Title: Widespread methylation quantitative trait loci and their role in schizophrenia risk

Short Title: Rampant methylation QTLs in schizophrenia risk

One-sentence summary: Most genetic variants associated with DNA methylation levels, and implicated schizophrenia GWAS variants in the human brain.

Authors:

Kira A. Perzel Mandell[1,2], Nicholas J. Eagles[1], Richard Wilton[3], Amanda J. Price[1,2], Stephen A. Semick[1], Leonardo Collado-Torres[1], Ran Tao[1], Shizhong Han[1,4], Alexander S. Szalay[3,5], Thomas M. Hyde[1,4,6], Joel E. Kleinman[1,4], Daniel R. Weinberger[1,2,4,6,7], Andrew E. Jaffe[1,2,4,7,8,9,+]

1. Lieber Institute for Brain Development, Johns Hopkins Medical Campus, Baltimore, MD 21205, USA
2. Department of Genetic Medicine, Johns Hopkins University School of Medicine (JHSOM), Baltimore, MD 21205, USA
3. Department of Physics and Astronomy, Johns Hopkins University, Baltimore, MD, USA
4. Department of Psychiatry and Behavioral Sciences, JHSOM, Baltimore, MD, USA
5. Department of Computer Science, JHSOM, Baltimore, MD, USA
6. Department of Neurology, JHSOM, Baltimore, MD, USA
7. Department of Neuroscience, JHSOM, Baltimore, MD, USA.
8. Department of Mental Health, Johns Hopkins Bloomberg School of Public Health (JHBSPH), MD 21205, USA
9. Department of Biostatistics, JHBSPH, Baltimore, MD, USA.

\+ andrew.jaffe@libd.org

Abstract:

DNA methylation (DNAm) regulates gene expression and may represent gene-environment interactions. Using whole genome bisulfite sequencing, we surveyed DNAm in a large sample (n=344) of human brain tissues. We identify widespread genetic influence on local methylation levels throughout the genome, with 76% of SNPs and 38% of CpGs being part of methylation quantitative trait loci (meQTLs). These associations can further be clustered into regions that are differentially methylated by a given SNP, highlighting putative functional regions that explain much of the heritability associated with risk loci. Furthermore, some CpH sites associated with genetic variation. We have established a comprehensive, single base resolution view of association between genetic variation and genomic methylation, and implicate schizophrenia GWAS-associated variants as influencing the epigenetic plasticity of the brain.

Main Text:

Introduction

DNA methylation (DNAm) plays an important role in the epigenetic regulation of gene expression. It varies throughout development and among tissue types, and has been thought to be a high-fidelity representation of the interaction between genes and environment. While some variation in DNAm can be attributed to developmental and exogenous factors, including diet (*1*) and cigarette smoking (*2*), Davies et al. (*3*) identified some inter-individual variation that is consistent across tissue types. This provided evidence that DNA sequence drives DNA methylation levels, at sites known as methylation quantitative trait loci (meQTLs). Inter-individual DNAm differences have since been confirmed by twin studies (*4*, *5*). Initial studies found methylation association with sequence variants at specific loci (*6*). These putative epigenetic associations extend beyond the role of genetic variation in ablating CpG dinucleotides across evolution through deamination of the cytosine base in this genomic context (*7*).

Genome-wide studies are necessary to fully understand the extent and genomic properties of meQTLs. However, so far most large-scale studies have used microarray technologies that only measure a small proportion of CpGs (*8*–*10*). The largest study to date to test associations between genotype and DNAm used MBD-seq, a method lacking single base pair resolution (*11*). Yet even with limited resolution, these initial studies have found that local genetic influence on DNAm is extensive throughout the genome, and meQTLs are enriched at regulatory sites (*12*, *13*).

Currently, a major puzzle in the field of functional genomics is understanding the molecular effects of genetic risk loci and variants identified by genome-wide association studies (GWAS)

for many common disorders and traits. This is particularly challenging in tissues like brain that are difficult to access or model, leaving little clarity into genetic mechanisms behind psychiatric disorders such as schizophrenia. While schizophrenia is highly heritable (*14*), and GWAS have identified a growing number of significant loci (*15*, *16*), only few loci have been functionally resolved (*17*). Genome-wide gene expression QTL (eQTL) approaches (*18*, *19*), and their extensions (*20*–*22*), have prioritized variants and associated genes, but many genomic loci fail to associate with nearby gene expression. In contrast, associating schizophrenia risk variants with a stable epigenetic mark like DNAm provide clues for potential epigenetic mechanisms of action (*23*, *24*). Indeed, previous meQTL maps using microarray technologies implicated a larger number of SCZD risk loci than eQTL maps, while only measuring a fraction of the methylome (*10*). DNAm itself may further reflect the cumulative effects of environmental exposures across the lifespan (*25*), and may represent a surrogate of "E" in GxE interactions that contribute to risks for many disorders (*26*) that can further act as a mediator of genetic risk on gene expression.

Unlike microarray technologies, whole genome bisulfite sequencing (WGBS) has the advantage of measuring cytosine methylation at single base pair resolution, as well as measuring CpH (H = A,T, or C) DNA methylation levels (in addition to CpGs). While CpH sites are generally unmethylated in somatic tissues, neurons in the human brain have uniquely high levels of CpHm (*27*). By leveraging this technology, we have created the most extensive genomic meQTL map in human postmortem brain tissue to date, and use this information to fine-tune our understanding of the molecular mechanisms of genetic and epigenetic risk for schizophrenia .

Results

Components of global variation in large-scale WGBS datasets

We performed whole genome bisulfite sequencing (WGBS) to gain a comprehensive view of genetic influence on DNAm in the adult human brain using two brain regions: the hippocampus and the dorsolateral prefrontal cortex (DLPFC). These regions have been prominently implicated in the pathogenesis of many psychiatric disorders, particularly schizophrenia (*28*). After data processing and quality control (see Methods), we analyzed 165 DLPFC samples and 179 hippocampal samples from a total of 183 adult donors aged 18 to 96 years (161 donors had data from both regions, **Table S1**). Data were generated across two large diagnosis- and region-balanced batches. We assessed 29,401,795 CpG sites across the epigenome, with an average post-processing coverage of 17.3 reads per CpG site. 78% of sites were highly (>80%) methylated while a minority, were lowly or unmethylated (8% are <20%). While the technical effects of measuring DNAm levels using microarrays is well-established (*29*), particularly in human brain tissue (*10*), corresponding assessments using WGBS data have been limited due to available comparisons being relatively small studies. We therefore assessed the contributions of different biological and technical variables on genome-wide CpG DNAm levels measured with WGBS.

First we performed principal component analysis (PCA) across the raw DNAm levels of the million most variable CpGs. The top principal components were associated with quantitative/genotype-defined ancestry (PC1: 6.5% variance explained, **Figure S1**), estimated neuronal fraction (PC2: 3.34%), processing batch (PC4: 1.37%) and brain region (PC5: 0.84%). The major batch effects resulted from inclusion of ENCODE "blacklist" regions (*30*), which have

been reported to cause problems with mapping and alignment in high-throughput sequencing data, particularly epigenomic data. These processing issues are likely further exacerbated in WGBS data, where the bisulfite treatment results in lower complexity libraries depleted of cytosines, which presumably relates to the influence of blacklist regions and ancestry on DNAm levels. Cytosines in these black-listed regions were therefore removed from reported site-specific analysis results. Another increasingly-common step in WGBS data processing involves "smoothing" local CpG DNAm levels within each sample to improve precision and borrow strength across nearby CpGs (*31*). Smoothing reordered the top components of variation (**Figure S2**), and resulted in the top component of variation representing both batch and estimated neuronal fraction (both PC1 and PC2, explaining 13.9% and 10% of variance, respectively). Previous analyses of Illumina microarray-derived adult homogenate DLPFC data suggested that estimated neuronal fraction was the largest component of (CpG) DNAm level variation (*10*). Microarray technology implicitly produces somewhat smooth DNAm levels for a large fraction of probes that target multiple CpG sites. While the effects of brain region were further magnified with smoothing, the effects of quantitative ancestry were dampened, and became associated with PC4 (1% variance explained) rather than PC1 of raw DNAm levels (6% explained variance; **Figure 1(A-E)**).

We further complemented these global analyses using site-specific variance components analysis, estimating the percentage of smoothed DNAm level variance explained by technical and biological components at each autosomal CpG, excluding the blacklist (N = 26,416,185, using ANOVA, see Methods, **Figure 1F**, **Data S1**). The factors that explained the largest components of site-specific variation were technical batch (median: 9.6% variance explained, interquartile range 3.4-19%), and related flow cell (7.9%, 6.3-9.6%) and instrument (2.9%,

2.1-3.2%) variables, as well as the more biological brain region (1.5%, 0.3-5.9%), estimated neuronal fraction (1%, 0.2-3.8%) variables. Traditionally-considered confounders in postmortem human brain studies - including tissue pH and postmortem interval (PMI) - had very little influence on site-specific DNAm levels using WGBS (in line with previous microarray-based analyses (*10*)). For example, pH and PMI explained more than 1% of variance across only approximately 5% of measured sites. Other technical variables hypothesized to influence DNAm levels like the sequencing alignment rates and bisulfite conversion rates (estimated with λ spike-in sequences, see Methods) showed little influence in this analysis. Overall, there was extensive residual variation of DNAm levels for the majority of sites beyond these technical and biological variables.

Local genetic variation has strong effects on CpG DNA methylation levels

We hypothesized that a large component of this residual DNAm variation was likely captured by local genetic sequence. We therefore performed genome-wide methylation quantitative trait locus (meQTL) analyses (see Methods) on smoothed DNA methylation levels in each brain region separately (across 29,401,795 CpG sites). In the DLPFC, we computed meQTLs between each of these CpGs and the subset of common SNPs within 20kb upstream and downstream, which identified 341,597,218 significant SNP-CpG pairs (at FDR < 0.01, **Data S2**), representing 5,971,724 (76%) of the tested SNPs and 11,155,961 (38%) of tested CpGs. Given high genomic correlation among both CpGs and SNPs, we performed the same analysis with a set of 535,859 LD-independent SNPs ($R^2 < 0.2$) to reduce the potential effects of linkage disequilibrium (LD) inflating these statistics. This sensitivity analysis found a similar number of CpGs identified as meQTLs (8,390,092 CpGs, 29%) with similar properties. Most SNPs associated with methylation levels at many nearby CpG sites (mean = 57 CpGs, median = 43

CpGs), and the methylation-associated SNPs had varying genomic widths of effect in this local window ranging from 1bp to the full 40kb (mean = 14.5kb , median = 12.7 kb). Effect sizes were generally small, with a mean of 2.6% change in methylation level per allele (IQR: 1.4%-3.1%), but ranging up to 47%. In both analyses we found that SNPs that disrupt a CpG dinucleotide (i.e. a variant at the C or G, which would ablate the capacity for methylation) have a slightly but significantly lower width of effect and a slightly higher number of correlated CpGs, meaning they have a higher effect density. This may be attributed to the fact that CpGs tend to cluster in the genome. Additionally, despite tested SNPs being LD-independent in the second analysis, we find that half of CpGs associate with more than one SNP, with a mean and median of 2, in line with previous observations in gene expression data (*18*, *32*). Using information from the Roadmap Epigenome (*33*), most genetically-associated CpGs were in quiescent genomic regions, and depleted for enhancer regions, in human brain (**Figure 2A**).

Analogous results were observed in hippocampus samples, yielding 360,079,758 significant pairs that represented 11,262,526 CpGs (38%) and 6,109,195 SNPs (78%, **Data S3**). Analyses on LD-independent SNPs yielded a similar proportion of SNPs (403,373 SNPs, 77%) implicating a similar number of CpGs as seen in the DLPFC (8,566,898). Effect sizes were similarly small, with a mean of 2.6% change in methylation level per allele. Hippocampal meQTLs have similar width of effect as those in DLPFC with a mean of 15,661 bp and an average of 60 CpGs associated with a SNP. These analyses suggested that the global properties of meQTLs were highly similar across brain regions.

We performed a series of secondary analyses to better characterize the determinants of such extensive genetic regulation of DNA methylation. First, due to the mixed ethnicities of our

samples, and the potentially large differences in allele frequencies between ancestry groups (*34*), we ran post-hoc meQTL analysis on a subset of meQTLs identified in full genome analysis in the DLPFC, separating samples into two groups by self-reported race. African Americans (AAs) made up 67% of our total sample, and thus were more likely to drive the results. Using significant meQTLs on chromosome 1, analyses using only African American samples (N=112) showed that 99.97% of the meQTLs were directionally consistent with the full analysis, with 97% marginally significant ($p < 0.05$) and 83% genome-wide significant (FDR < 0.01) in the smaller sample size and an overall sharing of $\pi_1 = 0.997$. In the European ancestry samples (N=53), 97% of meQTLs were directionally consistent, with 57% marginally significant and 21% genome-wide significant with an overall sharing of $\pi_1 = 0.854$. These decreased proportions compared to AA-specific analyses at least partially relate to the smaller sample size and resulting decreased statistical power (**Figure S3**). We also found that in general, differences in minor allele frequencies across ancestry groups did not associate with differences in meQTL effect magnitude (**Figure S4**), indicating that differences in ethnicity composition of our samples were likely not driving our combined ancestry analyses above. We further explored the robustness of the selected meQTL window size (20kb) using heritability analysis (see Methods) on the methylome (*35*) with different window sizes (20kb, 100kb, 500kb). DNA methylation levels were highly heritable using a 20kb window size, with 38% of tested CpG sites showing significant heritability (FDR < 0.01). These heritability results were further consistent with the above meQTL analyses, with similar proportion of CpGs showing genetic association (39% were meQTLs) and 95% of significantly heritable CpGs were meQTLs (and conversely, 86% of meQTL CpGs were heritable). Larger window sizes in heritability analysis actually identified fewer CpGs with significantly heritable methylation, implying that most genetic control of methylation acts in *cis* and confirming that our meQTL testing window was comprehensive. We

therefore identified widespread genetic control of CpG methylation levels. Understanding the details of this landscape may help elucidate the functional significance of SNPs highlighted by GWAS.

Widespread meQTLs among schizophrenia risk variants

DNA methylation previously has been shown to play a role in mediating genetic risk for neuropsychiatric (and other common) disorders (*10*, *36*–*38*), but all previous meQTL analyses have utilized microarray, not sequencing-based, methylome data. We performed extensive meQTL analyses on SNPs associated with genetic risk for schizophrenia in these large WGBS datasets. We specifically performed chromosome-scale meQTL analysis using each of the "index" SNPs for loci associated with schizophrenia from the most recent GWAS study of schizophrenia, i.e. PGC2+CLOZUK (*15*). We assessed index SNPs with high-quality genotype data in each region - 152 SNPs in DLPFC and 153 in the hippocampus. Each SNP was tested against every CpG site in the genome, considering a distance of <250kb *cis* and everything else *trans*. In DLPFC we found 25,382 significant (FDR < 0.01, **Table S2**) SNP-CpG pairs, representing 147 SNPs and 25,303 CpGs, showing that most PGC loci contain SNPs that associate with local DNA methylation levels (as only 107 SNP-CpG pairs were in *trans*). Schizophrenia risk-associated SNPs on average associated with 172 CpGs (median = 104), and in this *cis* window had an average genomic width of effect of 177kb (median = 147kb).

We further performed functional validation of these associations using corresponding gene expression data. Using RNA-seq data from the same regions and donors (see Methods), we assessed whether methylation at these CpGs correlated with neighboring expression levels. Using previous eQTL analyses on these same PGC loci (*18*, *39*), we assessed the mediation of

expression by mCpG (see Methods). Eleven of 127 loci had a correlation between gene expression and the methylation with which they are associated. Importantly, 10 of these associated with at least one CpG that mediated expression by at least 25%. The same analyses on the exon and junction levels picked up subtler effects, detecting 18 and 27 loci mediating expression levels via methylation, respectively. We found that overall, methylation mediation was most potent on the exon level (median = 40%), then the junction level (median = 32%), and least potent on the full gene level (median = 23%), in line with the putative role of DNAm in promoting gene splicing (*40*).

The same meQTL analysis was performed in the hippocampus WGBS data, revealing 48,023 significant SNP-CpG pairs (**Table S3**), representing 139/153 tested SNPs. There were 15,119 *trans*-meQTLs, many more than in DLPFC. Within the subset of significant DLPFC meQTLs, hippocampal meQTLs had an overall sharing of π_1 = 0.97, indicating that our findings are very consistent between brain regions.

These results indicated that meQTL effects, at least in the context of GWAS associations with schizophrenia, have much broader effects than traditionally considered, and much wider than the 20kb window examined at the full genome level. In order to see if schizophrenia-associated meQTLs are comparable to non-disease-associated meQTLs, we took 5000 random SNPs representing all levels of MAF and ran meQTL analysis with a 250kb window. Again we found that the majority of SNPs (93%) are meQTLs. We also find that neither MAF nor population-MAF differences associate with any meQTL characteristics. Interestingly, we found that these random meQTLs had significantly lower width of effect than schizophrenia-associated meQTLs in both regions (DLPFC p = 8.9e-5, Hippocampus p = 1.1e-11), and a significantly

fewer number of affected CpGs in hippocampus (p = 0.002). This combined with the chromatin state enrichment analysis below may indicate that these PGC-meQTLs are particularly functional, and potentially involved in disease processes, as opposed to just being standardly representative of the whole genome.

Risk-associated meQTL effects cluster in the genome

We then proceeded to cluster our meQTL CpGs into differentially methylated regions (DMRs) for better functional characterization. Using a CpG-specific t-statistic cutoff of 5 (see Methods), these sites could be clustered into 1277 DMRs (**Figure 3**, **Table S4**). The majority of SCZD index SNPs had such DMRs, and most had more than one (mean = 9.5, median = 6). The overall span of effect for each SNP was much larger than the 20kb *cis* window we tested above for meQTL analyses across the full genome, ranging up to 240 Mb on a single chromosome, with a median of 95 kb (mean = 17.5 Mb). Using Roadmap Epigenome (*33*) data, these SCZD risk-associated DMRs were enriched over the background of genome-wide LD-independent meQTLs for transcriptional and weak transcriptional chromatin signatures (**Figure 2B**). They were also comparatively depleted for weak repressive polycomb and quiescent chromatin signatures. Overall, these DMRs were in or near genes enriched for GO terms related to synapse and membrane potential (**Figure S5**). 20 DMRs overlapped with psychENCODE enhancers, and 142 overlapped with promoter regions. The genes connected to these promoters were enriched for GO terms related to acetylcholine, ion channels, and neurotransmitters.

Overall, results of clustering meQTL CpGs were quite similar between regions. In the hippocampus there were 1408 DMRs (**Table S5**), more than half (853) of DMRs directly

overlapped with a DMR in the DLPFC. Furthermore, 95.1% of DLPFC-identified DMRs showed marginal p < 0.05 significance among hippocampal meQTL effects (based on the average p-value each DMR). In general, allelic association with methylation in the hippocampus correlated with association in DLPFC DMRs (r^2 = 0.69, **Figure S6**). A strong majority of DMRs were located inside introns. Again we found that most SNPs have multiple associated DMRs (mean = 10, median = 7) and have a wide total genomic width (mean = 15,447,206 bp, median = 122,887 bp). Only 16 of these DMRs contain the actual GWAS index SNP, suggesting that these effects are more than just local consequences of genomic variations. Overall the genes represented by these DMRs have enrichment for GO terms related to synapses, membrane potential, and inositol triphosphate (IP3), a second messenger signalling molecule. In both regions, a handful of DMRs had correlation to expression of some gene (see Methods), and a few correlated to many genes, but most of these gene-DMR pairs were on different chromosomes, making results difficult to interpret. Most pairs were negatively correlated though, which fits with the traditional understanding of the suppressive effect of methylation on gene expression (*41*, *42*).

These SCZD risk-associated DMRs were further used as input to partitioned LDSC analysis (see Methods) to illuminate their clinical relevance (*43*). As a baseline, we ran partitioned LDSC analysis on the LD blocks of the schizophrenia GWAS loci used in meQTL analysis. These GWAS loci explained 15% of the additive genetic heritability explained by SNPs (h^2_{SNP}), with a 10-fold enrichment over the full genome (p = 8e-16), in line with being the top ranked loci in the GWAS. We then considered two sets of DMRs defined by two different statistical thresholds: a more liberal t-statistic > 3.5 cutoff (which corresponded roughly to controlling an FDR < 0.01 in *cis* meQTL analysis), termed $DMRs^{3.5}$ and the subset of entirely-contained DMRs defined by t

> 5 (from above), termed $DMRs^{5}$. The $DMRs^{3.5}$ were generally larger and more distant from each index SNP than the $DMRs^{5}$, with the majority of $DMRs^{3.5}$ distal (in *trans*). We further divided these DMR sets into those *cis* ($DMRs^{5}_{cis}$ and $DMRs^{3.5}_{cis}$) and *trans* ($DMRs^{5}_{trans}$ and $DMRs^{3.5}_{trans}$) relative to the PGC loci, i.e. those that were within the GWAS LD blocks, and those outside these blocks (**Figure 4**). First, by comparing the heritability estimates from *cis* versus *trans* DMRs at both cutoffs (i.e. $DMRs^{5}_{cis}$ versus $DMRs^{5}_{trans}$),we found that the majority of schizophrenia heritability and enrichment was driven by *cis* regions. For example, among the $DMRs^{3.5}$, the subset that were *cis* ($DMRs^{3.5}_{cis}$) explained 12% of h^2_{SNP}, with a 156-fold enrichment over the whole genome (p = 1.3e-14), and were further highly enriched compared to the background of the overall GWAS-significant LD blocks. Approximately 80% of all *cis* h^2_{SNP} of the GWAS significant loci (LD blocks) were captured by $DMRs^{3.5}_{cis}$ (12% versus 15%), even though only they contained 3% of loci sequence (1.65Mb vs 56.5Mb). In contrast, $DMRs^{3.5}_{trans}$ only explained 1.7% of h^2_{SNP}, and were not significantly enriched for schizophrenia risk (p = 0.14); **Figure 4**). Despite representing a very small portion of the genome (658kb), the more stringent *cis* DMRs still explained 8.7% of h^2_{SNP}, with very strong enrichment (243-fold, p = 1e-10). At only 48kb, the stringent *trans* DMRs were not wide enough to effectively detect enrichment, and only explained 0.5% of h^2_{SNP}. These results together suggest that the majority of significant schizophrenia genetic risk specifically localizes among small subsets of genomic regions associated with proximal/nearby DNAm levels.

Genetic regulation of CpH DNAm levels in homogenate brain tissue

While non-CpG (CpH) DNA methylation predominantly occurs in neuronal cells in the human brain (*27*), we could nevertheless observe detectable levels in homogenate/bulk tissue (which

contains 20-40% neuronal cells (*44*)). We analyzed 64,806,159 CpH sites in the DLPFC and 34,909,109 CpH sites in the hippocampus, after filtering to only sites which had at least moderate coverage and non-zero methylation levels across samples (see Methods). These numbers of observable sites in homogenate tissue from adult donors were much larger than in prenatal donors, as CpH methylation occurs in post-mitotic neurons (*45*), and comparable to smaller studies of neuronal nuclei sorted with the NeuN antibody and subjected to WGBS (*46*). We first performed full genome meQTL analysis on CpH sites, and found robust presence of CpH-meQTLs in the DLPFC, with 25,584,299 SNP-CpH pairs representing 5,805,754 SNPs, 947,073 of which were not significant meQTLs for CpG sites (**Data S4**). These CpH-associated SNPs further had CpG sites nearby, including in the testing window, suggesting potentially independent or complementary effects of CpH and CpG genetic associations. Unlike widespread CpG associations to genotype, there were far fewer unique CpH sites associated with genotype - only 976,094 CpHs associated with genotypes, corresponding to just 1.5% of tested sites. Generally, genetic control on CpH methylation appeared to have a narrower effect than on CpG methylation, with each SNP associating with a mean of 4 CpH sites over a mean width of 12,570 bp. The effect sizes of genotypes on methylation levels were much larger than they were for CpGs, with a mean of 27% change in methylation level per allele, and more than half (57%) of these CpH sites were inside genes. The landscape of CpH meQTLs in the hippocampus were similar to DLPFC, identifying 25,043,471 SNP-CpH pairs, representing 5,853,364 SNPs and 781,490 CpHs (**Data S5**). A large majority (90%) - but not all - of these SNPs and 63% of these CpH sites were also meQTLs in the DLPFC. Similarly, CpH-meQTLs had much larger effect sizes (mean = 29%) and most represented CpH sites (58%) were inside genes. These effects in each brain region presumably represent neuronal-specific genetic regulation of DNAm levels.

We also performed more focused CpH-meQTL analyses on the PGC SNPs described above and found 1444 significant *cis*-meQTLs and 48 *trans*-meQTLs in the DLPFC (**Table S6**). Again, a majority of PGC SNPs were represented (141/152). Some of these CpH sites were near CpG-meQTLs, but many were not (mean distance = 120kb, median = 2798bp), suggesting potential independent effects of genotype on different sequence contexts of DNAm. Like with CpGs associated with genotype, CpHs meQTLs were also enriched for transcriptional and weak transcriptional chromatin states over full genome CpH-meQTLs, and depleted for repressor polycomb and quiescent states (Figure 2C) (*33*). Most CpHs were inside genes that were subsequently enriched for neuronal GO terms related to neurons, synapses, and channels, further validating the neuronal contribution of CpH DNAm levels. We similarly observed much larger effect sizes of risk alleles in CpHs compared to CpGs in line with genome-wide analyses above, with a mean of 27% compared to 2% respectively. In hippocampus, we found 1588 *cis*-CpH-meQTLs and 92 *trans*-CpH-meQTLs (**Table S7**), representing 148/153 tested SNPs. Similar to all previous analyses we see that these sites are mostly inside genes and have much larger effect sizes than CpGs. The genes represented by these CpHs are enriched for GO terms related to neuronal anatomy, synapses and IP3. Again, distance to the nearest CpG-meQTL is highly variable, ranging from 1 to 4,217,747 bp (mean = 24,374, median = 2,761). Results were overall similar between both brain regions, and 1219 CpH-meQTLs were in common between both regions, though again, there were unique associations across regions.

Age associations to DNAm levels

While there was extensive evidence of meQTLs in our WGBS data, there were a subset of CpG sites that showed high percentages of variance explained by age (**Figure 1F**). We therefore

more formally modeled methylation over age in both brain regions, as DNA methylation has been shown to globally accumulate with age (*47*, *48*). We found extensive association with age, across approximately 2 million CpG sites in each region (at FDR < 0.05, **Data S6**) . The majority of these sites were age-associated in both regions, with a sizable fraction of sites showing some regional specificity (700,000 sites in the DLPFC and 800,000 sites in hippocampus). The majority (94%) of sites increase in methylation with age, with half of sites in promoter regions, and a quarter in CpG islands or shores. Only 9% of genes represented by these differentially methylated promoters had significant correlation to gene expression levels in these samples (see Methods). In contrast, there was very little CpH association with age, with only 5,136 and 445 significant sites in hippocampus and DLPFC respectively (at FDR < 0.05). These results suggest that CpH methylation may be more stable across adulthood and aging after establishment in postnatal life.

Given the large extents of meQTL- and age-associated sites, we asked whether any CpG sites showed dynamic meQTL effects across the adult lifespan. Despite age being associated with methylation at many sites throughout the genome, we found there were practically no changes in meQTL effects across the adult lifespan (i.e. statistical interaction between age and genotype), and, if anything, sites that were differentially methylated by age were depleted ($p < 2.2e-16$) for being associated with local genetic variation (i.e. being meQTLs).

Minimal illness-state associated differential methylation levels

We lastly modeled methylation differences between patients with schizophrenia and neurotypical controls. These associations are typically more subtle - fewer with smaller effect sizes - than age or genotype effects in microarray data (*10*) and more likely to represent cohort-

or dataset-specific findings (*49*). In these WGBS data, we found very few FDR-significant CpG sites - none in DLPFC and 70 in hippocampus. This is not surprising based on previous studies and the high multiple testing burden, almost two orders of magnitude more than microarray platforms. We saw similar lack of signal at CpH sites, with no significant hits in DLPFC and 1293 in hippocampus, with most (70%) of the hippocampal hits being in or nearby genes. These results suggest rather subtle effects of schizophrenia diagnosis on the methylome, particularly in the contexts of much stronger genotype- and age-associated effects.

Discussion

Here we present one of the most comprehensive whole genome bisulfite sequencing (WGBS) studies to date, particularly in human brain tissue, to better understand technical and biological factors that contribute to genome-wide DNA methylation levels at both CpG and CpH sites. We first demonstrated, at a single base pair resolution, that meQTLs are highly abundant throughout the entire genome at a breadth and scope previously unseen. Not only can common SNPs associate with CpG methylation, but they also uniquely and independently associate with CpH methylation levels in adult neurons. Furthermore, we demonstrated clinical relevance of these single base resolution meQTL maps to identify the functional significance of loci identified by GWAS in human brain. Using schizophrenia as an example, we found DNA methylation associations to nearly every genome-wide significant variant that clustered into many local differentially methylated regions (DMRs) that explained significant proportions of disease heritability.

Due to the expense and computational intensity, WGBS is challenging for epigenomic studies. With our large scale study, we were able to identify the effects of technical and potential biological variables on our data. This has been less well characterized than microarray studies, and we found that batch and ancestry cause much variance in the data, and their effects are exacerbated and alleviated, respectively, by the smoothing process. We also found that ENCODE blacklist regions are unreliable in WGBS data, due to the increased difficulty of alignment (*30*).

Previous studies have identified a genomic presence of meQTLs, but not at a single base pair resolution. Our findings are largely consistent with previous work, in that meQTLs are indeed extensive throughout the genome, and that most of their regulation occurs locally. However, while earlier estimates reported that 15% of CpGs were under genetic control (*11*), we greatly increased this fraction to 38%. Like Smith et al. (*8*), we showed that overlap was generally high between the two brain regions we surveyed, though clearly, there are differences as well. Studies have also found that functional meQTLs are enriched for active chromatin states (*11*) and that meQTLs appear to impact alternative splicing (*12*), further agreeing with our results and supporting the idea that schizophrenia risk associated loci are functional meQTLs. With our large sample and high genomic breadth, we are able to expand on all of these earlier findings at an in-depth genomic level.

These results further implicate DNAm as perhaps the most proximal molecular correlate of DNA sequence variation. The most comprehensive eQTL resource constructed in brain tissue, using over 1400 individuals, identified that approximately 25% of common genetic variants associated with nearby gene expression levels (*50*) and our meQTL maps here implicated three times as

many SNPs (76%) with a much smaller number of donors. Similarly, the recent GTEx v8 eQTL efforts - performed across 838 donors and 17,382 RNA-seq samples across 49 tissues - implicated 43% of tested SNVs with gene expression in at least one tissue.

These meQTLs further refined our understanding of the functional significance of schizophrenia genetic risk loci. By leveraging WGBS data combined with genotype data from the same samples, we identified molecular phenotypes associated with individual risk variants. This process could more generally filter GWAS findings to regions of the genome that could impart functional consequences of these risk variants. We found that regions that are differentially methylated by risk-associated genotypes explained most of the heritability imparted by the genome-wide significant schizophrenia risk loci, despite spanning a much smaller fraction of the genome (1.6Mb vs 56.5Mb). We also found that for some of these risk loci which have been previously identified as eQTLs (*39*), DNA methylation acts as a mediator of eQTL effects, refining the potential mechanism by which genetic risk variants may affect brain function. We note that the strongest mediation effects were seen among exons, indicating that differences in methylation may be key to alternative splicing, as has been previously hypothesized (*51*). While our data do not show that DNAm mediates expression for the majority of the meQTLs, this must be viewed with some caution. Our brain samples represent a moment in time in the lifespan of any given brain, and the data are from bulk tissue. At different life stages, perhaps in specific cell populations, mediation effects may be more prominent, particularly in the developing brain (*28*).

WGBS also gives the unique ability to examine CpH methylation, an often overlooked mark, particularly in brain. We found that DNAm levels at specific CpH sites were also associated with

genetic variation, which presumably reflected neuron-specific genetic regulation of DNAm levels. It is interesting that the genetic control of CpH methylation seems to have a much larger effect size than that on CpGm, particularly given the fact that the fraction of neurons in our homogenate tissue were uniquely driving these associations. This mark is particularly interesting to examine in relation to psychiatric disorders because it is specific to neurons, so we can point to the cell type of interest at these sites. Understanding which CpHs are under control of risk loci even further refines our understanding of the risk loci's functions because of this. It is also interesting that despite CpHs being abundant throughout the genome, most meQTL-CpHs are inside genes, possibly further pointing to functional significance. Large-scale analyses in sorted neuronal cell populations can further refine these associations, particularly in different subpopulations of neurons (i.e. inhibitory and excitatory) (*52*).

Overall we have established a comprehensive landscape of genetic control of genomic methylation in the human brain. Based on previous findings that many meQTLs are stable across tissue types, a large fraction of this meQTL map could apply to other tissues and cell types. It is clear that genotype has a robust role in determining local methylation levels, not only at CpG sites but at CpH sites as well. These findings can further be applied to understand the functional significance of genetic risk loci identified in GWAS.

Methods

Study Samples

Brain specimens were donated through the Offices of the Chief Medical Examiners of the District of Columbia and of the Commonwealth of Virginia, Northern District to the NIMH Brain Tissue Collection at the National Institutes of Health in Bethesda, MD, according to NIH Institutional Review Board guidelines (Protocol #90-M-0142). Audiotaped informed consent was obtained from legal next-of-kin on every case. Details of the donation process are described previously (*53*, *54*). All adult neurotypical controls were free from psychiatric and/or neurologic diagnoses and substance abuse according to DSM-IV, and had toxicology screening to exclude for acute drug and alcohol intoxication/use at time of death.

Data Generation

Genomic DNA was extracted from 100 mg of pulverized dorsolateral prefrontal cortex (DLPFC, corresponding to BA46/9) or hippocampus tissue (dissected as previously described (*39*)) with the phenol-chloroform method. DNA was subjected to bisulfite conversion followed by sequencing library preparation using the TruSeq DNA methylation kit from Illumina. Lambda DNA was spiked in prior to bisulfite conversion to assess its rate, and we used 20% PhiX to better calibrate Illumina base calling on these lower complexity libraries. Resulting libraries were pooled and sequenced on an Illumina HiSeq X Ten sequencer with paired end 150bp reads (2x150bp), targeting 90Gb per sample. This corresponds to 30x coverage of the human genome as extra reads were generated to account for the addition of PhiX.

Data Processing

The raw WGBS data was processed using FastQC to control for quality of reads, Trim Galore to trim reads and remove adapter content (*55*), Arioc for alignment to the GRCh38.p12 genome (obtained from ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/001/405/GCA_000001405.27_GRCh38.p12/GCA_000001405.27_GRCh38.p12_assembly_structure/Primary_Assembly/assembled_chromosomes/) (*56*), duplicate alignments were removed with SAMBLASTER (*57*), and filtered with samtools (*58*) to exclude all but primary alignments with a MAPQ >= 5. We used the Bismark methylation extractor to extract methylation data from aligned, filtered reads (*59*). We then used the bsseq R/Bioconductor package (v1.18) to process and combine the DNA methylation proportions across the samples for all further manipulation and analysis (*31*). After initial data metrics were calculated, the methylation data for each sample was locally smoothed using BSmooth with default parameters for downstream analyses. CpG results were filtered to those not in blacklist regions (DLPFC N = 26,155,085, Hippocampus N = 26,301,249), and those which had coverage >= 3. CpHs were filtered to sites which had >3 coverage and non-zero methylation in at least half the samples. Due to an unidentifiable primary source of variance, 11 samples in the DLPFC were dropped before analysis.

Assessment of technical and biological variation

Principal component analyses (PCA) were performed on the 1e6 most variable autosomal CpG sites using the prcomp() function in R. We calculated the percentage of variance explained by biological and technical variables using the anova() and lm() functions in R.

meQTL Analysis

We used R package Matrix eQTL (*60*) in all meQTL analyses. For full genome analysis, we set the maximum *cis* SNP to "gene" distance to 20kb. We approximated the p-value equivalent to FDR = 0.01 and used this as the p-value cutoff. We used only SNPs which were common (with minor allele frequencies, MAF > 5%) across the donors in each dataset separately that were in Hardy-Weinberg equilibrium (at p > 1e-6) with high non-missingness (>90% present), leading to analysis of 7,897,043 SNPs in the DLPFC and 7,865,986 SNPs in the hippocampus. The model adjusted for 28 covariates, which were the top 28 principle components of the methylation data. For PGC analyses, we set the *cis* distance to 250kb, and considered everything else in trans. We set the p-value cutoff to 1 so that we had statistics for every SNP-CpG pair in this analysis. meQTL interaction with age, neuronal composition, and MDS1 was assessed using the modelLINEAR_CROSS parameter. meQTLs were then organized into DMRs by using the bumphunter R/Bioconductor package (v1.30) (*61*) function regionFinder, to create clusters of adjacent meQTLs which all had an association statistic of t >= 5. These were filtered to DMRs containing at least two adjacent CpGs. We used the cleaningY() function from the jaffelab package (*62*) version 0.99.20 to regress out adjustment covariates to visualize the DNAm levels in subsequent plots.

Heritability Analysis

We estimated the SNP-heritability of DNAm for each CpG site using the GCTA software (*35*). We removed seven samples of DLPFC and eight samples of HIPPO so that all pairs of retained samples (DLPFC 158, HIPPO 171) had a genetic relatedness less than 0.025 and were included for heritability estimation. Genetic relationship matrix (GRM) was calculated using SNPs around each CpG site at three different window sizes (40 kb, 200 kb, and 1 Mb). We included the same set of covariates as we used in meQTL analysis in heritability estimation.

Functional Significance Analysis

We annotated our data using Gencode v. 29 on hg38. We performed gene ontology and gene set enrichment using clusterProfiler (v3.12) (*63*) with a p-value cutoff of 0.01 and q-value cutoff of 0.05. We performed LD score regression as described by Finucane et al (*43*) and with data from recent GWAS (*64*, *65*). To assess mediation of gene expression, we identified SNPs which were both eQTLs (*39*) and meQTLs, and which had some correlation (cor > 0.3) between gene expression and methylation levels. For every CpG-gene pair generated by this, we modeled the effect of genotype on expression, then added in the effect of methylation, and examined the difference in the effect size/coefficient for genotype.

Age and Diagnosis differential methylation modeling

Differential methylation analyses for both diagnosis and age were performed using linear regression modelling, accounting for sex, estimated neuronal fraction, batch, and the top 3 MDS components from genotype data. The regression analyses above were formed using limma (v3.30) (*66*, *67*) which employed empirical Bayes and returned moderated T-statistics, which were used to calculate p-values and estimate the false discovery rate (FDR, via Benjamini-Hochberg approach (*68*)).

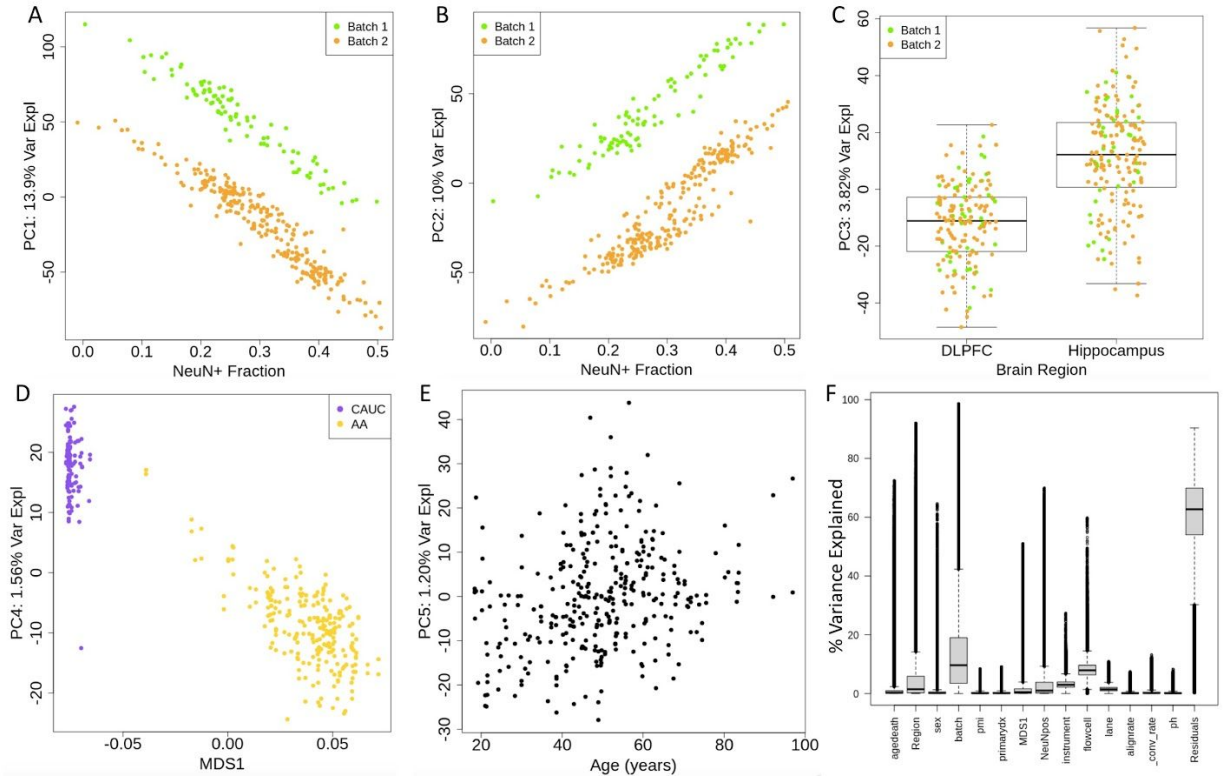

Figure 1: Variance in smoothed methylation data, post-QC. PCA was performed on all sites excluding the sex chromosomes and ENCODE's blacklist. **(A,B)** We find that the top principal components of smoothed methylation data associate with both batch and neuronal composition. **(C)** We see that the third principal component is associated with brain region, and no longer associated with batch. **(D)** In smoothed methylation data, ethnicity is reduced to the 4th principal component. **(E)** Age associates with the fifth principal component. **(F)** Variance explained was analyzed using ANOVA by each individual CpG site. We see that brain region and batch effects explain a large deal of variance, while biological factors such as PMI and pH explain very little.

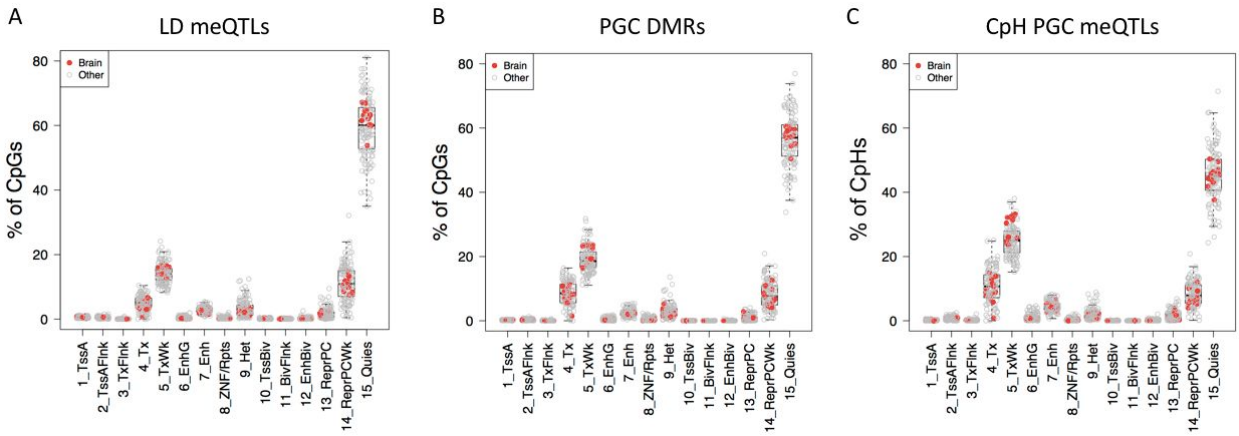

Figure 2: Chromatin State of meQTLs and DMRs. We assessed the chromatin state of sites identified as meQTLs and DMRs using data from various tissues from the Roadmap Epigenomics Project. Data for brain tissues are highlighted in red. Chromatin states are defined as follows. 1: Active TSS 2: Flanking Active TSS 3: Transcription at gene 5' and 3' 4: Strong transcription 5: Weak transcription 6: Genic enhancers 7: Enhancers 8: ZNF genes & repeats 9: Heterochromatin 10: Bivalent/Poised TSS 11: Flanking Bivalent TSS/Enhancer 12: Bivalent Enhancer 13: Repressed PolyComb 14: Weak Repressed PolyComb 15: Quiescent/Low **(A)** In genome-wide meQTLs, assessing a set of LD-independent SNPs, we see that the vast majority of meQTL-CpGs are in quiescent chromatin regions. **(B)** We see that compared to genomic SNPs, CpGs associated with PGC schizophrenia risk SNPs are enriched for regions of active transcription and depleted for regions of quiescent chromatin. **(C)** We see that CpH sites associated with PGC schizophrenia risk SNPs are often in regions of active transcription.

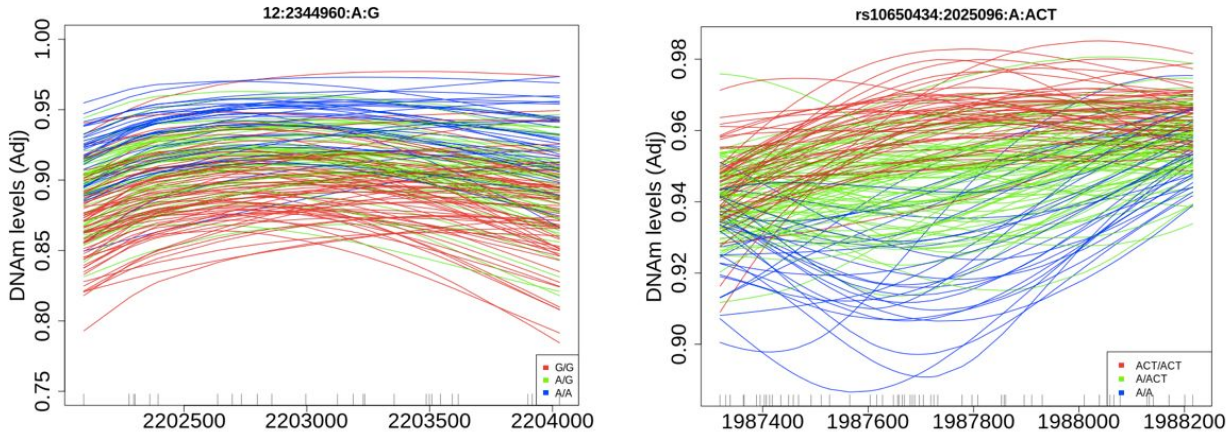

Figure 3: Schizophrenia-risk associated DMRs. Two examples of regions where methylation levels are associated with genotype at a schizophrenia risk associated phenotype.

Category	SNPs (%)	h2 (%)	Enrichment (fold)	P-value (Enrichment)	Width (bp)	P-value (Coefficient)
SCZD Loci	1.491%	15.0%	10.1	8.01E-16	56462978	3.32E-18
$DMR^{3.5}_{cis}$	0.077%	12.1%	156.7	1.30E-14	1658085	4.19E-17
$DMR^{3.5}_{trans}$	0.234%	1.7%	7.5	0.147	5427037	0.073142
DMR^{5}_{cis}	0.036%	8.7%	243.6	1.00E-10	657860	4.36E-12
DMR^{5}_{trans}	0.003%	0.5%	181.2	0.036	48414	0.0173215

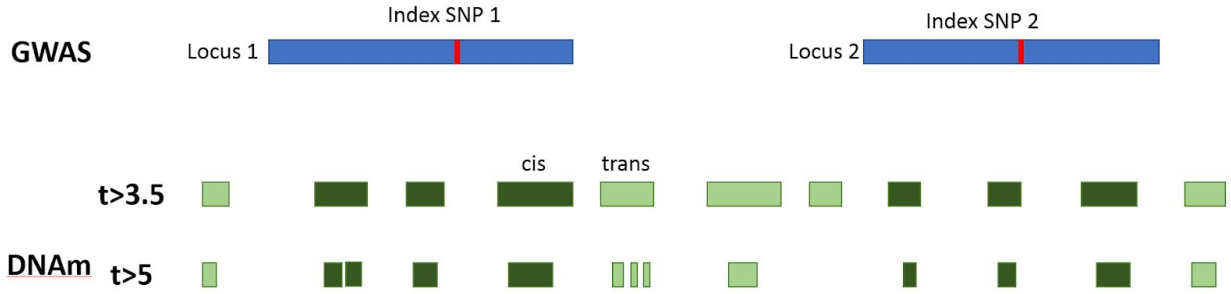

Figure 4: LDSC results for schizophrenia heritability. (Top) LDSC analysis outputs for each category of DMRs explained in Figure 4. Results showed that most of the enrichment for schizophrenia heritability in our DMR sites was in *cis*. (Bottom) Visual description of LDSC: we performed LDSC analysis on the GWAS-identified loci as a background, and two sets of DMRs, one with a cutoff of t > 3.5 and one with a more stringent cutoff of t > 5. We further divided these DMR sets into *cis* DMRs - those within the GWAS loci - and *trans* DMRs - those outside the GWAS loci.

References:

1. B. T. Heijmans *et al.*, Persistent epigenetic differences associated with prenatal exposure to famine in humans. *Proc Natl Acad Sci USA*. 105, 17046–17049 (2008).

2. L. P. Breitling, R. Yang, B. Korn, B. Burwinkel, H. Brenner, Tobacco-smoking-related differential DNA methylation: 27K discovery and replication. *Am. J. Hum. Genet.* 88, 450–457 (2011).

3. M. N. Davies *et al.*, Functional annotation of the human brain methylome identifies tissue-specific epigenetic variation across brain and blood. *Genome Biol.* 13, R43 (2012).

4. Z. A. Kaminsky *et al.*, DNA methylation profiles in monozygotic and dizygotic twins. *Nat. Genet.* 41, 240–245 (2009).

5. J. T. Bell, T. D. Spector, DNA methylation studies using twins: what are they telling us? *Genome Biol.* 13, 172 (2012).

6. E. L. Meaburn, L. C. Schalkwyk, J. Mill, Allele-specific methylation in the human genome: implications for genetic studies of complex disease. *Epigenetics*. 5, 578–582 (2010).

7. J. Sved, A. Bird, The expected equilibrium of the CpG dinucleotide in vertebrate genomes under a mutation model. *Proc Natl Acad Sci USA*. 87, 4692–4696 (1990).

8. A. K. Smith *et al.*, Methylation quantitative trait loci (meQTLs) are consistently detected across ancestry, developmental stage, and tissue type. *BMC Genomics*. 15, 145 (2014).

9. H. Schulz *et al.*, Genome-wide mapping of genetic determinants influencing DNA methylation and gene expression in human hippocampus. *Nat. Commun.* 8, 1511 (2017).

10. A. E. Jaffe *et al.*, Mapping DNA methylation across development, genotype and schizophrenia in the human frontal cortex. *Nat. Neurosci.* 19, 40–47 (2016).

11. J. L. McClay *et al.*, High density methylation QTL analysis in human blood via next-generation sequencing of the methylated genomic DNA fraction. *Genome Biol.* 16, 291 (2015).

12. M. Gutierrez-Arcelus *et al.*, Tissue-specific effects of genetic and epigenetic variation on gene regulation and splicing. *PLoS Genet.* 11, e1004958 (2015).

13. A. Hoffmann, M. Ziller, D. Spengler, The future is the past: methylation qtls in schizophrenia. *Genes (Basel)*. 7 (2016), doi:10.3390/genes7120104.

14. P. V. Gejman, A. R. Sanders, J. Duan, The role of genetics in the etiology of schizophrenia. *Psychiatr. Clin. North Am.* 33, 35–66 (2010).

15. A. F. Pardiñas *et al.*, Common schizophrenia alleles are enriched in mutation-intolerant genes and in regions under strong background selection. *Nat. Genet.* 50, 381–389 (2018).

16. Schizophrenia Working Group of the Psychiatric Genomics Consortium, S. Ripke, J. T. Walters, M. C. O'Donovan, Mapping genomic loci prioritises genes and implicates synaptic biology in schizophrenia. *medRxiv* (2020), doi:10.1101/2020.09.12.20192922.

17. N. Schrode *et al.*, Synergistic effects of common schizophrenia risk variants. *Nat. Genet.* 51, 1475–1485 (2019).

18. A. E. Jaffe *et al.*, Developmental and genetic regulation of the human cortex transcriptome illuminate schizophrenia pathogenesis. *Nat. Neurosci.* 21, 1117–1125 (2018).

19. M. Fromer *et al.*, Gene expression elucidates functional impact of polygenic risk for schizophrenia. *Nat. Neurosci.* 19, 1442–1453 (2016).

20. Z. Zhu *et al.*, Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat. Genet.* 48, 481–487 (2016).

21. A. Gusev *et al.*, Transcriptome-wide association study of schizophrenia and chromatin activity yields mechanistic disease insights. *Nat. Genet.* 50, 538–548 (2018).

22. E. R. Gamazon *et al.*, A gene-based association method for mapping traits using reference transcriptome data. *Nat. Genet.* 47, 1091–1098 (2015).

23. T. Huan *et al.*, Genome-wide identification of DNA methylation QTLs in whole blood highlights pathways for cardiovascular disease. *Nat. Commun.* 10, 4267 (2019).

24. E. R. Gamazon *et al.*, Enrichment of cis-regulatory gene expression SNPs and methylation quantitative trait loci among bipolar disorder susceptibility variants. *Mol. Psychiatry*. 18, 340–346 (2013).

25. A. Baccarelli, V. Bollati, Epigenetics and environmental chemicals. *Curr. Opin. Pediatr.* 21, 243–251 (2009).

26. A. Dempfle *et al.*, Gene-environment interactions for complex traits: definitions, methodological requirements and challenges. *Eur. J. Hum. Genet.* 16, 1164–1172 (2008).

27. J. U. Guo *et al.*, Distribution, recognition and regulation of non-CpG methylation in the adult mammalian brain. *Nat. Neurosci.* 17, 215–222 (2014).

28. R. Birnbaum, D. R. Weinberger, Genetic insights into the neurodevelopmental origins of schizophrenia. *Nat. Rev. Neurosci.* 18, 727–740 (2017).

29. J.-P. Fortin, T. J. Triche, K. D. Hansen, Preprocessing, normalization and integration of the Illumina HumanMethylationEPIC array with minfi. *Bioinformatics*. 33, 558–560 (2017).

30. H. M. Amemiya, A. Kundaje, A. P. Boyle, The ENCODE blacklist: identification of problematic regions of the genome. *Sci. Rep.* 9, 9354 (2019).

31. K. D. Hansen, B. Langmead, R. A. Irizarry, BSmooth: from whole genome bisulfite

sequencing reads to differentially methylated regions. *Genome Biol.* 13, R83 (2012).

32. F. Aguet *et al.*, The GTEx Consortium atlas of genetic regulatory effects across human tissues. *BioRxiv* (2019), doi:10.1101/787903.

33. Roadmap Epigenomics Consortium *et al.*, Integrative analysis of 111 reference human epigenomes. *Nature*. 518, 317–330 (2015).

34. R. Kosoy *et al.*, Ancestry informative marker sets for determining continental origin and admixture proportions in common populations in America. *Hum. Mutat.* 30, 69–78 (2009).

35. J. Yang, S. H. Lee, M. E. Goddard, P. M. Visscher, GCTA: a tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* 88, 76–82 (2011).

36. Y. Liu *et al.*, Epigenome-wide association data implicate DNA methylation as an intermediary of genetic risk in rheumatoid arthritis. *Nat. Biotechnol.* 31, 142–147 (2013).

37. A. P. Feinberg, The key role of epigenetics in human disease prevention and mitigation. *N. Engl. J. Med.* 378, 1323–1334 (2018).

38. N. T. Ventham *et al.*, Integrative epigenome-wide analysis demonstrates that DNA methylation may mediate genetic risk in inflammatory bowel disease. *Nat. Commun.* 7, 13507 (2016).

39. L. Collado-Torres *et al.*, Regional Heterogeneity in Gene Expression, Regulation, and Coherence in the Frontal Cortex and Hippocampus across Development and Schizophrenia. *Neuron*. 103, 203-216.e8 (2019).

40. S. Shukla *et al.*, CTCF-promoted RNA polymerase II pausing links DNA methylation to splicing. *Nature*. 479, 74–79 (2011).

41. A. Bird, Perceptions of epigenetics. *Nature*. 447, 396–398 (2007).

42. P. A. Jones, S. M. Taylor, Cellular differentiation, cytidine analogs and DNA methylation. *Cell*. 20, 85–93 (1980).

43. H. K. Finucane *et al.*, Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.* 47, 1228–1235 (2015).

44. J. Guintivano, M. J. Aryee, Z. A. Kaminsky, A cell epigenotype specific model for the correction of brain cellular heterogeneity bias and its application to age, brain region and major depression. *Epigenetics*. 8, 290–302 (2013).

45. K. A. Perzel Mandell *et al.*, Characterizing the dynamic and functional DNA methylation landscape in the developing human cortex. *Epigenetics*, 1–13 (2020).

46. A. J. Price *et al.*, Divergent neuronal DNA methylation patterns across human cortical development reveal critical periods and a unique role of CpH methylation. *Genome Biol.* 20, 196 (2019).

47. J.-P. Issa, Age-related epigenetic changes and the immune system. *Clin. Immunol.* 109, 103–108 (2003).

48. M. F. Fraga, R. Agrelo, M. Esteller, Cross-talk between aging and cancer: the epigenetic language. *Ann. N. Y. Acad. Sci.* 1100, 60–74 (2007).

49. A. E. Jaffe, J. E. Kleinman, Genetic and epigenetic analysis of schizophrenia in blood-a no-brainer? *Genome Med.* 8, 96 (2016).

50. D. Wang *et al.*, Comprehensive functional genomic resource and integrative model for the human brain. *Science*. 362 (2018), doi:10.1126/science.aat8464.

51. G. Lev Maor, A. Yearim, G. Ast, The alternative role of DNA methylation in splicing regulation. *Trends Genet.* 31, 274–280 (2015).

52. A. Kozlenkov *et al.*, A unique role for DNA (hydroxy)methylation in epigenetic regulation of human inhibitory neurons. *Sci. Adv.* 4, eaau6190 (2018).

53. B. K. Lipska *et al.*, Critical factors in gene expression in postmortem human brain: Focus on studies in schizophrenia. *Biol. Psychiatry*. 60, 650–658 (2006).

54. A. Deep-Soboslay *et al.*, Reliability of psychiatric diagnosis in postmortem research. *Biol. Psychiatry*. 57, 96–101 (2005).

55. F. Krueger, TrimGalore: A wrapper around Cutadapt and FastQC to consistently apply adapter and quality trimming to FastQ files, with extra functionality for RRBS data, (available at https://github.com/FelixKrueger/TrimGalore).

56. R. Wilton, X. Li, A. P. Feinberg, A. S. Szalay, Arioc: GPU-accelerated alignment of short bisulfite-treated reads. *Bioinformatics*. 34, 2673–2675 (2018).

57. G. G. Faust, I. M. Hall, SAMBLASTER: fast duplicate marking and structural variant read extraction. *Bioinformatics*. 30, 2503–2505 (2014).

58. H. Li *et al.*, The Sequence Alignment/Map format and SAMtools. *Bioinformatics*. 25, 2078–2079 (2009).

59. F. Krueger, S. R. Andrews, Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics*. 27, 1571–1572 (2011).

60. A. A. Shabalin, Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics*. 28, 1353–1358 (2012).

61. A. E. Jaffe *et al.*, Bump hunting to identify differentially methylated regions in epigenetic epidemiology studies. *Int. J. Epidemiol.* 41, 200–209 (2012).

62. L. Collado-Torres, A. E. Jaffe, *jaffelab: Commonly used functions by the Jaffe lab* (GitHub,

GitHub, 2018).

63. G. Yu, L.-G. Wang, Y. Han, Q.-Y. He, clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS*. 16, 284–287 (2012).

64. Brainstorm Consortium *et al.*, Analysis of shared heritability in common disorders of the brain. *Science*. 360 (2018), doi:10.1126/science.aap8757.

65. L. F. Rizzardi *et al.*, Neuronal brain-region-specific DNA methylation and chromatin accessibility are associated with neuropsychiatric trait heritability. *Nat. Neurosci.* 22, 307–316 (2019).

66. M. E. Ritchie *et al.*, limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* 43, e47 (2015).

67. B. Phipson, S. Lee, I. J. Majewski, W. S. Alexander, G. K. Smyth, Robust hyperparameter estimation protects against hypervariable genes and improves power to detect differential expression. *Ann. Appl. Stat.* 10, 946–963 (2016).

68. Y. Benjamini, Y. Hochberg, Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*. 57, 289–300 (1995).

Acknowledgements:

The authors would like to express their gratitude to our colleagues whose tireless efforts have led to the donation of postmortem tissue to advance these studies: the Office of the Chief Medical Examiner of the District of Columbia; the Office of the Chief Medical Examiner for Northern Virginia, Fairfax Virginia; and the Office of the Chief Medical Examiner of the State of Maryland, Baltimore, Maryland. We would also like to acknowledge Llewellyn B. Bigelow, MD, for his diagnostic expertise. This project was supported by The Lieber Institute for Brain Development and by NIH grants R01MH112751 and T32GM781437. Finally, we are indebted to the generosity of the families of the decedents, who donated the brain tissue used in these studies.

Supplementary Material

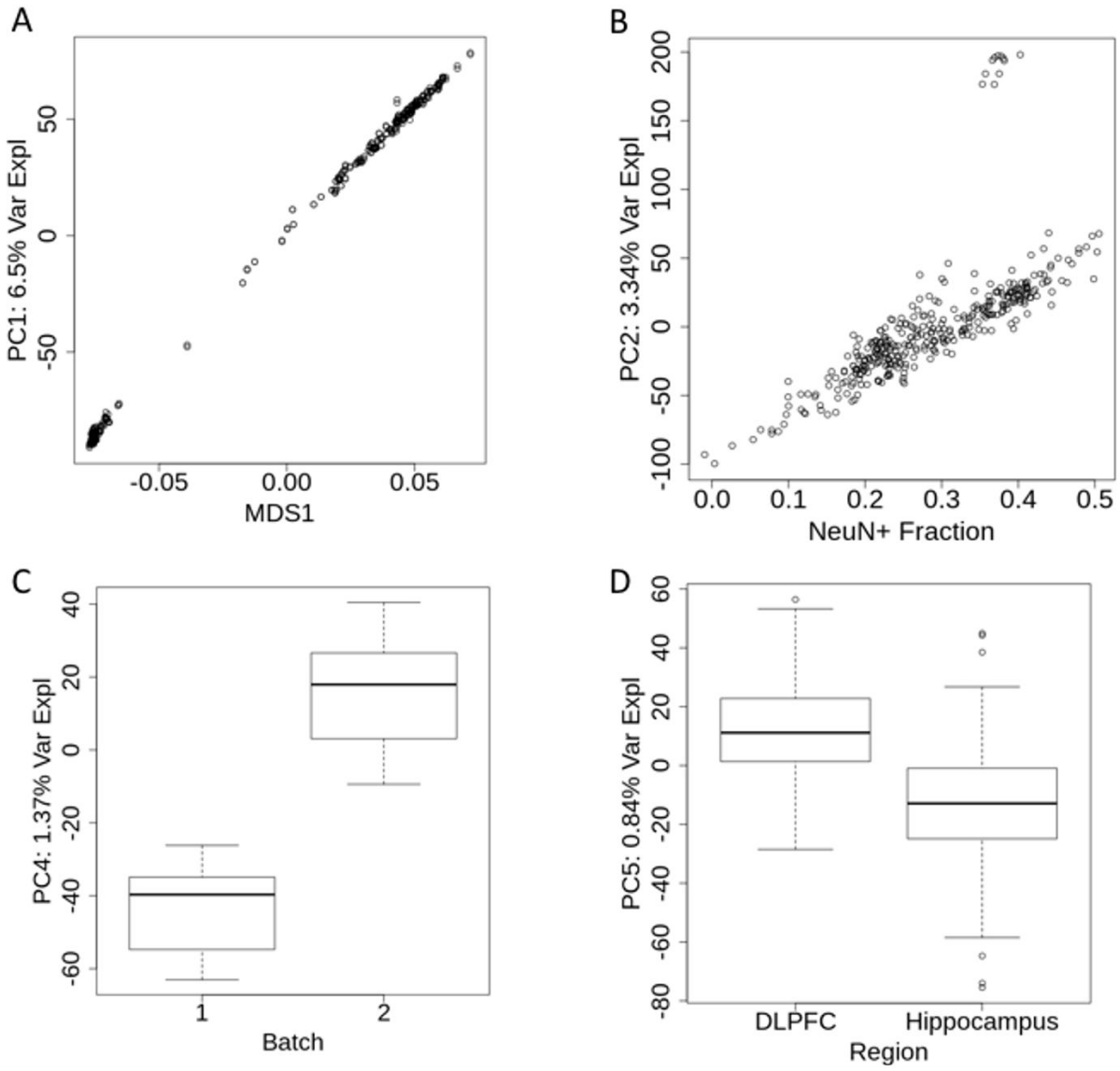

Figure S1: PCA on raw methylation values, excluding sex chromosomes. Principal components of variance are plotted against their most associated factor. **(A)** The top component of variance in raw methylation data corresponds with the top multidimensional scaling (MDS1) component of the genotype data, which is a representation of ethnicity. **(B)** The second principal component aligns with estimated neuronal fraction. Here we also see 11 DLPFC samples with

unexplained variance, so they are dropped from analysis. **(C)** The third principal component strongly associates with processing batch. **(D)** The fourth principal component correlates with brain region.

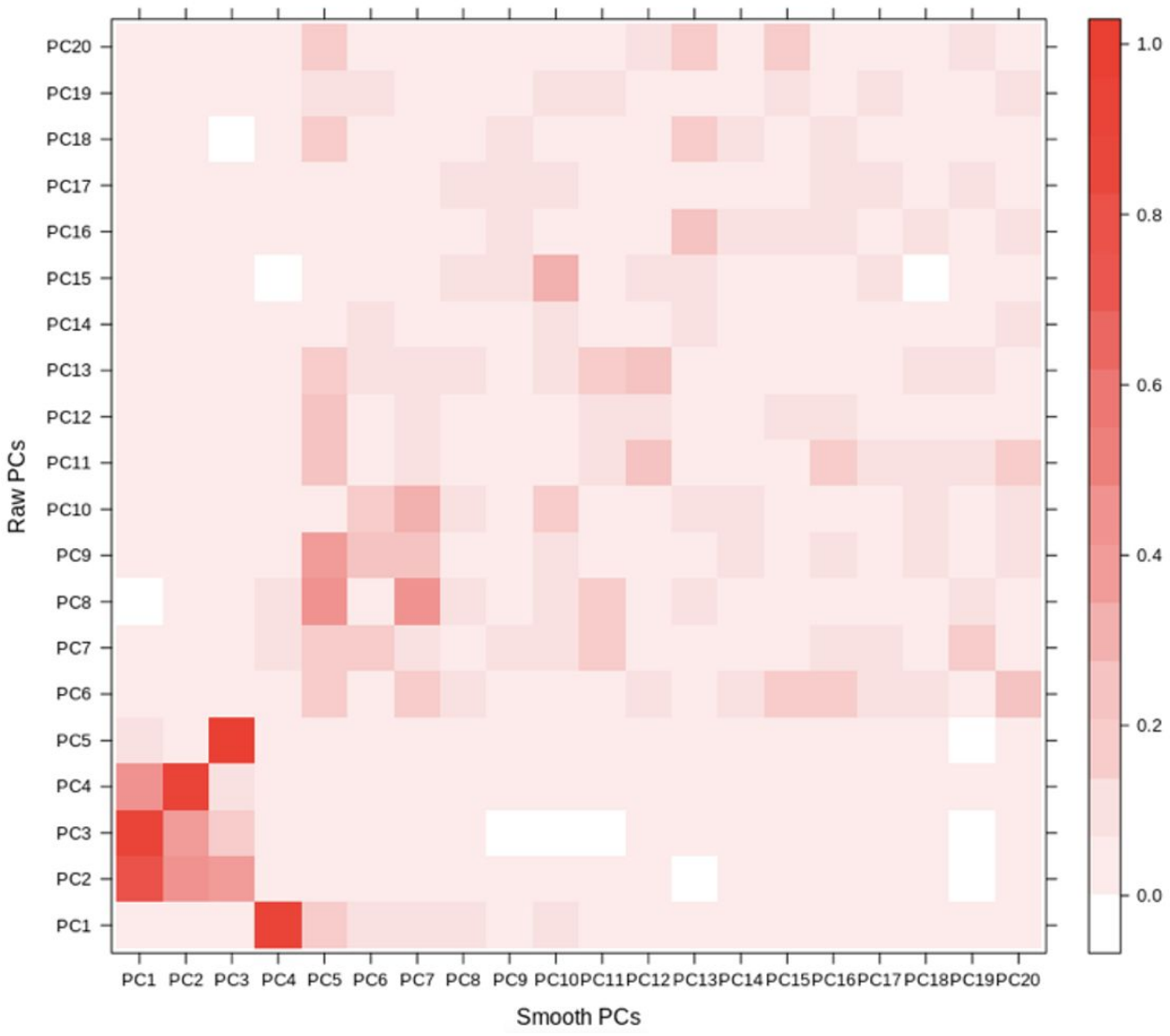

Figure S2: Comparison between top 20 PCs of raw and smooth methylation data. We see that smoothing methylation data alters the principal components of variants, mitigating the effects of ethnicity and increasing the effects of batch and cell composition.

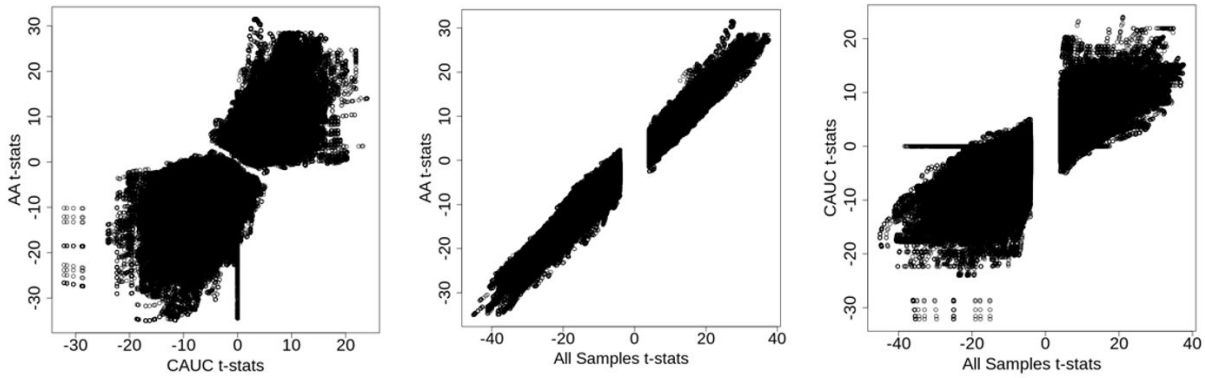

Figure S3: Ethnicity Differences in meQTL t-statistics. For meQTLs identified on chromosome 1, we ran post-hoc meQTL analysis on the samples divided by self-reported ethnicity. **(A)** There are some differences between meQTL strengths between ethnicities, but they are largely correlated and directionally consistent. **(B)** meQTL statistics in AA only were very similar to the overall findings, likely because the majority of our samples were AA. **(C)** meQTL statistics compared to t-statistics from all samples.

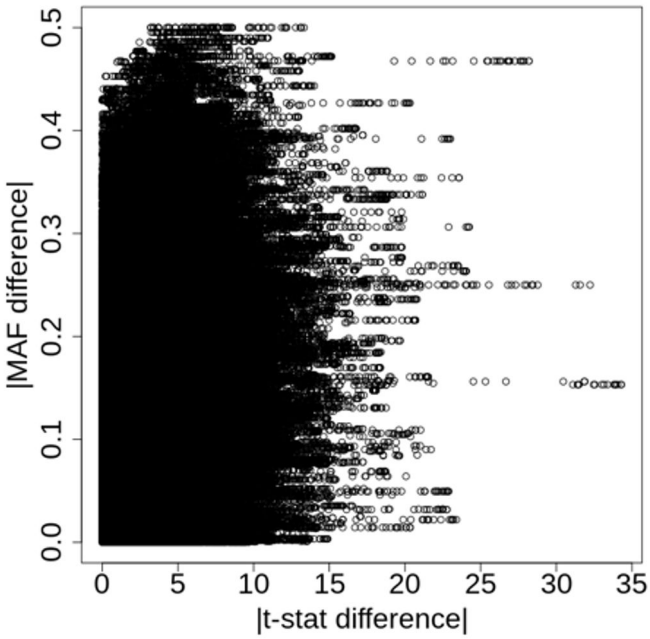

Figure S4: **Lack of association between population MAF differences and meQTL strength.** Though some differences in meQTL strength can be seen between ethnicities, meQTL strength does not associate with SNP MAF or MAF differences between ethnic populations.

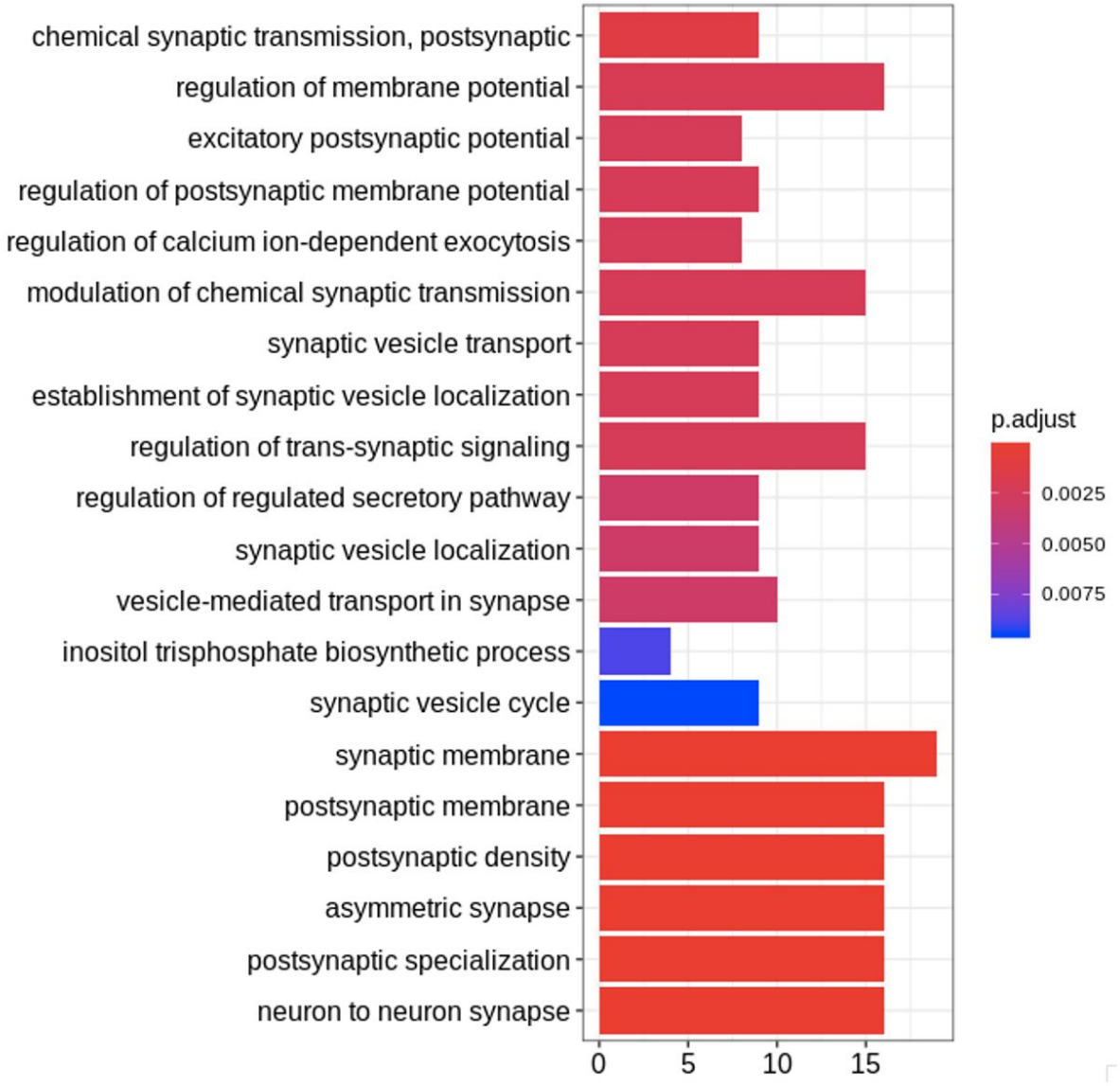

Figure S5: GO Enrichment of schizophrenia risk associated DMRs. We find that DMRs associated with schizophrenia risk variants are in or near genes enriched for synapse related GO terms.

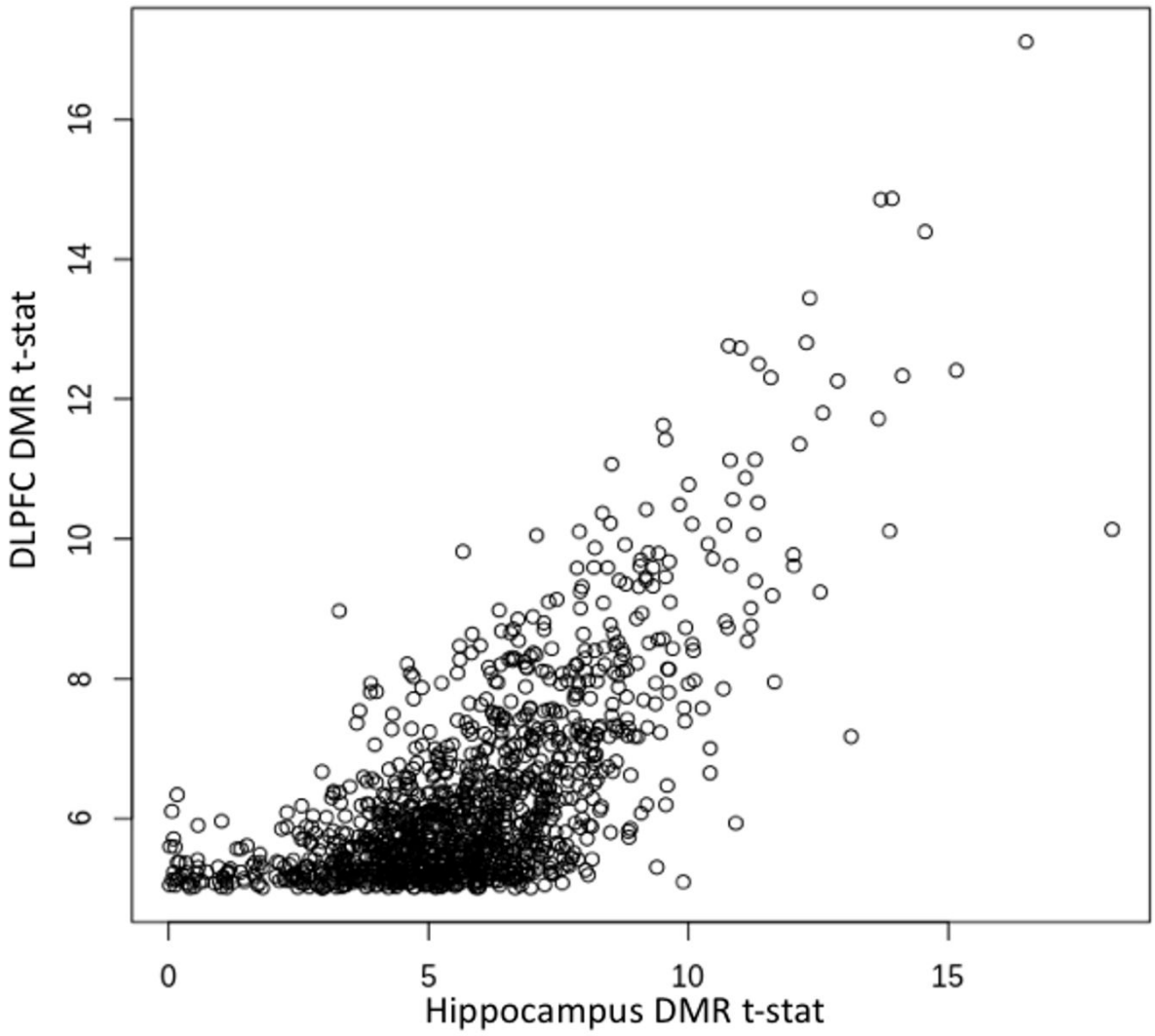

Figure S6: Comparison of Hippocampus association statistics to the statistics of DMRs identified in DLPFC. meQTL association statistics are highly correlated between brain regions.

Supplementary Tables

Table S1: Sample Demographic Data

Table S2: DLPFC CpG meQTLs to SCZD risk SNPs. snps: SNP identifier; cpg: location of CpG meQTL site (hg38); statistic: t-statistic for meQTL association between SNP and CpG; pvalue: P-value for meQTL association; FDR: Benjamin-Hochberg corrected p-value; beta: regression coefficient, or the change in DNAm levels per minor allele copy; snpChr: chromosome of genome where the SNP lies (hg38); snpPos: position on the chromosome where the SNP lies (hg38); snpRsNum: rs number of the genetic variant; snpCounted:the counted/minor allele of the meQTL variant (`beta` is relative to this); snpAlt: the reference allele for the variant; disruptCpG: whether the SNP disrupts any CpG dinucleotides (even one different than the meQTL); methChr: chromosome where the CpG lies (hg38); methPos: position where the CpG lies (hg38); distMethToSnp: genomic distance between SNP and CpG.

Table S3: HIPPO CpG meQTLs to SCZD risk SNPs. See Table S2 for field descriptions.

Table S4: DLPFC DMRs formed by SCZD "index" SNPs. chr: chromosome where the differentially methylated region (DMR) lies (hg38); start: start position of the DMR (hg38); end: end position of the DMR (hg38); value: mean meQTL t-statistic within the DMR; area: sum of the meQTL t-statistics within the DMR; cluster, indexStart, and indexEnd are internal indices for rows of the methylation matrix; L: number of CpGs in DMR; clusterL: number of nearby CpGs considered for DMR finding; width: genomic width of DMR; snpPos: position of the schizophrenia index SNP used to find DMR; disruptCpG: whether the SNP disrupts any CpG

dinucleotides; distStart: distance of SNP to start of DMR; distEnd: distance of SNP to end of DMR; location: location of DMR relative to genes

Table S5: HIPPO DMRs formed by SCZD "index" SNPs. See Table S4 for field descriptions.

Table S6: DLPFC CpH meQTLs to SCZD risk SNPs. See Table S2 for field descriptions.

Table S7: HIPPO CpH meQTLs to SCZD risk SNPs. See Table S2 for field descriptions.

Supplementary Datasets

Data S1: Variance of smoothed DNA methylation levels explained by technical and biological variables across all tested CpGs: https://mandell-wgbs-meqtl.s3.us-east-2.amazonaws.com/suppData_smooth_variance_explained_annotated.csv.gz

Data S2: meQTLs across DLPFC using smoothed DNAm values. Statistics from all significant SNP-CpG pairs are provided: https://mandell-wgbs-meqtl.s3.us-east-2.amazonaws.com/suppData_smooth_dlpfc_CpG_meqtls_fdr.csv.gz

Data S3: meQTLs across HIPPO using smoothed DNAm values. Statistics from all significant SNP-CpG pairs are provided: https://mandell-wgbs-meqtl.s3.us-east-2.amazonaws.com/suppData_smooth_hippo_CpG_meqtls_fdr.csv.gz

Data S4: CpH-meQTLs across DLPFC. Statistics from all significant SNP-CpH pairs are provided: https://mandell-wgbs-meqtl.s3.us-east-2.amazonaws.com/suppData_smooth_dlpfc_CpH_meqtls_fdr.csv.gz

Data S5: CpH-meQTLs across HIPPO. Statistics from all significant SNP-CpH pairs are

provided:

https://mandell-wgbs-meqtl.s3.us-east-2.amazonaws.com/suppData_smooth_hippo_CpH_meqtls_fdr.csv.gz

Data S6: Associations between smoothed DNAm values and age across the DLPFC for all significant sites:

https://mandell-wgbs-meqtl.s3.us-east-2.amazonaws.com/suppData_smooth_dlpfc_CpG_age.csv.gz