Published in final edited form as:
Hum Genet. 2015 January ; 134(1): 77–87. doi:10.1007/s00439-014-1494-5.

Refinement of Schizophrenia GWAS Loci using Methylome-wide Association Data

Gaurav Kumar[a], **Shaunna L. Clark**[a], **Joseph L. McClay**[a], **Andrey A. Shabalin**[a], **Daniel E. Adkins**[a], **Linying Xie**[a], **Robin Chan**[a], **Srilaxmi Nerella**[a], **Yunjung Kim**[c], **Patrick F. Sullivan**[c,b], **Christina M. Hultman**[b], **Patrik K.E. Magnusson**[b], **Karolina A. Aberg**[a], and **Edwin JCG van den Oord**[a,*]

[a]Center for Biomarker Research and Personalized Medicine, School of Pharmacy, Virginia Commonwealth University, Richmond, VA, USA

[b]Department of Medical Epidemiology and Biostatistics, Karolinska Institute, Stockholm, Sweden

[c]Department of Genetics, University of North Carolina at Chapel Hill, NC, USA

Abstract

Recent genome-wide association studies (GWAS) have made substantial progress in identifying disease loci. The next logical step is to design functional experiments to identify disease mechanisms. This step, however, is often hampered by the large size of loci identified in GWAS that is caused by linkage disequilibrium (LD) between SNPs. In this study, we demonstrate how integrating methylome-wide association study (MWAS) results with GWAS findings can narrow down the location for a subset of the putative casual sites. We use the disease schizophrenia as an example. To handle “data analytic” variation we first combined our MWAS results with two GWAS meta-analyses (N=32,143 and 21,953), that had largely overlapping samples but different data analysis pipelines, separately. Permutation tests showed significant overlapping association signals between GWAS and MWAS findings. This significant overlap justified prioritizing loci based on the concordance principle. To further ensure that the methylation signal was not driven by chance, we successfully replicated the top three methylation findings near genes *SDCCAG8*, *CREB1* and *ATXN7* in an independent sample using targeted pyrosequencing. In contrast to the SNPs in the selected region, the methylation sites were largely uncorrelated explaining why the methylation signals implicated much smaller regions (median size 78bp). The refined loci showed considerable enrichment of genomic elements of possible functional importance and suggested specific hypotheses about schizophrenia etiology. Several hypotheses involved possible variation in transcription factor binding efficiencies.

Corresponding author: Dr. Gaurav Kumar. Center for Biomarker Research and Personalized Medicine, School of Pharmacy, McGuire Hall, Room 219. P. O. Box 980533. Richmond, VA 23298-0581, gkumar@vcu.edu, Phone: +1 (804) 628-6910, Fax: +1 (804) 628-3991.

Conflict of interest

The authors declare no competing financial interests.

Data

MWAS data has been deposited in dbGaP http://www.ncbi.nlm.nih.gov/projects/gap/cgibin/study.cgi?study_id=phs000608.v1.p1. GWAS meta-analysis results are available from http://www.people.vcu.edu/~ejvandenoord/ and https://pgc.unc.edu/Results.php

Keywords

MWAS; fine-mapping; association study; MBD-seq; Schizophrenia

Introduction

Schizophrenia (SCZ) is a major public health problem (Collins et al. 2011) that ranks ninth in global burden of illness (Murray and Lopez 1996). Patients experience hallucinations, delusions and impairment of sensory processing and higher cognitive function such as reasoning and planning (Uhlhaas and Singer 2010). Onset is typically in adolescence or early adulthood and the course of illness is characterized by exacerbations, remissions, and relapses. It is characterized by substantially increased morbidity and mortality, with a projected lifespan of about 15 years less than the general population (Harris and Barraclough 1998).

Of a large set of prenatal risk factors, having a first-degree relative with SCZ is one of the major risk factors (Sullivan et al. 2003), with genetic factors accounting for the majority of this familial risk (Murray et al. 2003). Although SCZ genetics has proven difficult, recent mega/meta GWAS studies have made substantial progress in identifying specific disease loci (Aberg et al. 2013a; Genome-wide association study identifies five new schizophrenia loci 2011; Ripke et al. 2013; Shi et al. 2009). The next logical step is to design functional experiments to identify neurobiological disease mechanisms. This step, however, is hampered by the large size of the loci implicated by GWAS, caused by linkage disequilibrium (LD) between SNPs. For example, the average size of the 22 loci reported in Ripke *et. al.* (Ripke et al. 2013) was 447kb with the largest locus spanning over 7 Mb. Clearly, the possibility to refine these putative causal loci would greatly expedite our ability to design targeted functional experiments.

Convergent genomic approaches that integrate different kinds of data may reduce platform specific errors and increase confidence in the robustness of the findings when multiple lines of evidence converge to the same biological factors (Niculescu et al. 2000). While our results will have these desirable properties, in this paper we focus on the ability of whole methylome data to refine disease loci. Multiple scenarios are conceivable where findings from GWAS and methylome-wide associations studies (MWAS) may implicate similar loci. For example, similar to SNPs, methylation in critical sites can inhibit the binding of transcription factor to their recognition elements (Prendergast and Ziff 1991), resulting in gene silencing. In contrast to LD between SNPs, correlations among methylation sites tend to be much more localized (Aberg et al. 2012). Therefore, combining results from MWAS with results from GWAS may help to refine GWAS implicated regions for further analysis.

The most comprehensive method to interrogate the methylome involves the use of next-generation sequencing (NGS) after bisulfite conversion of unmethylated cytosines. However, this is currently not economically feasible with the sample sizes required for MWAS (Rakyan et al. 2011). As a cost-effective alternative, we first captured the methylated DNA fragments and then sequenced only this methylation-enriched portion of the genome (Serre et al. 2010) (see reference (Aberg et al. 2012) for discussion on the merits of MBD-seq) in 1,459 subjects. Next, association test were performed on a methylome-wide scale (Aberg et al. 2014).

The MWAS data was combined with GWAS data. Even if the same data is used, differences in data analyses (e.g. quality control approach, software and methods) will accumulate to produce non-perfect correlations between GWAS test statistics/*p*-values. When focusing on the top results (i.e. a restriction of range), these non-perfect correlations will result in different lists of prioritized loci. To handle this “data analytical” variation we first combined our MWAS results with two GWAS meta-analyses separately (N=32,143 and 21,953). Those GWAS studies had largely overlapping samples but different data analysis pipeline. Next, we simply prioritized the loci present in both top lists for further analyses. Thus, highly prioritized findings observed only in one MWAS-GWAS combination were excluded from further analysis.

Two tests were performed to ensure the prioritized regions were not false positive findings. First, we performed permutation tests to demonstrate that GWAS and MWAS findings have significant overlapping association signals. Second, in contrast to the two GWAS meta-analyses that were based on large samples, the MWAS was performed in a more modest sample size. We therefore replicated the three top methylation findings in independent samples using a different technology.

Methods

Data sets

Schizophrenia genome-wide association results—The study first analysed Swedish cases and controls and then conducted a meta-analysis with the first wave of PGC (Psychiatric Genomics Consortium) results for schizophrenia (Ripke et al. 2013). The combined number of subjects was 32,143 and the total number of SNPs (imputed on the 1000 genome reference panel) was 9,898,078. We refer to this dataset as GWAS-1. The second SCZ meta-analysis referred to as GWAS-2, containing 21,953 subjects of European descent (Aberg et al. 2013a). In this study 1,085,772 SNPs genotyped and imputed SNPs were available. Though, the two meta-analysis, GWAS-1 and GWAS-2, had largely overlapping samples but different data analysis pipeline. The third SCZ dataset refered as GWAS-SW contain only swedish samples after removing smples of other European ancestry from GWAS-1.

Schizophrenia methylome-wide association results—For details about the MWAS see Aberg *et al.* (Aberg et al. 2014). In summary, whole blood samples for the case and controls were collected. Cases were identified from the hospital discharge register and controls were separately selected at random from the national population registers in Sweden as a part of larger study (Ripke et al. 2013). We sequenced the methyl-CpG enriched genomic fraction and obtained an average of 68.0 million (SD=26.8) reads for 759 SCZ cases and 738 controls. We then estimated how many sequenced fragments covered each of the 26,752,702 autosomal CpGs in the reference genome (hg19/ GRCh37) to quantify methylation at each site. Extensive quality control was performed on reads, samples and sites. We also performed data reduction by combining correlated coverage estimates of adjacent CpGs into "blocks". This left 4,344,016 blocks for 1459 subjects. To control for potential confounders and improve power in the MWAS, we regressed out several laboratory variables, age/sex, and the first seven principal components (PCs).

Analyses

Testing for partly overlapping association signals—Integrating GWAS and MWAS results assumes that some of the findings overlap between the two approaches. To study this, for both GWAS-1 and GWAS-2, we mapped SNPs to the methylation blocks. SNPs may have good *p*-values because they tag effects of other SNPs in the region (i.e. LD) that are the causal variants. Therefore, when mapping methylation sites to SNPs, we can consider a broader region than just the location of the genotyped SNP itself. Specifically, we assumed an "LD" flanking region of 10kb around the genotyped SNP. Because of this flanking region we can now match methylation sites to SNPs, even if they do not overlap with the genotyped SNP but are merely located within 10kb. This is a reasonable approach because the causal variant could potentially still be at the methylation site. The advantage of using this flanking region is that we can now match many more methylation sites to SNPs, therefore improving the genome-wide coverage of the combined MWAS and GWAS analysis. Note that because we do not alter the size of the methylation site that is used to refine the association signal, this flanking method does not affect the fine mapping resolution.

Next, we used 10,000 permutations to test whether top SNPs from the GWAS analysis were also more likely to be among the top blocks from the MWAS. Our test statistic was an "information ratio" calculated as the observed number of GWAS association signals among the top MWAS findings divided by the expected number of GWAS association signals in the MWAS top results under the null hypothesis assuming no overlap. When defining what constitutes a significant association signal we avoided thresholds commonly used in traditional multiple testing approaches. This is because loci that may not reach genome-wide significance for a given platform may still represent true effects and emerge as a top finding when combined with results from different platforms. Instead, we explored combinations of four "empirical" percentile thresholds (1st, 5th, 10th and 25th) to define the top results in the GWAS and MWAS.

Concordance based prioritization—Results from the permutation tests described in the previous section are shown in Figure 1. They suggested that there was an overlap between loci among the top results in the MWAS and in the each of the GWAS studies. Thus, selecting the top 1% of both the MWAS and GWAS-1 suggested an almost 1.05 fold enrichment of GWAS findings among the top MWAS findings with a significant permutation *p*-value of 0.031 (Figure 1A). Similarly, selecting the top 5% of MWAS and GWAS-2 suggested an enrichment of almost 1.1 with a permutation *p*-value of 7×10^{-4} (Figure 1B).

Next we used a simple algorithm to prioritize the overlapping loci for further study. Based on the results from the permutation testing we selected all sites in the top 1% of combined analysis of MWAS and GWAS-1 and also top 5% of combined analysis of MWAS and

GWAS-2. Next, to ensure robustness of our results across GWAS studies, we focused on loci selected by both analyses.

Targeted replication of selected methylation sites—In contrast to the GWAS results that were based on large samples and showed robust associations across different analyses, the MWAS involved a more modest sample size. To validate the top MWAS findings selected based on prioritization, we replicated the three top findings in an independent sample drawn from the same (Swedish) population using targeted pyrosequencing of bisulfite converted DNA. The genomic DNA was bisulfite converted using Epitect 96 (Qiagen, Germantown, MD). The bisulfite converted DNA was used as input material for PyroMark PCR and PyroMark pyrosequencing following the standard protocols provided by the vendor (Qiagen, Germantown, MD). The assays were designed using the Pyromark Assay Design software. Table S1 provides primer sequences. The laboratory evaluations of the assays included checks for unspecific binding and DNA assay with known methylation levels, as previously described (Aberg et al. 2014). Furthermore, to ensure consistency between plates, each plate included negative control and two controls with known methylation levels. To test for association between SCZ and methylation in the pyrosequencing data, SCZ disease status (affected/unaffected) was regressed on percent methylation at each CpG site. Age, sex and plate indicator variables were included as covariates to control for possible age and sex differences, and potential batch effects.

Annotation—We annotated MWAS blocks implicating SNPs in GWAS studies with genomic and epigenomic (histone tags from the blood cell-line GM12878) features using UCSC genome browser data (www.genome.ucsc.edu) (Karolchik et al. 2014). SNP annotations were mapped using snp135CodingDbSnp data and dbNSFP2.1(Liu et al. 2013) data downloaded from UCSC genome browser and www.dbnsfp.houstanbioinformatic.org, respectively.

Results

Results from our permutation tests, shown in Figure 1, suggested that there was an overlap between loci among the top results in the MWAS and in the each of the GWAS studies. For example, selecting the top 1% of both the MWAS and GWAS-1 suggested an almost 1.05 fold enrichment of GWAS findings among the top MWAS findings with a permutation *p*-value of 0.031 (Figure 1A). Similarly, selecting the top 5% of MWAS and GWAS-2 suggested an enrichment of almost 1.1 with a permutation *p*-value of 7×10^{-4} (Figure 1B).

The significant overlap found between the MWAS results and each of the two GWAS justified combining the results from the different platforms to further prioritize findings. Starting with the top overlapping sites, defined as the sites that overlap with both GWAS and have the best *p*-values in the MWAS, we selected all methylation sites in the entire region implicated by the GWAS. Next, to find the loci where the methylation signals were also significant we performed a Bonferroni correction on all methylation signals *p*-values in the GWAS implicated region. This process was repeated until we encountered overlapping sites that did not result in methylation blocks that passed Bonferroni correction anymore.

For illustrative purposes, we plotted *p*-values for the top three sites in Figure 2. The figures show that GWAS *p*-values remain small across a fairly large region. As implicated by the bottom panels of Figure 3, this pattern may be explained by the substantial LD in the regions, which makes it difficult to pinpoint the location of a potential casual disease locus. In contrast, the methylation *p*-values in Figure 2 are much more localized. The reason is that the absence of long range correlations between the methylation sites in this region (see top panels in Figure 3 that indicate the absence of correlation between methylation blocks). The median size of the methylation sites was 78bp suggesting considerable refinement of the GWAS implicated disease locus.

Table 1 reports all concordance based prioritized loci implicated by each of the two GWAS overlapping MWAS results. The Swedish samples used in the MWAS as well as the Swedish samples in the GWAS meta-analyses are all part of the same study (see (Ripke et al. 2013). To study whether concordance is also found within samples from the same study, Table 1 also reports the GWAS findings from just these Swedish samples (GWAS-SW). The sample sizes for GWAS-SW are much smaller (5,001 cases and 6,243 controls) compared to the sample sizes for the full meta-analysis (13,833 cases and 18,310 controls) resulting in reduced statistical power. Thus, the *p*-values for the GWAS SNPs are expected to be higher. Nevertheless, GWAS SNPs from the Swedish samples implicating our top three MWAS blocks all had *p*-values less than 0.05. Furthermore, with the possible exception of SNP rs1568853 KIF5C/LYPD6B (*p*-value 0.56) all other SNPs have *p*-values that reach or are close to nominal significance. Thus, among our concordance based prioritized loci the relationship between GWAS and MWAS seems to hold within samples from the same Swedish study as well.

To further confirm that these methylation peaks were not false positive findings, we replicated the top three sites in independent samples using targeted pyrosequencing of bisulfite converted DNA. We did not attempt to replicate the SNP finding as the replication sample would inevitably be much smaller than the two GWAS analyses (N=32,143 and 21,953) that both implicated the SNPs. Table 2 shows that the negative control (originally reported in our recent MWAS (Aberg et al. 2014)) did not overlap between MWAS and GWAS, and did not replicate (*p*-values 0.32 and 0.52). In contrast all three sites selected through concordance based prioritization replicated with *p*-values in the 10^{-5} to 10^{-10} range and effect sizes (Cohen's D) of about half a standard deviation. Thus, the methylation peaks overlapping GWAS findings seen in Figure 2 are unlikely caused by chance.

The top 3 finding identified with the concordance based prioritization approach implicated the genes *SDCCAG8*, *CREB1* and *ATXN7* respectively. All three genes have previously been implicated in schizophrenia. The top prioritized region was located on chromosome 1, (genomic coordinates: 243,493,888–243,493,966 and *p*-value 1.80×10^{-6}). The methylation block overlaps with an exon of the gene *SDCCAG8.* The region is conserved among eutherian mammals and contains predicted transcription factor binding sites. Epigenetic annotations suggest binding of histone-lysine N-methyltransferase EZH2 and histone 3, trimethylated at lysine 36 (H3K36m3) to this region.

The second methylation block (chr 2, genomic coordinates: 208,461,619–208,461,657 and *p-value* 2.73×10^{-6}) overlapped both an exon (genomic coordinates: 208,461,637-208,470,284) and partial intron of *CREB1*. The block also partially overlapped with AluSx repeats of class SINE and family Alu, with the remainder of the block showing sequence conservation among eutherian mammals and predicted transcription factor binding sites. Epigenetic annotations suggest binding of histone-lysine N-methyltransferase EZH2; histone 3 trimethylation at lysine 36 (H3K36m3) and histone 3 dimethylation at lysine 79 (H3K79m2).

The third methylation block (chr 3, genomic coordinates: 63,980,718 -63,980,830 and *p-value* 4.37×10^{-6}) partly overlapped with an intron of *ATXN7* and a DNase cluster indicating active chromatin in several cell types. The block also partially overlapped with repeat class LINE and transcription factor binding sites. This block was also associated with epigenetic features such as histone-lysine N-methyltransferase EZH2; histone 3 trimethylation at lysine 36 (H3K36m3) and histone 4 monomethylation at lysine 20 (H4K20m1).

The other results in Table 1 suggest that the significant findings identified by our concordance based prioritization overlap with several functional genomic and epigenetic features. For example, methylation signals overlap with *RERE*, *KIF5C*, *SRPK2* and a second site in *CREB1*. There is a methylation signal in the *BTN2A1* gene located in its promoter region, upstream of the transcription start site of the gene. Similarly, active chromatin state and conserved regions are highlighted by several methylation blocks. Histone modification such as histone 3 trimethylation at lysine 36 (H3K36m3) and EZH2 binding sites suggests epigenetic features as a theme of the overlapping signals.

Discussion

In summary, our study shows how MWAS results can be used to refine GWAS implicated disease loci. The significant overlap between top results from MWAS and GWAS studies provided the justification for combining the two types of studies. In contrast to the substantial LD between significant GWAS SNPs, the disease associated methylation sites did not show long range correlations in the overlapping regions. The median size of the methylation sites was merely 78bp suggesting considerable refinement of the GWAS implicated disease locus. These methylation signals were unlikely detected by chance as they replicated in independent samples. The methylation signals in our top prioritization sites implicated regions that showed considerable enrichment of genomic elements of possible functional importance. For example, four of the nine methylation refined regions were transcription factor binding sites. As both SNPs and methylation can inhibit the binding of transcription factor to their recognition elements (Prendergast and Ziff 1991), this may also explain why these regions were implicated by both MWAS and GWAS. The refinement we obtained could potentially allow us to design functional experiments to identify neurobiological disease mechanisms.

SNPs that create or destroy CpGs, called CpG-SNPs, are common constituting about 30% of all SNPs (Shoemaker et al. 2010). A simple explanation for the overlap between MWAS and GWAS findings is that allele frequencies between cases and controls for such CpG-SNPs imply methylation differences. For two reasons, however, it is unlikely that our top methylation sites merely tag allele frequency differences between cases and controls of such SNPs. First, only one of our top methylation blocks comprised a common (MAF > 0.05) CpG-SNP. Second, the statistical power is generally too low to detect methylation differences caused by CpG-SNPs. FigureS1 shows that only in extreme scenarios where the CpG-SNP explains a substantial proportion of the methylation variation (e.g. 5%) and the sample size is in the range of our methylation study, would there be sufficient power to detect the effect. Thus, rather than merely tagging CpG-SNP allele frequency differences between cases and controls, other mechanistic explanations need to be considered to explain the overlap.

Each of the genes at our top 3 findings, *SDCCAG8* (Genome-wide association study identifies five new schizophrenia loci 2011; Ripke et al. 2013), *CREB1* (Aberg et al. 2014) *and ATXN7* (Greenwood et al. 2013), has been previously implicated in SCZ. The protein encoded by cAMP (cyclic adenosine monophosphate) response element binding protein (*CREB1*) is a transcription factor involved in regulating gene expression in the brain as part of cAMP signaling cascades (Montminy 1997), and is a critical component of memory-related synaptic plasticity (Kandel 2012). In Table S2, we show that the MWAS block at *CREB1* overlapped with several transcriptions factor binding sites, including for *NKX2-2* that may be involved in the morphogenesis of the central nervous system (Fancy et al. 2004).

Serologically defined colon cancer antigen 8 (*SDCCAG8*, also known as *CCCAP*, *BBS16* and other aliases) encodes a centrosome-associated protein, which may be involved in centromsome organization during interphase and mitosis (Andersen et al. 2003; Kenedy et al. 2003). While the association evidence for *SDCCAG8* involvement in SCZ is relatively strong (*p*-value 2.53×10^{-8}, see Table 3 Ripke *et. al*.) (Ripke et al. 2013), its function as a centrosome-associated protein does not imply an obvious etiological mechanism and previous commentary on this topic has been scant (Hamshere et al. 2013). Exploratory functional genomics strategies may be needed to suggest new hypotheses for the role of *SDCCAG8* in SCZ etiology. Potential targets for future characterization, however, are the non-synonymous SNPs that lead to non-conservative amino acid changes in the critical parts of the *SDCCAG8* protein. The methylation block overlaps with missense SNPs (+/− 250bp) that are predicted to be damaging (probability of alleles affecting the molecular function) by both SIFT (Kumar et al. 2009) and PolyPhen2 (Adzhubei et al. 2010). These functional SNPs, however, are rare (MAF < 0.05). We also observed that our methylated CpG block completely overlaps with the 10th exon of *SDCCAG8* and is only 28bp from the 5′ and 3′ splice sites. DNA methylation is known to affect splicing (Gelfman et al. 2013) and splice factors are recruited to certain histone modifications, including H3K36Me3 modification (Luco et al. 2010). Binding of this modified histone has been observed at the genomic region implicated at *SDCCAG8*, as indicated in Table 1. Together, these observations suggest altered transcript splicing as a potential risk mechanism at this locus.

The final gene among our top 3 findings was ataxin 7 (*ATXN7*). This gene is best known for its causative role in spinocerebellar ataxin type 7 (*SCA7*), which presents with retinal degeneration and visual loss, demetia, hypoacusia, severe hypotonia, and auditory hallucinations (Benton et al. 1998). Notably, the latter symptoms are core features of SCZ. We observed that methylation block at *ATXN7* overlaps with transcription factor binding sites (Table S2) where *RELA* (NF-kappa-B) binds. NF-kappa-B protein mediates the regulation of immune response and its abnormal expression is associated with Autism spectrum condition (ASC) (Young et al. 2011), characterized as having an inflammatory component.

In contrast to our top three findings that replicated in independent sample using a different technology, the other reported top results are more speculative. Statistical tools exist to estimate the prediction error in this set (Hastie et al. 2001) but are very difficult to implement in this specific case as they require access to the raw data of the two GWAS studies. Furthermore, non-standard features of our prioritization method that involves two separate analyses that each combine two different data sets, further complicates the estimation. Because of their more speculative nature, we only briefly summarize the other top findings. *RERE* and *KIF5C* as risk loci underlying shared genetic effects in five major psychiatric disorders (Identification of risk loci with shared effects on five major psychiatric disorders: a genome-wide analysis 2013). The methylation block implicating *KIF5C* partially overlaps with the transcription start site where transcription factor *FOXD1* binds (Table S2). *BTN3A3/ BTN2A1* that belongs to the cluster of immunoglobulin superfamily located in major histocompatibility complex (MHC) region and contain methylated CpGs in the upstream region. The MHC region near *BTN3A3*/*BTN2A1* is strongly associated (*p-value* 7.96×10^{-9}) with psychiatric disorders (Identification of risk loci with shared effects on five major psychiatric disorders: a genome-wide analysis 2013). We observed different types of histone modification associated with our methylated CpG blocks. This observation is consistent with a well-known mechanistic link between histone modification and DNA methylation, for example H3K9me3 is required for DNA methylation (Henckel et al. 2009; Rothbart et al. 2012).

The pathogenic processes for SCZ likely involve the brain. Our methylation data, however, were obtained using DNA extracted from whole blood. Although the use of blood rather than brain methylome data likely reduced the number of overlapping GWAS and MWAS signals, we did find significant overlap. This can be explained by several mechanisms. None of these mechanisms assume that methylation in blood directly affects disease susceptibility, although this is in principle possible because blood provides a biological environment for other tissues including brain. First, factors that increase disease susceptibility may leave biomarker signatures in blood. For example, we previously found evidence for methylation biomarker signatures in genes involved in immune response (Aberg et al. 2014). The altered methylation in blood may be of no functional relevance and the actual disease causing process may involve different mechanisms that could partially be mediate by genetic factors. Indeed, GWAS studies have consistently implicated the MHC region that harbors many genes affecting immune response (Sullivan et al. 2012). Second, the methylation status of sites in blood may mirror the corresponding sites in blood. Although epigenetic differences are often associated with cell differences and critical for differentiation, correlated methylation profiles across tissues are fairly common (Christensen et al. 2009). These mirror sites occur because peripheral tissues may reveal methylation marks predating or resulting from the epigenetic reprogramming events affecting germ line and embryogenesis (Efstratiadis 1994). In addition, a variety of factors can affect methylation levels in brain (Kerkel et al. 2008; Sutherland and Costa 2003) have been shown to leave corresponding changes in blood (Aberg et al. 2013b) as well.

A limitation of our approach is that not every GWAS signal has a methylation component and not every methylation signal has a GWAS component. For example, as environmental effects cannot alter sequence variation, these phenomena cannot be detected with GWAS studies. Thus, the value of methylation data to refine GWAS signals is limited to a subset of loci.

In this study, we show how integrating MWAS data with GWAS findings can be used to narrow down the location for a subset of the putative causal sites. Results suggested specific hypotheses about SCZ etiology that could be further explored using exploratory functional genomics strategies. A number of these hypotheses involved in transcription factor binding efficiencies at the implicated sites.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This study was supported by the National Institute of Mental Health (grant 1R01MH097283). The present study is part of a larger project entitled '*A Large-Scale Schizophrenia Association Study in Sweden*' that is supported by grants from NIMH (MH077139) and the Stanley Medical Research Institute. Institutions involved in this project are: Karolinska Institute, Icahn School of Medicine at Mount Sinai, University of North Carolina at Chapel Hill, Virginia Commonwealth University, Broad Institute, and the US National Institute of Mental Health.

Library construction and next generation sequencing was performed by EdgeBio Gaithersburg, MD.

References

Aberg KA, et al. A comprehensive family-based replication study of schizophrenia genes. JAMA Psychiatry. 2013a; 70:573–581.10.1001/jamapsychiatry.2013.288 [PubMed: 23894747]

Aberg KA, et al. Methylome-wide association study of schizophrenia: identifying blood biomarker signatures of environmental insults. JAMA Psychiatry. 2014; 71:255–264.10.1001/jamapsychiatry.2013.3730 [PubMed: 24402055]

Aberg KA, et al. MBD-seq as a cost-effective approach for methylome-wide association studies: demonstration in 1500 case--control samples. Epigenomics. 2012; 4:605–621.10.2217/epi.12.59 [PubMed: 23244307]

Aberg KA, et al. Testing two models describing how methylome-wide studies in blood are informative for psychiatric conditions. Epigenomics. 2013b; 5:367–377.10.2217/epi.13.36 [PubMed: 23895651]

Adzhubei IA, et al. A method and server for predicting damaging missense mutations. Nat Methods. 2010; 7:248–249.10.1038/nmeth0410-248 [PubMed: 20354512]

Andersen JS, Wilkinson CJ, Mayor T, Mortensen P, Nigg EA, Mann M. Proteomic characterization of the human centrosome by protein correlation profiling. Nature. 2003; 426:570–574.10.1038/nature02166 [PubMed: 14654843]

Benton CS, de Silva R, Rutledge SL, Bohlega S, Ashizawa T, Zoghbi HY. Molecular and clinical studies in SCA-7 define a broad clinical spectrum and the infantile phenotype. Neurology. 1998; 51:1081–1086. [PubMed: 9781533]

Christensen BC, et al. Aging and environmental exposures alter tissue-specific DNA methylation dependent upon CpG island context. PLoS Genet. 2009; 5:e1000602.10.1371/journal.pgen.1000602 [PubMed: 19680444]

Collins PY, et al. Grand challenges in global mental health. Nature. 2011; 475:27–30.10.1038/475027a [PubMed: 21734685]

Efstratiadis A. Parental imprinting of autosomal mammalian genes. Curr Opin Genet Dev. 1994; 4:265–280. [PubMed: 8032205]

Fancy SP, Zhao C, Franklin RJ. Increased expression of Nkx2.2 and Olig2 identifies reactive oligodendrocyte progenitor cells responding to demyelination in the adult. CNS Mol Cell Neurosci. 2004; 27:247–254. [pii].

Gelfman S, Cohen N, Yearim A, Ast G. DNA-methylation effect on cotranscriptional splicing is dependent on GC architecture of the exon-intron structure. Genome Res. 2013; 23:789–799.10.1101/gr.143503.112 [PubMed: 23502848]

Genome-wide association study identifies five new schizophrenia loci. Nat Genet. 2011; 43:969–976.10.1038/ng.940 [PubMed: 21926974]

Greenwood TA, et al. Genome-wide linkage analyses of 12 endophenotypes for schizophrenia from the Consortium on the Genetics of Schizophrenia. Am J Psychiatry. 2013; 170:521–532.10.1176/appi.ajp.2012.12020186 [PubMed: 23511790]

Hamshere ML, et al. Genome-wide significant associations in schizophrenia to ITIH3/4, CACNA1C and SDCCAG8, and extensive replication of associations reported by the Schizophrenia PGC. Mol Psychiatry. 2013; 18:708–712.10.1038/mp.2012.67 [PubMed: 22614287]

Harris EC, Barraclough B. Excess mortality of mental disorder. Br J Psychiatry. 1998; 173:11–53. [PubMed: 9850203]

Hastie, T.; Tibshirani, R.; Friedman, J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer Verlag; New York: 2001.

Henckel A, Nakabayashi K, Sanz LA, Feil R, Hata K, Arnaud P. Histone methylation is mechanistically linked to DNA methylation at imprinting control regions in mammals. Hum Mol Genet. 2009; 18:3375–3383.10.1093/hmg/ddp277 [PubMed: 19515852]

Identification of risk loci with shared effects on five major psychiatric disorders: a genome-wide analysis. Lancet. 2013; 381:1371–1379.10.1016/S0140-6736(12)62129-1 [PubMed: 23453885]

Kandel ER. The molecular biology of memory: cAMP, PKA, CRE, CREB-1, CREB-2, and CPEB. Mol Brain. 2012; 5:14.10.1186/1756-6606-5-14 [PubMed: 22583753]

Karolchik D, et al. The UCSC Genome Browser database: 2014 update. Nucleic Acids Res. 2014; 42:D764–770.10.1093/nar/gkt1168 [PubMed: 24270787]

Kenedy AA, Cohen KJ, Loveys DA, Kato GJ, Dang CV. Identification and characterization of the novel centrosome-associated protein CCCAP. Gene. 2003; 303:35–46. S0378111902011411 [pii]. [PubMed: 12559564]

Kerkel K, et al. Genomic surveys by methylation-sensitive SNP analysis identify sequence-dependent allele-specific DNA methylation. Nat Genet. 2008; 40:904–908.10.1038/ng.174 [PubMed: 18568024]

Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat Protoc. 2009; 4:1073–1081.10.1038/nprot.2009.86 [PubMed: 19561590]

Liu X, Jian X, Boerwinkle E. dbNSFP v2.0: a database of human non-synonymous SNVs and their functional predictions and annotations. Hum Mutat. 2013; 34:E2393–2402.10.1002/humu.22376 [PubMed: 23843252]

Luco RF, Pan Q, Tominaga K, Blencowe BJ, Pereira-Smith OM, Misteli T. Regulation of alternative splicing by histone modifications. Science. 2010; 327:996–1000.10.1126/science.1184208 [PubMed: 20133523]

Montminy M. Transcriptional regulation by cyclic AMP. Annu Rev Biochem. 1997; 66:807–822.10.1146/annurev.biochem.66.1.807 [PubMed: 9242925]

Murray CJ, Lopez AD. Evidence-based health policy--lessons from the Global Burden of Disease Study. Science. 1996; 274:740–743. [PubMed: 8966556]

Murray, RM.; Jones, PB.; Susser, E.; Van Os, J.; Cannon, M., editors. The Epidemiology of Schizophrenia. Cambridge: Cambridge University Press; 2003.

Niculescu AB 3rd, Segal DS, Kuczenski R, Barrett T, Hauger RL, Kelsoe JR. Identifying a series of candidate genes for mania and psychosis: a convergent functional genomics approach. Physiol Genomics. 2000; 4:83–91. 4/1/83 [pii]. [PubMed: 11074017]

Prendergast GC, Ziff EB. Methylation-sensitive sequence-specific DNA binding by the c-Myc basic region. Science. 1991; 251:186–189. [PubMed: 1987636]

Rakyan VK, Down TA, Balding DJ, Beck S. Epigenome-wide association studies for common human diseases. Nat Rev Genet. 2011; 12:529–541.10.1038/nrg3000 [PubMed: 21747404]

Ripke S, et al. Genome-wide association analysis identifies 13 new risk loci for schizophrenia. Nat Genet. 2013; 45:1150–1159.10.1038/ng.2742 [PubMed: 23974872]

Rothbart SB, et al. Association of UHRF1 with methylated H3K9 directs the maintenance of DNA methylation. Nat Struct Mol Biol. 2012; 19:1155–1160.10.1038/nsmb.2391 [PubMed: 23022729]

Serre D, Lee BH, Ting AH. MBD-isolated Genome Sequencing provides a high-throughput and comprehensive survey of DNA methylation in the human genome. Nucleic Acids Res. 2010; 38:391–399.10.1093/nar/gkp992 [PubMed: 19906696]

Shi J, et al. Common variants on chromosome 6p22.1 are associated with schizophrenia. Nature. 2009; 460:753–757.10.1038/nature08192 [PubMed: 19571809]

Shoemaker R, Deng J, Wang W, Zhang K. Allele-specific methylation is prevalent and is contributed by CpG-SNPs in the human genome. Genome Res. 2010; 20:883–889.10.1101/gr.104695.109 [PubMed: 20418490]

Sullivan PF, Daly MJ, O'Donovan M. Genetic architectures of psychiatric disorders: the emerging picture and its implications. Nat Rev Genet. 2012; 13:537–551.10.1038/nrg3240 [PubMed: 22777127]

Sullivan PF, Kendler KS, Neale MC. Schizophrenia as a complex trait: evidence from a meta-analysis of twin studies. Arch Gen Psychiatry. 2003; 60:1187–1192.10.1001/archpsyc.60.12.1187 [PubMed: 14662550]

Sutherland JE, Costa M. Epigenetics and the environment. Ann N Y Acad Sci. 2003; 983:151–160. [PubMed: 12724220]

Uhlhaas PJ, Singer W. Abnormal neural oscillations and synchrony in schizophrenia. Nat Rev Neurosci. 2010; 11:100–113.10.1038/nrn2774 [PubMed: 20087360]

Young AM, Campbell E, Lynch S, Suckling J, Powis SJ. Aberrant NF-kappaB expression in autism spectrum condition: a mechanism for neuroinflammation. Front Psychiatry. 2011; 2:27.10.3389/fpsyt.2011.00027 [PubMed: 21629840]

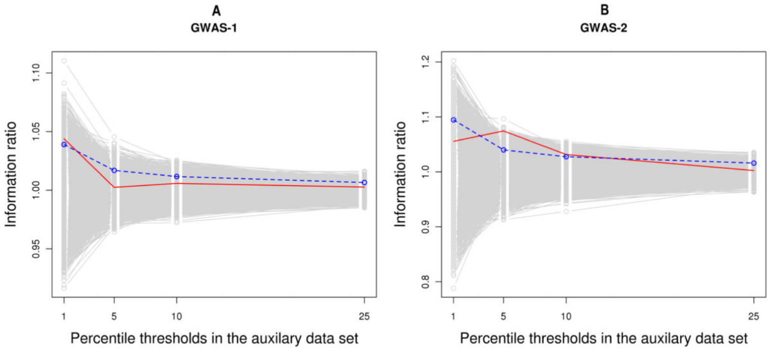

Figure 1. Permutation test results

This figure plots the “Information ratio”, calculated as the observed top 5% of CpG blocks overlapping at four “empirical” percentile thresholds (1st, 5th, 10th and 25th) in the auxiliary datasets GWAS-1 and GWAS-2. The information ratio for each threshold shown in red, while the blue dotted line indicates upper 95% percentile of the permutation in our significance threshold.

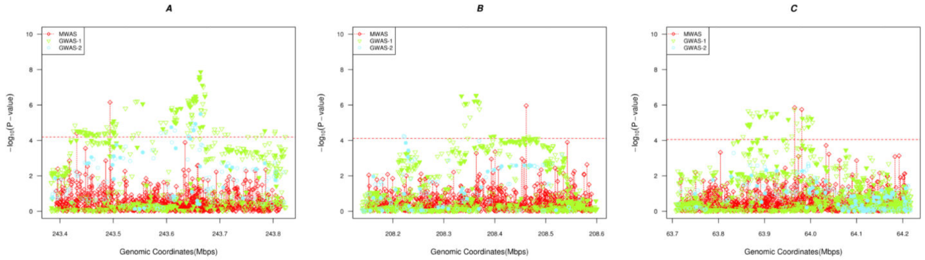

Figure 2. Regional Plot

Fine-map of top three MWAS blocks implicated in the GWAS datasets. MWAS block (CpG block), GWAS-1 SNPs and GWAS-2 SNPs are marked by different colours and shapes as red diamond, green-yellow triangle and cyan-blue circle, respectively. Filled triangles and circles represent a CpG-SNP in respective datasets. The horizontal dotted red line suggests significance of blocks after Bonferroni correction. Left (A), middle (B) and right (C) figures in this panel represent genomic regions near genes *SDCCAG8*, *CREB1 and ATXN7*, respectively.

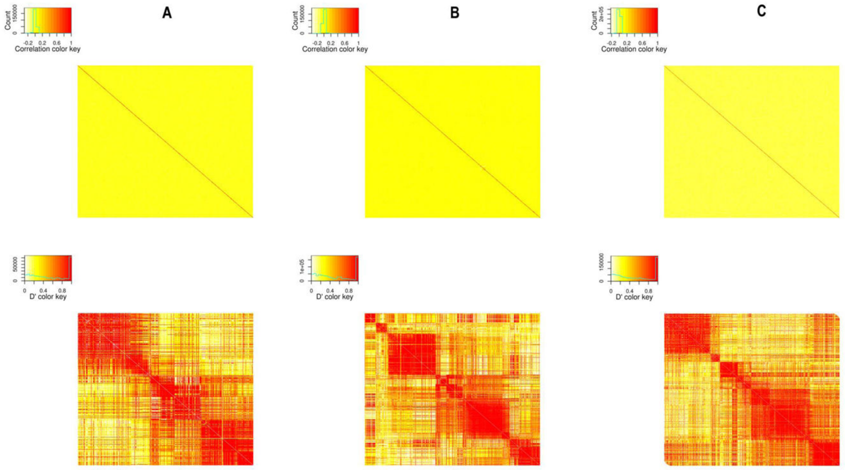

Figure 3. Block correlations and LD association in the region of our top 3 findings

Top panels show blocks correlations in the region of our top MWAS findings. Lower panels show the D’ between SNPs caused by linkage disequilibrium (LD) in the corresponding regions for each of the top findings.

Table 1

Top results showing GWAS SNPs implicated by MWAS (CpG) blocks with *p-value* < 0.01.

Chr	MWAS start coordinate	MWAS end coordinate	SNP	Gene name	MWAS *p*-value	GWAS-1 *p*-value	GWAS-2 *p*-value	GWAS-SW *p*-value	CpG -SNP	Genomic Features	Epigenetic Features	Functional SNP
1	8478545	8478577	rs301798	*RERE*	9.77E-06	4.02E-06	1.15E-05	6.87E-02	0	conservation; intron	EZH2, H3K27ac, H3K36m3, H3K4m1, H3K4m2, H3K4m3, H3K79m2, H3K9ac, H4K20m1	
1	243493888	243493966	rs2992632	*SDCCAG8*	1.80E-06	1.77E-05	1.10E-03	1.81E-02	0	conservation; exon; intron; spliceSite_flank_2bp; tfbsConsSites	EZH2; H3K36m3	Missense: 5 cds-synon: 1 Damaging: 1
2	149876935	149877120	rs1568853	*KIF5C/LYPD6B*	1.80E-03	8.24E-04	5.44 E-05	5.55E-01	0	conservation; DNase_Cluster; intron; tfbsConsSites;	EZH2; H2AZ; H3K27ac; H3K4m1; H3K4m2; H3K4m3; H3K9ac	-
2	208400393	208400492	rs2551641	*CREB1*	8.17E-04	3.81E-04	3.84E-03	1.43E-02	0	DNase_Cluster; Intron; repeats	EZH2; H3K27ac; H3K36m3; H3K4m1; H3K4m2; H3K4m3; H3K79m2; H3K9ac; H3K9m3; H4K20m1	-
2	208457511	208457511	rs1045780	*CREB1/METTL21A*	2.53E-03	1.04E-04	2.5E-03	3.59E-02	1	Intron; repeats	EZH2; H3K36m3; H3K79m2; H3K9m3	-
2	208461619	208461657	rs2551931	*CREB1/METTL21A*	2.73E-06	8.41E-05	2.63E-03	3.71E-02	0	conservation; exon; intron; repeats; spliceSite_flank2bp, tfbsConsSites	EZH2; H3K36m3; H3K79m2	Missense: 1 cds-synon: 4 Damaging:3
3	63980718	63980830	rs11922435	*ATXN7/LOC100507062/PSMD6*	4.37E-06	7.41E-06	4.78E-03	1.02E-03	0	DNase_Cluster; intron; repeats; tfbsConsSites	EZH2; H3K36m3; H4K20m1	-
6	26457220	26457353	rs3799380	*BTN3A3/BTN2A1*	1.46E-03	7.29E-06	2.39E-04	1.54E-01	0	upstream_8k	H2AZ; H3K27ac; H3K36m3; H3K4m1; H3K4m2; H3K4m3; H3K79m2; H3K9ac	-
7	104871507	104871603	rs2299319	*SRPK2*	3.14E-03	9.71E-07	1.23E-03	8.76E-04	0	DNase_Cluster; intron; repeats	H3K36m3; H3K4m1; H3K79m2; H4K20m1	-

Chromosome (“Chr”), MWAS start and end coordinates for each CpG are given for the human reference genome (hg19/ GRCh37). “SNP” implicated by CpG blocks in GWAS-1 GWAS-2 and GWAS-SW datasets. Association *p*-values for MWAS (CpG) blocks, GWAS-1, GWAS-2 and GWAS-SW datasets are shown. “Gene name” indicates genes within +/− 20 Kb flank of the CpG block. “CpG-SNP” associated with CpG blocks (+/− 250bp) indicates that a substitution of a C or G allele as that site could cause a CpG to be created or destroyed at that SNP location. Genomic and Epigenetic features describe the attributes of CpG blocks (+/−250bp).

Table 2

Replication of selected top sites.

Gene	Chr	Position (bp)	n	Beta	T-value	*p*-value	Cohen's D
SDCCAG8	1	243493888	358	−0.069	−4.56	5.18E-06	−0.49
		243493893	355	−0.064	−4.39	1.12E-05	−0.47
CREB1	2	208561648	1100	−0.047	−6.34	2.33E-10	−0.39
		208561657	1086	−0.048	−6.46	1.03E-10	−0.41
ATXN7	3	63980776	370	−0.067	−5.29	1.17E-07	−0.57
Negative Control							
PARK2	6	162054935	333	0.043	0.64	5.20E-01	0.02
		162054989	330	0.019	1.00	3.20E-01	−0.06

"Gene" = gene in which DMR (differentially methylated region) is located; "Chr" = chromosome; "Position" = co-ordinate on human reference genome (hg19/ GRCh37); "n" = number of samples, with methylation measurements for that CpG; "Beta" = regression coefficient; "T-value" = test statistic value. "*p*-value"= probability value of obtaining test statistic. "Cohen's D"= measure of the effect size.