Supplementary material to:

Independent methylome-wide association studies of schizophrenia detect consistent case-control differences

Swedish schizophrenia cohort 2

Data processing and quality control 2

Determining the significance of the cumulative MWAS signal by resampling 3

Gene Ontology 4

Identification of CpGs with concordance between blood-brain 5

Three array-based large-scale methylation datasets for schizophrenia 5

Table S1. Demographic overview of the study samples 7

Table S2. Enrichment testing of overlapping biological features 8

Table S3. Enriched Gene Ontology terms 8

Table S4. Comparison between cell-type-corrected MWAS and previous results 9

Figure S1. Scree test 10

Figure S2. Distribution of lambdas 11

Figure S3. Cluster plot of significantly enriched Gene Ontology terms 12

Figure S4. Cumulative MWAS signal 13

References 14

Swedish schizophrenia cohort

Our primary study sample included existing methylome-wide sequencing data from 759 schizophrenia cases and 738 controls,[1] which is a subset of individuals from a large-scale schizophrenia association study sample in Sweden.[2] A demographical overview is presented in **Table S1**. Cases with schizophrenia were identified via the Swedish Hospital Discharge Register. Population controls, who had never received a discharge diagnosis of schizophrenia, were selected at random from the national population registers and then group matched to the cases on age, gender and county of residence. All procedures were approved by ethical committees in Sweden and in the USA, and all subjects provided written informed consent (or legal guardian consent and subject assent).

Data processing and quality control

We performed thorough quality control of samples, reads, and CpG[3].

The existing study sample included 1,459 individuals with methylation data available. Using genotype information from previous GWAS studies and sequence variants called from the methylation data we searched for potential sample swaps. For 11 individuals the two data types did not match and it could not be determined if the sample swap had occurred in the methylation data or in the genotype data. Therefore, these individuals were excluded from further analysis. This left a sample of 1,448 subjects for statistical analysis.

Akin to filtering rare SNPs (SNPs with low minor allele frequency), we excluded rarely methylated sites. As these sites are unmethylated in most individuals they may create false positive MWAS findings due to low power or statistical problems associated with analyzing sparse data. This left 18,793,496 CpGs for MWAS, which corresponds to 67.3%% of all common CpGs in the human genome. Each methylation profile was sequenced with an average of 67.6 million (SD=26.2 million) reads per sample. Methylation scores were calculated by estimating the number of fragments covering the CpG using a non-parametric estimate of the

fragment size distribution[4]. These scores provide a relative measure of the amount of methylation for each individual at that specific locus. The average CpG score[4] across the methylation profiles was 2.57 (SD=1.07) with an average nonCpG-to-CpG score ratio[5] of 0.02 (SD=0.008). Thus, the average signal at the tested CpGs is sufficiently strong and the background noise level is low.

Determining the significance of the cumulative MWAS signal by resampling

To study the significance of the combined MWAS signals from associated methylation sites, we used the 'ramwas7riskScoreCV' function in RaMWAS. This function uses elastic-nets[6-8] as implemented in the R Glmnet package to predict case-control status. Elastic-nets are akin to multiple regression analysis but are suitable for our scenario where the number of predictors is much larger than the number of observations. Elastic nets were fitted by setting the alpha parameter to zero (i.e., ridge regression that retains all predictive sites in the model). To avoid over-fitting, *k*-fold cross-validation is used[9]. That is, the sample was randomly partitioned into k=10 equal sized subsamples. Of the k subsamples, $k-1$ are used as a "training" set to fit the elastic net and obtain weights for each CpG. The weights are then used to estimate schizophrneia disease status from the methylation data in the remaining "test" set. By alternating the subjects used in the training and test set, estimates are obtained for all subjects in the study. RaMWAS repeats the entire cycle of CpG selection through MWAS followed by estimation of weights using elastic-nets for each of the *k*-folds. Because both the selection of CpG sites and estimation of their weights is repeated for every fold and not affected by the participants in the test set, we obtain unbiased predictions of the disease status for each subject. By testing whether these predictions are significantly correlated with actual schizophrenia status, we performed an "in sample replication" of the cumulative MWAS signal.

Gene Ontology

We collected level 5 Gene Ontology (GO) terms using Bioconductor package GO.db (version 3.7.0) and extracted their gene annotation associations (http://geneontology.org/gene-associations/goa_human.gaf.gz file date: 2018-07-24 09:10). To prevent biased estimation of term enrichment, genes of a single gene family that were highly concentrated (i.e. tandemly arrayed genes) in terms of genomic location (e.g. immunoglobulins, olfactory receptors) were condensed into broader gene clusters such that each cluster had minimum interval of 50 kb to the next nearest family member. The assembled level 5 GO database was then use for enrichment analysis. These analyses used circular permutations that properly control the Type I error in the presence of correlated sites. Furthermore, as the permutations are performed on a CpG level they also properly account for gene size, as genes with more CpGs are more likely to be among the top results in the permutations.

Specifically, we first mapped the top MWAS CpGs to genes (Ensembl gene annotations GRCh37, release 91: ftp://ftp.ensembl.org/pub/grch37/release-91/) using the Bioconductor GRanges package. CpGs were allowed to map to multiple independent genes if their genomic position overlapped multiple unique gene annotations. After mapping, we performed 100,000 circular permutations at the CpG level. For each permutation, a two by two table was created by cross classifying whether or not the genes were among the top MWAS findings versus whether or not the gene was in the tested GO term. Each gene was counted only once when creating this table (thus, if there were three CpGs in the gene, this was counted as 1 and not as 3). Cramér's V (sometimes referred to as Cramér's phi) was used as the test statistic to measure whether genes from the GO term were overrepresented among the top MWAS genes. *P* values were calculated as the proportion of permutations that yielded a value equal to or greater than that of Cramér's V observed in the empirical data. To correct for multiple testing we controlled the family-wise error rate at the 0.05 level. For this purpose we performed 100,000 permutations

and determined the threshold that resulted in one or more significant GO terms in 5% of the 100,000 permutations. As the distribution of the permutation test statistics can vary somewhat across terms, they were standardized prior to correcting for multiple testing. In addition to controlling the family wise error rate we calculated the false discovery rate. For a more liberal threshold, we report enriched terms (*P* value < 0.01) all containing at least three overlapping genes at false discovery rate of 0.25 (*q* value ≤ 0.25).

Finally, one challenge for enrichment analysis in databases of biological pathways is that many pathways share a large number of common gene members. Therefore we used the Louvain Method for community detection[10] as implemented in igraph[11] to cluster significantly enriched terms based on the gene members in which they share, to help visualize nested/correlated GO terms.

Identification of CpGs with concordance between blood-brain

Overlapping CpGs from the analyses between the MBD-seq MWAS and Montano/Hannon2 were queried using BECon[12] (https://redgar598.shinyapps.io/BECon/) to obtain mean correlations between blood and brain for each site. CpGs with a modest correlation between blood and brain ($r \geq |0.2|$) were annotated to identify the genes used in our Gene Ontology analyses that were implicated by at least one blood-brain concordant CpG. The number of blood-brain concordant loci is reported per GO term in **Table S3a-b**. Note, as the BECon tool was developed using 450K array data, it was infeasible to search for blood-brain concordant CpGs only within the MBD-seq dataset.

Three array-based large-scale methylation datasets for schizophrenia

Three array-based methylation datasets for schizophrenia were generated using the Infinium Human Methylation450 BeadChip (Illumina). The datasets are:

[Montano] A study by Montano et al. [13] included DNA from blood from 689 schizophrenia cases and 645 controls. These samples originated from three multisite consortia: the Consortium on

the Genetics of Endophenotypes in Schizophrenia[14], the Project Among African-American to Explore Risks for Schizophrenia[15], and the Multiplex Multigenerational Family Study of Schizophrenia[16]. A demographical overview is presented in **Table S1**. Diagnostic assessment was performed using Diagnostic Interview for Genetic Studies along with medical records and schizophrenia diagnosis were set according to the Diagnostic and Statistical Manual for Mental Disorders 4th edition (DSM-IV) criteria. Written informed consent was obtained from all participants. The study was approved by relevant institutional review boards in the USA.

[Hannon-1] This methylation dataset, presented in a study by Hannon et al.,[17] included DNA from blood from 353 schizophrenia cases and 322 controls. These samples originated from the University College of London case-control cohort.[18] A demographical overview is presented in **Table S1**. Diagnostic assessment was performed with the clinical International Classification of Disease 10th edition (ICD-10) diagnosis for schizophrenia. In addition, research diagnostic criteria diagnosis were confirmed using interviews with the Schedule for Affective Disorders and Schizophrenia – Lifetime version (SADA-L).[19] All participants gave informed consent. The study was approved by both local and multiregional ethical committees in the UK.

[Hannon-2] Also this methylation dataset was originally presented by Hannon et al.[17] The sample included methylation profiles from blood from 414 schizophrenia cases and 433 controls from the Aberdeen case-control sample.[20] A demographical overview is presented in **Table S1**. Diagnostic assessments for schizophrenia were performed with ICD-10 and met criteria for DSM-IV. All participants gave informed consent. The study was approved by both local and multiregional ethical committees in the UK.

Table S1. Demographic overview of the study samples

Study Sample	Cases				Controls				Race
	N	Age [a]		Sex [b]	N	Age [a]		Sex [b]	
MBD-seq MWAS dataset									
Primary	744	53.1	11.55	55.1	704	55.0	11.64	54.6	Caucasian - collected in Sweden.
Array datasets									
Montano [c]	689	37.7	-	69.2	645	39.5	-	42.3	African American (37.4%/65.0% for cases/controls) & non-African American - collected in the US.
Hannon-1 [d]	353	43.7	14.63	72.0	322	36.8	14.65	44.1	Caucasian - collected in the UK.
Hannon-2 [e]	414	44.2	14.10	68.4	433	44.9	12.16	73.7	Mixed races, cases are matched with controls - collected in Scotland.

[a] Mean and standard deviations are given. [b] Percentage males. [c] Information obtained from eTable 1 in Montano et al. [13] Standard deviation of age was not reported in the original publication. [d] Information calculated from data available in Gene Expression Omnibus accession number GEO:GSE80417. Missing age information for 2/18 cases/controls. [e] Information calculated from data available in Gene Expression Omnibus accession number GEO:GSE84727. Missing age information for 154/28 cases/controls.

Table S2. Enrichment testing of overlapping biological features

Tested genomic feature	Background	Odds Ratio	P
Gene	Rest of the genome (all genes excluded)	1.15	<0.00001
Exon	Rest of the genes	0.88	0.9999
Intron	Rest of the genes	1.08	<0.00001
3’ UTR	Rest of the genes	0.95	0.9992
5’ UTR	Rest of the genes	0.99	0.9129
8 kb upstream of transcription start site (potential promoter)	Rest of the genome (all 8kb upstream excluded)	1.18	<0.00001
CG island in Dnase cluster	Rest of the DNase clusters	0.87	0.9994
CG island not in Dnase cluster	Rest of the DNase clusters	0.86	1.0000
CG island in gene	Rest of the genes	0.77	1.0000
CG island not in gene	Rest of the genes	0.75	1.0000
CG island shore in gene (2kb)	Rest of the genes	0.94	0.9993
CG island shore not in gene (2kb)	Rest of the genes	0.97	0.9654
Dnase cluster in gene	Rest of the genes	1.01	0.6393
Dnase cluster not in gene	Rest of the genes	1.04	0.4017
Enhancer in conserved	Rest of the conserved regions	1.24	0.1164
ncRNA	Rest of the genome (all ncRNA excluded)	1.12	<0.00001
Repeat	Rest of the genome (all Repeat excluded)	1.26	<0.00001
Splice site in conserved	Rest of the conserved regions	1.10	0.2415
Splice site in gene	Rest of the genes	0.96	0.9063
TFBS in conserved	Rest of the conserved regions	1.01	0.2595
TFBS in gene	Rest of the genes	0.88	1.0000
TFBS not in conserved	Rest of the conserved regions	1.01	0.5830
TFBS not in gene	Rest of the conserved regions	0.88	1.0000
Conserved in gene	Rest of the genes	0.97	0.9905
Conserved not in gene	Rest of the genes	0.92	0.9765

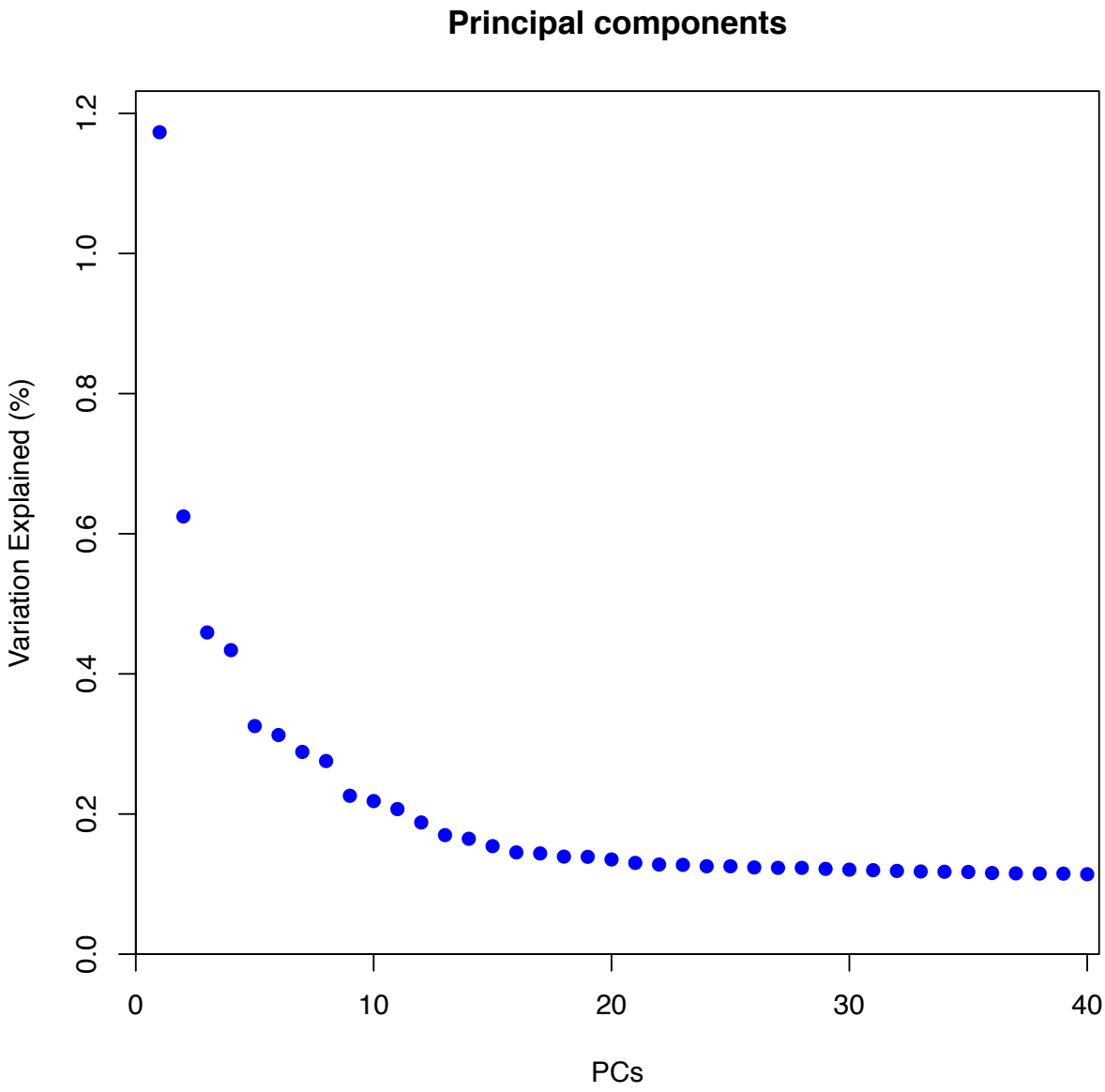
Table S3. Enriched Gene Ontology terms

Please see separate excel file. Enriched level 5 Gene Ontology terms for **a)** the MBD-seq MWAS results of suggestive significance (P < 1e-5), **b)** the overlap between the top 5% MBD-seq and top 1% Montano MWAS findings, and **c)** the overlap between the top 5% of MBD-seq and top 5% of Hannon-2 MWAS findings. To correct for multiple testing we controlled the family-wise error rate at the 0.05 level and for a more liberal threshold, we report enriched terms (P value < 0.01) all containing at least three overlapping genes at false discovery rate of 0.25 (q value ≤ 0.25). "Blood/Brain Concord." shows genes implicated by at least one CpG with modest or better inter-individual correlation ($r \geq |0.2|$) between blood and brain that are part of the GO term enrichment.

Table S4. Comparison between cell-type-corrected MWAS and previous results

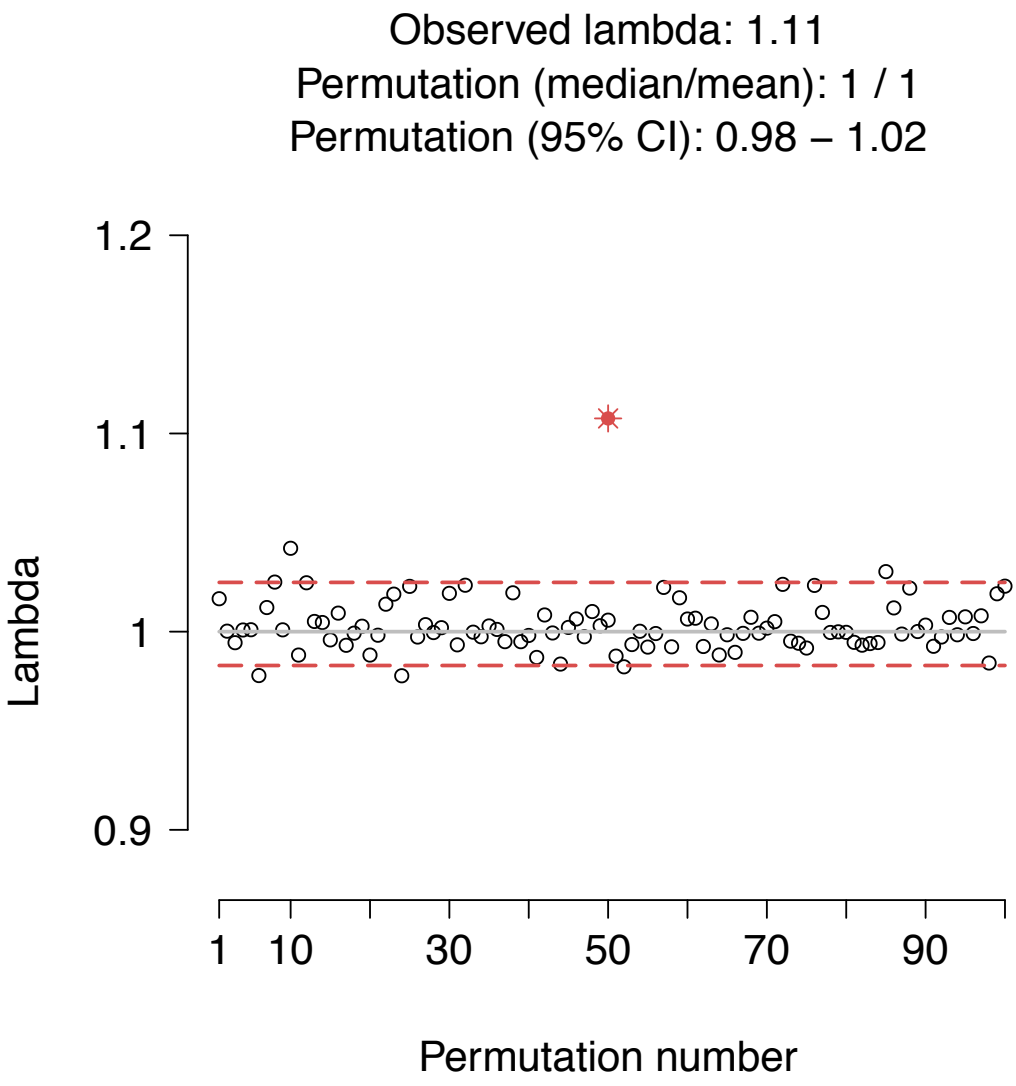
Please see separate excel file. Sites in the cell-type-corrected MWAS with P values < 1.00×10^{-5} with corresponding results from the previous analysis. Please note, due to improved alignment algorithms and differences in analysis strategy[1] not all sites in the cell-type-corrected MWAS have corresponding information in the previous analysis.

Figure S1. Scree test



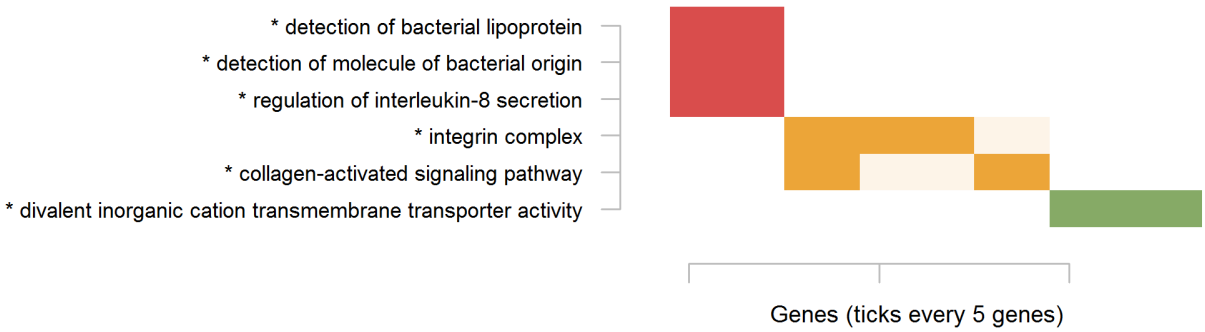
Scree test showing the percent variance explained (y-axis) by the first 40 principal components (PCs; x-axis) observed in the MBD-seq dataset after controlling for relevant covariates

Figure S2. Distribution of lambdas



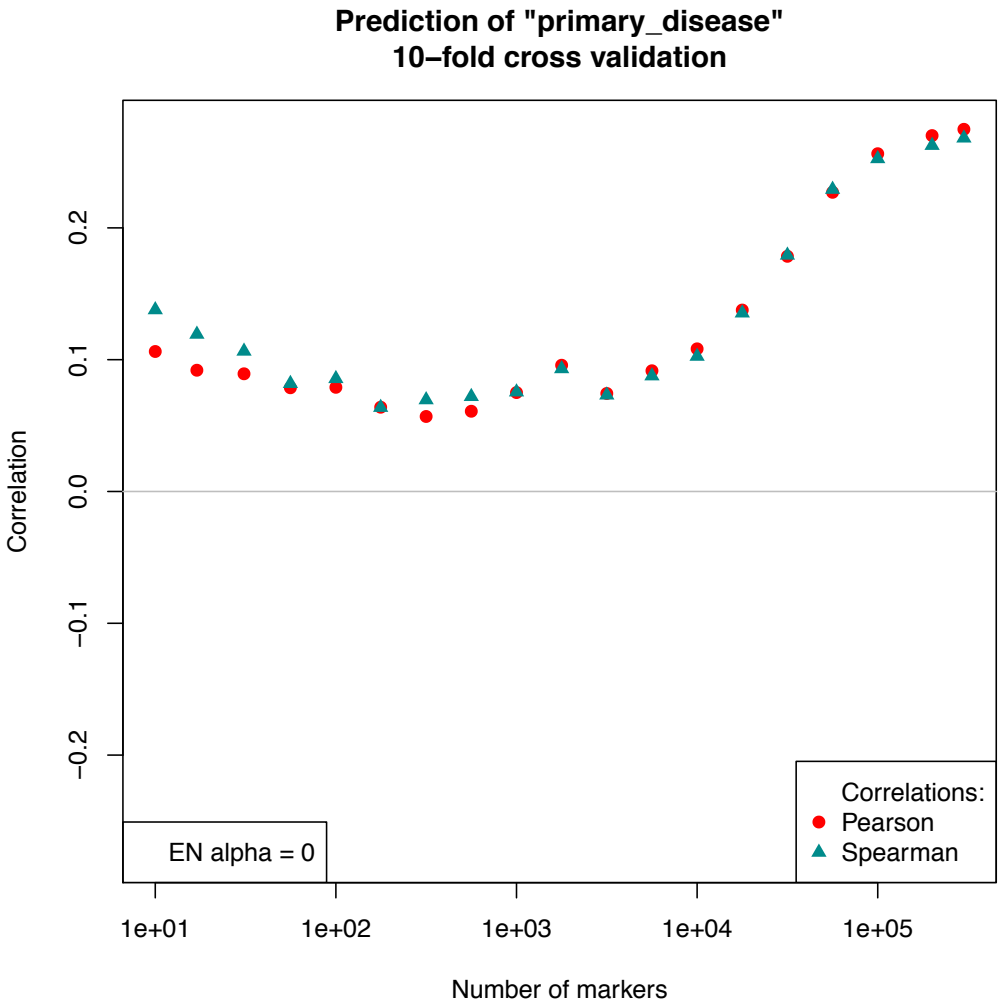
Lambdas from 100 MWAS of permuted case-control status vs. the lambda observed in the original MBD-seq dataset

Figure S3. Cluster plot of significantly enriched Gene Ontology terms



The associated loci included 388 genes that were enriched ($P < 0.01$, $q < 0.25$, minimum 3 gene overlap) for six Gene Ontology (GO) terms that segregated into three clusters (see **Table S3** for full statistics).

Figure S4. Cumulative MWAS signal



The correlation (y-axis) between the methylation-predicted case-control status and actual disease status is shown for the number of MWAS top markers (x-axis) included in the prediction. The cumulative effect (correlation) detected by this approach steadily increases with the inclusion of additional markers and reaches a plateau at ~100,000 markers. Thus, the steady increase shows that different associated sites contribute (partly) unique information. However, this should not be interpreted as if all included markers have an independent (uncorrelated) effect, or any effect at all, but rather that the majority of independent effects are represented among the top 100,000 markers. The observation of many markers with effects is in agreement with an observed lambda slightly above 1 that could not be explained by statistical artifacts as the permuted MWAS lambdas showed that the test statistic followed the theoretical null distribution.

References

1. Aberg KA, McClay JL, Nerella S, et al. Methylome-wide association study of schizophrenia: identifying blood biomarker signatures of environmental insults. *JAMA Psychiatry* Mar 2014;71(3):255-264.
2. Bergen SE, O'Dushlaine CT, Ripke S, et al. Genome-wide association study in a Swedish population yields support for greater CNV and MHC involvement in schizophrenia compared with bipolar disorder. *Mol Psychiatry* Sep 2012;17(9):880-886.
3. Aberg KA, McClay JL, Nerella S, et al. MBD-seq as a cost-effective approach for methylome-wide association studies: demonstration in 1500 case-control samples. *Epigenomics* Dec 2012;4(6):605-621.
4. van den Oord EJ, Bukszar J, Rudolf G, Nerella S, McClay JL, Xie LY, Aberg KA. Estimation of CpG coverage in whole methylome next-generation sequencing studies. *BMC Bioinformatics* Feb 12 2013;14(1):50.
5. Shabalin AA, Hattab MW, Clark SL, Chan RF, Kumar G, Aberg KA, van den Oord E, Birol I. RaMWAS: Fast Methylome-Wide Association Study Pipeline for Enrichment Platforms. *Bioinformatics* Feb 12 2018.
6. Friedman J, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *J Stat Softw* 2010;33(1):1-22.
7. Simon N, Friedman J, Hastie T, Tibshirani R. Regularization Paths for Cox's Proportional Hazards Model via Coordinate Descent. *J Stat Softw* Mar 2011;39(5):1-13.
8. Tibshirani R, Bien J, Friedman J, Hastie T, Simon N, Taylor J, Tibshirani RJ. Strong rules for discarding predictors in lasso-type problems. *J R Stat Soc Series B Stat Methodol* Mar 2012;74(2):245-266.
9. Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York: Springer Verlag; 2001.
10. Vincent DB, Jean-Loup G, Renaud L, Etienne L. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment* 2008;2008(10):P10008.
11. Csardi G, Nepusz T. The igraph software package for complex network research. *InterJournal Complex Systems* 2006;1695.
12. Edgar RD, Jones MJ, Meaney MJ, Turecki G, Kobor MS. BECon: a tool for interpreting DNA methylation findings from blood in the context of brain. *Translational Psychiatry* 08/01/online 2017;7:e1187.
13. Montano C, Taub MA, Jaffe A, et al. Association of DNA Methylation Differences With Schizophrenia in an Epigenome-Wide Association Study. *JAMA Psychiatry* May 01 2016;73(5):506-514.
14. Calkins ME, Dobie DJ, Cadenhead KS, et al. The Consortium on the Genetics of Endophenotypes in Schizophrenia: model recruitment, assessment, and endophenotyping methods for a multisite collaboration. *Schizophr Bull* Jan 2007;33(1):33-48.

15. Aliyu MH, Calkins ME, Swanson CL, Jr., et al. Project among African-Americans to explore risks for schizophrenia (PAARTNERS): recruitment and assessment methods. *Schizophr Res* Oct 2006;87(1-3):32-44.
16. Gur RE, Nimgaonkar VL, Almasy L, et al. Neurocognitive endophenotypes in a multiplex multigenerational family study of schizophrenia. *Am J Psychiatry* May 2007;164(5):813-819.
17. Hannon E, Dempster E, Viana J, et al. An integrated genetic-epigenetic analysis of schizophrenia: evidence for co-localization of genetic associations and differential DNA methylation. *Genome Biol* Aug 30 2016;17(1):176.
18. Datta SR, McQuillin A, Rizig M, et al. A threonine to isoleucine missense mutation in the pericentriolar material 1 gene is strongly associated with schizophrenia. *Mol Psychiatry* Jun 2010;15(6):615-628.
19. Spitzer R, Endicott, J. The schedule for affective disorder and schizophrenia, lifetime version. 3rd ed. New York: New York Stste Psychatric Institute; 1977.
20. International Schizophrenia Consortium. Rare chromosomal deletions and duplications increase risk of schizophrenia. *Nature* 2008;455:237-241.