A Functional Genomics Pipeline to Identify High-Value CpG Dinucleotides Related to Allergy or Asthma in the Human Methylome

Andréanne Morin[1,17], Emma E. Thompson[1,17], Britney A. Helling[1,17], Lyndsey E. Shorey-Kendrick[2], Pieter Faber[3], Tebeb Gebretsadik[4], Leonard B. Bacharier[5], Meyer Kattan[6], George T. O'Connor[7], Katherine Rivera-Spoljaric[8], Robert A. Wood[9], Kathleen C. Barnes[10], Rasika A. Mathias[11], Matthew C. Altman[12], Kasper Hansen[13], Cindy T. McEvoy[14], Eliot R. Spindel[2], Tina Hartert[4], Daniel J. Jackson[15], James E. Gern[15], Chris G. McKennan[16,18,*], and Carole Ober[1,18,*] on behalf of program collaborators for Environmental Influences on Child Health Outcomes and Children's Respiratory and Environmental Workgroup

[1] Department of Human Genetics, University of Chicago, Chicago, IL 60637

[2] Division of Neuroscience, Oregon National Primate Research Center, Oregon Health and Science University, Beaverton, OR 97006

[3] Genomics Core, University of Chicago, Chicago, IL 60637

[4] Department of Medicine, Vanderbilt University School of Medicine, Nashville, TN 37232

[5] Department of Pediatrics, Monroe Carell Jr Children's Hospital at Vanderbilt University Medical Center, Nashville, TN 37203

[6] Department of Pediatrics, Columbia University Medical Center, New York, NY 10032

[7] Pulmonary Center, Boston University School of Medicine, Boston, MA 02118

[8] Department of Pediatrics, Washington University, St. Louis, MO 63110

[9] Department of Pediatrics, Johns Hopkins University, Baltimore, MD 21287

[10] Department of Medicine, University of Colorado Denver, Aurora, CO 80045

[11] Department of Medicine, Johns Hopkins University, Baltimore, MD 21224

[12] Systems Immunology Division, Benaroya Research Institute Systems, Seattle, WA 98101; Department of Medicine, University of Washington, Seattle, WA 98195

[13] Department of Biostatistics, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, MD 21205

[14] Department of Pediatrics, Oregon Health and Science University, Portland, OR 97239

[15] Department of Pediatrics, University of Wisconsin School of Medicine and Public Health, Madison, WI 53792-9988

[16] Department of Statistics, University of Pittsburgh, Pittsburgh, PA 15260

[17] A.M., E.E.T., and B.A.H. contributed equally to this work.

[18] C.M.K. and C.O. jointly supervised this work.

*Corresponding authors:

Carole Ober, PhD
Department of Human Genetics, University of Chicago
920 E. 58th St, Chicago, IL 60637
Email: c-ober@genetics.uchicago.edu

Christopher Gordon McKennan, PhD
Department of Statistics, University of Pittsburgh
1826 Wesley W. Posvar Hall
230 S Bouquet St, Pittsburgh, PA 15260
Email: CHM195@pitt.edu

Abstract

DNA methylation of cytosines at CpG dinucleotides is a widespread epigenetic mark that is responsive to environmental exposures and associated with a wide range of human pathologies. However, compared to genetic variation, genome-wide epigenetic variation has been relatively unexplored due to the limited representation of variable CpGs on commercial high-throughput platforms. To address this gap, we developed a pipeline to identify high-value asthma- and allergy-associated CpGs among the >28 million CpG sites in the human genome. We focused on epithelial cells – sentinels in the airway that respond to inhaled microbes, pollution and allergens and mediate their downstream effects on asthma and allergic disease risk. We combined whole-genome bisulfite sequencing in these cells from children with and without allergic asthma and *in silico* evidence of functionality to identify high-value CpGs for a custom DNA methylation array. Compared to commercial arrays, the Asthma&Allergy Custom array was enriched for CpGs with

intermediate methylation levels, which are more likely to be correlated with the expression of their nearest or target genes and associated with allergic sensitization in children of diverse ancestries compared to CpGs on the commercial array. Our study revealed signature features of functional CpGs and unveiled a wealth of DNA methylation variation at functional CpG sites that are missed by existing high-throughput arrays, indicating that inter-individual variation in DNA methylation may contribute more to disease risk than previously appreciated.

The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Author Contributions: A.M., E.E.T., B.A.H. performed all data analyses; A.M., E.E.T., B.A.H., C.G.M., and C.O. designed the study and wrote the manuscript; P.F. oversaw the laboratory assays; L.E.S-K., T.G., L.B.B., M.K., G.T.O., K.R.S., R.A.W., C.T.M., E.R.S., T.H., D.J.J., and J.E.G. provided clinical samples and information and consulted on study design; M.C.A. provided gene expression data; K.C.B., R.A.M., and K.H. consulted on study design. All authors read, commented on and approved the manuscript.

Competing Interest Statement: L.B.B. reports personal fees from GlaxoSmithKline Genentech/Novartis, Merck, DBV Technologies, Teva, Boehringer Ingelheim, AstraZeneca, WebMD/Medscape, Sanofi/Regeneron, Vectura, Circassia, Elsevier, Kinaset, and Vertex outside the submitted work. G.T.O. has received consulting fees from Dicerna and Menarini unrelated to this work. R.A.W. has received grants from Aimmune, Astellas, DBV, Genentech, Novartis, Regeneron, and Sanofi. K.C.B. is employed by Tempus Labs. M.C.A. has received consulting fees from Regeneron outside the submitted work. D.J.J. reports grants and personal fees from GlaxoSmithKline and Regeneron and personal fees from Pfizer, Sanofi, AstraZeneca, and Vifor Pharma, outside the submitted work. J.E.G. is a paid consultant for AstraZeneca and Meissa Vaccines Inc. and has stock options in Meissa Vaccines Inc.

Classification: Biological Sciences, Genetics

Keywords: Allergic asthma, epigenetics, transcriptomics.

Significance Statement

DNA methylation at CpG dinucleotides is a widespread epigenetic mark associated with both exposures and diseases, and with the expression of genes, in specific cell types. Yet, nearly all previous studies have been limited to the CpGs on commercial arrays, which represent <5% of the human methylome. We performed whole-genome bisulfite sequencing in airway epithelial cells from children with and without allergic asthma and used functional annotations of gene regulatory elements to design a custom Asthma&Allergy DNA methylation array. We identified signature features of high-value, functional CpGs that are associated with allergic sensitization and robust to ancestry. Our findings suggest that a large portion of the "functional" human methylome has been overlooked in previous epigenome-wide association studies.

Main Text

Introduction

Epigenetics refers to modifications of DNA molecules that do not alter the DNA sequence but play important roles in regulating gene expression. Environmental exposures can directly modify epigenetic marks in the human genome (1), and epigenetic responses can mediate the effects of exposures on gene expression (2) and disease risk (3). Thus, the epigenome may contribute directly to disease risk or be sites of gene-environment interactions, providing both complementary data and mechanistic insights, respectively, to genome-wide association studies (GWAS). The most common epigenetic mark in the human genome is methylated cytosines at CpG dinucleotides and the availability of high-throughput array-based platforms to measure DNA methylation has led to an explosion of epigenome-wide association studies (EWAS) (4). The most used commercial array, the InfinityMethylationEPIC BeadChip (Illumina, Inc., San Diego, CA), interrogates up to 850,000 CpGs, but this comprises <5% of CpGs in the human genome. Moreover, the selection of CpGs on the EPIC array was agnostic with respect to disease or tissue

types. As a result, findings to date may represent just the tip of the iceberg with an abundance of yet unknown functional CpG sites contributing to disease risk.

The goal of our study was to develop and validate a custom DNA methylation array enriched for CpGs in airway epithelial cells that are likely to be functional and associated with asthma and allergic diseases in populations of diverse ancestries. We focused on airway epithelial cells because of their critical role in responding to inhaled microbes, pollution and allergens and mediating their downstream effects on asthma and allergic disease risk (5, 6). An overview of our study design is shown in **Figure 1**.

Results

Identifying Differentially Methylated Regions in Whole-Genome Bisulfite Sequences

We performed whole-genome bisulfite sequencing (WGBS) in airway epithelial cell DNA from 20 African American children (10 with allergic asthma, 10 without asthma or allergies; 11 years old) from the Urban Environment and Childhood Asthma (URECA) cohort (7) and 19 European American young adults (9 with allergic asthma, 10 without asthma or allergies; 18-20 years old) from the Childhood Origins of Asthma (COAST) cohort (8). After quality control (QC) (see **Methods**), analyses for differentially methylated regions (DMRs) between the asthma/allergy cases and non-asthma/non-allergy controls were performed in the African American sample, the European American sample, and the combined sample (see **Methods** for additional details). Overall, we identified 16,611 DMRs covering 199,473 CpGs (**Table S1**).

Selecting High-Value CpGs for an Asthma&Allergy Custom DNA Methylation Array

Our goal was to develop a custom array with 50,000 probes for CpGs that would complement and serve as a “booster” for the EPIC array for studies of asthma and allergic diseases. Our selection

criteria involved three steps (**Figure S1**; also see **Methods**). Briefly, we first identified a set of CpGs based on prior evidence of association with asthma or allergic diseases (atopic dermatitis/eczema, allergic rhinitis/hay fever, and food allergy) from three categories of studies. The first category included the 199,473 CpGs within DMRs from the WGBS study described above, the second included 19,057 CpGs associated with asthma or allergic diseases in 17 previous EWAS (**Table S2**), and the third included 570,350 CpGs at 140 GWAS loci (**Table S3**). We then removed CpGs on the EPIC array, in ENCODE blacklisted regions (9), and those that overlapped with a common single nucleotide polymorphism (SNP) in 1000 Genomes European (CEU) or African (YRI) populations. To further prioritize the remaining 696,225 CpGs, in the second step we considered their overlap with functional annotations, requiring that CpGs within DMRs overlap with at least three annotations or other category of prior evidence, CpGs from prior EWAS overlap with at least one annotation or other category of prior evidence, and CpGs at GWAS loci overlap with at least four annotation categories or other category of prior evidence. After removing duplicates, 92,024 high-value CpGs remained; 53,700 passed the Illumina design step and, in the third step, we pruned to 44,047 by selecting CpGs from all categories with the highest number of functional annotations. After manufacture and array QC, 37,863 CpGs targeted by 42,192 probes remained (**Table S4**).

The distributions of CpGs on the Custom and EPIC arrays are shown by annotation category in **Figure 2A**. Of note, 90% of the Custom array CpGs overlapped with a transcription factor binding site (TFBS), and 94% overlapped with a predicted enhancer, compared to 56% and 54% of the EPIC array CpGs, respectively (Fisher exact test [FET] $p<10^{-16}$ for all tests). Moreover, compared to EPIC CpGs, Custom CpGs were enriched in introns (52.8% vs. 47.2%) and exons (24.0% vs. 22.1%) and depleted in intergenic regions (23.1% vs. 30.7%) and 5'UTRs (6.8% vs. 7.4%) (FET $p<10^{-4}$ in all analyses) **(Figure 2B)**. The CpGs on the Custom array

proportionally represented DMR-CpGs from the three WGBS studies and from the three categories of prior evidence (**Figure S2**).

Evaluating methylation patterns of CpGs on the Custom Array

To assess the reproducibility of methylation measures of CpGs on the Custom array, we compared β values in airway epithelial cell DNA in the WGBS to the same CpGs on the Custom and the EPIC arrays for the subset of individuals with all three (19 of the URECA children) or two (17 of the COAST young adults) measures. We included CpGs with at least 10 WGBS reads that overlapped with an array CpG that passed QC. This resulted in 502,229 CpGs for the WGBS and EPIC array comparison in URECA and 24,744 and 20,861 CpGs for the WGBS and Custom array comparison in URECA and COAST, respectively. The β distribution plots were similar and highly correlated between CpGs measured by both WGBS and the arrays (Spearman's $p<2.2x10^{-16}$ for each) (**Figure S3**). Notably, however, compared to the EPIC array, CpGs on the Custom array were depleted for hypermethylated CpGs and enriched for intermediate methylated (IM) CpGs (β values 20-80%).

To more broadly assess EPIC and Custom array DNA methylation patterns, we compared the distributions of CpG β values across different tissues using available data for the EPIC array in DNA from airway epithelial cells (this study), airway smooth muscle cells (10), and buccal swabs, placenta cells, and cord blood (11), and for the Custom array in DNA from nasal epithelial cells and nasal lavage cells from URECA children, and buccal cells, cord blood cells, or placenta tissue from infants in the Vitamin C to Decrease the Effects of Smoking in Pregnancy on Infant Lung Function (VCSIP) cohort (11). The β distributions were similar across cell sources for CpGs on the EPIC array (**Figure 3A**) but showed varying patterns between cell types on the Custom array (**Figure 3B**). Whereas an average of 68% (range: 60-79%) of CpGs on the EPIC

array were either hypomethylated (0-20%) or hypermethylated (80-100%) in all six cell types, most on the Custom array (66% in nasal epithelial cells and 52% on average across cell types) were IM CpGs with depletions of CpGs with β values between both 0-10% (24% in nasal epithelial cells and 29% on average) and 90-100% (10% in nasal epithelial cells and 20% on average). In fact, the β value distributions of CpGs on the EPIC array in nine GTEx tissues are remarkably similar to those reported here (**Figure S4**) (12). We confirmed that differences between the two arrays were attributable to the selection criteria we used for designing the array. Using the same criteria, high-value CpGs on the EPIC array also showed a depletion of both hypomethylated and hypermethylated CpGs and an enrichment for IM CpGs (**Figure S5**).

EWAS of Allergic Sensitization in 280 Multi-Ancestry Children using the Custom and EPIC Arrays

Our ultimate goal was to develop a DNA methylation array that could detect asthma- or allergy-associated differential methylation that is missed by the content on the EPIC array. Allergic sensitization (AS) is an IgE response to allergens (13) and a crucial step in the development of allergic diseases and common phenotypes of asthma (14-16). We conducted an EWAS of AS in airway epithelial cell DNA collected from 280 11-year old URECA children using the Custom and the EPIC arrays. We used the percent of positive skin prick tests (14 airborne and oral allergens) as a quantitative measure of AS (see **Methods**). The demographic and clinical description of the URECA children is shown in **Table S5** and **Figure S6**.

Following QC of the array (see **Methods**), we performed two EWAS of AS in the same URECA children using CpGs on the EPIC or Custom array. In each analysis, we included as covariates sex, percent epithelial cells, the first 3 ancestry principal components (PCs), and latent factors to adjust for additional unwanted variation (17, 18) (see **Methods**). Using a q-value

threshold of 0.05, the EWAS of AS revealed 1,805 differentially methylated CpGs (DMCs) using the EPIC array and 193 DMCs using the Custom array (**Figure 4A-B**). Overall, the Custom array was enriched for AS DMCs compared to the EPIC array (0.50% vs. 0.23%; FET $p<2.2x10^{-16}$ compared to all CpGs on the array). Among the DMCs on the EPIC array, 295 (16%) were high-value EPIC CpGs, a significant enrichment compared to 3.5% of high-value CpGs among all CpGs on the array (FET $p<2.2x10^{-16}$). In contrast to the differences in CpGs overall (**Figure 3A**), the β distributions of DMCs from the two arrays were similar (**Figure 4A-B**; middle panels), showing a near complete depletion of both very hypomethylated and hypermethylated CpGs and an enrichment of IM CpGs among AS DMCs in airway epithelial cells.

The DMCs on both arrays were distributed across the autosomes (**Figure S7A-B**), DMCs on the Custom array showed spikes of association signals compared to the sparse distribution of DMCs on the EPIC array. In fact, 50% of DMCs on the Custom array are within 100bp of the next nearest DMC compared to only 3% of DMCs on the EPIC array (**Figure 4D**). Regional clustering of DMCs provides greater confidence in and internal validation of EWAS results.

Most of the functional annotation categories were represented by DMCs proportionally to all CpGs on each array (**Figure 2A and 2C**; see **Figure 2** legend for all statistical test results). However, DMCs from both arrays were depleted at transcription start sites and in areas of open chromatin compared to all CpGs on the arrays. Among the primary criteria (prior evidence), CpGs in DMRs were modestly enriched for DMCs compared to all CpGs on the Custom array. Similarly, DMCs in prior EWAS studies of asthma and allergic diseases were modestly enriched on the Custom array but significantly enriched on the EPIC array. In contrast, CpGs at GWAS loci for asthma and allergic diseases were not enriched among DMCs on the Custom array and only modestly enriched on the EPIC array compared to all CpGs on the array. The genomic locations of the DMCs were proportional to all CpGs on the arrays (**Figure 2B and 2D**).

Because AS is an important step in the development of childhood asthma, we next asked whether AS DMCs were also associated with allergic asthma. Considering only the AS DMCs on the EPIC array, 1,155 (64%) were also associated with allergic asthma at q-value <0.05 (1,805 tests). Among the AS DMCs on the Custom array, 115 (60%) were also associated with allergic asthma at q-value <0.05 (193 tests). The effect sizes were significantly correlated between AS DMCs and allergic asthma DMCs (r=0.68). Among the CpGs that were DMCs for both AS and allergic asthma, all showed the same direction of effect; among all AS DMCs, 93% showed the same direction of effect with allergic asthma (r=0.84) (**Figure 4E**).

Validating the Custom Array in the INSPIRE Cohort

The URECA cohort is diverse with respect to ancestry but includes <1% non-Hispanic White children (7). Therefore, to both replicate results of the EWAS in URECA children and assess the performance of the Custom array in a primarily non-Hispanic White population, we studied 5- to 7-year-old children from the Infant Susceptibility to Pulmonary Infections and Asthma Following RSV Exposure (INSPIRE) study (19). INSPIRE infants were enrolled following delivery into this population-based longitudinal study in central Tennessee. An EWAS was performed using the Custom array and nasal epithelial cell DNA from 474 children with measures of AS (proportion of positive SPT to 13 inhaled allergens). The demographic and clinical characteristics of the INSPIRE children are shown in **Figure S6** and **Table S5**. See **Methods** for further descriptions of the sample, allergens tested, and processing and QC of the array data.

The AS EWAS model in INSPIRE included as covariates sex, parent-reported race or ethnicity, DNA concentration, and latent factors to adjust for unwanted variation (17) (see **Methods**). At a q-value threshold of <0.05 based on 37,261 tests, 85 CpGs on the Custom array were DMCs (0.2% of CpGs). The β distribution of the DMCs (**Figure 4C**) and the spikes of associations (**Figure S7**) were similar to those observed in URECA. Among the AS DMCs in

INSPIRE, 38% were also allergic asthma DMCs (q-value<0.05 based on 85 tests), all with the same direction of effect. Among all AS DMCs in INSPIRE, 93% had the same direction of effect with allergic asthma (**Figure 4F)**. Moreover, among the AS DMCs in URECA, 25 (13%) were also AS DMCs in INSPIRE, which was a highly significant enrichment ($p<2.2x10^{-16}$), and all had concordant directions of effect. The effect sizes of the 253 CpGs that were DMCs in either the URECA or INSPIRE AS EWAS were highly correlated (r=0.61; $p<2.2x10^{-16}$) (**Figure 4G**). The high reproducibility of EWAS results in two cohorts demonstrates that the high-value CpGs on the Custom array identify AS DMCs that are robust to ancestry, ascertainment strategies, age at sampling, and geography.

Regional association plots combining results of the three AS EWAS for the 10 most significant DMCs in the EPIC and Custom EWAS in URECA are presented in **Supplementary Materials**, along with three exemplary regions and their associations with gene expression (**Supplementary Materials**, pages 2-9).

Evaluating the Biological Significance of DMCs in Airway Epithelial Cells

Because nearly all CpGs on the Custom array were selected to overlap with a TFBS or enhancer mark, we expected these CpGs to be correlated with the expression of genes in airway epithelial cells more often than CpGs on the EPIC array (20). To test this, we used gene expression data (16) in the same cells as those used for DNA methylation studies in 249 of the URECA children and defined two sets of genes from among the 15,551 genes detected as expressed in these cells: the nearest gene to each CpG and the promoter capture Hi-C (pcHi-C)-defined target gene in airway epithelial cells (see **Methods**). The latter identifies putative enhancers that physically interact with target gene promoters and account for the 3-dimensional structure of the genome that allows for distal regulatory elements to interact with and regulate the activity of promoters over large distances (e.g., up to 1 Mb or more), which are often not the nearest gene. Among the

nearest expressed genes to DMCs on the Custom (98 genes) and EPIC (1,449 genes) arrays, 63 were the nearest gene on both arrays. Among the pcHi-C target genes for DMCs on the Custom (245 genes) or EPIC (1,155 genes) arrays, 95 were target genes on both arrays. Thus, although no CpGs overlapped between the arrays, 101 of nearest or pcHi-C target genes were shared on both arrays (**Table S6**). The 318 nearest or pcHi-C target genes on the Custom array that were expressed in airway epithelial cells were enriched in 16 pathways (FDR <0.05), including "Th1 and Th2 cell differentiation," "JAK-STAT signaling," "Th17 cell differentiation," and "Viral protein interaction with cytokine and cytokine receptor" (**Table S7**). In contrast, the 2,366 nearest or pcHi-C target genes on the EPIC array that were expressed in airway epithelial cells were not enriched for any pathways (smallest FDR = 0.13). These findings further demonstrate that the CpGs on the Custom array are enriched in pathways relevant to asthma and allergic disease.

For each CpG-gene pair, we tested for correlation between DNA methylation and expression levels to identify expression quantitative methylation (eQTM) CpGs with their nearest and pcHi-C target gene(s). We used a linear model that included sex, percent epithelial cells, and three ancestry PCs as covariates (FDR <0.05). Significantly more CpGs on the Custom array were eQTMs with their nearest gene (23%) or pcHi-C target gene (16%) compared to CpGs on the EPIC array (11% and 9% respectively; FET $p<2.2x10^{-16}$ in both analyses). The high-value EPIC CpGs were also enriched for eQTMs compared to all EPIC CpGs (20% and 12%, respectively; FET $p<2.2x10^{-16}$ in both analyses). Although more AS DMCs that were eQTMs on both arrays, significantly more were on the Custom compared to the EPIC array (nearest gene 35% vs. 20% [FET p = 0.0019] and pcHi-C 22% vs 15% [FET p = 0.0082], respectively) (**Table S8**). The β distributions for DMCs on the Custom and EPIC arrays that were eQTMs were further enriched for IM CpGs and depleted of CpGs at the extremes of the distribution (**Figure 4** and **Figure S8**).

Examples of correlations between methylation and gene expression levels for DMCs are shown in **Supplementary Materials** (pages 4-10) for three representative EWAS loci.

Discussion

The epigenome plays a critical role in regulating gene expression in a context-specific manner, such as in the presence of environmental exposures or disease states. DNA methylation, for example, is responsive to disease promoting exposures in both in *in vitro* cell models (10, 21-23) and *ex vivo* cells from individuals who are healthy and with disease (1, 24-27). Yet the DNA methylome has been underexplored and largely limited to the CpGs on commercial arrays. As a result, it is unknown whether the 800,000 or fewer CpGs interrogated in nearly all EWAS studies to date include CpGs most relevant to any specific exposure, disease, or tissue type. Here, we addressed this question by designing a custom Allergy&Asthma DNA methylation array with 37,863 high-value CpGs that are not included on the commercial EPIC array. Our study design allowed us to make important observations regarding features of high-value CpGs and to propose a pipeline that can be used to prioritize CpGs from among the more than 28 million in the human genome in future studies.

Although we did not select CpGs based on methylation levels, those that passed through our pipeline were significantly enriched for IM CpGs (β values 20-80%). The CpGs that were DMCs and eQTMs were further enriched for IM CpGs on both the EPIC and Custom arrays, suggesting that this one feature is a signature of high-value, variable and functional CpGs in the genome. In contrast, the non-IM CpGs may be more functionally constrained and less responsive to environmental perturbations. Our results are supported by previous WGBS studies showing that IM CpGs were more likely to be tissue specific, to vary between individuals, and to play important roles in gene regulation (20, 28). Our study further revealed that CpGs with intermediate levels of methylation are also more likely to be associated with AS, an important

clinical phenotype that reflects both an immune response to past allergen exposures (13) and a risk factor for the development of asthma and allergic diseases (14-16). Indeed, AS DMCs were highly enriched for allergic asthma DMCs. Although the CpGs on the custom array were overall enriched for IM CpGs, AS DMCs were further enriched, with a near complete depletion of very hypomethylated (0-20% methylation) and very hypermethylated (80-100%) CpGs. This important observation was further supported by EWAS results using the EPIC array. Whereas EPIC array CpGs are enriched for hypomethylated and hypermethylated CpGs in all tissues, the EPIC DMCs were depleted for CpGs at both extremes and enriched for IM CpGs, similar to the DMCs on the custom Allergy&Asthma array. We further showed that the 26,905 CpGs on the EPIC array that passed our filtering pipeline were also enriched for IM CpGs and enriched among AS DMCs: 16.3% of the DMCs in the AS EWAS were among the high-value EPIC CpGs compared to 3.4% of all CpGs on the EPIC array. These results, combined with previous studies (20, 28), suggest that filtering CpGs on the EPIC array to include just IM CpGs (β=20-80%) should enrich for functional CpGs and increase power to detect DMCs associated with exposure or disease outcomes. Our data support this suggestion. Using all CpGs on the EPIC array, we identified 1,805 DMCs (0.23% of all CpGs). However, when we include only IM CpGs in the EWAS, we identified 2,182 DMCs (0.70% of all IM CpGs), more than three times the proportion of DMCs among tested CpGs. Finally, DMCs that were eQTMs were further enriched for IM CpGs (**Figure 4** and **Figure S8**), supporting their role in gene regulation.

In addition to the enrichment of IM CpGs on the Custom compared to the EPIC array, the patterns of associations with AS on the two arrays were notable. Whereas DMCs on the EPIC are more sparsely distributed throughout the genome, the DMCs on the Custom show spikes of association, analogous to GWAS peaks. The spatial organization among DMCs on the Custom provides confirmatory evidence for their veracity, which is not available for CpGs on the EPIC

array (**Figure 4D**). Another distinguishing feature of the DMCs on the Custom array were their enrichment in exons and depletions in intergenic regions compared to DMCs on the EPIC array. This is consistent with results of an earlier study that observed enrichments of IM CpGs in exons, where exon expression at those sites correlated with DNA methylation (20).

The CpGs included on the custom Allergy&Asthma array were selected based on prior evidence of associations with asthma or allergic diseases using three classes of evidence: within a DMR in our WGBS study, in a previous published EWAS, or at published GWAS loci. Overall, 62% of CpGs on the custom array were selected from DMRs compared to 1.6% of all CpGs on the EPIC array, yet the DMR CpGs were enriched among DMCs on both the custom (77.2%) and EPIC (3.9%) arrays. Because we excluded all CpGs on the EPIC from the Custom array, few previous EWAS CpGs were on the Custom array, although we still observed a modest enrichment of EWAS CpGs among DMCs on the Custom array (1.9% overall vs. 4.1% of DMCs) and a highly significant enrichment among DMCs on the EPIC array (4.9% overall vs. 21.7% of DMCs). These represent replications of results from previous EWAS of asthma and allergic phenotypes. In contrast to these two categories of prior evidence, CpGs at asthma and allergy GWAS loci were very modestly enriched among EPIC array DMCs but not enriched, and even slightly depleted, among the Custom array DMCs. In a WGBS in 25 different cell and tissue types, Elliott et al. (20) showed that IM CpGs were more likely to be allele independent than to show allele-specific methylation. Busche et al. (28) conducted a WGBS study in adipose and blood cells from twins and concluded that non-genetic effects were mainly tissue specific and located in gene regions, consistent with the features of CpGs on the Custom array, whereas CpGs with genetic effects were shared across tissues and more likely to be intergenic, consistent with features of CpGs on the EPIC array. They concluded that non-shared environments accounted for most of the variance in methylation levels. Our findings of a lack of enrichment of DMCs at

GWAS loci, but enrichments for IM CpGs and for AS DMCs in exons suggest that inter-individual variation for IM CpGs on the Custom array may be the set of more plastic CpGs in the human genome and better reflect the effects of environmental exposures than of genetics. Nonetheless, our study still revealed examples of AS DMCs at important GWAS loci (e.g., *IRF1* on chromosome 5, *CLEC16A* on chromosome 16, and *HDAC7*/*VDR* on chromosome 12) (**Supplementary Results**; pages 5-8).

In summary, our study revealed high-value functional CpGs in airway epithelial cells that are not interrogated on the EPIC array. Our pipeline identified 92,024 high-value CpGs after excluding 26,905 high-value CpGs on the EPIC array. Therefore, among the >28 million CpG sites in the genome, less than 0.5% are likely to be functional, or high value, in any particular tissue or disease/environment context. The Custom array, which included about half of the non-EPIC high-value CpGs in airway epithelial cells, indicated that associations with AS and allergic asthma were robust to race/ethnicity, ascertainment, age, and geography. We suggest that similar pipelines can be used to identify high-value CpGs for other disease groups or specific environmental exposures leading to the discovery of currently invisible portions of the epigenome that are relevant to human health.

Materials and Methods

Cohorts Included in WGBS or EWAS Studies

Three longitudinal birth cohort studies were included in these studies. All studies were approved by the Institutional Review Boards of each of the institutions recruiting subjects; for each cohort, one parent or guardian provided informed consent for their child's participation. Informed consent was obtained from participants in the COAST cohort after their 18th birthday.

The URECA study is an observational birth cohort study initiated in 2005 in Baltimore, Boston, New York City, and St. Louis under the NIAID-funded Inner City Asthma Consortium (7). Either the pregnant mother or the father of their unborn child had a history of asthma, allergic rhinitis, or eczema. Asthma was assessed at age 10 according to a definition that considered symptoms, diagnosis by a health care provider, and measurements of pulmonary function, as described (16). Skin prick testing was performed at age 10 (255 subjects) or age 7 (25 subjects) and included the following allergens: mouse epithelia, dog epithelia, *Dermatophagoides fainae* (mite), *Dermatophagoides pteronyssinus* (mite), cat hair, rat epithelia, American/German cockroach mix, German cockroach, *Alternaria tenuis* (mold), *Aspergillus* mix, ragweed mix, tree pollen (oak or birch), *Penicillium Notatum/Pennicillium Chrysogenum*, and Timothy grass. Nasal brushings were obtained at age 11 years, as described (16). Twenty African American children were selected for the WGBS studies (10 with asthma and allergic disease [3 females, 7 males], 10 without asthma or allergic disease [3 females, 7 males]), and 280 unrelated children were included in the EWAS using both the Custom and EPIC arrays. The characteristics of the 280 URECA participants are described in **Table S5**.

The COAST study is an observational birth cohort study initiated in 1998 in Madison, Wisconsin (8). Pregnant women were recruited in the third trimester of pregnancy if the mother or the father had a history of asthma or allergic diseases. Asthma was assessed beginning at age 6 years. Children were diagnosed with asthma if they fulfilled at least one of the following criteria: (1) physician-diagnosed asthma; (2) frequent albuterol use for coughing or wheezing episodes as prescribed by a physician; (3) use of a prescribed daily controller medication; (4) an implemented step-up plan, including use of albuterol or inhaled corticosteroids during illness as prescribed by a physician; (5) use of prednisone for an asthma exacerbation. Nasal brushings were obtained at ages 18-20 years, as described (29). Twenty participants of European American ancestry were

selected for the WGBS studies (10 with asthma and allergic disease [5 males, 5 females], 10 without asthma or allergic disease [5 males, 5 females]).

The INSPIRE study is a population-based observational birth cohort of healthy infants in central Tennessee (19). Children were followed longitudinally for the outcomes of asthma and allergic diseases. Flocked nasal swabs were obtained at age 5-6 years and samples were stored at –80°C in RNA lysis buffer. RNA and DNA were isolated from the swabs using the Qiagen AllPrep DNA/RNA kit. Subjects were tested for the following thirteen allergens: dog, cat, *Dermatophagoides pteronyssinus* and *Dermatophagoides farinae* mix, American/German cockroach, *Penicillium Notatum*/*Penicillium Chrysogenum*, *Alternaria Tenuis*, *Cladosporum Herbarum*, *Aspergillus* mix, Ragweed mix, Eastern 6 Tree mix, K-O-T Grass mix, Maple/Box Elder mix, and Weed mix. The characteristics of the 474 unrelated INSPIRE children included in the EWAS are described in **Table S5**.

Cohorts included in the Cross-Tissue Comparison Studies

In addition to studying nasal epithelial cell DNA with the Custom array in URECA, we included results from studies of the Custom array in DNA from nasal lavage cells (n=96), buccal cells (n=96), placenta (n=96) and cord blood (96). The nasal lavage cell DNA was isolated from URECA subjects around age 13 (mean = 13.4 years, range = 12.8 - 14.9 years) using the QIAamp DNA Micro Kit. The buccal, placenta and cord blood cell DNA was collected from 96 participants in the VCSIP cohort (11). Placental DNA was extracted from powdered tissue under liquid nitrogen, using the QIAcube for automated nucleic acid extraction. Cord blood DNA was extracted from 200µL of whole blood using the QIAamp DNA Mini Kit and the QIAcube, and buccal DNA was extracted from buccal swabs following proteinase K digestion using the Maxwell 16 Blood DNA Extraction Kit and the Maxwell 16 Instrument for automated nucleic acid extraction.

Whole Genome Bisulfite Sequencing and DMR Studies

DNA (250 ng) from participants in the URECA and COAST cohorts was transferred to the University of Chicago Genomics Facility. Bisulfite conversion and sequencing were conducted using the EZ DNA Methylation-Gold Kit library prep and the Swift Biosciences Accel-NGS Methyl-seq DNA library kit, respectively. Samples were sequenced to a minimum of 330 million reads on the Illumina NovaSEQ6000 (S4 flowcell). Adapters were removed from sequence reads using trimgalore (30) prior to mapping the reads using bismark (version 0.18.2) (9) and the hg19 reference assembly. Bismark was further used to remove duplicate reads and call methylation values. Prior to DMR analyses, CpGs were removed if they overlapped with a common SNP (minor allele frequency [MAF] >0.05) in 1000 genomes CEU or YRI populations (31) or in Blacklisted regions (32). Variant calling from WGBS was performed using the biscuit algorithm, which also identified sample swaps and contamination. The variant calls from WGBS were compared to either array-based genotypes (COAST) or whole genome-sequence-based genotypes (URECA). All variant calls matched with an accuracy of >95%, except for one mismatched sample in COAST. The latter sample was excluded from further analysis, leaving 19 samples for analysis (9 allergic asthmatics and 10 non-asthma/non-allergic controls).

DMR analyses were conducted in the African American sample (n=20), the European American sample (n=19), and the combined African American and European American sample (n=39). The methylation data were then smoothed using BSmooth (33), and DMRs were called using the bsseq package (version 1.14) (33). Only CpGs covered by at least 10 reads in 80% of the cases and controls (AA only, EA only, and combined n=39) were included. T-statistics cutoffs were based on 5% quantiles, and a maximum gap of 300bp was required between CpGs to define a cluster, as recommended by BSmooth. To maximize the number of DMRs, we required three or more CpGs per DMR and a minimum of 5% difference in methylation levels between the allergy

asthma cases and non-allergy, non-asthmatic controls. The union of DMRs between analyses was assessed using the *reduce* function from the GenomicRanges R package (version 1.30).

Selection of CpGs for Custom Array

We first identified CpGs with a high likelihood of being associated with asthma and allergic disease (**Figure S1**). In the first step, we prioritized regions with prior evidence of association with asthma or allergic diseases (atopic dermatitis/eczema, allergic rhinitis/hay fever, and food allergy) from three categories of studies. The first category included the 199,473 CpGs within the DMRs from the WGBS. The second category included CpGs from previous DNA methylation studies (EWAS) of asthma or allergic diseases. For this, we conducted a literature search for array-based studies of DNA methylation in asthma and allergic disease. We identified 15 studies that conducted 25 EWAS of asthma- or allergy-related phenotypes, five of which were conducted in respiratory epithelium and 10 in blood (**Table S2**). We also included CpG sites for five additional genes from two candidate-gene DNA methylation studies of food allergies; *FOXP3* (34) and *IL4*, *IL5, IL10* and *INFG* (35). In total, 19,057 unique DMCs were identified. The third category included CpGs located within the 140 GWAS loci defined in two recent large studies of adult-onset and childhood-onset asthma (36) and allergic diseases (asthma, hay fever and eczema) (37) in UK Biobank subjects (38), which included nearly all the loci reported in other GWAS in multi-ancestry populations. In total, the GWAS loci covered a total of 570,350 CpG motifs. We also included CpGs for the *MALT1* gene that was significantly associated with peanut allergy in a genome-wide interaction study (39). We next removed duplicate CpGs from among the three categories of prior studies, CpGs on the EPIC array, CpGs in ENCODE blacklist regions (9), and those in which the cytosine nucleotide overlapped with common SNPs (MAF > 5%) in 1000 Genomes CEU or YRI populations (**Table S4**). A total of 696,225 CpGs remained for consideration in the second step.

To further prioritize the CpGs, in the second step we considered their overlap with six functional annotations: 1) ENCODE (40) TFBSs from all cell types; 2-4) ROADMAP Epigenetics (41) transcriptional start sites, poised enhancers and active enhancers from smooth muscle (E078, E076, E103, E111), epithelial (E055, E056, E059, E061, E058), and blood cells (E062, E034, E045, E044, E043, E039, E041, E042, E040, E037, E048, E038, E047, E029, E050, E032, E046); 5) ATAC-seq in human cultured bronchial epithelial cells exposed to rhinovirus or a vehicle control from asthmatic and non-asthmatic individuals (42), and 6) pcHi-C from *ex vivo* human bronchial epithelial cells (42). We required that CpGs at DMRs overlapped with at least three functional annotations or prior evidence (GWAS or EWAS), CpGs from previous EWASs overlapped with at least one functional annotation or prior evidence (GWAS or DMR), and CpGs at GWAS loci overlapped with at least four functional annotations or prior evidence (EWAS or DMR). Lastly, we selected all remaining CpGs that were within both a GWAS locus and a DMR.

Of the 92,024 resulting high-value CpGs identified, 53,700 were amenable to the Illumina design algorithm, and 44,047 CpGs targeted by 49,999 probes were selected by us for manufacturing. Fourteen percent of the CpGs failed manufacturing, resulting in a Custom array with 38,541 CpGs targeted by 43,605 probes. After removing probes with unknown origins and those without genomic coordinates, 37,863 CpGs passed array QC. The locations of this final set of CpGs, their selection among the three primary criteria, and their overlap with the annotation categories are shown in Dataset S1. Finally, we removed probes that included a common SNP (MAF≥5%) within 3 bp of the CpG interrogation site, probes that failed the minfi QC, and probes on the X chromosome, leaving 37,256 CpGs following both array and processing QC.

RNA-sequencing studies in URECA

Protocols for processing samples for RNA-seq in nasal epithelial cells from the URECA children have been described (16). Gene expression data were available in 249 of the children with DNA methylation data. In these data, 15,643 genes were detected as expressed.

Estimation of genetic ancestry

Ancestry PCs were estimated in the URECA children using a set of 3,534 SNPs that were genotyped in URECA and in reference panels from the 1000 Genomes Project (1KG; n=156)(31) and the Human Genome Diversity Project (HGDP; n=52). European, West African, and East Asian reference samples were randomly selected from CEU (n=52), YRI (n=52), JPT (n=26), and CHB (n=26) samples, respectively, in the phase 3 1KG reference panel (31). Ancestry PCs were calculated using PC-Air (43). The first two ancestry PCs for the URECA children in our study are shown in **Figure S9**.

Processing Custom and EPIC array methylation data in URECA

DNA methylation was assessed using the Illumina Allergy&Asthma Custom BeadChip or the Illumina Infinium MethylationEPIC BeadChip (Illumina, San Diego, CA) following bisulfite conversion at the University of Chicago Genomics Facility. Methylation data from both arrays were processed using minfi v1.29.1 (44). The annotation and manifest packages for the Allergy&Asthma array can be found here: https://github.com/hansenlab/IlluminaHumanMethylationAllergymanifest.

For the EPIC array, probes that failed (detection $P < 0.01$ in at least 25% of samples), overlapped with known SNPs with MAF of at least 5% in African American or European Americans, mapped to the X or Y chromosomes, overlapped ENCODE blacklist regions, or mapped to multiple locations in a bisulfite-converted genome were removed. Raw probe values were background corrected using preprocessIllumina(bg.correct="TRUE", normalize="no"), and

quantile normalization was performed using ENmix (v 1.30.01) (https://doi.org/10.1186/s13148-021-01207-1), followed by SWAN (45) normalization. Samples that failed sex checks using the getSex function in minfi were removed. DNA concentration, collection site, array and plate showed batch effects by principal components analysis (PCA) (46). The effects of collection site, array, and plate were removed using ComBat (47); DNA concentration was removed using linear regression. We estimated unobserved variation in the data correlated with our phenotype of interest (18) and included these latent factors in the model (18). The first three ancestry PCs were also included as covariates to capture the effects of admixture in the sample. After QC, we retained 789,290 (91.1%) of CpGs on the EPIC array for analysis.

For the Custom array, probes that failed (detection $P < 0.01$ in at least 25% of samples), contained a SNP with MAF of at least 5% in African American or European Americans within 3bp of the CpG interrogation site, mapped to the X chromosome, or were missing genomic coordinates were removed. Raw probe values were background corrected using preprocessIllumina(bg.correct="TRUE", normalize="no"), and quantile normalization was performed using ENmix. Additional probe-type normalization using SWAN was not applied because the dye intensities for the methylated and unmethylated type I and type II probes had very different distributions. DNA concentration, collection site, array, and plate showed batch effects by PCA. The effects of site, plate, and chip were removed using ComBat, and DNA concentration was removed using linear regression. We estimated unobserved variation in the data that was not correlated with the phenotype of interest and included these latent factors in the model. To account for regional differences in the density of Custom array CpGs on the estimation of latent factors, we used the partition.params function in FALCO (17) (https://github.com/chrismckennan) to partition groups of CpGs into independent units (default size = 100 kb). The first three ancestry PCs were included as covariates to capture the effects of

admixture in the sample for analyses in URECA. After QC, we retained 37,256 (98.1%) CpGs on the Custom array for analyses.

For comparison purposes, we filtered the CpGs on the EPIC array through the same pipeline described above for selecting CpGs for the Custom array; we refer to these as "high-value" EPIC CpGs. Of the 789,290 CpGs on the EPIC array, 26,905 (3.4%) were considered high-value, similar in number to the 37,863 CpGs that passed QC on the Custom array (Dataset S2).

Processing Custom array methylation data in INSPIRE

The same pipeline used in URECA was used to process INSPIRE Custom methylation data. Twenty-one probes failed *P*-value detection and were removed from the analysis. We estimated sex using the X chromosome CpGs and removed five samples with discrepancies between these classifications and reported sex. DNA concentration, collection year, plate, and array were identified as having batch effects by PCA (**Table S9**). The effects of collection year, plate, and array were removed using ComBat, and DNA concentration was removed using linear regression. Because DNA concentration was correlated with other variables, it was also included as a covariate in the model, along with sex, parent-reported race/ethnicity, and latent factors estimated using the same methods described above for URECA. A total of 37,261 CpGs passed processing QC and were included in the analysis.

EWAS in the URECA and INSPIRE cohorts

For each EWAS we used the following models:

URECA: DNAm ~ proportion positive SPTs + sex + %epithelial cells + $AncPCs_{1-3}$ + LFs_{1-n}

INSPIRE: DNAm ~ proportion positive SPTs + sex + race/ethnicity + DNA concentration + LFs_{1-n}

For the URECA EWAS, we included 14 and 5 latent variables for the Custom (17) and EPIC (18) arrays, respectively. Cell counts (300 cells/slide) were performed on each sample (16). Because we did not have information on cell proportion or genetic ancestry for INSPIRE, we included sex, parent-reported race/ethnicity, DNA concentration, and 21 latent factors as covariates to adjust for cell composition, as well as other unwanted variation in the EWAS. Analyses were performed in R (version 4.1.0) using limma v 3.50.0 (48, 49). To control the false discovery rate, we used a q-value threshold of 0.05. The URECA EWAS sample included 19 of the 20 African American children in the WGBS studies and the 96 children shown in the cross-tissue density plots of β distributions (**Figure 3**).

To assess the overlap of AS DMCs with allergic asthma DMCs, we tested the AS DMCs for association with allergic asthma in URECA and INSPIRE. For these studies, we defined allergic asthma cases as children with a diagnosis of asthma at age 10 (URECA) or age 6 (INSPIRE) and sensitization to at least one allergen (N=62 and 22, respectively). Controls were defined as children without a diagnosis of asthma (ever) and without any positive skin prick tests (n=76 and 271, respectively). We calculated q-values based on the 193 AS DMCs in URECA and the 85 AS DMCs in INSPIRE. The overlaps of AS and allergic asthma cases and controls are shown in **Figure S10**.

Correlations of methylation level with expression of nearest gene and target promoter capture Hi-C gene

Correlations between methylation levels at each CpG with the expression levels of the nearest gene and the pcHi-C target genes (42) were assessed using linear regression in both the full data

set (all CpGs that passed QC on each array) and in the DMCs. Comparisons were only assessed for DMCs falling in "capture end regions" (± 1kb) that interacted with gene promoters (42). Sex, ancestry PCs, and epithelial cell composition were included as covariates in the model. The nearest gene was annotated using the GenomicRanges package v1.46.1 in R with the gene list from the R biomaRt package v2.50.2 using ensemble GRCh37. A false discovery rate of 5% was used to assess significance.

Pathway analysis

Pathway over-representation testing was performed with Advaita Bio's iPathwayGuide (https://www.advaitabio.com/ipathwayguide) using KEGG Pathways (release 100.0+/11-12, Nov 2021). For the Custom and the EPIC arrays, the 318 and 2,366 genes, respectively, that were either the nearest or pcHi-C target gene to a DMC, were entered as significant against a background of all nearest genes to CpGs on each array that are expressed in NECs (n=14,049).

Acknowledgments

Research reported in this publication was supported by the Environmental influences on Child Health Outcomes (ECHO) program, Office of The Director, National Institutes of Health, under Award Numbers U2COD023375 (Coordinating Center), U24OD023382 (Data Analysis Center), U24OD023319 (PRO Core), UH3 OD023282 (COAST, INSPIRE, URECA) and UH3 OD023288 (VCSIP); and grants UM1 AI114271, UM1 AI160040, U19 AI62310, U19 AI095227, UL1 TR002243 P51 OD011092 and R01 HL104608.

The authors wish to thank our ECHO colleagues, the medical, nursing, and program staff, and the children and families participating in the ECHO cohorts. We also acknowledge the contribution of the following ECHO program collaborators:

ECHO Components—Coordinating Center: Duke Clinical Research Institute, Durham, NC: Smith PB, Newby KL.

References

1. M. Prunicki *et al.*, Air pollution exposure is linked with methylation of immunoregulatory genes, altered immune cell profiles, and increased blood pressure in children. *Sci Rep* **11**, 4067 (2021).
2. Y. Li *et al.*, DNA methylation and transcriptome aberrations mediated by ERalpha in mouse seminal vesicles following developmental DES exposure. *Proc Natl Acad Sci U S A* **115**, E4189-E4198 (2018).
3. E. W. Tobi *et al.*, DNA methylation as a mediator of the association between prenatal adversity and risk factors for metabolic disease in adulthood. *Sci Adv* **4**, eaao4364 (2018).
4. M. Li *et al.*, EWAS Atlas: a curated knowledgebase of epigenome-wide association studies. *Nucleic Acids Res* **47**, D983-D988 (2019).
5. S. T. Holgate, The sentinel role of the airway epithelium in asthma pathogenesis. *Immunol Rev* **242**, 205-219 (2011).
6. C. A. Akdis, The epithelial barrier hypothesis proposes a comprehensive understanding of the origins of allergic and other chronic noncommunicable diseases. *J Allergy Clin Immunol* **149**, 41-44 (2022).
7. J. E. Gern *et al.*, The Urban Environment and Childhood Asthma (URECA) birth cohort study: design, methods, and study population. *BMC Pulm Med* **9**, 17 (2009).
8. R. F. Lemanske, Jr., The childhood origins of asthma (COAST) study. *Pediatr Allergy Immunol* **13**, 38-43 (2002).
9. F. Krueger, S. R. Andrews, Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics* **27**, 1571-1572 (2011).
10. E. E. Thompson *et al.*, Cytokine-induced molecular responses in airway smooth muscle cells inform genome-wide association studies of asthma. *Genome Med* **12**, 64 (2020).
11. L. E. Shorey-Kendrick *et al.*, Impact of vitamin C supplementation on placental DNA methylation changes related to maternal smoking: association with gene expression and respiratory outcomes. *Clin Epigenetics* **13**, 177 (2021).
12. M. Oliva *et al.*, Genetic regulation of DNA methylation across tissues reveals thousands of molecular links to complex traits. *Research Square* https://doi.org/10.21203/rs.3.rs-492596/v1 (2021).
13. R. van Ree, L. Hummelshoj, M. Plantinga, L. K. Poulsen, E. Swindle, Allergic sensitization: host-immune factors. *Clin Transl Allergy* **4**, 12 (2014).

14. S. Illi *et al.*, The pattern of atopic sensitization is associated with the development of asthma in childhood. *J Allergy Clin Immunol* **108**, 709-714 (2001).
15. D. J. Jackson, J. E. Gern, R. F. Lemanske, Jr., The contributions of allergic sensitization and respiratory pathogens to asthma inception. *J Allergy Clin Immunol* **137**, 659-665; quiz 666 (2016).
16. M. C. Altman *et al.*, Endotype of allergic asthma with airway obstruction in urban children. *J Allergy Clin Immunol* 10.1016/j.jaci.2021.02.040 (2021).
17. C. McKennan, Factor analysis in high dimensional biological data with dependent observations. *arXiv:2009.11134v1* (2020).
18. C. McKennan, D. L. Nicolae, Accounting for unobserved covariates with varying degree of estimability in high dimensional experimental data. *Biometrika* **106**, 823-840 (2019).
19. E. K. Larkin *et al.*, Objectives, design and enrollment results from the Infant Susceptibility to Pulmonary Infections and Asthma Following RSV Exposure Study (INSPIRE). *BMC Pulm Med* **15**, 45 (2015).
20. G. Elliott *et al.*, Intermediate DNA methylation is a conserved signature of genome regulation. *Nat Commun* **6**, 6363 (2015).
21. A. Pacis *et al.*, Bacterial infection remodels the DNA methylation landscape of human dendritic cells. *Genome Res* **25**, 1801-1811 (2015).
22. J. Nicodemus-Johnson *et al.*, Genome-wide methylation study identifies an IL-13-induced epigenetic signature in asthmatic airways. *Am J Respir Crit Care Med* **193**, 376-385 (2016).
23. M. M. Soliai *et al.*, Multi-omics colocalization with genome-wide association studies reveals a context-specific genetic mechanism at a childhood onset asthma risk locus. *Genome Med* **13**, 157 (2021).
24. A. Cardenas *et al.*, Controlled human exposures to diesel exhaust: a human epigenome-wide experiment of target bronchial epithelial cells. *Environ Epigenet* **7**, dvab003 (2021).
25. E. S. Wan *et al.*, Cigarette smoking behaviors and time since quitting are associated with differential DNA methylation across the human genome. *Hum Mol Genet* **21**, 3073-3082 (2012).
26. K. M. Bakulski, J. Dou, N. Lin, S. J. London, J. A. Colacino, DNA methylation signature of smoking in lung cancer is enriched for exposure signatures in newborn and adult blood. *Sci Rep* **9**, 4576 (2019).

27. S. Sikdar *et al.*, Comparison of smoking-related DNA methylation between newborns from prenatal exposure and adults from personal smoking. *Epigenomics* **11**, 1487-1500 (2019).
28. S. Busche *et al.*, Population whole-genome bisulfite sequencing across two tissues highlights the environment as the principal source of human methylome variation. *Genome Biol* **16**, 290 (2015).
29. S. Basnet *et al.*, CDHR3 Asthma-Risk Genotype Affects Susceptibility of Airway Epithelium to Rhinovirus C Infections. *Am J Respir Cell Mol Biol* **61**, 450-458 (2019).
30. A. Kechin, U. Boyarskikh, A. Kel, M. Filipenko, cutPrimers: A New Tool for Accurate Cutting of Primers from Reads of Targeted Next Generation Sequencing. *J Comput Biol* **24**, 1138-1143 (2017).
31. 1000 Genomes Project Consortium *et al.*, A global reference for human genetic variation. *Nature* **526**, 68-74 (2015).
32. H. M. Amemiya, A. Kundaje, A. P. Boyle, The ENCODE Blacklist: Identification of Problematic Regions of the Genome. *Sci Rep* **9**, 9354 (2019).
33. K. D. Hansen, B. Langmead, R. A. Irizarry, BSmooth: from whole genome bisulfite sequencing reads to differentially methylated regions. *Genome Biol* **13**, R83 (2012).
34. L. Paparo *et al.*, Epigenetic features of FoxP3 in children with cow's milk allergy. *Clin Epigenetics* **8**, 86 (2016).
35. R. Berni Canani *et al.*, Differences in DNA methylation profile of Th1 and Th2 cytokine genes are associated with tolerance acquisition in children with IgE-mediated cow's milk allergy. *Clin Epigenetics* **7**, 38 (2015).
36. M. Pividori, N. Schoettler, D. L. Nicolae, C. Ober, H. K. Im, Shared and distinct genetic risk factors for childhood-onset and adult-onset asthma: genome-wide and transcriptome-wide studies. *Lancet Respir Med* **7**, 509-522 (2019).
37. A. Johansson, M. Rask-Andersen, T. Karlsson, W. E. Ek, Genome-wide association analysis of 350 000 Caucasians from the UK Biobank identifies novel loci for asthma, hay fever and eczema. *Hum Mol Genet* **28**, 4022-4041 (2019).
38. C. Bycroft *et al.*, The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203-209 (2018).
39. A. Winters *et al.*, The MALT1 locus and peanut avoidance in the risk for peanut allergy. *J Allergy Clin Immunol* **143**, 2326-2329 (2019).

40. Encode Project Consortium *et al.*, Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature* **583**, 699-710 (2020).
41. Roadmap Epigenomics Consortium *et al.*, Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317-330 (2015).
42. B. A. Helling *et al.*, Altered transcriptional and chromatin responses to rhinovirus in bronchial epithelial cells from adults with asthma. *Commun Biol* **3**, 678 (2020).
43. M. P. Conomos, M. B. Miller, T. A. Thornton, Robust inference of population structure for ancestry prediction and correction of stratification in the presence of relatedness. *Genet Epidemiol* **39**, 276-293 (2015).
44. M. J. Aryee *et al.*, Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics* **30**, 1363-1369 (2014).
45. J. Maksimovic, L. Gordon, A. Oshlack, SWAN: Subset-quantile within array normalization for illumina infinium HumanMethylation450 BeadChips. *Genome Biol* **13**, R44 (2012).
46. J. T. Leek *et al.*, Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat Rev Genet* **11**, 733-739 (2010).
47. W. E. Johnson, C. Li, A. Rabinovic, Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8**, 118-127 (2007).
48. B. Phipson, S. Lee, I. J. Majewski, W. S. Alexander, G. K. Smyth, Robust Hyperparameter Estimation Protects against Hypervariable Genes and Improves Power to Detect Differential Expression. *Ann Appl Stat* **10**, 946-963 (2016).
49. M. E. Ritchie *et al.*, limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* **43**, e47 (2015).

Data Sharing Plans

The gene expression data used this study are available on the Gene Expression Omnibus (GEO) repository (accession number GSE145505). The DNA methylation array data used in the EWAS study are available in GEO (accession number in progress). The whole-genome bisulfite sequencing data are available in dbGaP (in progress).

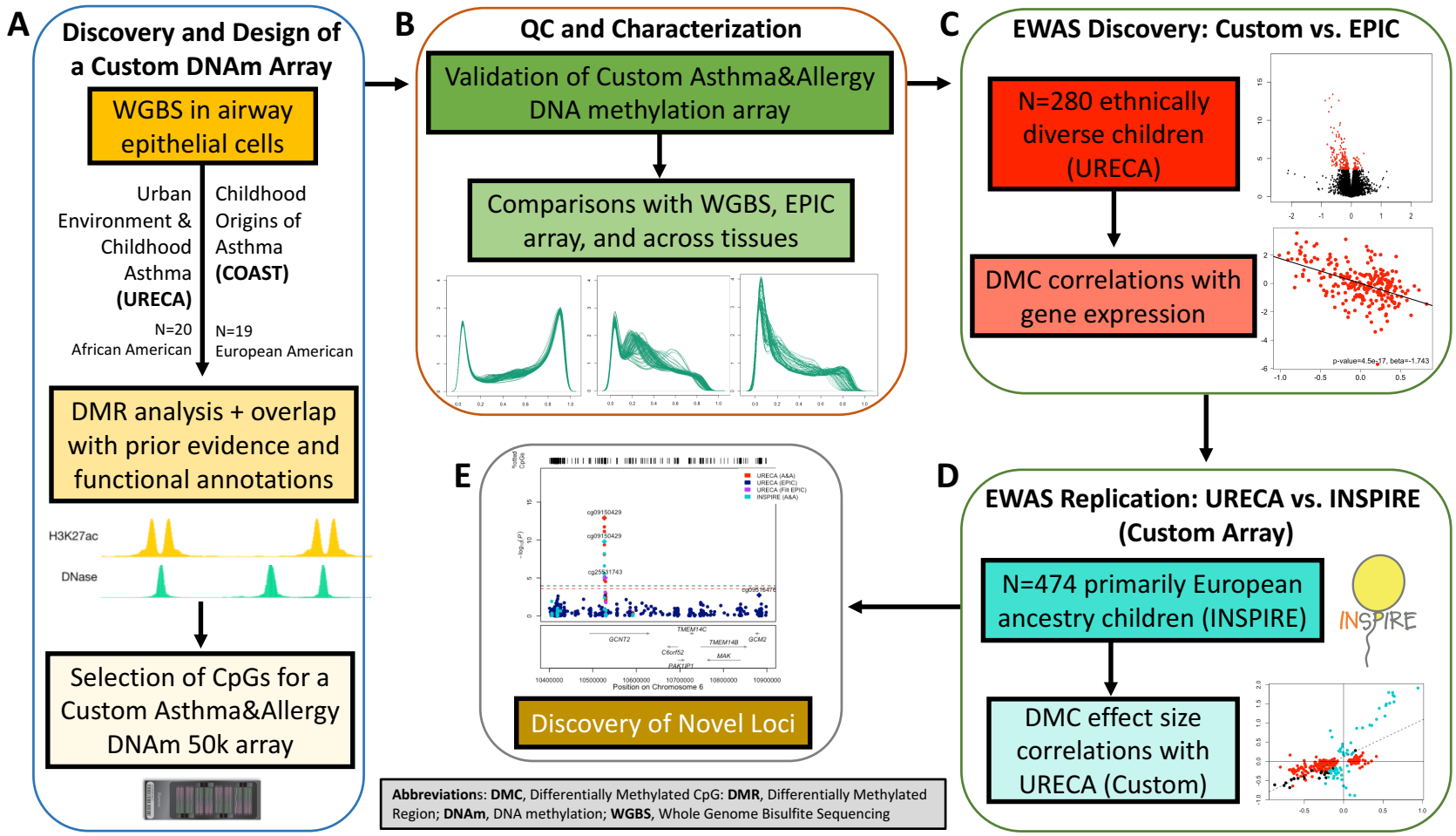

Figure 1. Overview of study design. (A) Whole-genome bisulfite sequencing (WGBS) and differential methylation were performed in airway epithelial cell DNA from birth cohorts comprised of African American children or European American young adults (half of each with allergic asthma and half without asthma or allergies). To select CpGs for the Custom array, we identified CpGs based on

prior evidence of association with asthma or allergic disease from three sources (see text and **Supplementary Figure 1**). The CpGs were then prioritized based on their overlap with functional annotations to ultimately design a Custom array with 42,192 probes for 37,863 CpGs. **(B)** Airway epithelial cell DNA was hybridized to both the Custom array and the EPIC array for comparison, and β value distributions on the arrays were compared to the WGBS data and across tissues using the Custom and EPIC array. **(C-D)** We validated the Custom array by performing an EWAS of allergic sensitization (AS), examined correlations with the gene expression in the same cells, and replicated findings in a second cohort and with allergic asthma. **(E)** The combined results of the Custom and EPIC arrays in the discovery cohort and of the Custom array in the replication cohort for known and novel loci were examined in regional association plots.

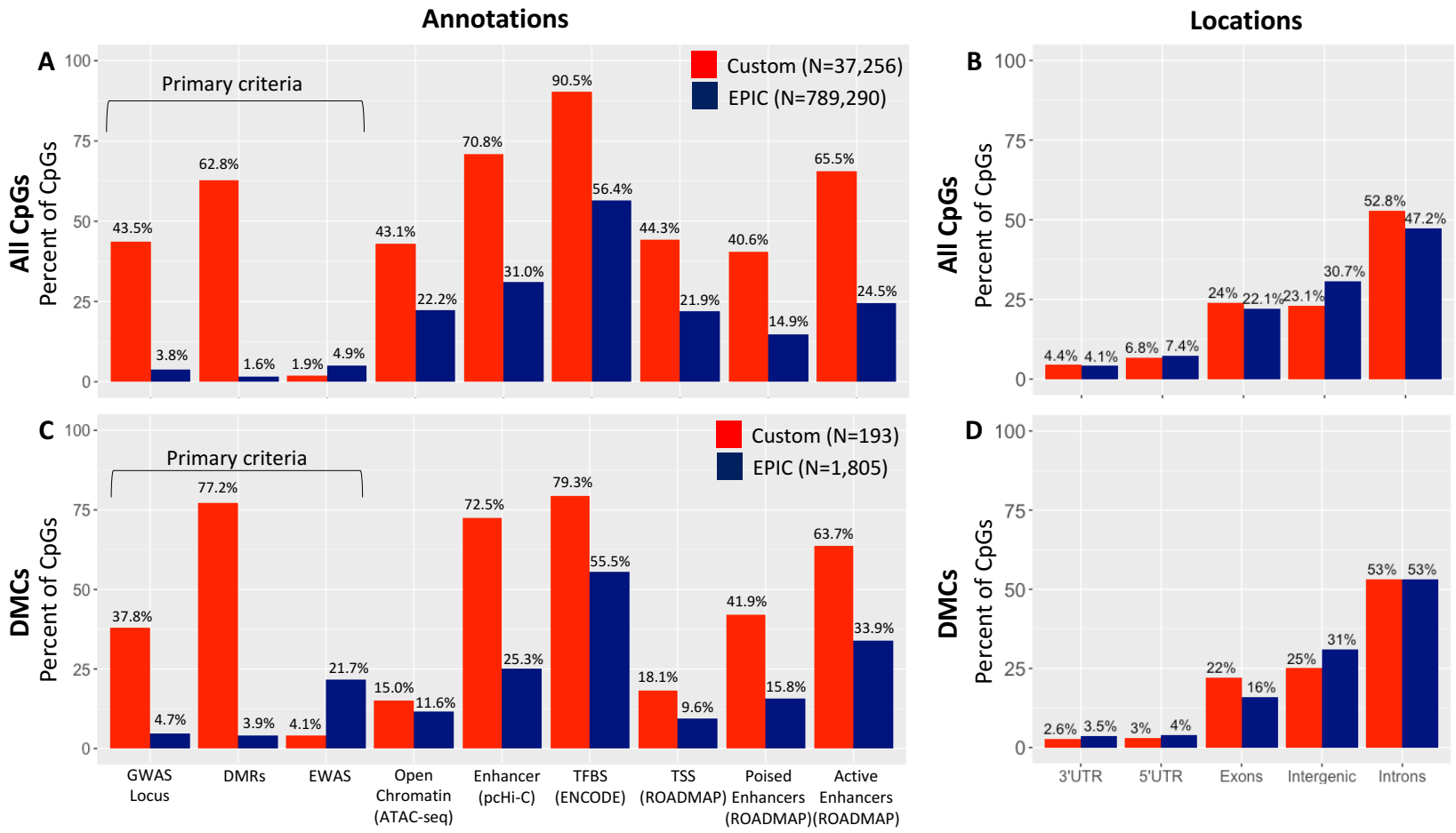

Figure 2. Proportions of Custom and EPIC CpGs by primary criteria, functional annotation category, and genomic location. All comparisons were performed with a Fisher's Exact Test. **(A)** Proportions of Custom CpGs that passed processing QC in URECA on each array by three primary criteria and six functional annotation categories. CpGs on the Custom array were significantly enriched

($p<2.2x10^{-16}$) in all categories compared to the EPIC array with the exception of prior EWAS, in which they were depleted ($p<2.2x10^{-16}$). (**B**) The distributions of CpGs by genomic location. Compared to the EPIC, CpGs on the Custom array were enriched in introns and exons and depleted in intergenic regions and 5'UTRs ($p<10^{-4}$), but did not differ in any other categories. (**C**) Proportions of Custom and EPIC DMCs from each array by three primary criteria and six functional annotation categories. Compared to all CpGs, DMCs from both arrays were depleted at transcription start sites and in areas of open chromatin (Custom $p=1.14x10^{-7}$ and $p=4.11x10^{-9}$; EPIC $p<2.2x10^{-16}$ for both, respectively). DMRs were marginally enriched for DMCs compared to all CpGs on the Custom array ($p=0.066$) but significantly enriched on the EPIC array ($p=1.54x10^{-6}$). DMCs in prior EWAS studies of asthma and allergic diseases were modestly enriched on the Custom array (FET $p=0.059$) and significantly enriched on the EPIC array ($p<2.2x10^{-16}$). In contrast, CpGs at GWAS loci for asthma and allergic diseases were not enriched among DMCs on the Custom array ($p=0.32$) and only modestly enriched on the EPIC array ($p=0.042$). (**D**) The distribution of DMCs by genomic location. The distribution of DMCs on both arrays did not differ from the distributions of all CpGs.

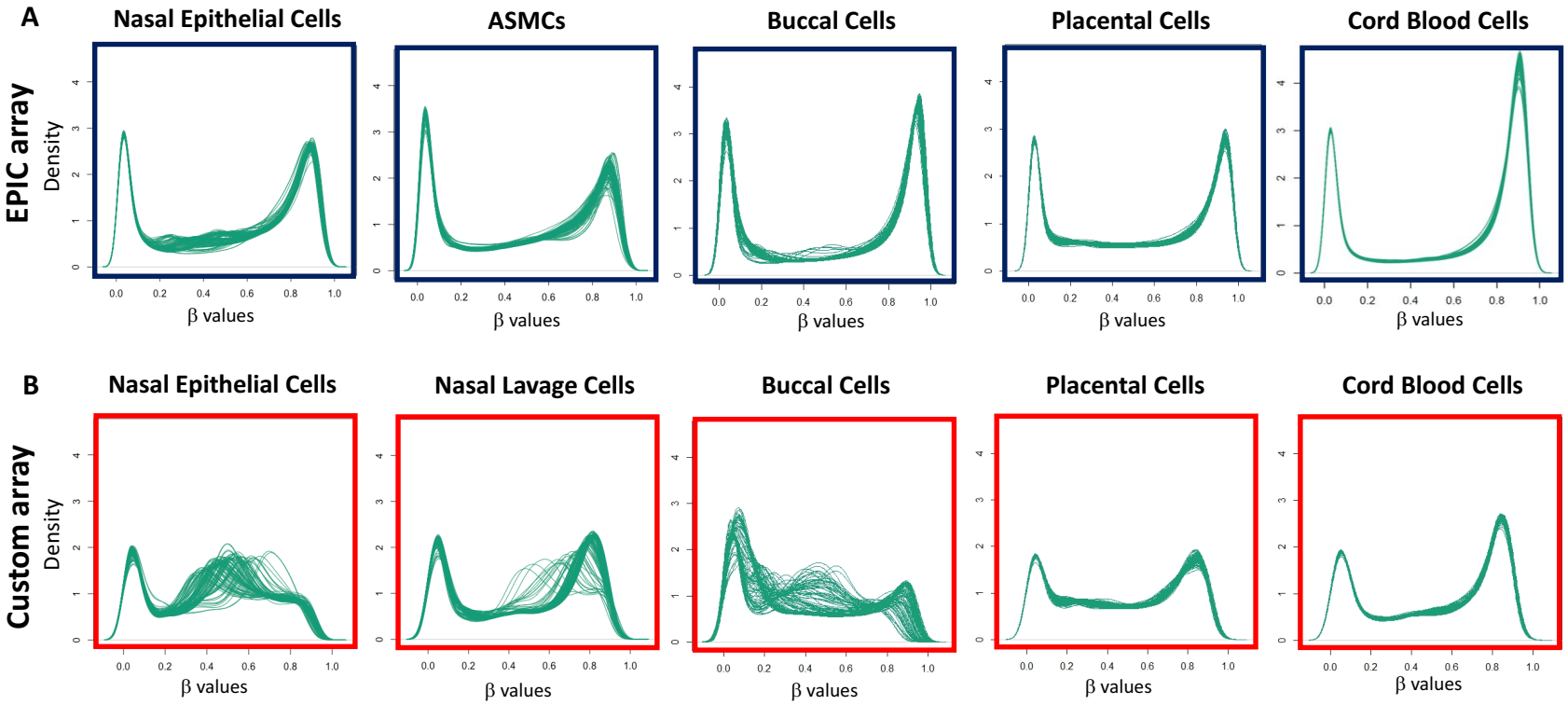

Figure 3. Cross-tissue comparisons of methylation levels on the EPIC and Custom arrays. Density distribution plots of methylation levels, measured as β values, in 96 individuals after quantile (Custom) or quantile + SWAN (EPIC) normalizations (see **Methods**). The x-axis shows the proportion of methylation at CpG sites on the EPIC (**A**) and Custom (**B**) arrays for each individual.

All DNA methylation data were processed using the same pipeline. The nasal epithelial cell DNA (EPIC and Custom) and nasal lavage cell DNA (Custom) were from the same randomly selected 96 URECA children with both the EPIC and Custom arrays in nasal epithelial cells. The buccal, placenta, and cord blood cells were from infants in the VCSIP Study (see **Methods**).

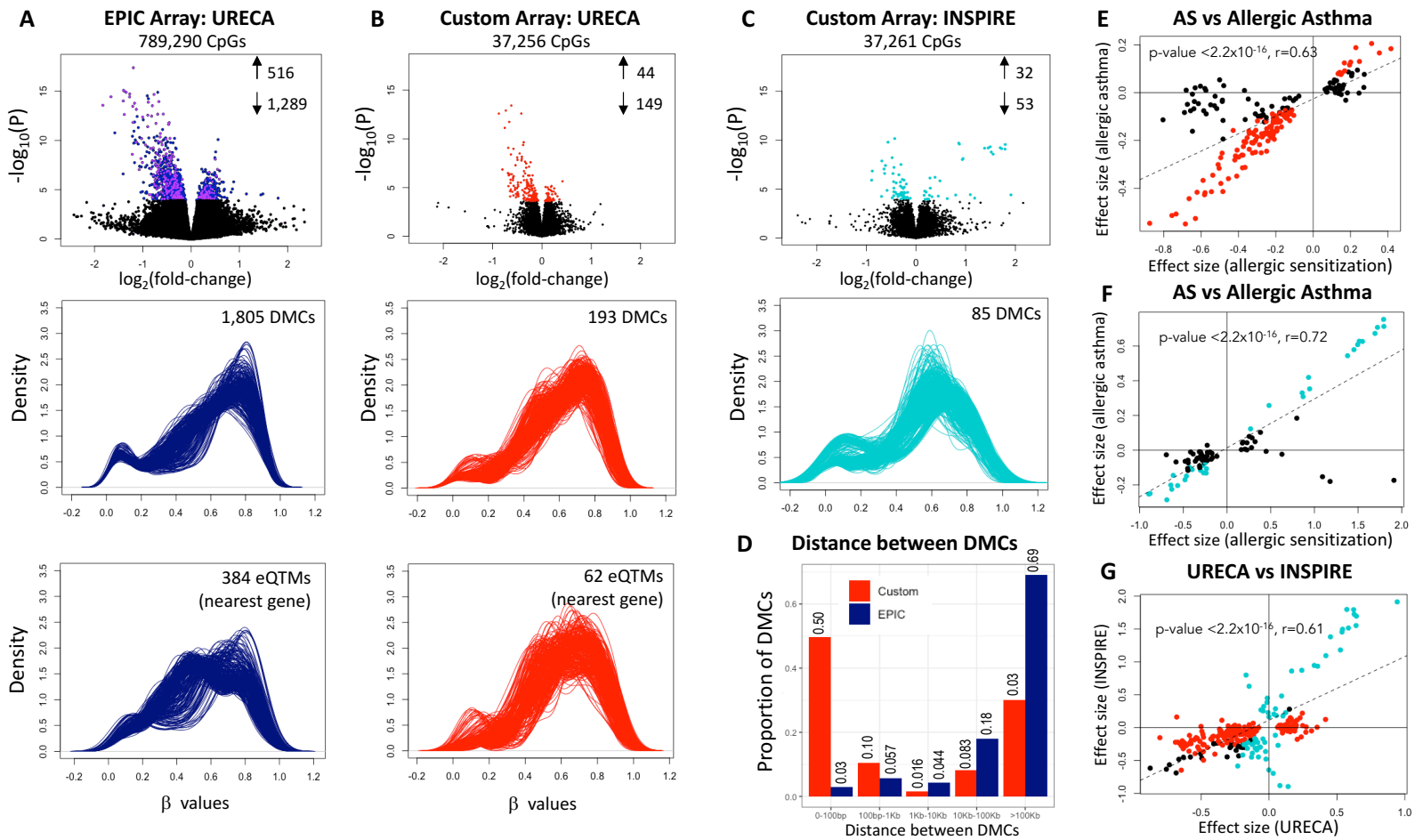

Figure 4. Asthma EWAS results in URECA and INSPIRE children using the EPIC and Custom arrays. (A-C) Upper panel: Volcano plots showing the $\log_2$-fold change in methylation proportion (M-values) with and without allergic sensitization (x-axis) and the $-\log_{10}$(P-value) from the EWAS (y-axis). Significant DMCs at a q-value threshold of 0.05 are shown are shown in blue or purple

(EPIC and filtered EPIC, respectively) in URECA, in red (Custom) in URECA or turquoise (Custom) in INSPIRE. The number of DMCs that were hypermethylated or hypomethylated in the children with AS are shown as up and down arrows, respectively in the figures. **(A-C)** Middle panel: Density plots of the DMCs in each EWAS. (**A-B**) Lower panel: Density plots of CpGs that are eQTMs for their nearest genes. **D**) Distributions of DMCs to the next nearest DMC on the Custom (red) and EPIC (blue) arrays. The distance to nearest DMC is shown on the x-axis; the proportion of DMCs in each distance bin is shown on the y-axis. (**E-F**) Correlation of effect sizes (log fold change) of AS DMCs and allergic asthma in URECA and INSPIRE. The effect size of DMCs for AS (x-axis) and allergic asthma (y-axis) are shown. Red or turquoise dots were associated with allergic asthma at a q-value threshold of 0.05 in URECA or INSPIRE, respectively; black dots are AS-only DMCs. (**G)** Correlation of effect sizes (log fold change) between DMCs identified in the INSPIRE AS EWAS (x-axis) and the URECA AS EWAS (y-axis). Red or turquoise points are AS DMCs only in URECA or INSPIRE, respectively, at a q-value threshold of 0.05; black dots are AS DMCs in both at a q-value threshold of 0.05. The dashed lines in E-G are the correlation lines.