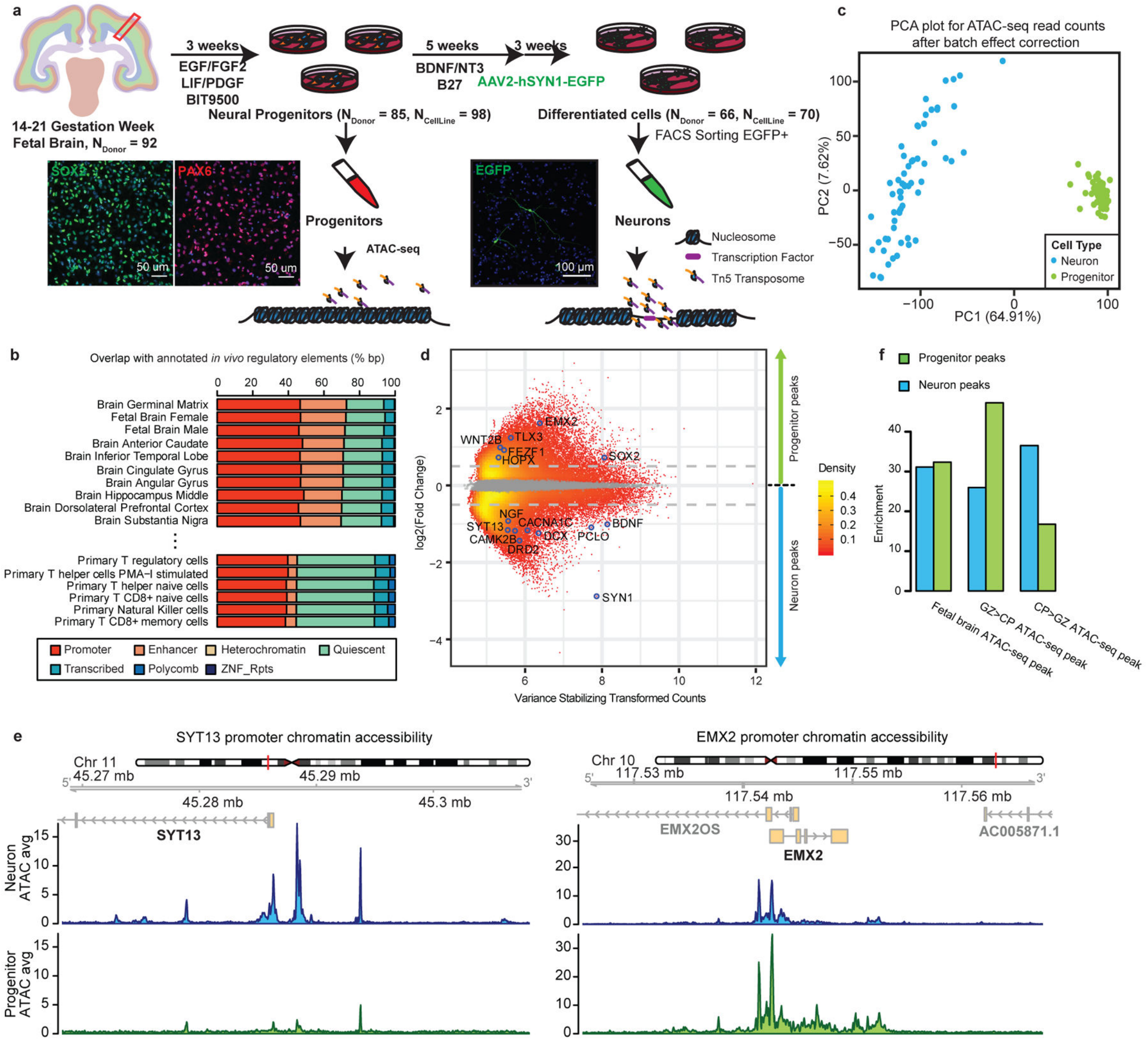


Figure 1:

Profiling genome-wide chromatin accessibility in progenitors and neurons.

(a) Schematic cartoon of experimental design. SOX2 and PAX6 immunolabeled neural progenitors (*left*), showing over 90% of cells were radial glia. EGFP labeled differentiated neurons (*right*), showing expected neuronal morphology.

(b) Percentage of accessible peak base pairs (bp) detected in these cultures overlapped with epigenetically annotated regulatory elements from multiple tissues. From top to bottom, tissues ordered by bp percentage overlap with enhancers and promoters.

(c) PCA plot of ATAC-seq read count after batch effect correction colored by cell types, showing two separate clusters for progenitors and neurons.

(d) MA plot for differentially accessible peaks between progenitors and neurons. All peaks can be found in Supplementary Table 1.

(e) ATAC-seq coverage plot (average normalized read counts) for promoters of neuron expressed gene SYT13, showing higher accessibility in neurons than progenitors (LFC= −1.16, FDR=3.28e-35). ATAC-seq coverage plot (average normalized read counts) for promoter of progenitor expressed gene EMX2, showing higher accessibility in progenitors than neurons (LFC=1.62, FDR=1.12e-32).

(f) Enrichment of neuron/progenitor peaks with differentially accessible peaks from fetal brain tissue[15]. GZ: neural progenitor-enriched region encompassing the ventricular zone (VZ), subventricular zone (SVZ), and intermediate zone (IZ); CP: the neuron-enriched region containing the subplate (SP), cortical plate (CP), and marginal zone (MZ).

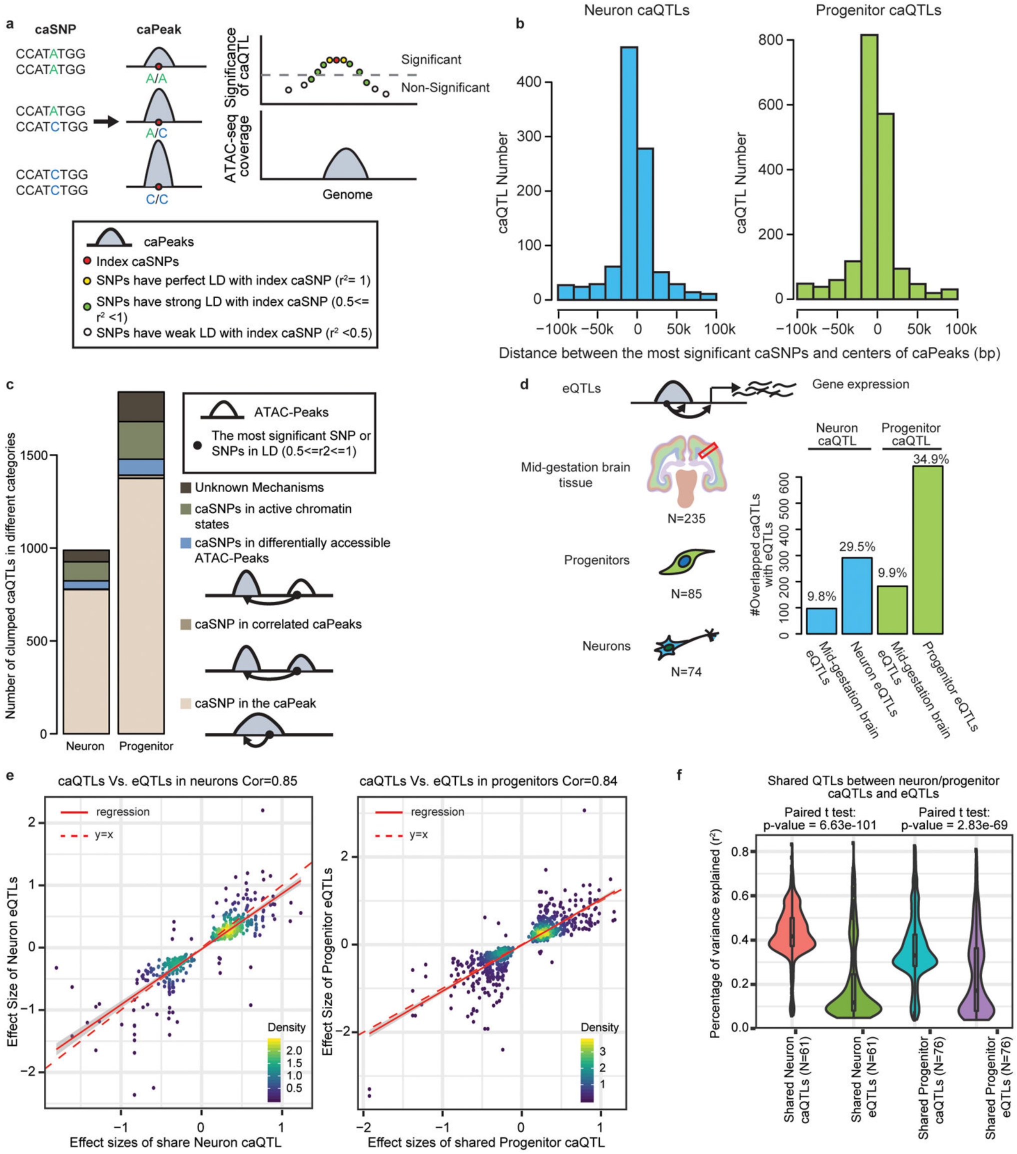


Figure 2:

Chromatin accessibility quantitative trait loci (caQTL) in progenitors and neurons.

(a) caQTL schematic.

(b) Number of the most significant caSNPs relative to the distance from the center of the caPeaks (*left*: neuron caQTLs; *right*: progenitor caQTLs). The most significant caQTLs for each caPeak can be found in Supplementary Table 3.

(c) Numbers of caQTLs in different functional categories.

(d) Schematic cartoon of fetal cortical[14] and cell-type specific eQTLs[26] (*Left*). Percentage of neuron/progenitor caQTLs with shared effects in fetal cortical and cell-type specific eQTLs (All shared caQTLs and eQTLs can be found in Supplementary Table 4.).

(e) For the most significant caSNP for each caPeak, correlation of effect sizes between shared caQTLs and eQTLs in neurons (*left*) and progenitors (*right*).

(f) Comparison of percent variance explained (r^2) for shared caQTLs and eQTLs (subset to the same sample size) in neurons and progenitors. We found 500 (e)caSNP-caPeak-eGene combinations in neurons and 1,025 (e)caSNP-caPeak-eGene combinations in progenitors. We observed higher percent variance explained for caQTLs than eQTLs in both neurons and progenitors. P values are estimated by the two-sided paired t-test. The center of the box is median of the data, the bounds of the box are 25th percentile and 75th percentile of the data, and the whisker boundary is 1.5 times the IQR. Maximum and minimum are the maximum and minimum of the data.

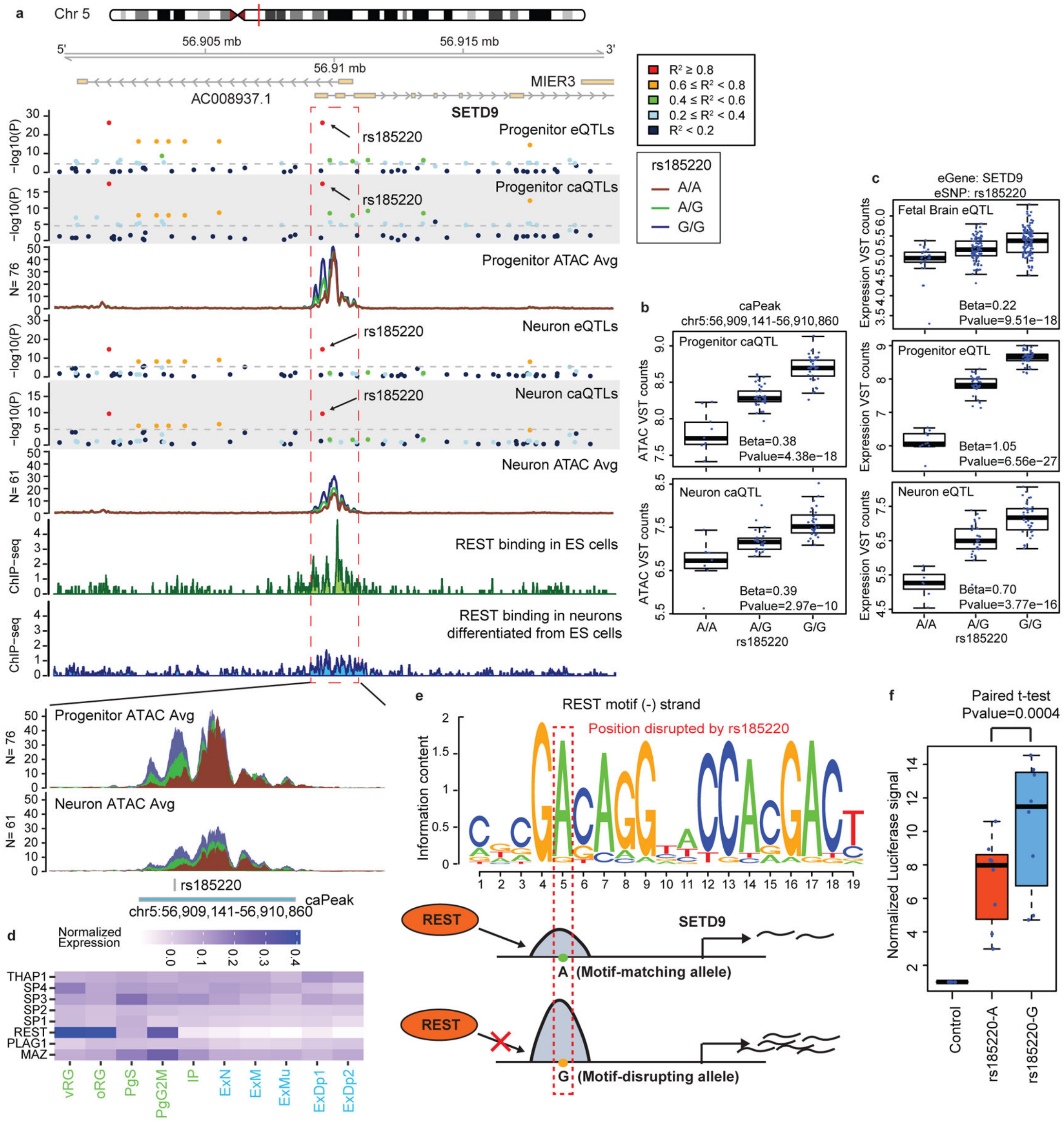


Figure 3:

Fine-mapping and regulatory mechanism underlying eQTLs.

(a) Co-localization of caQTL and eQTL for SETD9. ChIP-seq data of REST binding in H1 cells and neurons differentiated from H1 cells[8,9].

(b) The association between rs185220 and chromatin accessibility of the labeled peak in progenitors (*top*, N=76) and neurons (*bottom*, N=61). P values are estimated by a linear mixed effects model (Methods) using a two-sided test.

(c) The association between rs185220 and expression of *SETD9* in the mid-gestation cortex (*top*, N=235), progenitors (*middle*, N=85) and neurons (*bottom*, N=74). P values are estimated by a linear mixed effects model (Methods) using a two-sided test.

(d) The expression of TFs in which motifs are disrupted by rs185220. vRG: ventricular Radial Glia; oRG: outer Radial Glia; PgS: Cycling progenitors (S phase); PgG2M: Cycling progenitors (G2/M phase); IP: Intermediate progenitors; ExN: Migrating excitatory; ExM:

Maturing excitatory; ExM-U: Maturing excitatory upper enriched; ExDp1: Excitatory deep layer 1; ExDp2: Excitatory deep layer 2.

(e) The motif Logo of *REST*, where the red box shows the position disrupted by rs185220. Schematic cartoon of proposed mechanism for rs185220 regulating chromatin accessibility and gene expression.

(f) The box plot for luciferase signal for alleles of rs185220 in progenitors (N=8). P value is from a two-sided paired t-test.

(For box plots in (b) (c) and (f), the center of the box is median, the bounds of the box are 25th percentile and 75th percentile of the data, and the whisker boundary is 1.5 times the IQR. Maximum and minimum are the maximum and minimum of the data.)

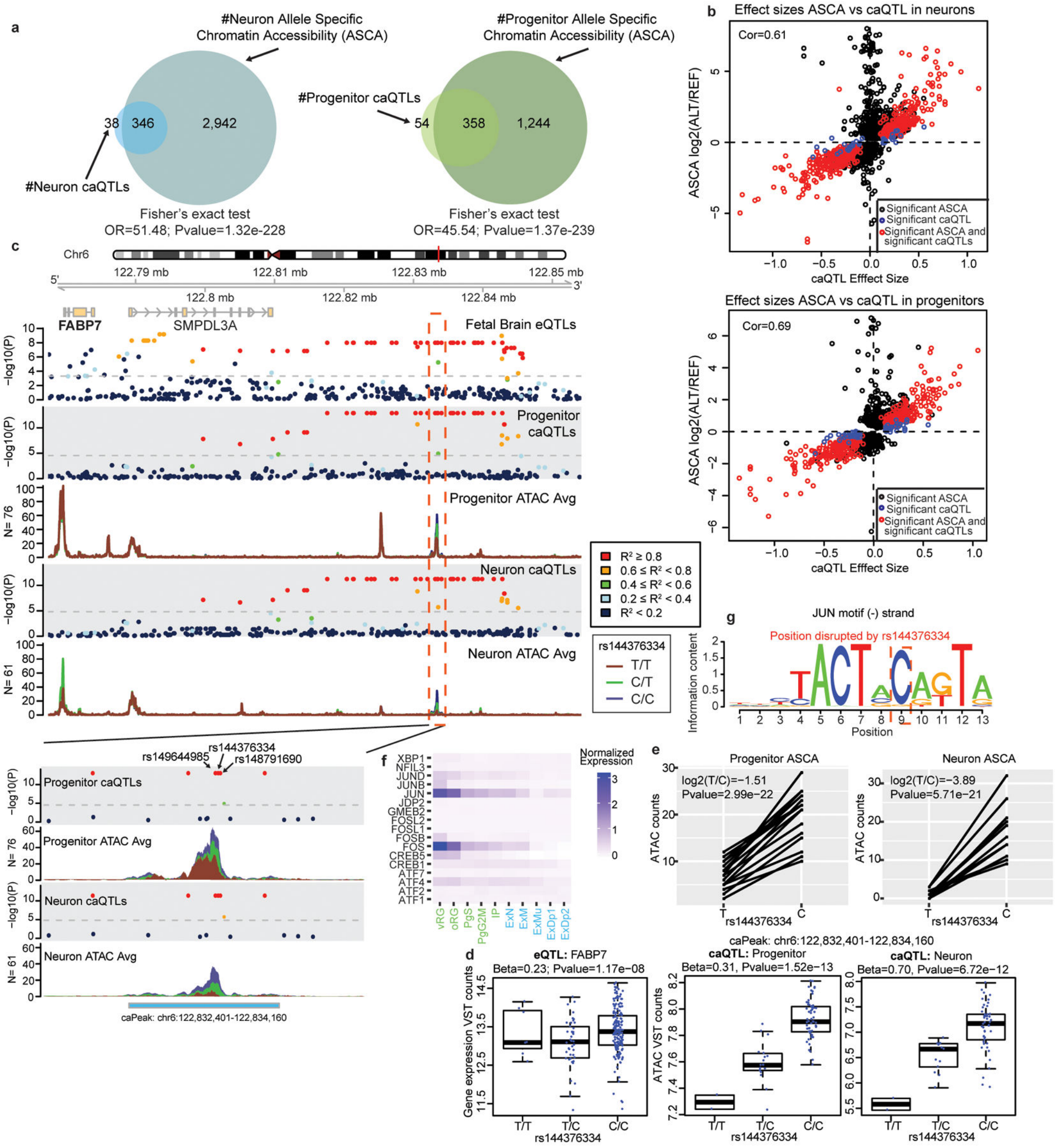


Figure 4:

Allele Specific Chromatin Accessibility (ASCA).

(a) Numbers of shared/non-shared significant caQTLs and significant ASCA in neurons (*left*) and progenitors (*right*). All significant ASCA in neurons and progenitors can be found in Supplementary Table 6.

(b) Correlation of effect sizes for caQTL and ASCA from (A) neurons (*top*) and progenitors (*bottom*).

(c) Co-localization of caQTL and ASCA in progenitors and neurons as well as mid-gestation cortical eQTL for *FABP7*.

(d) Association between rs144376334 and expression of *FABP7* in mid-gestation cortex (*left*, N=235), chromatin accessibility of the labeled peak in progenitors (*middle*, N=85) and neurons (*right*, N=74). P values are estimated using a linear mixed effects model with a two-sided test (Methods). The center of the box is median of the data, the bounds of the box are

25th percentile and 75th percentile of the data, and the whisker boundary is 1.5 times the IQR. Maximum and minimum are the maximum and minimum of the data.

(e) ASCA detected at rs144376334 in progenitors (*left*) and neurons (*right*). P values are estimated using the negative binomial generalized linear model from DESeq2 with a two-sided test (Methods).

(f) The expression of TFs whose motifs are disrupted by rs144376334.

(g) The motif logo of JUN with boxed position disrupted by rs144376334.

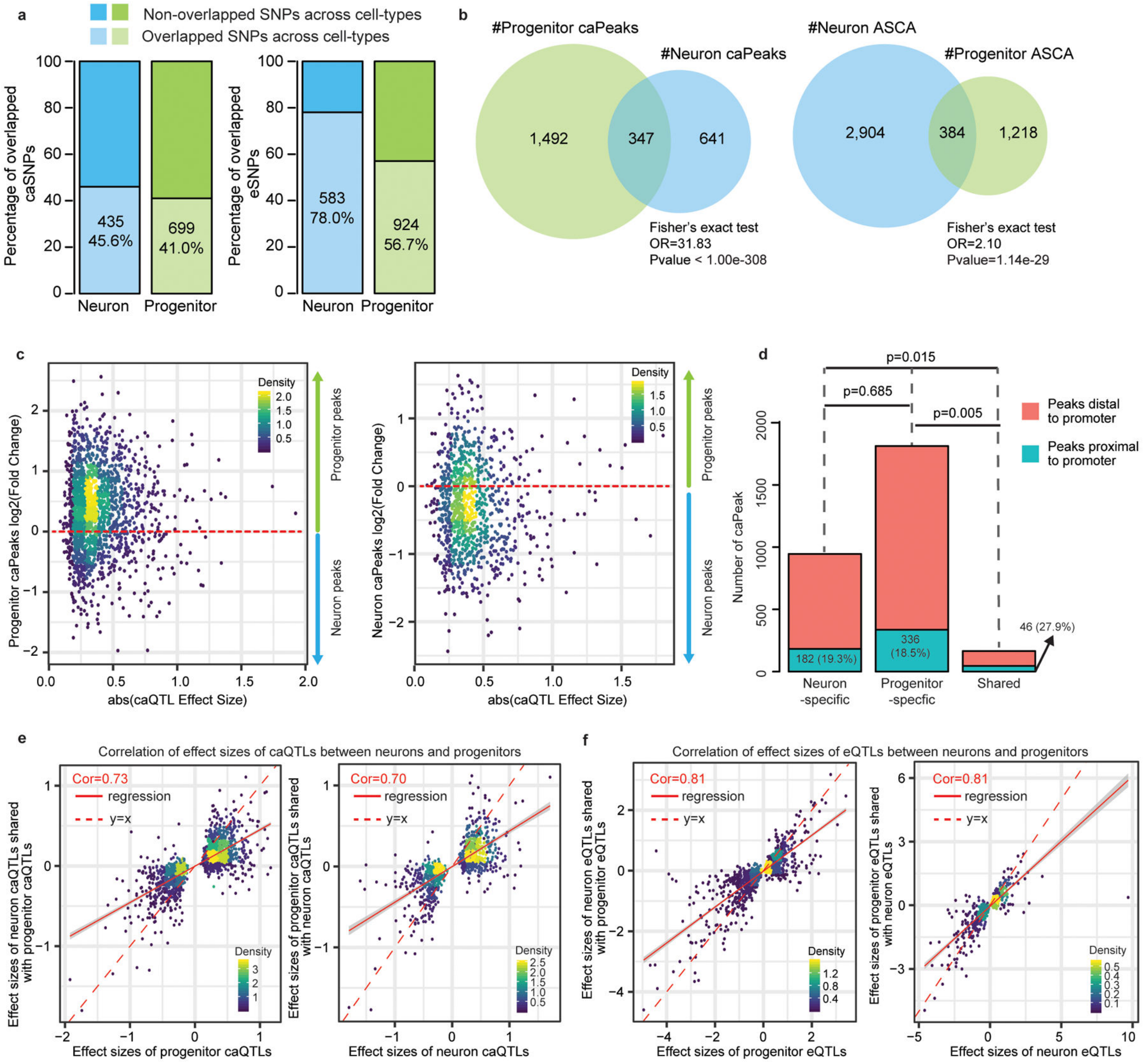


Figure 5:

Cell-type specificity of caQTLs.

(a) Percentage of caQTLs (*left*)/eQTLs (*right*) with shared effects in neurons and progenitors.

(b) Numbers of overlapped/non-overlapped caPeaks (*left*; P-value estimated to be less than the floating point precision value) and ASCA (*right*) between neurons and progenitors. P values are estimated by two-sided fisher's exact tests.

(c) Differential accessibility of progenitor caPeaks (*left*) and neuron caPeaks (*right*).

(d) Numbers of caPeaks distal to promoters or proximal to promoters for neuron-specific caQTLs, progenitor-specific caQTLs and shared caQTLs between neurons and progenitors. P values are estimated by the two-sided two-proportions z-test.

(e) Correlations of effect sizes of caQTLs between neurons and progenitors (*left*: progenitor caQTLs vs. the same caSNP-caPeak pairs in neurons; *right*: neuron caQTLs vs. the same caSNP-caPeak pairs in progenitors).

(f) Correlations of effect sizes of eQTLs between neurons and progenitors (*left*: progenitor eQTLs vs. the same eSNP-eGene pairs in neurons; *right*: neuron eQTLs vs. the same eSNP-eGene pairs in progenitors).

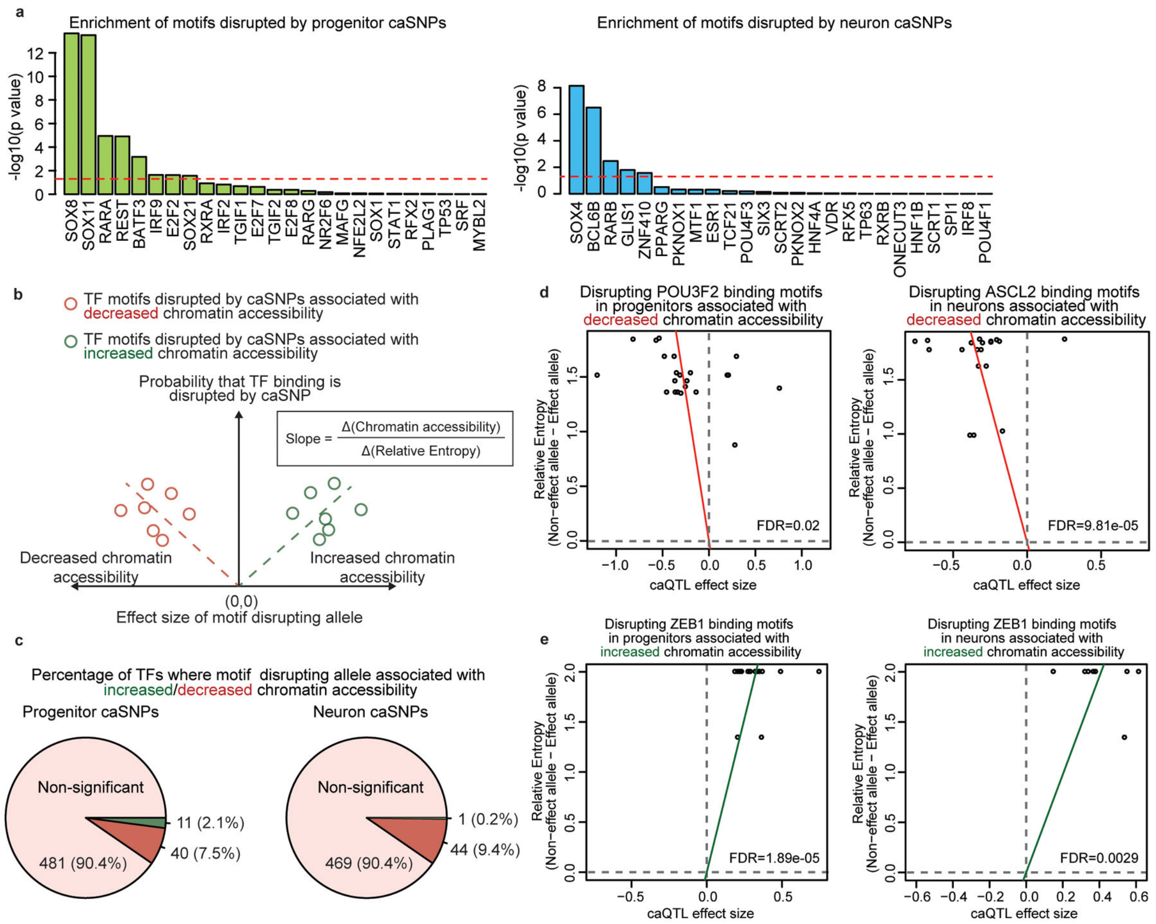


Figure 6:

Prediction of disrupted transcription factor (TF) binding due to genetic variation.

(a) Enrichment of caSNP-disrupted motifs in accessible peaks in progenitors or in neurons.

(b) Schematic of TF motifs disrupted by caSNPs associated with decreased/increased chromatin accessibility.

(c) Numbers of TFs where the motif-disrupting allele was associated with increased/decreased chromatin accessibility in progenitors (*left*) and neurons (*right*). For most TFs, the motif-disrupting allele was associated with decreased chromatin accessibility in progenitors and neurons.

(d) Examples of TF motifs disrupted by caSNPs associated with decreased chromatin accessibility in progenitors (*POU3F2*; *left*) and neurons (*ASCL2*; *right*).

(e) Disrupting *ZEB1* (a transcriptional repressor) binding motif was associated with increased chromatin accessibility in progenitors and neurons.

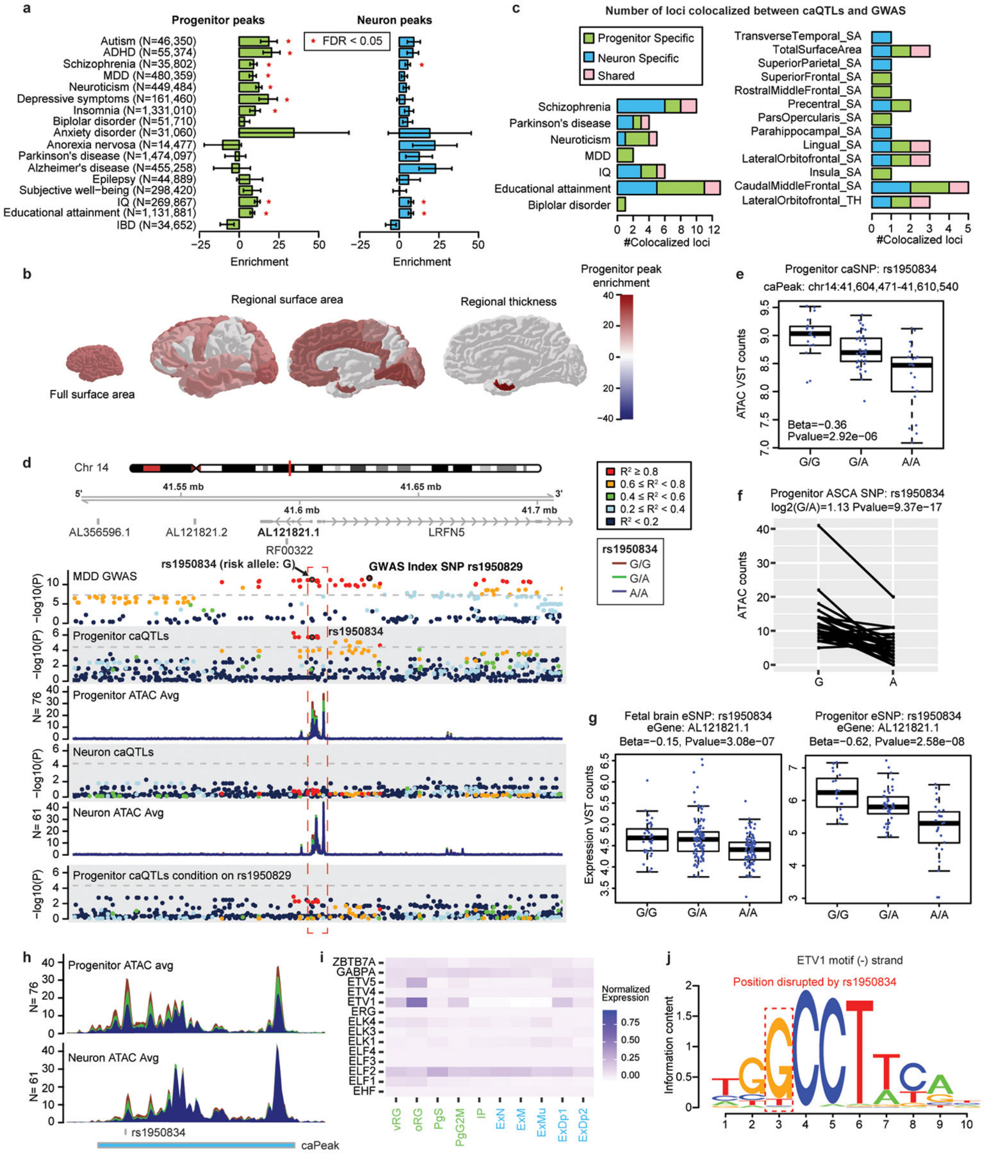


Figure 7:

Cell-type specific caQTLs lead to regulatory mechanisms underlying GWAS loci.

(a) Partitioned heritability enrichment. P values are estimated from LD score regression (two-sided test) and corrected by FDR (Methods). Data are presented as mean values +/− SE.

(b) Partitioned heritability enrichment demonstrated a significant (FDR < 0.05) enrichment of heritability for surface area of the full cortex and other subregions within progenitor peaks.

(c) Numbers of colocalizations between caQTLs and GWAS loci.

(d) A colocalized locus between progenitor-specific caQTL and MDD GWAS.

(e) Association between rs1950834 and chromatin accessibility of the labeled peak in progenitors (N=76). P values are estimated by the mixed effects linear model using a two-sided test.

(f) ASCA of rs1950834 in progenitors. P values are estimated by the negative binomial generalized linear models from DESeq2 using a two-sided test (Methods).

(g) Association between rs1950834 and expression of lncRNA AL12182.1 in fetal brain (*left*, N=235) and progenitors (*right*, N=85). P values are estimated by the linear mixed effects model with a two-sided test.

(h) Zoomed in plot of caPeaks colored by genotype at rs1950834.

(i) The expression of TFs in which motifs are disrupted by rs1950834.

(j) The motif logo of ETV1 where the boxed region is disrupted by rs1950834.

(For box plots in (e) and (g), the center of the box is median of the data, the bounds of the box are 25th percentile and 75th percentile of the data, and the whisker boundary is 1.5 times the IQR. Maximum and minimum are the maximum and minimum of the data.)