Allele-skewed DNA modification in the brain: relevance to schizophrenia GWAS

Sarah A. Gagliano[1,3#], Carolyn Ptak[1#], Denise Y.F. Mak[1,2#], Mehrdad Shamsi[1,2], Gabriel Oh[1], Joanne Knight[1,3,4,7], Paul C. Boutros PC[2,5,6*], Arturas Petronis[1, 3, 4, 5*]

1. Campbell Family Mental Health Research Institute, Centre for Addiction and Mental Health, 250 College Street, Toronto, Ontario M5T1R8 Canada
2. Informatics and Biocomputing Program, Ontario Institute for Cancer Research, 661 University Avenue Suite 510, Toronto, Ontario, M5G 0A3 Canada
3. Institute of Medical Science, University of Toronto, Ontario, Canada
4. Department of Psychiatry, University of Toronto, Toronto, Ontario, Canada
5. Department of Pharmacology & Toxicology, University of Toronto, Toronto, ON, Canada
6. Department of Medical Biophysics, University of Toronto, Toronto, ON, Canada
7. Biostatistics Division, Dalla Lana School of Public Health, University of Toronto, Toronto, Ontario, Canada

# - equal contribution

*Correspondence: Art Petronis (art.petronis@camh.ca) and Paul Boutros (paul.boutros@oicr.on.ca)

Abstract

Numerous recent studies have suggested that phenotypic effects of DNA sequence variants can be mediated or modulated by their epigenetic marks, such as allele-skewed DNA modification (ASM). We performed a comprehensive search of ASM effects in human *post-mortem* brain and sperm samples (total n=256) from individuals with major psychosis and controls using Affymetrix SNP microarray. Depending on the phenotypic category of the brain samples, 1.4-7.5% of interrogated SNPs exhibited ASM effects. Next we investigated ASM in the context of genetic studies of schizophrenia and detected that brain ASM-SNPs were significantly overrepresented among the schizophrenia GWAS sub-threshold SNPs. Brain ASM-SNPs showed a much stronger enrichment in schizophrenia GWAS compared to 17 large GWAS of non-psychiatric diseases and traits arguing that ASM effects are at least partially tissue specific. Germline and control brain ASM-SNP studies supported a causal association between ASM and schizophrenia. Finally, significantly higher proportions of ASM-SNPs, relative to non-ASM-SNPs, were detected at the loci exhibiting epigenetic signatures of enhancers, promoters, and were overrepresented within transcription factor binding regions and DNase I hypersensitive sites. All these findings collectively indicate that ASM-SNPs should be prioritized in GWAS follow-up studies.

Main Text

Genome-wide association studies (GWAS) have become the key approach for identifying DNA variants that influence complex traits. As a rule, even well-powered GWAS detect only a small number of genome-wide significant SNPs (5×10^{-8} is the accepted p-value significance threshold in GWAS[1]), while thousands of SNPs fall into the grey zone of sub-threshold p-values. It has been demonstrated

that SNPs that do not reach GWAS significance (even those with p-values as high as 0.1) account for a considerable proportion of predisposition to disease and contribute to "missing" heritability.[2] Separation of biological signal from noise within this sub-threshold zone requires hundreds of thousands of additional individuals,[3,4] but these are expensive to collect and may not even be feasible for rare phenotypes. Genetic - epigenetic interactions may provide new insights for prioritization of sub-threshold GWAS SNPs. SNPs that exhibit allele-skewed modification (ASM; "modification" is used to denote various types of epigenetic cytosine marks, not only methylation) are particularly interesting, because differential modifications may provide a mechanism for DNA polymorphisms to affect regulation of protein-coding genes and non-coding regions.[5,6] ASM refers to differential DNA modification of alleles in the vicinity of the polymorphic DNA site. Systematic analysis of ASM-SNPs in GWAS has been proposed numerous times,[5,6] but its testing has been hampered by lack of comprehensive lists of tissue-specific ASM-SNPs and poor overlap between platforms used for ASM detection (*e.g.* Illumina 27k or 450k microarrays) and for GWAS (*e.g.* Affymetrix SNP 6.0 microarrays).

To explore the relationship between GWAS sub-threshold SNPs and ASM, we identified brain ASM-SNPs from control individuals and individuals affected by schizophrenia (SCZ) or bipolar disorder (BPD). Frozen prefrontal cortex (Brodmann area 10) tissues from *post-mortem* control subjects (n=76), BPD (n=67), and SCZ (n=65) individuals were obtained from the Stanley Medical Research Institute and the Harvard Brain Tissue Resource Center. The overwhelming majority of the individuals were of European origin: n=74 in the control group, n=65 in BPD and n=64 in SCZ (as verified by Principal Component Analysis). We also investigated sperm samples from BPD individuals (n=24) and controls (n=24). Germline samples were collected in accordance with the ethical standards at the Centre for Addiction and Mental Health (Toronto), and informed consent from each donor was obtained. Germ cells were isolated from the semen using two-layer discontinuous gradient separation protocol, ISolate® (Irvine Scientific). DNA samples from brain tissues and sperm were extracted using

standard phenol-chloroform method. Additional demographic data for all of the brain and sperm samples are summarized in Table S1.

To make our ASM-SNP analysis applicable to GWAS, we used Affymetrix SNP 6.0 microarrays - one of the most common GWAS platforms, which has also been thoroughly validated for ASM mapping.[7–9] Each of the 256 DNA samples from brain and sperm was interrogated twice on this platform. First, we performed regular SNP genotyping following the manufacturer's instructions. Second, we investigated allelic differences in DNA modification by enriching the unmodified DNA fraction using DNA modification- sensitive restriction enzymes.[10] For each sample, 500 ng of genomic DNA was separately digested with three restriction enzymes: *HpaII*, *HinP1I*, and *HpyCH4IV* (New England Biolabs). The three digests per sample were then pooled in equal amounts, and adaptors were ligated onto the ends of DNA fragments. To eliminate the fragments containing modified cytosines between the restriction enzyme targets, ligation products were additionally digested with *McrBC* (New England Biolabs). Samples were then PCR-amplified using primers complementary to the adaptor sequences, fragmented with DNAseI (EpiCentre), labelled (GeneChip DNA labelling reagent, Affymetrix), and hybridized to Affymetrix SNP 6.0 microarrays. Since DNA fragments longer than 2 kbp cannot be efficiently amplified under standard PCR conditions, the interrogated ASM effects were limited to <2 kbp regions surrounding a SNP. We used R (v2.12.1)[11] and the crlmm R package (v1.8.11) to background correct, normalize and summarize (via RMA) the SNP probes, and to make genotype calls. Datasets were normalized separately, as were genotyping and DNA modification arrays. For each SNP, we obtained a genotype call and a DNA modification level for each allele.

Depending on the cohort being examined, we removed 260,900-365,830 (28.8-40.4%) SNPs that failed our quality control tests. Such SNPs were missing annotation information, not found on autosomes or sex chromosomes, diverged from Hardy-Weinberg equilibrium (HWE), exhibited low minor allele frequency (MAF) or limited genotype confidence call rates. There was no particular

sample with unusually low genotype call confidence rates, and therefore all samples were included in the analysis. A threshold of $p < 10^{-10}$ was used to filter SNPs that failed HWE, and the vast majority of SNPs were in even stronger agreement with HWE: 97% of SNPs with $p > 10^{-10}$ also exhibited $p > 10^{-7}$. Batch effects were examined using Principal Component Analysis, and no outliers were found based on visual inspection of the first two principal components.

To identify ASM SNPs we investigate the relationship between modification and genotype. To optimize ASM-SNP detection, we performed a large-scale simulation study, testing five statistical models: Pearson's correlation, Spearman's rank-order correlation, mutual information (MI), piecewise linear regression (PWL), and analysis of variance (ANOVA). To compare these statistical methods, we generated a simulated dataset of 100 paired samples with 1,000,000 SNPs representative of our samples following the probability distribution of the genotype values and the mean plus standard deviation of the DNA modification levels to represent ternary and continuous variables, respectively. A row-wise dependency was added for 100,000 SNPs by adding a random signal-intensity offset uniformly sampled from the set {0.25, 0.5, 0.75}. We varied an adjusted p-value threshold and compared the number of dependencies found (Figure S1).

We used receiver operation characteristic (ROC) curves (Figure S2) to visualize the overall performance of each measure across all FDR adjusted p-value thresholds.[12] ANOVA and PWL were the top two performers in terms of sensitivity and specificity and showed nearly identical ROC curves with a true positive rate ~50% when false positive rate was ~10%. We chose PWL, which with the use of multiple CPU cores, is the ideal choice for large-scale data. Furthermore, PWL is a two-step linear regression model that provides a pattern of dependency between the genotypes AA to AB (slope 1) and AB to BB (slope 2). Statistically significant allelic DNA modification differences were identified as ASM-SNPs if they had at least one non-zero slope (*i.e.* either AA *vs.* AB or AB *vs.* BB) with $q < 0.01$.

We classified SNPs with FDR < 0.01 as ASM-SNPs and removed SNPs in linkage disequilibrium with one another ($r^2 > 0.25$). We generated four ASM-SNP lists: 1,374 ASM-SNPs were detected in control brains (1.31% of all SNPs investigated); 2,921 in SCZ brains (2.79%); 1,313 in BPD brains (1.25%); and 7,744 in all brain samples (7.40%). Examples of an ASM-SNP and non-ASM-SNP are shown in Figure 1. All pairwise overlaps among the four ASM-SNP lists were significantly higher than expected by chance alone ($p < 2.2 \times 10^{-16}$, hypergeometric test; Fig. 2).

Our primary goal was to evaluate whether brain ASM-SNPs could help prioritize sub-threshold GWAS SNPs in psychiatric diseases. To test this hypothesis we split the SCZ GWAS[3] (n= 81,080 SCZ cases and controls; referred henceforth as the 81k SCZ GWAS) SNPs into ten bins by p-value ($p \leq 0.1$; $0.1 < p \leq 0.2$; $0.2 < p \leq 0.3$; … $0.9 < p \leq 1$). Enrichment of ASM-SNPs in GWAS p-value bins was assessed using the hypergeometric test relative to a background of interrogated SNPs lacking ASM effects (non-ASM SNPs). Both GWAS and ASM-SNP lists were pruned to ensure our observations were not confounded by correlated SNPs. Pruning was implemented in PLINK[13], and based on an r^2 of 0.25, window size of 500kbp with a shift of 5 SNPs between windows. It was conducted using the LD structure from the HapMap Project CEU (European) samples (Phase 2, release 23)[14] to reflect the European-derived ASM-SNP lists. (HapMap data files were downloaded from the PLINK website.)

All four brain ASM-SNP lists showed significant enrichment in the GWAS $p \leq 0.1$ bin, but not in any of the other nine bins (Fig. 3a; Table S2). The strongest evidence for ASM-SNP enrichment was detected for the "all brains-ASM-SNPs" (OR 1.31, 95% CI 1.23-1.38; $p = 2.03 \times 10^{-19}$, hypergeometric test; Fig. 3a). No enrichment was observed in random sets of SNPs (Table S2). This pattern held when the $p \leq 0.1$ bin was further divided into five sub-bins between p-values of 0 and 0.05, where the strongest enrichment of ASM-SNPs in the SCZ GWAS SNPs with $p \leq 0.01$ was detected (Fig. 3b; Table S3). For example, the number of observed SCZ ASM-SNPs with GWAS $p \leq 0.01$ sub-bin was 76% larger than expected by chance (181 *vs.* 103; OR 1.55, 95% CI 1.2-2.1; $p = 2.85 \times 10^{-13}$, hypergeometric test).

By contrast, the number of ASM-SNPs with GWAS $0.04 < p \leq 0.05$ sub-bin was indistinguishable from that expected by chance alone (Table S3).

The same trend applies to the lower than $p \leq 0.01$ sub-bins: the more significant p-value sub-bin, the higher the proportion of ASM-SNPs. For example, SCZ ASM-SNPs in the SCZ GWAS SNPs with $p<10^{-7}$ exhibited an odds ratio of 7.3, while the odds ratio was 1.4 for $0.001 \leq p < 0.01$ (Fig. S3).

To demonstrate that ASM-SNP analysis can identify sub-threshold GWAS SNPs more likely to be disease-associated, we analyzed a 52k-individual SCZ GWAS,[15] which was a subset of the 81k-individual SCZ GWAS. We categorized sub-threshold ($5 \times 10^{-8} < p < 0.1$) GWAS SNPs in the 52k SCZ study as either ASM-SNPs or non-ASM-SNPs and tested how these two categories perform in the 81k SCZ GWAS. In the quantile-quantile plot of the p-values in the 81k SCZ GWAS, the ASM-SNPs showed a steeper slope compared to the non-ASM SNPs (observed p-values *vs.* expected p-values; Fig. 4). This demonstrates the potential of ASM-SNPs to prioritize GWAS candidates.

Next, we assessed whether brain ASM-SNPs can also be useful for candidate prioritization in other GWA studies. We tested the same four brain ASM-SNP lists in 17 additional large GWAS(>15 thousand individuals; range n = 15-170k individuals, with a mean of 89.7k, SD ± 45.7k; Table S4) conducted from 2010 onwards for non-psychiatric diseases and normal traits. Generally, if the same study conducted more than one GWAS on correlated traits, we selected the largest. After Bonferroni correction for multiple testing (720 tests were performed in total: 4 ASM-SNP lists, 10 GWAS p-value bins, and 18 GWAS), only three GWAS (in addition to SCZ) demonstrated $p < 0.01$ for at least one of the four ASM-SNP lists: high density lipoprotein, platelet count, and coronary heart disease (Fig. 5, Fig. 6 and Fig. S4; Table S4; Table S5). SCZ and cardiovascular traits exhibit strong association,[16] which partially may be determined by common genetic and epigenetic factors.

As in all epigenetic studies, the association between epigenetic factors and phenotypes raises a question about the direction of effect. Brain ASM-SNPs may be predisposing to SCZ or may be induced

by pathological processes in the brain, by treatment, or by other disease-associated factors. Two lines of evidence supported (but did not prove) a causal relationship between ASM and SCZ. First, ASM-SNPs detected in psychiatric disease-free (control) brains were enriched in the SCZ GWAS $p \leq 0.1$ bin (Fig. 3a; Table S2), indicating that ASM effects are present before disease onset. Second, we performed an ASM-SNP analysis in sperm samples from BPD males (n = 24) and age-matched controls (n = 24) as outlined above. There was significant overlap between the sperm ASM-SNP lists and the all brain ASM-SNP list. Forty one (56%) of the BPD-sperm ASM-SNPs were also all brain ASM-SNPs, and 134 (49%) of the control-sperm ASM-SNPs were also all brain ASM-SNPs ($p<2.2 \times 10^{-16}$ for both comparisons; hypergeometric test). Enrichments in the SCZ GWAS $p \leq 0.1$ bin were detected for ASM-SNP from both control-sperm list (51 *vs.* 37; OR 1.49, 95% CI 1.08-2.02; $p = 6.9 \times 10^{-3}$; hypergeometric test) and BPD-sperm list (20 *vs.* 10; OR 2.38, 95% CI 1.36-4.01; $p = 7.7 \times 10^{-4}$; hypergeometric test) but not in any other bin (Table S2). This finding, although not sufficiently robust to withstand multiple-testing correction, suggests a causal relationship between ASM-SNPs and SCZ.

To further elucidate the roles of ASM-SNPs in disease, we explored functional features of the genomic regions in which they are located. Splice sites and nonsynonymous SNPs were taken from the UCSC Genome Browser.[17] Splice site boundaries were defined as a window of 5 bp upstream and 5 bp downstream of a splice site. Nonsynonymous variants (coding SNPs that fall into one of the following categories: stop-gained/nonsense, missense, stop-lost, frameshift or in-frame indel) were defined as a single base pair. Cis eQTLs were defined as single base pairs from the GTEx[18–21] and UK Brain Expression Consortium[22]. DNase hypersensitivity clusters, available through the ENCODE Project,[23] have been created from all available cell types by uniform processing and merging replicates; peaks were defined by an FDR 1% threshold. UCSC Genes was available from the UCSC Genome Browser.[17] Three histone marks (H3K4Me1, H3K4Me3, H3K27Ac) and transcription factor binding sites were based on regions identified by chromatin immunoprecipitation followed by sequencing (ChiP-seq).

The "peaks" data available on UCSC Genome Browser[17] were used: regions of statistically significant signal enrichment, where scores associated with each enriched interval is the mean signal value across the interval. miRNA targets were identified by TargetScan.[24] Transcription start sites were single base pairs as defined by Gencode.[25] For all these annotations, we considered functional data available regardless of cell type, where applicable. Functional genomic characterization of ASM-SNPs was performed by comparing frequencies of functional elements within ASM-SNPs to the ones within SNPs that did not exhibit ASM, using the hypergeometric test.

A higher proportion of ASM-SNPs in the SCZ GWAS $p \leq 0.1$ bin, relative to non-ASM-SNPs in the same bin, was detected at the loci exhibiting epigenetic marks of enhancers (H3K4Me1 0.65 *vs.* 0.44, $p = 3 \times 10^{-93}$; H3K27ac 0.61 *vs*. 0.39, $p = 1 \times 10^{-103}$) and promoters (H3K4Me3 0.48 *vs*. 0.24, $p = 7 \times 10^{-134}$) as well as within transcription factor binding sites (0.23 *vs*. 0.08; $p = 4.5 \times 10^{-109}$) and DNase I hypersensitive sites (0.41 *vs.* 0.15; $p = 2 \times 10^{-205}$)(Table 1). It is important to note, however, that these epigenetic marks investigated were not brain-specific. Well characterized functional brain genome functional data, which is being generated through the psychENCODE project[26] will be particularly useful in brain ASM-GWAS studies.

Two recent articles also explored the additive value of ASM in SCZ GWAS and arrived at conclusions very similar to ours [27,28]. Hannon et al. performed ASM (which they called mQTL - methylation quantitative trait loci) analysis on a fetal brain collection and detected significant enrichment of SCZ GWAS hits amongst fetal brain mQTLs. Jaffe et al. showed that about half of SCZ GWAS significant or sub-threshold ($p<10^{-4}$) SNPs (or other SNPs in the LD block) exhibited mQTL effects in the prefrontal cortex. Since the two studies used a different platform (Illumina HumanMethylation450 microarray) and different design, direct comparison with our ASM findings is not straight-forward; however, the converging results of the three studies indicate that ASM effects are associated with disease risk.

In summary, we detected that brain ASM-SNPs show enrichment in the sub-threshold GWAS SNP bins and exhibit a gradient of the effect size: the smaller the p-value of sub-threshold SNPs, the larger the proportion of ASM-SNPs. If ASM was not associated with disease risk SNPs, the distribution of ASM-SNPs across different GWAS p value bins was not deviating from the one expected by chance only. Furthermore, brain ASM-SNP enrichment was the strongest in SCZ GWAS compared to the non-brain phenotypes, and germline ASM-SNPs exhibited enrichment in the sub-threshold SCZ GWAS SNP bins suggesting a causal association between at least some ASM-SNPs and disease. Finally, SCZ ASM-SNPs are overrepresented in the functional regions of the genome, such as promoters and enhancers. It can be hypothesized that epigenetic differences at regulatory elements are central to the etiopathogenic effects of SNPs, and this mechanism accounts for differential regulation of disease genes. The key element of this model is that both DNA sequence- and epigenetic-variation is necessary, and their synergistic effects at the gene regulatory regions generate a *bona fide* disease risk factor. Further ASM studies would benefit from high precision DNA modification mapping using bisulfite sequencing and differentiation between cytosine methylation, hydroxymethylation and other modifications.

Accession Codes

Microarray data have been deposited in the Gene Expression Omnibus (GEO) under accession number GSE71443.

Acknowledgements

This work was supported by the Canadian Institutes of Health Research (MOP-199170, MOP-119451 and MOP-77689), the US National Institutes of Health (MH088413, DK085698), the Krembil Foundation and Brain Canada to AP. AP is the Tapscott Chair in Schizophrenia Studies, University of

Toronto, Canada. This study was also supported by funding provided by the Ontario Government to the Ontario Institute for Cancer Research (PCB). PCB is a CIHR New Investigator. JK was funded by charitable donors at the CAMH Foundation: http://www.supportcamh.ca. JK holds the Joanne Murphy Professor in Behavioural Science. SAG was funded by the Peterborough K.M. Hunter Graduate Studentship, and also through the Armstrong Family via the CAMH Foundation. SAG is a fellow of CIHR STAGE Training Grant in Genetic Epidemiology and Statistical Genetics (GET-101831). The authors declare no competing financial interests.

Web Resources

UK Brain Expression Consortium website www.braineac.org

GTEx website http://www.ncbi.nlm.nih.gov/projects/gap/eqtl/index.cgi

PLINK website resource download page http://pngu.mgh.harvard.edu/~purcell/plink/res.shtml

Supplemental Data

Supplemental Data include four figures and five tables.

References

1. Pe'er, I., Yelensky, R., Altshuler, D., and Daly, M.J. (2008). Estimation of the multiple testing burden for genomewide association studies of nearly all common variants. Genet. Epidemiol. *32*, 381–385.

2. International Schizophrenia Consortium, Purcell, S.M., Wray, N.R., Stone, J.L., Visscher, P.M., O'Donovan, M.C., Sullivan, P.F., and Sklar, P. (2009). Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature *460*, 748–752.

3. Schizophrenia Working Group of the Psychiatric Genomics Consortium (2014). Biological insights from 108 schizophrenia-associated genetic loci. Nature *511*, 421–427.

4. Jostins, L., Ripke, S., Weersma, R.K., Duerr, R.H., McGovern, D.P., Hui, K.Y., Lee, J.C., Philip Schumm, L., Sharma, Y., Anderson, C.A., et al. (2012). Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. Nature *491*, 119–124.

5. Tycko, B. (2010). Mapping allele-specific DNA methylation: a new tool for maximizing information from GWAS. Am. J. Hum. Genet. *86*, 109–112.

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

7. Hellman, A., and Chess, A. (2007). Gene body-specific methylation on the active X chromosome. Science *315*, 1141–1143.

8. Kerkel, K., Spadola, A., Yuan, E., Kosek, J., Jiang, L., Hod, E., Li, K., Murty, V.V., Schupf, N., Vilain, E., et al. (2008). Genomic surveys by methylation-sensitive SNP analysis identify sequence-dependent allele-specific DNA methylation. Nat. Genet. *40*, 904–908.

9. Schalkwyk, L.C., Meaburn, E.L., Smith, R., Dempster, E.L., Jeffries, A.R., Davies, M.N., Plomin, R., and Mill, J. (2010). Allelic skewing of DNA methylation is widespread across the genome. Am. J. Hum. Genet. *86*, 196–212.

10. Schumacher, A., Kapranov, P., Kaminsky, Z., Flanagan, J., Assadzadeh, A., Yau, P., Virtanen, C., Winegarden, N., Cheng, J., Gingeras, T., et al. (2006). Microarray-based DNA methylation profiling: technology and applications. Nucleic Acids Res. *34*, 528–542.

11. R Core Development Team (2008). A language and environment for statistical computing. R Found. Stat. Comput.

12. Fawcett, T. (2006). An Introduction to ROC Analysis. Pattern Recogn Lett *27*, 861–874.

13. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J., Sklar, P., de Bakker, P.I.W., Daly, M.J., et al. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. *81*, 559–575.

14. Altshuler, D.M., Gibbs, R.A., Peltonen, L., Altshuler, D.M., Gibbs, R.A., Peltonen, L., Dermitzakis, E., Schaffner, S.F., Yu, F., Peltonen, L., et al. (2010). Integrating common and rare genetic variation in diverse human populations. Nature *467*, 52–58.

15. Schizophrenia Psychiatric Genome-Wide Association Study (GWAS) Consortium (2011). Genome-wide association study identifies five new schizophrenia loci. Nat. Genet. *43*, 969–976.

16. Correll CU, Robinson DG, Schooler NR, and et al (2014). Cardiometabolic risk in patients with first-episode schizophrenia spectrum disorders: Baseline results from the raise-etp study. JAMA Psychiatry *71*, 1350–1363.

17. Meyer, L.R., Zweig, A.S., Hinrichs, A.S., Karolchik, D., Kuhn, R.M., Wong, M., Sloan, C.A., Rosenbloom, K.R., Roe, G., Rhead, B., et al. (2013). The UCSC Genome Browser database: extensions and updates 2013. Nucleic Acids Res. *41*, D64–D69.

18. Montgomery SB, S.M., Gutierrez-Arcelus M, Lach RP, Ingle C, Nisbett J, Guigo R, Dermitzakis ET (2010). Transcriptome genetics using second generation sequencing in a Caucasian population. Nature *464*, 773–777.

19. Schadt, E.E., Molony, C., Chudin, E., Hao, K., Yang, X., Lum, P.Y., Kasarskis, A., Zhang, B., Wang, S., Suver, C., et al. (2008). Mapping the Genetic Architecture of Gene Expression in Human Liver. PLoS Biol *6*, e107.

20. Gibbs JR, van der B.M., Hernandez DG, Traynor BJ, Nalls MA, Lai SL, Arepalli S, Dillman A, Rafferty IP, Troncoso J, Johnson R, Zielke HR, Ferrucci L, Longo DL, Cookson MR, Singleton AB (2010). Abundant Quantitative Trait Loci Exist for DNA Methylation and Gene Expression in Human Brain. PLoS Genet. *6*, e1400952.

21. Stranger, B.E., Forrest, M.S., Dunning, M., Ingle, C.E., Beazley, C., Thorne, N., Redon, R., Bird, C.P., de Grassi, A., Lee, C., et al. (2007). Relative impact of nucleotide and copy number variation on gene expression phenotypes. Science *315*, 848–853.

22. Trabzuni, D., Ryten, M., Walker, R., Smith, C., Imran, S., Ramasamy, A., Weale, M.E., and Hardy, J. (2011). Quality control parameters on a large dataset of regionally dissected human control brains for whole genome expression studies. J Neurochem *119*, 275–282.

23. The ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57–74.

24. Lewis, B.P., Burge, C.B., and Bartel, D.P. (2005). Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. Cell *120*, 15–20.

25. Harrow, J., Frankish, A., Gonzalez, J.M., Tapanari, E., Diekhans, M., Kokocinski, F., Aken, B.L., Barrell, D., Zadissa, A., Searle, S., et al. (2012). GENCODE: the reference human genome annotation for The ENCODE Project. Genome Res *22*, 1760–1774.

26. Akbarian, S., Liu, C., Knowles, J.A., Vaccarino, F.M., Farnham, P.J., Crawford, G.E., Jaffe, A.E., Pinto, D., Dracheva, S., Geschwind, D.H., et al. (2015). The PsychENCODE project. Nat. Neurosci. *18*, 1707–1712.

27. Hannon, E., Spiers, H., Viana, J., Pidsley, R., Burrage, J., Murphy, T.M., Troakes, C., Turecki, G., O'Donovan, M.C., Schalkwyk, L.C., et al. (2015). Methylation QTLs in the developing brain and their enrichment in schizophrenia risk loci. Nat. Neurosci. *advance online publication*,.

28. Jaffe, A.E., Gao, Y., Deep-Soboslay, A., Tao, R., Hyde, T.M., Weinberger, D.R., and Kleinman, J.E. (2015). Mapping DNA methylation across development, genotype and schizophrenia in the human frontal cortexM. Nat. Neurosci. *advance online publication*,.

Table 1. Overrepresentation of ASM-SNPs in the functional regions of the genome. ASM-SNPs in SCZ GWAS $p \leq 0.1$ and are found in numerous functional regions of the genome more than expected by chance alone (uncorrected hypergeometric test p-values). In order to retain all group-specific ASM-SNPs, we investigated the union of the four brain ASM-SNP lists. There were 2,351 ASM-SNPs/122,186 non-ASM-SNPs in SCZ GWAS $p \leq 0.1$. All SNPs are annotated in a binary fashion, indicating the presence or absence of a functional characteristic for the SNP itself.

FUNCTIONAL CHARACTERISTIC	SNPS WITH GWAS P ≤ 0.1		P
	ASM-SNPs	non-ASM-SNPs	
Splice site	0.0021	0.0009	0.0249
Non-synonymous variant	0.0030	0.0039	0.7032
DNase HS Clusters	0.4122	0.1483	2.02×10^{-205}
GTEx Consortium eQTLs	0.02850	0.0109	4.51×10^{-12}
UK Brain Consortium eQTLs	0.1268	0.0885	3.80×10^{-10}
UCSC Genes	0.5070	0.4207	3.06×10^{-17}
H3K4Me1	0.6508	0.4392	3.00×10^{-93}
H3K4Me3	0.4785	0.2419	7.22×10^{-134}
H3K27ac	0.6159	0.3931	1.27×10^{-103}
Txn Factor ChIP	0.2344	0.0821	4.54×10^{-109}
miRNA targets predicted by TargetScan	0	0	N/A
Gencode-Transcription start sites	0.0936	0.0835	0.0396

Figure Legends

Figure 1. Beeswarm plots of unmodified DNA signal intensity by genotype. Examples of (a) ASM-SNP - rs16835902 (C/G; chromosome 3) and (b) non-ASM-SNP - rs39704 (C/T; chromosome 3). Based on the sequences flanking the SNPs, the informative enzymes for rs16835902 and rs39704 were *HpaII* and *HpyCH4IV*, respectively.

Figure 2. Overlap of identified ASM-SNPs among brain cohorts: controls, BPD, SCZ and the all brain sample. All brain= ASM-SNPs identified in all the brains; SCZ= ASM-SNPs identified in the brains of schizophrenia individuals; control= ASM-SNPs identified in the control brains; BPD= ASM-SNPs identified in the brains of bipolar disorder individuals.

Figure 3. Distribution of ASM-SNPs in schizophrenia GWAS p-value bins. (a) ASM-SNPs detected in the brains of controls, SCZ and BPD patients are overrepresented in the sub-threshold $p \leq 0.1$ SCZ GWAS SNP group. SCZ GWAS p-value bins are plotted on the x-axis, $-\log_{10}$ p-values are on the y-axis. (b) Further division of the ASM-SNPs in GWAS $p \leq 0.05$ reveals the highest density of ASM-SNPs in the $p \leq 0.01$ sub-bin.

Figure 4. The quantile-quantile plot shows ASM-SNPs and non-ASM-SNPs with a $p \leq 0.1$ in the 52k SCZ GWAS. The observed quantiles were derived from the 81k SCZ GWAS p-values for the respective SNPs, while the expected quantiles were from a continuous uniform distribution of p-values. The steeper slope of the ASM-SNPs indicates that these SNPs have lower p-values in the 81k SCZ GWAS, where both the sample size and power is greater, compared to the non-ASM-SNPs. The plotted ASM-SNPs are those from all brains in the $p \leq 0.1$ bin of the 52k SCZ GWAS (n = 1,376) and the plotted non-ASM-SNPs are

those in the 52k SCZ GWAS $p \leq 0.1$ bin (n = 163,592 from the total of n = 1,252,902 SNPs tested in 52K SCZ GWAS).

Figure 5. Distribution of $-\log_{10}$ p-values (corrected for multiple testing) for 4 lists of brain ASM-SNPs interrogated in 18 large GWAS. Only GWAS SNP $p \leq 0.1$ bins are presented here. Sample size of each GWAS in thousands (k) is shown above each row of ASM-SNP p-values.

Figure 6. Odds ratios for the enrichment of ASM-SNPs in the $p \leq 0.1$ bin for various GWAS. The colored bars are those ASM-SNP lists that were significantly enriched in the GWAS $p \leq 0.1$ bin after correction for multiple testing ($p < 0.01$).