Case-control meta-analysis of blood DNA methylation and autism spectrum disorder

Shan V. Andrews, Johns Hopkins Bloomberg School of Public Health
Brooke Sheppard, Johns Hopkins Bloomberg School of Public Health
Gayle C. Windham, California Department of Public Health
Laura A. Schieve, Centers for Disease Control and Prevention
Diana E. Schendel, Aarhus University
Lisa A. Croen, Kaiser Permanente Division of Research
Pankaj Chopra, Emory University
Reid S. Alisch, University of Wisconsin Madison
Craig J. Newschaffer, Drexel University
Stephen Warren, Emory University

Only first 10 authors above; see publication for full author list.

Journal Title: Molecular Autism
Volume: Volume 9, Number 1
Publisher: BMC (part of Springer Nature) | 2018-06-28, Pages 40-40
Type of Work: Article | Final Publisher PDF
Publisher DOI: 10.1186/s13229-018-0224-6
Permanent URL: https://pid.emory.edu/ark:/25593/t3h30

Final published version: http://dx.doi.org/10.1186/s13229-018-0224-6

Copyright information:

© 2018 The Author(s).

This is an Open Access work distributed under the terms of the Creative Commons Attribution 4.0 International License (https://creativecommons.org/licenses/by/4.0/).

Accessed November 15, 2022 11:17 PM EST
Case-control meta-analysis of blood DNA methylation and autism spectrum disorder

Shan V. Andrews 1,2, Brooke Sheppard 1,2, Gayle C. Windham 3, Laura A. Schieve 4, Diana E. Schendel 5,6,7, Lisa A. Croen 8, Pankaj Chopra 9, Reid S. Alish 10, Craig J. Newschaffer 11,12, Stephen T. Warren 9,13,14, Andrew P. Feinberg 15,16, M. Daniele Fallin 2,15,17* and Christine Ladd-Acosta 1,2*†

Abstract

Background: Several reports have suggested a role for epigenetic mechanisms in ASD etiology. Epigenome-wide association studies (EWAS) in autism spectrum disorder (ASD) may shed light on particular biological mechanisms. However, studies of ASD cases versus controls have been limited by post-mortem timing and severely small sample sizes. Reports from in-life sampling of blood or saliva have also been very limited in sample size and/or genomic coverage. We present the largest case-control EWAS for ASD to date, combining data from population-based case-control and case-sibling pair studies.

Methods: DNA from 968 blood samples from children in the Study to Explore Early Development (SEED 1) was used to generate epigenome-wide array DNA methylation (DNAm) data at 485,512 CpG sites for 453 cases and 515 controls, using the Illumina 450K Beadchip. The Simons Simplex Collection (SSC) provided 450K array DNAm data on an additional 343 cases and their unaffected siblings. We performed EWAS meta-analysis across results from the two data sets, with adjustment for sex and surrogate variables that reflect major sources of biological variation and technical confounding such as cell type, batch, and ancestry. We compared top EWAS results to those from a previous brain-based analysis. We also tested for enrichment of ASD EWAS CpGs for being targets of meQTL associations using available SNP genotype data in the SEED sample.

Findings: In this meta-analysis of blood-based DNA from 796 cases and 858 controls, no single CpG met a Bonferroni discovery threshold of $p < 1.12 \times 10^{-7}$. Seven CpGs showed differences at $p < 1 \times 10^{-5}$ and 48 at $1 \times 10^{-4}$. Of the top 7, 5 showed brain-based ASD associations as well, often with larger effect sizes, and the top 48 overall showed modest concordance ($r = 0.31$) in direction of effect with cerebellum samples. Finally, we observed suggestive evidence for enrichment of CpG sites controlled by SNPs (meQTL targets) among the EWAS CpG hits, which was consistent across EWAS and meQTL discovery $p$ value thresholds.

(Continued on next page)
Conclusions: No single CpG site showed a large enough DNAm difference between cases and controls to achieve epigenome-wide significance in this sample size. However, our results suggest the potential to observe disease associations from blood-based samples. Among the seven sites achieving suggestive statistical significance, we observed consistent, and stronger, effects at the same sites among brain samples. Discovery-oriented EWAS for ASD using blood samples will likely need even larger samples and unified genetic data to further understand DNAm differences in ASD.

Keywords: DNA methylation, Epigenome, Autism spectrum disorders, Peripheral blood, Study to Explore Early Development, Simons Simplex Collection

Findings

The etiology of autism spectrum disorder (ASD) may involve epigenetic mechanisms. Indirect evidence supporting this hypothesis comes from the observation that children with Rett, Fragile X, and Angelman syndromes often show impaired communication and exhibit repetitive behaviors [1–3], two core domains affected in autism. All three of these syndromes are caused by epigenetic defects [4–9]. Additional evidence stems from genetic studies of rare variation in non-syndromic forms of ASD. Although these studies have primarily identified private variants associated with ASD, there is now strong evidence that the ASD-associated variants converge upon three biological pathways, one of which is chromatin remodeling [10–12]. Finally, there is direct evidence from case-control postmortem brain studies supporting epigenetic involvement in ASD. Several candidate gene-based studies have shown altered epigenetic states associated with autism [13–18]. Genome-scale screens have identified changes in DNA methylation (DNAm) at specific CpG sites [19, 20] as well as global changes in non-CpG methylation levels [21] in postmortem cerebral tissue from individuals with ASD relative to controls. Studies of cerebral cortex tissue has revealed genomic spreading of histone H3 lysine 4 methylation and histone H3 lysine 27 acetylation marks, away from the promoter region, among a subset of individuals with ASD compared to controls [22, 23].

Examination of the affected tissue, i.e., the brain, can provide important insights into potential mechanisms of disease etiology; however, there are considerable limitations with these types of studies. They suffer from severely small sample sizes, have historically had low genomic coverage, and often lack comprehensive unified clinical, demographic, and genomic data. Importantly, they are based on autopsy-derived tissue and do not reflect epigenetic marks in a living individual, are not at optimal developmental timing, and may be influenced by life experiences and cause of death. To overcome these barriers, complementary, large population-based autism epigenetic studies using accessible tissues, such as blood, from living individuals are needed. DNAm from peripheral tissues is useful to analyze for its potential to mimic DNAm signatures from the brain. But outside of brain concordance, it can also be useful for its ability reflect consequences of ASD, provided that sample collection is performed following ASD onset. Moreover, peripheral tissue DNAm signatures related to ASD may result from environmental exposures that simultaneously contribute to ASD. To date, three genome-scale epigenetic studies of autism in accessible peripheral tissues have been performed. One study of peripheral blood from 50 monozygotic twin pairs, including 6 pairs discordant for ASD at age 15, examined DNAm at over 27,000 CpG sites in promoter regions. The authors found suggestive evidence for epigenetic alterations associated with ASD and associated traits within families [24]. Similarly, an investigation of DNAm at CpG island regions in lymphoblastoid cell lines, obtained from 7 twin pairs including 3 discordant for ASD, found ASD-related DNAm changes at the RORA gene [25]. Both of these studies were limited by the small number of samples examined, lack of genome-scale coverage, and specific focus on twin pairs with a lack of extension to the general population. Ectoderm cell lineage-derived buccal cells, obtained from 47 ASD cases and 48 controls born to mothers aged 35 and older, have also shown suggestive epigenetic alterations associated with ASD [26]. While suggestive, it is unclear how these buccal-based epigenetic findings relate to a population sample and in a larger number of individuals. Thus, more research in accessible tissues from larger population-based, non-familial, samples is needed.

Here, we overcome previous limitations and perform the largest epigenome-scale examination of DNAm, to date, among two large US case-control studies of autism: the Study to Explore Early Development, phase I (SEED I) and the Simons Simplex Collection (SSC). Both measured DNAm at over 450,000 loci in childhood blood samples from either population-based cases and controls (SEED I) or discordant sibling pairs (SSC). Meta-analysis across both sets included 796 ASD cases and 858 controls. In addition to CpG-specific differential DNAm, we explored the set of blood-derived differentially methylated sites for their concordance in post-mortem brain tissue and their enrichment for genetically controlled CpG sites.
Methods

Study to Explore Early Development (SEED)

The Study to Explore Early Development is a multi-site case-control study with population-based ascertainment. In SEED phase 1, a total of 3899 families were recruited across 6 study sites (California, Colorado, Georgia, Maryland, North Carolina, and Pennsylvania) and classified into 3 groups according to a child's diagnosis: an autism spectrum disorder (ASD) group, a general population control group (POP), and a (non-ASD) developmental delay group. Details regarding participant recruitment, biospecimen collection, and final outcome classification have been previously described [27, 28]. Briefly, eligible children were born in one of the catchment areas between September 1, 2003, and August 31, 2006, which corresponded to being born in one of the catchment areas between September 1, 2003, and August 31, 2006, which corresponded to being born in one of the catchment areas [27]. This provided the SEED study. Informed consent was obtained on the ADI-R and ADOS, as described in detail elsewhere [34]. ASD was confirmed based on scores of the Autism Diagnostic Observation Schedule (ADOS) [31]. Other instruments provided additional measures of the core features of autism, as well as of intellectual ability (verbal and nonverbal), adaptive behavior, emotional and behavior problems, motor function, and language. A description of instruments employed can be found at https://sfari.org/ssc-instruments. A comprehensive family medical history was obtained that included the proband's prenatal and perinatal history, developmental milestones, immunizations, medications, dietary supplements, and common behavioral treatments. Emphasis was placed on common “comorbidities” including gastrointestinal complaints, sleep irregularities, and seizures. In addition, questions were asked about genetic, autoimmune, and psychiatric disorders in members of the extended family. Probands were excluded who were younger than 4 years of age or older than 18. Probands were also excluded for conditions that might compromise the validity of diagnostic instruments, such as nonverbal mental age below 18 months, severe neurological deficits, birth trauma, perinatal complications, or genetic evidence of fragile X or Down syndromes. A complete description of exclusion/inclusion criteria can be found at http://sfari.org. Measures of adaptive function, behavior-emotional problems, and symptoms of autism were examined in parents and siblings as well as probands.

Thus, the SSC represents a unique, well-described sample of able children and adolescents with relatively severe ASD, as indicated by ADI-R and ADOS Calibrated Severity Scores [38].

Reliability of data

To maximize the consistency of clinical observations across sites, each clinician was trained in administration of the ADOS and ADI-R to achieve research reliability as judged by expert clinicians. Most clinicians who had not previously received research training required 4–6 months of practice. Videotapes of interviews were exchanged to ensure that reliability requirements were met and maintained throughout the study. Error rates were very low, averaging less than 0.50 errors/1000 data points. Most errors could be corrected immediately, resulting in an unusually clean data set for a multisite study of this size. During each visit, a blood sample was collected from each study participant and DNA was extracted from blood cells, while plasma was stored for future use.

Simons Simplex Collection (SSC) A complete description of the SSC, which enrolled and collected biospecimens from children and adolescents aged 4–18 years, can be found elsewhere [37]. Briefly, a geneticist and a clinical psychologist were appointed as co-principal investigators at each site. Probands were evaluated with a battery of diagnostic measures, including the Autism Diagnostic Interview – Revised (ADI-R) [34] and the Autism Diagnostic Observation Schedule (ADOS) [31]. Other instruments provided additional measures of the core features of autism, as well as of intellectual ability (verbal and nonverbal), adaptive behavior, emotional and behavior problems, motor function, and language. A description of instruments employed can be found at https://sfari.org/ssc-instruments. A comprehensive family medical history was obtained that included the proband's prenatal and perinatal history, developmental milestones, immunizations, medications, dietary supplements, and common behavioral treatments. Emphasis was placed on common “comorbidities” including gastrointestinal complaints, sleep irregularities, and seizures. In addition, questions were asked about genetic, autoimmune, and psychiatric disorders in members of the extended family. Probands were excluded who were younger than 4 years of age or older than 18. Probands were also excluded for conditions that might compromise the validity of diagnostic instruments, such as nonverbal mental age below 18 months, severe neurological deficits, birth trauma, perinatal complications, or genetic evidence of fragile X or Down syndromes. A complete description of exclusion/inclusion criteria can be found at http://sfari.org. Measures of adaptive function, behavior-emotional problems, and symptoms of autism were examined in parents and siblings as well as probands.

Thus, the SSC represents a unique, well-described sample of able children and adolescents with relatively severe ASD, as indicated by ADI-R and ADOS Calibrated Severity Scores [38].

Reliability of data

To maximize the consistency of clinical observations across sites, each clinician was trained in administration of the ADOS and ADI-R to achieve research reliability as judged by expert clinicians. Most clinicians who had not previously received research training required 4–6 months of practice. Videotapes of interviews were exchanged to ensure that reliability requirements were met and maintained throughout the study. Error rates were very low, averaging less than 0.50 errors/1000 data points. Most errors could be corrected immediately, resulting in an unusually clean data set for a multisite study of this size. During each visit, a blood sample was collected from each study participant and DNA was extracted from blood cells, while plasma was stored for future use.
DNA methylation data quality control (QC) and processing

For the SEED samples, genomic DNA was isolated from 980 whole blood samples using the QIAasympohony midi kit (Qiagen). For each sample, 500 ng of DNA was bisulfite treated using the 96-well EZ DNA Methylation kit (Zymo Research). Samples were randomized within and across plates, and across two main processing dates to minimize batch effects, and run on the Illumina HumanMethylation450 BeadChip. Background correction and dye-bias equalization was performed using the function preprocessNoob() in the minfi R package [41]. We included 12 cross-plate duplicates for quality control purposes; pairwise correlation metrics for the duplicate samples ranged from 0.990 to 0.997 with a mean correlation value equal to 0.995. Samples were removed if they had low overall intensity (median unmethylated or methylated signal < 11) or had a detection p value > 0.01 in more than 1% of probes (N = 7), or if reported sex did not match predicted sex generated via the minfi function getSex() (N = 3). Probes were removed if they had a detection p value > 0.01 in more than 10% of samples (n = 702) and then if they had been previously identified as being ambiguously mapped [42] (n = 29,146). Following QC, the analytic data included DNAm for 455,664 sites on 970 samples. We further removed two samples who were missing a final outcome classification, leaving a total of 453 cases and 515 controls used for association analyses.

For the SSC samples, 500 ng of human genomic DNA was sodium bisulfite-treated for cytosine to thymine conversion using the EZ DNA Methylation Gold kit (Zymo Research). A total of 728 samples (from 364 families) were randomized within and across plates to minimize batch effects and run on the Illumina HumanMethylation450 BeadChip. Additional details have been previously described [43]. Similar quality control procedures as used for the SEED samples were used for the SSC samples. After background correction and dye-bias equalization, samples were removed for low overall intensity (median unmethylated or methylated signal < 11) or for detection p value > 0.01 in more than 1% of probes (N = 42). Probes were removed if they had a detection p value > 0.01 in more than 10% of samples (n = 483) and then if they had been previously identified as being ambiguously mapped (n = 29,213). These steps resulted in an analytic data set with 455,816 sites on 686 samples, consisting of 343 proband-sibling pairs.

Finally, for all SEED and SSC samples, we estimated cell type proportions for six different cell types (granulocytes, monocytes, CD4 T cells, CD8 T cells, B cells, and natural killer cells) using the estimateCellCounts() function in the minfi R package. Estimation incorporated reference data from 60 samples generated from 6 healthy adult men [44].

Genotype data quality control and processing

Whole genome genotyping data was available for 943 of the 970 SEED 1 samples which passed DNAm quality control steps. After genotype measurement using the Illumina HumanOmni1-Quad BeadChIp, standard quality control measures were applied, including removal of samples with < 95% SNP call rate, sex discrepancies, relatedness (Pi-hat > 0.2), or excess hetero- or homozygosity, and removal of markers with < 98.5% call rate, or that were monomorphic. Phasing was performed using SHAPEIT [45] followed by SNP imputation via the IMPUTE2 software [46], using 1000 Genomes Project samples as reference. Principal components to account for ancestry were determined via the EigenStrat program [47].

Epigenome-wide association testing and meta-analysis

For the SEED data, we used linear regression modeling of the M value (the ratio of methylated to total signal determined at every probe in every sample) [48] as a dependent variable and ASD status, sex, and surrogate variables (SVs) (described below) as independent variables. We implemented this model using the lmFit() function in the limma R package [49], separately for each of the DNAm probes that passed QC. For the SSC data, we implemented a generalized estimating equation (GEE) model using the gee() function in the gee R package [50] to account for the correlation inherent to the familial structure in the data. We used a fixed correlation structure of 0.5 for each sibling pair, and regressed M value onto ASD status, sex, and SVs.

To account for sources of technical and biological variability in our association analyses, we estimated surrogate variables (SVs) [51] in the cleaned SEED and SSC dataset to include as covariates in our downstream analyses. SVs have been shown to capture and adjust for differences related to batch effects and cell type proportions across samples in a wide variety of simulated settings [52], and to remove the effects of unwanted sources of technical and biological variation [51]. In order to explicitly address the strong confounding effect of sex resulting from the high degree of male bias in ASD diagnosis, we removed sex chromosomes, where DNAm values strongly correlate to sex, before SV estimation, and included sex along with ASD status in the model used for SV estimation. We then used a data-driven procedure individually in the SEED and SSC data to select the number of SVs to include in the association models.

First, to examine the relationship between each SV and known sources of technical and biological variation, we estimated the association between each estimated SV and cell type composition, principal components of genetic ancestry, and processing batch. We then generated a visual
representation of the degree of association with these variables using a heat map (Additional file 1: Figures S1a, S2a).

We next examined the influence of iterative inclusion of SVs as adjustment variables in our association regression models. To do this, we first ran a case-control association model with adjustment for the strongest estimated SV [51], then progressively included the next strongest SV in the analysis and continued this procedure until all estimated SVs were included. For each model, we recorded the inflation factor, or lambda, calculated via the estlambda() function from the GenABEL R package [53], and visualized the relationship between number of SVs adjusted for and lambda values (Additional file 1: Figures S1b, S2b). We chose the number of SVs to include in the model by considering both the number of SVs at which the estimated lambda values began to plateau and where the known potential confounders appeared to be captured by one or more SVs. We chose to include 19 SVs in the SEED association analysis and 14 in the SSC analysis.

After completing each association analysis, we then performed a meta-analysis using the METAL software [54] on the 445,068 probes that were present in both the SEED and SSC cleaned datasets. Our approach weighted individual study effects by sample size and also took into account the direction of effect. We also computed the false discovery rate (FDR) using the Benjamini-Hochberg method [55]. We also determined statistical power for this meta-analysis a priori using an estimation method specifically designed for epigenome-wide association studies [56].

**Comparison of blood epigenome-wide association studies (EWAS) hits to brain-based DNA methylation**

We sought to compare the consistency of top EWAS results from the blood-based meta-analysis to our previous analysis of post-mortem brain samples from ASD cases and controls [19]. These data consist of DNA methylation quantitative trait loci (meQTLs) from three brain regions: cerebellum, prefrontal cortex, and temporal cortex. For the CpG sites reaching suggestive levels of significance ($p < 1 \times 10^{-4}$) in the meta-analysis, we computed mean differences between cases and controls in each of these three brain regions. We then computed Pearson correlations and quadrant count ratios between the blood effect sizes and three lists of brain effect sizes. We computed quadrant count ratios as the sum of concordant effect sizes (both positive or both negative) minus the sum of discordant effect sizes, all divided by the total number of effect sizes being compared.

**Methylation quantitative trait loci (meQTL) query and meQTL target enrichment test**

We were interested in exploring the propensity of CpG sites that reached a level of suggestive significance in the EWAS meta-analysis to be significantly associated with nearby SNPs. We used joint DNA methylation and genotype data to define SNPs associated with DNA methylation, sometimes referred to as “methylation quantitative trait loci (meQTLs)”, and the CpG sites under genetic control, or “meQTL targets”. We then tested for enrichment of meQTL targets in the top ranked CpG sites from the meta-analysis.

In lieu of applying our SV selection method (see “Epigenome-wide association testing and meta-analysis”) to every SNP-CpG association test in the meQTL query, we conducted separate meQTL queries in each processing batch of the SEED data ($N_{Batch1} = 606; N_{Batch2} = 362$). In each batch, we first used a data-driven procedure we have described in detail previously [57] to select three key parameters for the meQTL query: the SNP minor allele frequency threshold for inclusion, the CpG variability threshold for inclusion, and the maximum distance between SNP and CpG site to be considered for analysis. Briefly, this procedure selects parameters to ensure 80% power to detect a 5% DNA methylation quantitative trait loci difference with each addition of the minor allele, at a Bonferroni-defined significance threshold. We then performed the meQTL query in each batch using the MatrixEQTL R package, adjusting for sex, the first five principal components to account for genetic ancestry, and the first two principal components derived from the cell composition estimates. We then defined SNP-CpG association pairs based on results that can gain 100% power in the parameter survey (“permissive”), 90% power (“intermediate”), and 80% power (“stringent”). If a SNP-CpG association pair was significant at a designated threshold in each batch, the CpG site was labeled a meQTL target under that threshold for the downstream enrichment analysis.

We tested for enrichment of meQTL targets in ASD-associated CpG sites. We examined this using two ASD EWAS meta-analysis $p$ value thresholds ($p < 1 \times 10^{-3}$ and $p < 1 \times 10^{-4}$) and the three meQTL $p$ value thresholds. In each enrichment test, we accounted for the two main features of CpG sites likely to affect results: the degree of variability in DNA methylation quantitative trait loci and the number of SNPs in the boundary considered. To do this, we binned each CpG site by decile according to these factors. For each EWAS/meQTL threshold scenario, we compared the proportion of meQTL targets among ASD-related CpGs to a null distribution of randomly selected CpGs, equal in count to the number of ASD-associated sites, matched on the same variability and nearby SNP decile. We defined a fold enrichment statistic as the count of meQTL targets in the ASD-associated CpGs divided by the mean proportion of meQTL targets from the null set, and an enrichment $p$ value as the number of null CpG sets with a count of meQTL targets that was equal to or exceeded the count in the ASD-associated CpG list.
Results

ASD EWAS meta-analysis in blood

We performed a meta-analysis over the 445,068 probes that were present in both the SEED (Table 1) and SSC (Additional file 2: Table S1) cleaned datasets. Figure 1a shows the range of p values and effect sizes detected in our meta-analysis. No CpG sites reached a Bonferroni level of significance, and effect sizes were modest (1.12 × 10⁻⁷; Fig. 1a). The genomic inflation factor (λ) was 1.03, with a slight separation from expectation at the tail (Fig. 1b). A total of 48 CpG sites met or exceeded a p value < 1 × 10⁻⁴, and 7 CpG sites (Table 2) reached a significance level of p < 1 × 10⁻⁵. We have provided a full list of summary statistics for both SEED and SSC for all 445,068 probes (Additional file 3: Table S2). Based on our analytic sample size, we had 80% power to detect a 3.8% DNAm difference between cases and controls at a Bonferroni level of significance.

Consistency of blood EWAS hits in brain

We considered the consistency of signal for the 48 blood-based CpGs with suggestive significance (p value < 1 × 10⁻⁴), among results from three different brain regions with data available from our previous analysis of post-mortem brain samples and ASD [19]. The cerebellum exhibited a moderate degree of concordance in effect size and direction (r = 0.31; QCR = 0.33), although prefrontal cortex (r = 0.02; QCR = 0.125) and temporal cortex (r = -0.10; QCR = -0.125) showed only minimal concordance (Additional file 4: Table S3, Additional file 5: Figure S3). When considering the seven CpG sites with more stringent blood-based p values < 1 × 10⁻⁵, the direction of effect was consistent for at least five of these seven in all three brain region results, with typically larger effect sizes (Table 3). The CpG site with the largest effect size in blood (cg09671955) displayed consistent, and larger, effect sizes in all three brain regions (Table 2, Additional file 4: Table S3).

meQTL target enrichment test

When considering all CpGs associated with ASD at a liberal p < 1 × 10⁻³ EWAS threshold, we found meQTL target enrichment at nominal significance (p<sub>enrichment</sub> = 0.041), but not at a significance threshold corrected for the six tests performed overall (Table 4). All other combinations of EWAS and meQTL p values tested displayed similar suggestive levels of significance (0.089 ≤ p<sub>enrichment</sub> ≤ 0.243) and a consistent direction of effect towards enrichment. Also, tests conducted for CpGs meeting the more stringent EWAS p value threshold (1 × 10⁻⁶) displayed a consistently greater effect size than their corresponding tests from the more liberal EWAS threshold.

Discussion

We report results from a large study investigating the relationship between ASD and DNAms. A case-control meta-analysis of peripheral blood samples from the Study to Explore Early Development and the Simons Simplex Collection revealed that none of the 455,068 CpG sites investigated were associated with ASD at a genome-wide significant threshold. However, 48 CpG sites reached suggestive significance levels at p < 1 × 10⁻⁴, including 7 CpGs at p < 1 × 10⁻⁵. Associations with ASD at these sites display moderate concordance with post-mortem brain sample results from a previous study and display suggestive evidence for enrichment of SNP-controlled CpG sites, or meQTL targets.

Blood-based DNAm signatures of ASD may reflect causal mechanisms that mimic signatures seen in the brain. However, case-control sampling, by design occurring after disease onset, can also reflect consequences of ASD or of environmental exposures. Nonetheless, given the potential involvement of epigenetic mechanisms in ASD [4–12], and the availability of blood samples from existing studies, this was an important project to pursue. In contrast to our null findings in blood samples, previous work using brain samples has shown specific DNAms to be associated with ASD [15, 19, 20]. The difference in tissue type—specifically that brain-based signatures better reflect ASD etiology—might explain the inconsistent results across these studies. Nonetheless, a previous study of ASD and DNAms in peripheral blood [24] using the earlier 27K Illumina array reported numerous differentially methylated sites. Our study,

Table 1: Demographic characteristics for SEED samples

|               | ASD (N = 453) | Control (N = 513) | p value |
|---------------|---------------|------------------|---------|
| Sex           |               |                  |         |
| M             | 368 (0.81)    | 273 (0.53)       | < 2.2E−16|
| F             | 85 (0.19)     | 242 (0.47)       |         |
| Processing batch |           |                  |         |
| 1             | 289 (0.64)    | 317 (0.62)       | 0.5136  |
| 2             | 164 (0.36)    | 198 (0.28)       |         |
| Cell composition |          |                  |         |
| Granulocytes  | 0.46 [0.38–0.52] | 0.47 [0.40–0.54] | 0.02    |
| CD4T          | 0.21 [0.16–0.26] | 0.20 [0.16–0.24] | 0.09    |
| B cell        | 0.12 [0.08–0.13] | 0.11 [0.08–0.13] | 0.01    |
| CD8T          | 0.11 [0.08–0.15] | 0.11 [0.08–0.13] | 0.58    |
| Monocytes     | 0.07 [0.05–0.09] | 0.07 [0.06–0.09] | 0.63    |
| NK            | 0.02 [0.00–0.03] | 0.02 [0.00–0.03] | 0.83    |
| Genetic ancestry |          |                  |         |
| European      | 214 (0.47)    | 324 (0.63)       | 1.02E−06|
| African       | 71 (0.16)     | 39 (0.08)        |         |
| Admixed or Asian |       |                  |         |
|                | 153 (0.34)    | 142 (0.28)       |         |
| Missing       | 15 (0.03)     | 10 (0.02)        |         |
despite a much higher sample size, did not observe associations at these same sites. However, the previous study’s reported differentially methylated sites were based on a ranking that integrated both degree of statistical significance and effect size; none of their single CpG associations achieved statistical significance at a Bonferroni correction level. Also, the previous study did not explicitly account for potential confounding by cell type proportions or address potential batch effects in the analytic pipeline, as this was not yet commonplace in EWAS pipelines at the time of their report. We used a rigorous, data-driven method to account for these factors and control the genome-wide type I error rate. It is also possible that true differentially methylated positions for ASD exist in blood but have smaller effects sizes than we were powered to detect. Recent EWAS discoveries have indeed found replicable very small effect sizes, particularly in environmental health [58].

Ideally, tissue-relevant samples from earliest in utero development would provide the greatest mechanistic insights. However, this is not typically available and our study is based instead on a case-control design using blood-based samples. Recent examples of blood-based epigenetic work in brain-based disorders have shown utility, particularly when focused on integration of genetic and epigenetic variation or when inferring biological pathways rather than discovery of specific CpGs [57, 59–63]. For example, DNAm under genetic control (meQTL targets) can inform genetic associations observed for ASD. Further, consideration of aggregated sets of CpGs associated with ASD, rather than single sites, can elucidate pathways of interest [57].

Table 2 CpG sites identified from meta-analysis as being suggestively associated ($1 \times 10^{-5}$) with ASD

| Probe ID     | CHR | Position | SEED p  | SEED mean diff | SSC p  | SSC mean diff | METAL p | FDR  | Nearest gene | Location |
|--------------|-----|----------|---------|----------------|--------|---------------|---------|------|--------------|----------|
| cg21151899   | 22  | 42,337,657| 1.74E-06| 0.78           | 2.75E-02| 0.54          | 3.82E-07| 0.16 | CENPM        | Intrinsic |
| cg03731974   | 16  | 86,531,598| 2.59E-04| 0.09           | 8.03E-04| 0.08          | 7.29E-07| 0.16 | FENDRR       | Intrinsic |
| cg09962502   | 2   | 96,971,189| 1.89E-04| 0.05           | 2.53E-03| 0.08          | 1.58E-06| 0.21 | SNRNP200     | Exonic   |
| cg1798266    | 1   | 1.53E+08  | 2.06E-05| 0.43           | 1.93E-02| 0.13          | 1.90E-06| 0.21 | PGLYRP4      | Exonic   |
| cg1716316    | 17  | 40,897,182| 1.28E-03| 0.09           | 1.09E-03| 0.13          | 4.94E-06| 0.44 | EZH1         | 111 bp   |
| cg16234726   | 14  | 1.02E+08  | 3.41E-04| 0.2            | 8.42E-03| 0.37          | 9.13E-06| 0.5  | DIO3         | 69 Kb    |
| cg09671955   | 1   | 1.69E+08  | 7.20E-03| 0.47           | 2.25E-04| 1.25          | 9.33E-06| 0.5  | CCDC181      | 65 bp    |

*Mean difference values computed as mean in cases – mean in controls

*Location within gene or distance to nearest gene

Fig. 1 Meta-analysis results of epigenome-wide association analysis for ASD in peripheral blood in SEED in SSC samples. a Volcano plot depicting meta-analysis $p$ value (log base 10 scale) on the y-axis and average of mean difference values in SEED and SSC samples weighted by sample size on the x-axis. b Quantile-quantile plot ($\lambda = 1.03$)
our own analysis, the most differential blood-based CpGs had consistent effect sizes and directions, although weaker, with brain-based results, particularly for cerebellum. Blood-based CpGs were also moderately enriched (though not to a statistically significant extent) for meQTL targets. These results suggest that blood DNAm can be reflective of DNAm in affected tissues and suggest genetic control of DNAm as a mechanism for this occurrence, at least in an ASD context. More precise evidence is needed, but given the easy accessibility of blood for DNAm measurements versus brain [61], the utility of blood-based DNAm research in ASD is worthy of additional consideration.

In summary, our results suggest a potential role of genetic factors in contributing to DNAm differences in ASD. This is consistent with a recently reported meta-analysis combining these results with an additional European case control sample. In that report, while ASD itself was not associated with DNAm at a genome-wide significance threshold, polygenic risk scores for ASD were associated with DNAm at particular CpGs [64]. In contrast, we show enrichment for meQTL targets (DNAm at least partially controlled by SNP variation) among CpGs with at least marginal association with ASD. We further characterize these ASD DNAm associations and compare to brain-based results. Together, our work and the report by Hannon et al. suggest more investigation of the relationship between genetic risk, epigenetic marks, and ASD is warranted in larger samples and across multiple tissues and designs. To this end, we have provided our full summary statistics and meta-analysis results. The need for greater sample sizes mimics the initial stages of genetic variation discovery in ASD, for which large mega-analyses are starting to pay dividends [65].

### Additional files

**Additional file 1** Figures S1-S2. Depiction of surrogate variable selection process for SEED (S1) and SSC (S2). Panel A: Heatmap indicating degree of association with known potential technical variables or confounders with estimated surrogate variables. Panel B: Inflation factor (lambda) calculated for progressively including surrogate variables in association models. The number of surrogate variables to include in the ultimate association testing model was to determine to be that which properly controlled the inflation factor and adequately captured known technical variables or confounders. See “Methods” for additional explanation. (PDF 19 kb)

**Additional file 2** Table S1. Demographic characteristics for samples in the SSC (S2) dataset. (XLSX 11 kb)

**Additional file 3** Table S2. Full summary statistics and meta-analysis results for all 445,608 CpG sites that were present in both the cleaned SEED and SSC datasets. (CSV 40707 kb)

**Additional file 4** Table S3. Concordance between suggestively associated (p < 1 × 10^{-5}) CpG sites in peripheral blood and their corresponding effect sizes in three brain regions. (CSV 3 kb)

**Additional file 5** Figures S3. Quadrant plots depicting concordance in effect sizes between suggestively associated (p < 1 × 10^{-5}) CpG sites in peripheral blood and three brain regions. A) Prefrontal cortex B) Temporal Cortex C) Cerebellum. Points in red indicate those sites with p < 1 × 10^{-5} in peripheral blood. (PNG 21 kb)
Abbreviations
ASD: Autism spectrum disorder; DNAm: DNA methylation; EWAS: Epigenome-wide association study; meQTL: Methylation quantitative trait loci; SEED: Study to Explore Early Development; SNP: Single nucleotide polymorphism; SSC: Simons Simplex Collection; SV: Surrogate variable

Funding
This project was supported by Centers for Disease Control and Prevention (CDC) Cooperative Agreements announced under the following RFAs: 01086, 02199, DD11-002, DD06-003, DD04-001, and DD09-002. The findings and conclusions in this report are those of the authors and do not necessarily represent the official position of the Centers for Disease Control and Prevention. The DNA methylation assays were supported by Autism Speaks Award #56396 awarded to D. G. The methylation data and preprocessing for the SEED and SSC methylation data were performed by NIAMS R01ES019001; R01ES017646. S. Andrews was supported by the Burroughs-Wellcome Trust training grant; Maryland, Genetics, Epidemiology and Medicine (MD-GEM). The SSC was supported by Simons Foundation (SFARI) award and NIH grant MH089606, both awarded to S.T. Warren.

Availability of data and materials
The SEED I data generated and analyzed for this study are not publicly available due to lack of explicit consent for such sharing in the written informed consents for SEED sites, according to the CDC IRB that governs the SEED network. The SEED data generated and analyzed for this study are not publicly available due to confidentiality and informed consent requirements. We have submitted the raw 450k methylation data from the SSC samples to the National Database for Autism Research (NDAR; ndarr.nih.gov), which can be found under collection #300.

Authors’ contributions
MDF and CL-A conceived the study, GCW, LAS, DES, LAC, MDF, and CIN led the participation and sample recruitment for SEED. STW, RSA, and PC generated the methylation data and performed initial analysis on all the SSC samples. CL-A and APF supervised methylation data collection. SVA performed quality control and preprocessing for the SEED and SSC methylation data. BS and CL-A performed all analyses, supervised by MDF, and CL-A. MDF, and APF obtained funding for the DNA methylation measurements in SEED. SVA, MDF, and CL-A wrote the manuscript. All authors contributed to interpretation of results and edited and reviewed the manuscript. All authors read and approved the final manuscript.

Ethics approval and consent to participate
This study was approved by the institutional review boards at each SEED site: SEED 1 recruitment was approved by the IRBs of each recruitment site: Institutional Review Board (IRB)-C, CDC Human Research Protection Office; Kaiser Foundation Research Institute (KFR) Kaiser Permanente Northern California IRB; Colorado Multiple IRB, Emory University IRB, Georgia Department of Public Health IRB, Maryland Department of Health and Mental Hygiene IRB, Johns Hopkins Bloomberg School of Public Health Review Board, University of North Carolina IRB and Office of Human Research Ethics, IRB of The Children’s Hospital of Philadelphia, and IRB of the University of Pennsylvania. All enrolled families provided written consent for participation. This methylation substudy was approved as an amendment of the Johns Hopkins Institutional Review Board (IRB) approval. For participants from the Simons simplex collection, parents consented and children assented as required by each local institutional review board, which included a coalition of clinics located at Michigan, Yale, Emory, Columbia, Vanderbilt, McGill Washington, and Harvard Universities (Children’s Hospital of Boston), and at the Universities of Washington, Illinois (Chicago), Missouri, UCLA, and the Baylor College of Medicine. To protect the privacy of participants, Global Unique Identifiers (GUIDs) were constructed from personal information using an algorithm devised in collaboration with scientists at the NIH [68]. Each clinic retained personal identifiers on site and transmitted de-identified GUIDs to a central database, as described previously [37].

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Publisher’s Note
Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Author details
1Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, 615 N. Wolfe Street, Baltimore, MD 21205, USA. 2Wendy Klag Center for Autism and Developmental Disabilities, Johns Hopkins Bloomberg School of Public Health, 615 N. Wolfe Street, W6509, Baltimore, MD 21205, USA. 3California Department of Public Health, Environmental Health Investigations Branch, 850 Marina Bay Parkway, Richmond, CA 94804, USA. 4National Center on Birth Defects and Developmental Disabilities, Centers for Disease Control and Prevention, MS E-86, 1600 Clifton Road, Atlanta, GA 30333, USA. 5Department of Public Health, Section of Epidemiology, Aarhus University, Aarhus, Denmark. 6Department of Economics and Business, National Centre for Register-based Research, Aarhus University, Aarhus, Denmark. 7Lundbeck Foundation Initiative for Integrative Psychiatric Research, IPSYCH, Aarhus, Denmark. 8Kaiser Permanente Division of Research, 2000 Broadway, Oakland, CA 94612, USA. 9Department of Human Genetics, Emory University School of Medicine, 615 Michael Street, Atlanta, GA 30322, USA. 10Department of Psychiatry, University of Wisconsin-Madison, 6001 Research Park Blvd, Madison, WI 53719, USA. 11Department of Epidemiology and Biostatistics, Drexel University School of Public Health, 3215 Market Street, Philadelphia, PA 19104, USA. 12J. Drexel Autism Institute, 3200 Market Street Suite 560, Philadelphia, PA 19104, USA. 13Department of Biochemistry, Emory University School of Medicine, 615 Michael Street, Atlanta, GA 30322, USA. 14Department of Pediatrics, Emory University School of Medicine, 615 Michael Street, Atlanta, GA 30322, USA. 15Center for Epigenetics, Johns Hopkins School of Medicine, 855 N. Wolfe Street, Baltimore, MD 21205, USA. 16Department of Medicine, Johns Hopkins School of Medicine, 855 N. Wolfe Street, Baltimore, MD 21205, USA. 17Department of Mental Health, Johns Hopkins Bloomberg School of Public Health, 624 N. Broadway, HH850, Baltimore, MD 21205, USA.

Received: 8 March 2018 Accepted: 21 June 2018

Published online: 28 June 2018

References
1. Jedele KB. The overlapping spectrum of Rett and Angelman syndromes: a clinical review. Semin Pediatr Neurol. 2007;14:108–17.
2. Mount RV, Charman T, Hastings RP, Reilly S, Cass H. Features of autism in Rett syndrome and severe mental retardation. J Autism Dev Disord. 2003;33:435–42.
3. Budimovic DB, Kaufmann WE. What can we learn about autism from studying fragile X syndrome? Dev Neurosci. 2011;33:379–94.
4. Amir RE, Van den Veyver IB, Wan M, Tran QC, Francke U, Zoghby HY. Rett syndrome is caused by mutations in X-linked MECP2, encoding methyl-CpG-binding protein 2. Nat Genet. 1999;23:185–8.
5. Wan M, Lee SS, Zhang X, Houwink-Manville I, Song HR, Amir RE, et al. Rett syndrome and beyond: recurrent spontaneous and familial MECP2 mutations at CpG hotspots. Am J Hum Genet. 1999;65:1520–9.
6. Sutcliffe JS, Nakao M, Christian S, Orstavik KH, Tommerup N, Ledbetter DH, et al. Deletions of a differentially methylated CpG island at the SNRPN gene define a putative imprinting control region. Nat Genet. 1994;8:52–8.
7. Kishino T, Lalande M, Wagstaff J. UBE3A/6-AP mutations cause Angelman syndrome. Nat Genet. 1997;15:70–3.
8. Bell MV, Hirst MC, Nakahori Y, MacKinnon RN, Roche A, Flint TJ, et al. Physical mapping across the fragile X hypermethylated and clinical expression of the fragile X syndrome. Cell. 1991;64:861–6.
9. Vincent A, Heitz D, Petit C, Kretz C, Oberlé I, Mandel JL. Abnormal pattern detected in fragile-X patients by pulsed-field gel electrophoresis. Nature. 1991;349:624–5.
10. Sanders SJ, He X, Willissey AJ, Ercan-Sencicek AG, Samocha KE, Cicek AE, et al. Insights into autism spectrum disorder genomic architecture and biology from 71 risk loci. Neuron. 2015;87:1215–33.
11. Pinto D, Delabry E, Menico D, Barbosa M, Merikangas A, Klei L, et al. Convergence of genes and cellular pathways dysregulated in autism spectrum disorders. Am J Hum Genet. 2014;94:671–84.
12. Krumm N, O’Roak BJ, Shendure J, Eichler EE. A de novo convergence of autism genetics and molecular neuroscience. Trends Neurosci. 2014;37:95–105.
specific epigenetic variation across brain and blood. Genome Biol. 2012;13:R43.

61. Bakulski KM, Halladay A, Hu WW, Mill J, Fallin MD. Epigenetic research in neuropsychiatric disorders: the “tissue issue”. Curr Behav Neurosci Rep. 2016;3:264–74.

62. Montano C, Taub MA, Jaffe A, Briem E, Feinberg JJ, Trygvadottir R, et al. Association of DNA methylation differences with schizophrenia in an epigenome-wide association study. JAMA Psychiatry. 2016;73:506–14.

63. 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. Transl Psychiatry. 2017;7:e1187.

64. Hannon E, Schendel D, Ladd-Acosta C, Grove J, IPSYCH-Broad ASD Group, Hansen CS, et al. Elevated polygenic burden for autism is associated with differential DNA methylation at birth. Genome Med. 2018;10:19.

65. Autism Spectrum Disorders Working Group of The Psychiatric Genomics Consortium. Meta-analysis of GWAS of over 16,000 individuals with autism spectrum disorder highlights a novel locus at 10q24.32 and a significant overlap with schizophrenia. Mol Autism. 2017;8:21.

66. Johnson SB, Whitney G, McAluliffe M, Wang H, McCriddy E, Rozenblit L, et al. Using global unique identifiers to link autism collections. J Am Med Inform Assoc. 2010;17:689–95.