A Functional Genomics Pipeline to Identify High-Value CpG Dinucleotides 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

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

This PDF file includes:

Main Text
Figures 1 to 4

Abstract

DNA methylation of cytosines at CpG dinucleotides is a widespread epigenetic mark; but genome-wide variation has been relatively unexplored due to the limited representation of variable CpGs on commercial high-throughput arrays. To explore this hidden portion of the epigenome, we combined whole-genome bisulfite sequencing with *in silico* evidence of function to design a custom array of high-value CpGs. We focused these studies in airway epithelial cells from children with and without allergic asthma because these cells mediate the effects of inhaled microbes, pollution and allergens on asthma and allergic disease risk. Epigenome-wide association studies (EWAS) of allergic sensitization revealed enrichments for CpGs with intermediate methylation levels that were more likely to be correlated with the expression of their target genes and associated with allergic sensitization in children of diverse ancestries compared to CpGs on commercial arrays. Overall, our study revealed signature features of functional CpGs and unveiled a wealth of DNA methylation variation at functional CpG sites that have been ignored in previous EWAS.

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

97 DNA methylation has led to an explosion of epigenome-wide association studies (EWAS) (4). The
98 most used commercial array, the InfinityMethylationEPIC BeadChip (Illumina, Inc., San Diego,
99 CA), interrogates up to 850,000 CpGs, but this comprises <5% of CpGs in the human genome.
100 Because the selection of CpGs on the EPIC array was agnostic with respect to disease or tissue
101 types, findings to date may represent just the tip of the iceberg with an abundance of yet
102 unknown functional CpG sites contributing to disease risk.

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

109
110
111 **Results**
112
113 **Identifying Differentially Methylated Regions in Whole-Genome Bisulfite Sequences**

114 We performed whole-genome bisulfite sequencing (WGBS) in airway epithelial cell DNA from 20
115 African American children (10 with allergic asthma, 10 without asthma or allergies; 11 years old)
116 from the Urban Environment and Childhood Asthma (URECA) cohort (7) and 19 European
117 American young adults (9 with allergic asthma, 10 without asthma or allergies; 18-20 years old)
118 from the Childhood Origins of Asthma (COAST) cohort (8). After quality control (QC) (see
119 **Methods**), analyses for differentially methylated regions (DMRs) between the asthma/allergy
120 cases and non-asthma/non-allergy controls were performed in the African American sample, the

121 European American sample, and the combined sample (see **Methods** for additional details).
122 Overall, we identified 16,611 DMRs that included 199,473 CpGs (**Supplementary Table 1**).

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

124 Our goal was to develop a custom array with 50,000 probes for CpGs that would complement
125 and serve as a "booster" for the EPIC array for studies of asthma and allergic diseases. Our
126 selection criteria involved three steps (**Supplementary Fig. 1**; also see **Methods**). Briefly, we first
127 identified a set of CpGs based on prior evidence of association with asthma or allergic diseases
128 (atopic dermatitis/eczema, allergic rhinitis/hay fever, and food allergy) from three categories of
129 studies. The first category included the 199,473 CpGs within DMRs from the WGBS study
130 described above, the second included 19,057 CpGs associated with asthma or allergic diseases in
131 17 previous EWAS (**Supplementary Table 2**), and the third included 570,350 CpGs at 140 GWAS
132 loci (**Supplementary Table 3**). We then removed CpGs on the EPIC array, in ENCODE blacklisted
133 regions (9), and those that overlapped with a common single nucleotide polymorphism (SNP) in
134 1000 Genomes European (CEU) or African (YRI) populations. To further prioritize the remaining
135 696,225 CpGs, in the second step we considered their overlap with functional annotations,
136 requiring that CpGs within DMRs overlap with at least three annotations or other category of
137 prior evidence, CpGs from prior EWAS overlap with at least one annotation or other category of
138 prior evidence, and CpGs at GWAS loci overlap with at least four annotation categories or other
139 category of prior evidence. After removing duplicates, 92,024 high-value CpGs remained; 53,700
140 passed the Illumina array design step and, in the third step, we pruned to 44,047 by selecting
141 CpGs from all categories with the highest number of functional annotations. After manufacture
142 and array QC, 37,863 CpGs targeted by 42,192 probes remained (**Supplementary Table 4**).

143 The distributions of CpGs on the Custom and EPIC arrays are shown by annotation
144 category in **Figure 2A**. Of note, 90% of CpGs on the Custom array overlapped with a
145 transcription factor binding site (TFBS), and 94% overlapped with a predicted enhancer,
146 compared to 56% and 54% of the EPIC array CpGs, respectively (Fisher exact test [FET] $p<10^{-16}$
147 for both comparisons). Moreover, compared to EPIC CpGs, Custom CpGs were enriched in
148 introns (52.8% vs. 47.2%) and exons (24.0% vs. 22.1%) and depleted in intergenic regions (23.1%
149 vs. 30.7%) and 5'UTRs (6.8% vs. 7.4%) (FET $p<10^{-4}$ for all comparisons) **(Figure 2B**). The CpGs on
150 the Custom array proportionally represented DMR-CpGs from the three DMR studies in WGBS
151 and from the three categories of prior evidence (**Supplementary Fig. 2**).

152 **Evaluating methylation patterns of CpGs on the Custom Array**

153 To assess the reproducibility of methylation measures of CpGs on the Custom array, we
154 compared β values in airway epithelial cell DNA in the WGBS to the same CpGs on the Custom
155 and the EPIC arrays for the subset of individuals with all three (19 URECA subjects) or two (17
156 COAST subjects) measures. We included CpGs with at least 10 WGBS reads that overlapped with
157 an array CpG that passed QC. This resulted in 502,229 CpGs for the WGBS and EPIC array
158 comparison in URECA and 24,744 and 20,861 CpGs for the WGBS and Custom array comparison
159 in URECA and COAST, respectively. The β distribution plots were similar and highly correlated
160 between CpGs measured by both WGBS and the arrays (Spearman's $p<2.2\times10^{-16}$ for each)
161 (**Supplementary Fig. 3**). Notably, however, compared to the EPIC array, CpGs on the Custom
162 array were depleted for hypermethylated CpGs and enriched for intermediate methylated (IM)
163 CpGs (β values 20-80%).

164 To more broadly assess EPIC and Custom array DNA methylation patterns, we compared
165 the distributions of CpG β values across different tissues using available data for the EPIC array
166 in DNA from airway epithelial cells (this study), airway smooth muscle cells (10), and buccal
167 swabs, placenta cells, and cord blood (11), and for the Custom array in DNA from nasal epithelial
168 cells and nasal lavage cells from URECA children, and buccal cells, cord blood cells, or placenta
169 tissue from infants in the Vitamin C to Decrease the Effects of Smoking in Pregnancy on Infant
170 Lung Function (VCSIP) cohort (11). The β distributions were similar across cell sources for CpGs
171 on the EPIC array (**Figure 3A**) but showed varying patterns between cell types on the Custom
172 array (**Figure 3B**). Whereas an average of 68% (range: 60-79%) of CpGs on the EPIC array were
173 either hypomethylated (0-20%) or hypermethylated (80-100%) in all six cell types, most CpGs on
174 the Custom array (66% in nasal epithelial cells and 52% on average across cell types) were IM
175 CpGs with depletions of β values between both 0-10% (24% in nasal epithelial cells and 29% on
176 average) and 90-100% (10% in nasal epithelial cells and 20% on average). In fact, the β value
177 distributions of CpGs on the EPIC array in nine GTEx tissues are remarkably similar to those
178 reported here (**Supplementary Fig. 4**) (12). We confirmed that differences between the two
179 arrays were attributable to the selection criteria we used for designing the array. Using the same
180 criteria, high-value CpGs on the EPIC array also showed a depletion of both hypomethylated and
181 hypermethylated CpGs and an enrichment for IM CpGs (**Supplementary Fig. 5**).

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

184 Our ultimate goal was to develop a DNA methylation array that could detect asthma- or allergy-
185 associated differential methylation that is missed by the content on the EPIC array. Allergic

186 sensitization (AS) is an IgE response to allergens (13) and a crucial step in the development of
187 allergic diseases and common phenotypes of asthma (14-16). We conducted an EWAS of AS in
188 airway epithelial cell DNA collected from 280 11-year old URECA children using the Custom and
189 the EPIC arrays. We used the percent of positive skin prick tests (14 airborne and oral allergens)
190 as a quantitative measure of AS (see **Methods**). The demographic and clinical description of the
191 URECA children is shown in **Supplementary Table 5** and **Supplementary Fig. 6**.

192 Following QC of the array (see **Methods**), we performed two EWAS of AS in the same
193 URECA children using CpGs on the EPIC or Custom array. In each analysis, we included as
194 covariates sex, percent epithelial cells, the first 3 ancestry principal components (PCs), and
195 latent factors to adjust for additional unwanted variation (17, 18) (see **Methods**). Using a q-
196 value threshold of 0.05, the EWAS of AS revealed 1,805 differentially methylated CpGs (DMCs)
197 using the EPIC array and 193 DMCs using the Custom array (**Figure 4A-B**), revealing an
198 enrichment for AS DMCs on the Custom compared to the EPIC array (0.50% vs. 0.23% of CpGs on
199 each array, respectively; FET $p<2.2x10^{-16}$). Among the DMCs on the EPIC array, 295 (16%) were
200 high-value EPIC CpGs, a significant enrichment compared to 3.5% of high-value CpGs among all
201 CpGs on the EPIC array (FET $p<2.2x10^{-16}$). In contrast to the different β distributions of CpGs on
202 each array (**Figure 3A**), the β distributions of DMCs from the two arrays were similar (**Figure 4A-**
203 **B**; middle panels), showing a near complete depletion of both hypomethylated and
204 hypermethylated CpGs and an enrichment of IM CpGs among AS DMCs in airway epithelial cells.

205 The DMCs on both arrays were distributed across the autosomes (**Supplementary Fig.**
206 **7A-B**), however, DMCs on the Custom array showed spikes of association signals compared to
207 the sparse distribution of DMCs on the EPIC array. In fact, 50% of DMCs on the Custom array are
208 within 100bp of the next nearest DMC compared to only 3% of DMCs on the EPIC array (**Figure**

209 **4D**). Regional clustering of DMCs on the Custom array provides greater confidence in and
210 internal validation of EWAS results.

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

221 Because AS is an important step in the development of childhood asthma, we next
222 asked whether AS DMCs were also associated with allergic asthma. Considering only the AS
223 DMCs on the EPIC array, 1,155 (64%) were also associated with allergic asthma at q-value <0.05
224 (1,805 tests). Among the AS DMCs on the Custom array, 115 (60%) were also associated with
225 allergic asthma at q-value <0.05 (193 tests). The effect sizes were significantly correlated
226 between AS DMCs and allergic asthma DMCs (r=0.68). Among the CpGs that were DMCs for
227 both AS and allergic asthma, all showed the same direction of effect; among all AS DMCs, 93%
228 showed the same direction of effect with allergic asthma (r=0.84) (**Figure 4E**).

229 **Validating the Custom Array in the INSPIRE Cohort**

230 The URECA cohort is diverse with respect to ancestry but includes <1% non-Hispanic White
231 children (7). Therefore, to both replicate results of the EWAS in URECA children and assess the
232 performance of the Custom array in a primarily non-Hispanic White population, we studied 5- to
233 7-year-old children from the Infant Susceptibility to Pulmonary Infections and Asthma Following
234 RSV Exposure (INSPIRE) study (19). INSPIRE infants were enrolled following delivery into this
235 population-based longitudinal study in central Tennessee. An EWAS was performed using the
236 Custom array and nasal epithelial cell DNA from 474 children with measures of AS (proportion of
237 positive SPT to 13 inhaled allergens). The demographic and clinical characteristics of the INSPIRE
238 children are shown in **Supplementary Fig. 6** and **Supplementary Table 5**. See **Methods** for
239 further descriptions of the sample, allergens tested, and processing and QC of the array data.

240 The AS EWAS model in INSPIRE included as covariates sex, parent-reported race or
241 ethnicity, DNA concentration, and latent factors to adjust for unwanted variation (17) (see
242 **Methods**). At a q-value threshold of <0.05 based on 37,261 tests, 85 CpGs on the Custom array
243 were DMCs (0.2% of CpGs). The β distribution of the DMCs (**Figure 4C**) and the spikes of
244 associations (**Supplementary Fig. 7**) were similar to those observed in URECA. Among the AS
245 DMCs in INSPIRE, 38% were also allergic asthma DMCs (q-value<0.05 based on 85 tests), all with
246 the same direction of effect. Among all AS DMCs in INSPIRE, 93% had the same direction of
247 effect with allergic asthma (**Figure 4F)**. Moreover, among the AS DMCs in URECA, 25 (13%) were
248 also AS DMCs in INSPIRE, which was a highly significant enrichment ($p<2.2x10^{-16}$), and all had
249 concordant directions of effect. Finally, the effect sizes of the 253 CpGs that were DMCs in
250 either the URECA or INSPIRE AS EWAS were highly correlated ($r=0.61$; $p<2.2x10^{-16}$) (**Figure 4G**).
251 The high reproducibility of EWAS results in two cohorts demonstrates that the high-value CpGs

252 on the Custom array identified AS DMCs that are robust to ancestry, ascertainment strategies,
253 age at sampling, and geography.

254 Regional association plots combining results of the three AS EWAS for the 10 most
255 significant DMCs in the EPIC and Custom EWAS in URECA are presented in **Supplementary**
256 **Materials**, along with three exemplary regions and their associations with gene expression
257 (**Supplementary Materials**, pages 4-10).

258 **Evaluating the Biological Significance of DMCs in Airway Epithelial Cells**

259 Because nearly all CpGs on the Custom array were selected to overlap with a TFBS or enhancer
260 mark, we expected these CpGs to be correlated with the expression of genes in airway epithelial
261 cells more often than CpGs on the EPIC array (20). To test this, we used gene expression data
262 (16) in the same cells as those used for DNA methylation studies in 249 of the URECA children
263 and defined two sets of genes from among the 15,551 genes detected as expressed in these
264 cells: the nearest gene to each CpG and the promoter capture Hi-C (pcHi-C)-defined target gene
265 in airway epithelial cells (see **Methods**). The latter identifies putative enhancers that physically
266 interact with target gene promoters and account for the 3-dimensional structure of the genome
267 that allows for distal regulatory elements to interact with and regulate the activity of promoters
268 over large distances (e.g., up to 1 Mb or more), which are often not the nearest gene. Among
269 the nearest expressed genes to DMCs on the Custom (98 genes) and EPIC (1,449 genes) arrays,
270 63 were the nearest gene to CpGs on both arrays. Among the pcHi-C target genes for DMCs on
271 the Custom (245 genes) or EPIC (1,155 genes) arrays, 95 were target genes of CpGs on both
272 arrays. Thus, although no CpGs overlapped between the arrays, 101 of nearest or pcHi-C target
273 genes were shared on both arrays (**Supplementary Table 6**). 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" (**Supplementary Table 7**). 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 of their nearest and pcHi-C target gene(s) to identify expression quantitative methylation (eQTM) CpGs. 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 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) (**Supplementary Table 8**). 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 **Supplementary Fig. 8**).

294 Examples of correlations between methylation and gene expression levels for DMCs are
295 shown in **Supplementary Materials** (pages 4-10) for three representative EWAS loci.
296
297 **Discussion**
298
299 The epigenome plays a critical role in regulating gene expression in a context-specific manner,
300 such as in specific cell types or in the presence of environmental exposures or disease states.
301 DNA methylation, for example, is responsive to disease promoting exposures in both in *in vitro*
302 cell models (10, 21-23) and *ex vivo* cells from individuals who are healthy and with disease (1,
303 24-27). Yet the DNA methylome has been underexplored and largely limited to the CpGs on
304 commercial arrays. As a result, it is unknown whether the 800,000 or fewer CpGs interrogated in
305 nearly all EWAS studies to date include CpGs most relevant to any specific exposure, disease, or
306 cell type. Here, we addressed this question by designing a custom Allergy&Asthma DNA
307 methylation array with 37,863 high-value CpGs that are not included on the commercial EPIC
308 array. Our study design allowed us to make important observations regarding features of high-
309 value CpGs and to propose a pipeline that can be used to prioritize CpGs from among the more
310 than 28 million in the human genome in future studies.

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

319 intermediate levels of methylation are also more likely to be associated with AS, an important
320 clinical phenotype that reflects both an immune response to past allergen exposures (13) and a
321 risk factor for the development of asthma and allergic diseases (14-16). Indeed, AS DMCs were
322 highly enriched for allergic asthma DMCs. Although the CpGs on the custom array were overall
323 enriched for IM CpGs, AS DMCs were further enriched, with a near complete depletion of very
324 hypomethylated (0-20% methylation) and very hypermethylated (80-100%) CpGs. This
325 important observation was further supported by EWAS results using the EPIC array. Whereas
326 EPIC array CpGs are enriched for hypomethylated and hypermethylated CpGs in all tissues, the
327 EPIC DMCs were depleted for CpGs at both extremes and enriched for IM CpGs, similar to the
328 DMCs on the custom Allergy&Asthma array. We further showed that the 26,905 CpGs on the
329 EPIC array that passed our filtering pipeline were also enriched for IM CpGs and enriched among
330 AS DMCs: 16.3% of the DMCs in the AS EWAS were among the high-value EPIC CpGs compared
331 to 3.4% of all CpGs on the EPIC array. These results, combined with previous studies (20, 28),
332 could justify filtering CpGs on the EPIC array to include just IM CpGs (β=20-80%) to enrich for
333 functional CpGs and increase power to detect DMCs associated with exposure or disease
334 outcomes. Our data support this suggestion. Using all CpGs on the EPIC array and a q-value
335 threshold of 0.05, we identified 1,805 DMCs (0.23% of all CpGs). However, when we include only
336 IM CpGs in the EWAS, we identified 2,182 DMCs (0.70% of all IM CpGs), more than three times
337 the proportion of DMCs among tested CpGs. Finally, DMCs that were eQTMs were further
338 enriched for IM CpGs (**Figure 4** and **Supplementary Fig. 8**), supporting their role in gene
339 regulation.

340 In addition to the enrichment of IM CpGs on the Custom compared to the EPIC array,
341 the patterns of associations with AS on the two arrays were notable. Whereas DMCs on the EPIC

342 are more sparsely distributed throughout the genome, the DMCs on the Custom show spikes of
343 association, analogous to GWAS peaks. The spatial organization among DMCs on the Custom
344 provides confirmatory evidence for their veracity, which is not available for CpGs on the EPIC
345 array (**Figure 4D**). Another distinguishing feature of the DMCs on the Custom array were their
346 enrichment in exons and depletions in intergenic regions compared to DMCs on the EPIC array.
347 This is consistent with results of an earlier study that observed enrichments of IM CpGs in exons,
348 where exon expression at those sites correlated with DNA methylation (20).

349 The CpGs included on the custom Allergy&Asthma array were selected based on prior
350 evidence of associations with asthma or allergic diseases using three classes of evidence: within
351 a DMR in our WGBS study, in a previous published EWAS, or at published GWAS loci. Overall,
352 62% of CpGs on the custom array were selected from DMRs compared to 1.6% of all CpGs on
353 the EPIC array, yet the DMR CpGs were enriched among DMCs on both the custom (77.2%) and
354 EPIC (3.9%) arrays. Because we excluded all CpGs on the EPIC from the Custom array, few
355 previous EWAS CpGs were on the Custom array, although we still observed a modest
356 enrichment of EWAS CpGs among DMCs on the Custom array (1.9% overall vs. 4.1% of DMCs)
357 and a highly significant enrichment among DMCs on the EPIC array (4.9% overall vs. 21.7% of
358 DMCs). These represent replications of results from previous EWAS of asthma and allergic
359 phenotypes. In contrast to these two categories of prior evidence, CpGs at asthma and allergy
360 GWAS loci were very modestly enriched among EPIC array DMCs but not enriched, and even
361 slightly depleted, among the Custom array DMCs. In a WGBS in 25 different cell and tissue types,
362 Elliott et al. (20) showed that IM CpGs were more likely to be allele independent than to show
363 allele-specific methylation. Busche et al. (28) conducted a WGBS study in adipose and blood cells
364 from twins and concluded that non-genetic effects were mainly tissue specific and located in

365 gene regions, consistent with the features of CpGs on the Custom array, whereas CpGs with
366 genetic effects were shared across tissues and more likely to be intergenic, consistent with
367 features of CpGs on the EPIC array. They concluded that non-shared environments accounted
368 for most of the variance in methylation levels. Our findings of a lack of enrichment of DMCs at
369 GWAS loci, but enrichments for IM CpGs and for AS DMCs in exons suggest that inter-individual
370 variation for IM CpGs on the Custom array may be the set of more plastic CpGs in the human
371 genome and better reflect the effects of environmental exposures than of genetics.
372 Nonetheless, our study still revealed examples of AS DMCs at important GWAS loci (e.g., *IRF1* on
373 chromosome 5, *CLEC16A* on chromosome 16, and *HDAC7*/*VDR* on chromosome 12)
374 (**Supplementary Results**; pages 6-9).

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

385
386
387 **Materials and Methods**
388
389 *Cohorts Included in WGBS or EWAS Studies*

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

394 The URECA study is an observational birth cohort study initiated in 2005 in Baltimore,
395 Boston, New York City, and St. Louis under the NIAID-funded Inner City Asthma Consortium (7).
396 Either the pregnant mother or the father of their unborn child had a history of asthma, allergic
397 rhinitis, or eczema. Asthma was assessed at age 10 according to a definition that considered
398 symptoms, diagnosis by a health care provider, and measurements of pulmonary function, as
399 described (16). Skin prick testing was performed at age 10 (255 subjects) or age 7 (25 subjects)
400 and included the following allergens: mouse epithelia, dog epithelia, *Dermatophagoides fainae*
401 (mite), *Dermatophagoides pteronyssinus* (mite), cat hair, rat epithelia, American/German
402 cockroach mix, German cockroach, *Alternaria tenuis* (mold), *Aspergillus* mix, ragweed mix, tree
403 pollen (oak or birch), *Penicillium Notatum/Pennicillium Chrysogenum*, and Timothy grass. Nasal
404 brushings were obtained at age 11 years, as described (16). Twenty African American children
405 were selected for the WGBS studies (10 with asthma and allergic disease [3 females, 7 males],
406 10 without asthma or allergic disease [3 females, 7 males]), and 280 unrelated children were
407 included in the EWAS using both the Custom and EPIC arrays. The characteristics of the 280
408 URECA participants are described in **Supplementary Table 5**.

409 The COAST study is an observational birth cohort study initiated in 1998 in Madison,
410 Wisconsin (8). Pregnant women were recruited in the third trimester of pregnancy if the mother
411 or the father had a history of asthma or allergic diseases. Asthma was assessed beginning at age
412 6 years. Children were diagnosed with asthma if they fulfilled at least one of the following

413 criteria: (1) physician-diagnosed asthma; (2) frequent albuterol use for coughing or wheezing
414 episodes as prescribed by a physician; (3) use of a prescribed daily controller medication; (4) an
415 implemented step-up plan, including use of albuterol or inhaled corticosteroids during illness as
416 prescribed by a physician; (5) use of prednisone for an asthma exacerbation. Nasal brushings
417 were obtained at ages 18-20 years, as described (29). Twenty participants of European American
418 ancestry were selected for the WGBS studies (10 with asthma and allergic disease [5 males, 5
419 females], 10 without asthma or allergic disease [5 males, 5 females]).

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

430 *Cohorts included in the Cross-Tissue Comparison Studies*

431 In addition to studying nasal epithelial cell DNA with the Custom array in URECA, we included
432 results from studies of the Custom array in DNA from nasal lavage cells (n=96), buccal cells
433 (n=96), placenta (n=96) and cord blood (96). The nasal lavage cell DNA was isolated from URECA
434 subjects around age 13 (mean = 13.4 years, range = 12.8 - 14.9 years) using the QIAamp DNA
435 Micro Kit. The buccal, placenta and cord blood cell DNA was collected from 96 participants in

436 the VCSIP cohort (11). Placental DNA was extracted from powdered tissue under liquid nitrogen,
437 using the QIAcube for automated nucleic acid extraction. Cord blood DNA was extracted from
438 200μL of whole blood using the QIAamp DNA Mini Kit and the QIAcube, and buccal DNA was
439 extracted from buccal swabs following proteinase K digestion using the Maxwell 16 Blood DNA
440 Extraction Kit and the Maxwell 16 Instrument for automated nucleic acid extraction.

441 *Whole Genome Bisulfite Sequencing and DMR Studies*

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

457 DMR analyses were conducted in the African American sample (n=20), the European
458 American sample (n=19), and the combined African American and European American sample
459 (n=39). The methylation data were then smoothed using BSmooth (33), and DMRs were called
460 using the bsseq package (version 1.14) (33). Only CpGs covered by at least 10 reads in 80% of
461 the cases and controls (AA only, EA only, and combined n=39) were included. T-statistics cutoffs
462 were based on 5% quantiles, and a maximum gap of 300bp was required between CpGs to
463 define a cluster, as recommended by BSmooth. To maximize the number of DMRs, we required
464 three or more CpGs per DMR and a minimum of 5% difference in methylation levels between
465 the allergy asthma cases and non-allergy, non-asthmatic controls. The union of DMRs between
466 analyses was assessed using the *reduce* function from the GenomicRanges R package (version
467 1.30).

468 *Selection of CpGs for Custom Array*

469 We first identified CpGs with a high likelihood of being associated with asthma and allergic
470 disease (**Supplementary Fig. 1**). In the first step, we prioritized regions with prior evidence of
471 association with asthma or allergic diseases (atopic dermatitis/eczema, allergic rhinitis/hay
472 fever, and food allergy) from three categories of studies. The first category included the 199,473
473 CpGs within the DMRs from the WGBS. The second category included CpGs from previous DNA
474 methylation studies (EWAS) of asthma or allergic diseases. For this, we conducted a literature
475 search for array-based studies of DNA methylation in asthma and allergic disease. We identified
476 15 studies that conducted 25 EWAS of asthma- or allergy-related phenotypes, five of which
477 were conducted in respiratory epithelium and 10 in blood (**Supplementary Table 2**). We also
478 included CpG sites for five additional genes from two candidate-gene DNA methylation studies
479 of food allergies; *FOXP3* (34) and *IL4*, *IL5, IL10* and *INFG* (35). In total, 19,057 unique DMCs were

480 identified. The third category included CpGs located within the 140 GWAS loci defined in two
481 recent large studies of adult-onset and childhood-onset asthma (36) and allergic diseases
482 (asthma, hay fever and eczema) (37) in UK Biobank subjects (38), which included nearly all the
483 loci reported in other GWAS in multi-ancestry populations. In total, the GWAS loci covered a
484 total of 570,350 CpG motifs. We also included CpGs for the *MALT1* gene that was significantly
485 associated with peanut allergy in a genome-wide interaction study (39). We next removed
486 duplicate CpGs from among the three categories of prior studies, CpGs on the EPIC array, CpGs
487 in ENCODE blacklist regions (9), and those in which the cytosine nucleotide overlapped with
488 common SNPs (MAF > 5%) in 1000 Genomes CEU or YRI populations (**Supplementary Table 4**). A
489 total of 696,225 CpGs remained for consideration in the second step.

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

502 Of the 92,024 resulting high-value CpGs identified, 53,700 were amenable to the
503 Illumina design algorithm, and 44,047 CpGs targeted by 49,999 probes were selected by us for
504 manufacturing. Fourteen percent of the CpGs failed manufacturing, resulting in a Custom array
505 with 38,541 CpGs targeted by 43,605 probes. After removing probes with unknown origins and
506 those without genomic coordinates, 37,863 CpGs passed array QC. The locations of this final set
507 of CpGs, their selection among the three primary criteria, and their overlap with the annotation
508 categories are shown in Supplementary Dataset 1. Finally, we removed probes that included a
509 common SNP (MAF≥5%) within 3 bp of the CpG interrogation site, probes that failed the minfi
510 QC, and probes on the X chromosome, leaving 37,256 CpGs following both array and processing
511 QC**.**

512 *RNA-sequencing studies in URECA*

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

516 *Estimation of genetic ancestry*

517 Ancestry PCs were estimated in the URECA children using a set of 3,534 SNPs that were
518 genotyped in URECA and in reference panels from the 1000 Genomes Project (1KG; n=156)(31)
519 and the Human Genome Diversity Project (HGDP; n=52). European, West African, and East Asian
520 reference samples were randomly selected from CEU (n=52), YRI (n=52), JPT (n=26), and CHB
521 (n=26) samples, respectively, in the phase 3 1KG reference panel (31). Ancestry PCs were
522 calculated using PC-Air (43). The first two ancestry PCs for the URECA children in our study are
523 shown in **Supplementary Fig. 9**.

524 *Processing Custom and EPIC array methylation data in URECA*

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

531 For the EPIC array, probes that failed (detection $P < 0.01$ in at least 25% of samples),
532 overlapped with known SNPs with MAF of at least 5% in African American or European
533 Americans, mapped to the X or Y chromosomes, overlapped ENCODE blacklist regions, or
534 mapped to multiple locations in a bisulfite-converted genome were removed. Raw probe values
535 were background corrected using preprocessIllumina(bg.correct="TRUE", normalize="no"), and
536 quantile normalization was performed using ENmix (v 1.30.01)
537 (https://doi.org/10.1186/s13148-021-01207-1), followed by SWAN (45) normalization. Samples
538 that failed sex checks using the getSex function in minfi were removed. DNA concentration,
539 collection site, array and plate showed batch effects by principal components analysis (PCA)
540 (46). The effects of collection site, array, and plate were removed using ComBat (47); DNA
541 concentration was removed using linear regression. We estimated unobserved variation in the
542 data correlated with our phenotype of interest (18) and included these latent factors in the
543 model (18). The first three ancestry PCs were also included as covariates to capture the effects
544 of admixture in the sample. After QC, we retained 789,290 (91.1%) of CpGs on the EPIC array for
545 analysis.

546 For the Custom array, probes that failed (detection $P < 0.01$ in at least 25% of samples),
547 contained a SNP with MAF of at least 5% in African American or European Americans within 3bp
548 of the CpG interrogation site, mapped to the X chromosome, or were missing genomic
549 coordinates were removed. Raw probe values were background corrected using
550 preprocessIllumina(bg.correct="TRUE", normalize="no"), and quantile normalization was
551 performed using ENmix. Additional probe-type normalization using SWAN was not applied
552 because the dye intensities for the methylated and unmethylated type I and type II probes had
553 very different distributions. DNA concentration, collection site, array, and plate showed batch
554 effects by PCA. The effects of site, plate, and chip were removed using ComBat, and DNA
555 concentration was removed using linear regression. We estimated unobserved variation in the
556 data that was not correlated with the phenotype of interest and included these latent factors in
557 the model. To account for regional differences in the density of Custom array CpGs on the
558 estimation of latent factors, we used the partition.params function in FALCO (17)
559 (https://github.com/chrismckennan) to partition groups of CpGs into independent units (default
560 size = 100 kb). The first three ancestry PCs were included as covariates to capture the effects of
561 admixture in the sample for analyses in URECA. After QC, we retained 37,256 (98.1%) CpGs on
562 the Custom array for analyses.

563 For comparison purposes, we filtered the CpGs on the EPIC array through the same
564 pipeline described above for selecting CpGs for the Custom array; we refer to these as "high-
565 value" EPIC CpGs. Of the 789,290 CpGs on the EPIC array, 26,905 (3.4%) were considered high-
566 value, similar in number to the 37,863 CpGs that passed QC on the Custom array
567 (Supplementary Dataset 2).

568 *Processing Custom array methylation data in INSPIRE*

569 The same pipeline used in URECA was used to process INSPIRE Custom methylation data.
570 Twenty-one probes failed *P*-value detection and were removed from the analysis. We estimated
571 sex using the X chromosome CpGs and removed five samples with discrepancies between these
572 classifications and reported sex. DNA concentration, collection year, plate, and array were
573 identified as having batch effects by PCA (**Supplementary Table 9**). The effects of collection
574 year, plate, and array were removed using ComBat, and DNA concentration was removed using
575 linear regression. Because DNA concentration was correlated with other variables, it was also
576 included as a covariate in the model, along with sex, parent-reported race/ethnicity, and latent
577 factors estimated using the same methods described above for URECA. A total of 37,261 CpGs
578 passed processing QC and were included in the analysis.

579 *EWAS in the URECA and INSPIRE cohorts*

580 For each EWAS we used the following models:

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

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

583 For the URECA EWAS, we included 14 and 5 latent variables for the Custom (17) and
584 EPIC (18) arrays, respectively. Cell counts (300 cells/slide) were performed on each sample (16).
585 Because we did not have information on cell proportion or genetic ancestry for INSPIRE, we
586 included sex, parent-reported race/ethnicity, DNA concentration, and 21 latent factors as
587 covariates to adjust for cell composition, as well as other unwanted variation in the EWAS.
588 Analyses were performed in R (version 4.1.0) using limma v 3.50.0 (48, 49). To control the false
589 discovery rate, we used a q-value threshold of 0.05. The URECA EWAS sample included 19 of the

590 20 African American children in the WGBS studies and the 96 children shown in the cross-tissue
591 density plots of β distributions (**Figure 3**).

592 To assess the overlap of AS DMCs with allergic asthma DMCs, we tested the AS DMCs for
593 association with allergic asthma in URECA and INSPIRE. For these studies, we defined allergic
594 asthma cases as children with a diagnosis of asthma at age 10 (URECA) or age 6 (INSPIRE) and
595 sensitization to at least one allergen (N=62 and 22, respectively). Controls were defined as
596 children without a diagnosis of asthma (ever) and without any positive skin prick tests (n=76 and
597 271, respectively). We calculated q-values based on the 193 AS DMCs in URECA and the 85 AS
598 DMCs in INSPIRE. The overlaps of AS and allergic asthma cases and controls are shown in
599 **Supplementary Fig. 10**.

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

602 Correlations between methylation levels at each CpG with the expression levels of the nearest
603 gene and the pcHi-C target genes (42) were assessed using linear regression in both the full data
604 set (all CpGs that passed QC on each array) and in the DMCs. Comparisons were only assessed
605 for DMCs falling in "capture end regions" (± 1kb) that interacted with gene promoters (42). Sex,
606 ancestry PCs, and epithelial cell composition were included as covariates in the model. The
607 nearest gene was annotated using the GenomicRanges package v1.46.1 in R with the gene list
608 from the R biomaRt package v2.50.2 using ensemble GRCh37. A false discovery rate of 5% was
609 used to assess significance.

610 *Pathway analysis*

611 Pathway over-representation testing was performed with Advaita Bio's iPathwayGuide

612 (https://www.advaitabio.com/ipathwayguide) using KEGG Pathways (release 100.0+/11-12, Nov

613 2021). For the Custom and the EPIC arrays, the 318 and 2,366 genes, respectively, that were

614 either the nearest or pcHi-C target gene to a DMC, were entered as significant against a

615 background of all nearest genes to CpGs on each arrays that are expressed in NECs (n=14,049).

616 **References**

617 1. M. Prunicki *et al.*, Air pollution exposure is linked with methylation of immunoregulatory
618 genes, altered immune cell profiles, and increased blood pressure in children. *Sci Rep*
619 **11**, 4067 (2021).
620 2. Y. Li *et al.*, DNA methylation and transcriptome aberrations mediated by ERalpha in
621 mouse seminal vesicles following developmental DES exposure. *Proc Natl Acad Sci U S A*
622 **115**, E4189-E4198 (2018).
623 3. E. W. Tobi *et al.*, DNA methylation as a mediator of the association between prenatal
624 adversity and risk factors for metabolic disease in adulthood. *Sci Adv* **4**, eaao4364
625 (2018).
626 4. M. Li *et al.*, EWAS Atlas: a curated knowledgebase of epigenome-wide association
627 studies. *Nucleic Acids Res* **47**, D983-D988 (2019).
628 5. S. T. Holgate, The sentinel role of the airway epithelium in asthma pathogenesis.
629 *Immunol Rev* **242**, 205-219 (2011).
630 6. C. A. Akdis, The epithelial barrier hypothesis proposes a comprehensive understanding
631 of the origins of allergic and other chronic noncommunicable diseases. *J Allergy Clin*
632 *Immunol* **149**, 41-44 (2022).
633 7. J. E. Gern *et al.*, The Urban Environment and Childhood Asthma (URECA) birth cohort
634 study: design, methods, and study population. *BMC Pulm Med* **9**, 17 (2009).
635 8. R. F. Lemanske, Jr., The childhood origins of asthma (COAST) study. *Pediatr Allergy*
636 *Immunol* **13**, 38-43 (2002).
637 9. F. Krueger, S. R. Andrews, Bismark: a flexible aligner and methylation caller for Bisulfite-
638 Seq applications. *Bioinformatics* **27**, 1571-1572 (2011).
639 10. E. E. Thompson *et al.*, Cytokine-induced molecular responses in airway smooth muscle
640 cells inform genome-wide association studies of asthma. *Genome Med* **12**, 64 (2020).
641 11. L. E. Shorey-Kendrick *et al.*, Impact of vitamin C supplementation on placental DNA
642 methylation changes related to maternal smoking: association with gene expression and
643 respiratory outcomes. *Clin Epigenetics* **13**, 177 (2021).

644 12. M. Oliva *et al.*, Genetic regulation of DNA methylation across tissues reveals thousands
645 of molecular links to complex traits. *Research Square* https://doi.org/10.21203/rs.3.rs-
646 492596/v1 (2021).
647 13. R. van Ree, L. Hummelshoj, M. Plantinga, L. K. Poulsen, E. Swindle, Allergic sensitization:
648 host-immune factors. *Clin Transl Allergy* **4**, 12 (2014).
649 14. S. Illi *et al.*, The pattern of atopic sensitization is associated with the development of
650 asthma in childhood. *J Allergy Clin Immunol* **108**, 709-714 (2001).
651 15. D. J. Jackson, J. E. Gern, R. F. Lemanske, Jr., The contributions of allergic sensitization
652 and respiratory pathogens to asthma inception. *J Allergy Clin Immunol* **137**, 659-665;
653 quiz 666 (2016).
654 16. M. C. Altman *et al.*, Endotype of allergic asthma with airway obstruction in urban
655 children. *J Allergy Clin Immunol* 10.1016/j.jaci.2021.02.040 (2021).
656 17. C. McKennan, Factor analysis in high dimensional biological data with dependent
657 observations. *arXiv:2009.11134v1* (2020).
658 18. C. McKennan, D. L. Nicolae, Accounting for unobserved covariates with varying degree
659 of estimability in high dimensional experimental data. *Biometrika* **106**, 823-840 (2019).
660 19. E. K. Larkin *et al.*, Objectives, design and enrollment results from the Infant
661 Susceptibility to Pulmonary Infections and Asthma Following RSV Exposure Study
662 (INSPIRE). *BMC Pulm Med* **15**, 45 (2015).
663 20. G. Elliott *et al.*, Intermediate DNA methylation is a conserved signature of genome
664 regulation. *Nat Commun* **6**, 6363 (2015).
665 21. A. Pacis *et al.*, Bacterial infection remodels the DNA methylation landscape of human
666 dendritic cells. *Genome Res* **25**, 1801-1811 (2015).
667 22. J. Nicodemus-Johnson *et al.*, Genome-wide methylation study identifies an IL-13-
668 induced epigenetic signature in asthmatic airways. *Am J Respir Crit Care Med* **193**, 376-
669 385 (2016).
670 23. M. M. Soliai *et al.*, Multi-omics colocalization with genome-wide association studies
671 reveals a context-specific genetic mechanism at a childhood onset asthma risk locus.
672 *Genome Med* **13**, 157 (2021).
673 24. A. Cardenas *et al.*, Controlled human exposures to diesel exhaust: a human epigenome-
674 wide experiment of target bronchial epithelial cells. *Environ Epigenet* **7**, dvab003 (2021).

675 25. E. S. Wan *et al.*, Cigarette smoking behaviors and time since quitting are associated with
676 differential DNA methylation across the human genome. *Hum Mol Genet* **21**, 3073-3082
677 (2012).
678 26. K. M. Bakulski, J. Dou, N. Lin, S. J. London, J. A. Colacino, DNA methylation signature of
679 smoking in lung cancer is enriched for exposure signatures in newborn and adult blood.
680 *Sci Rep* **9**, 4576 (2019).
681 27. S. Sikdar *et al.*, Comparison of smoking-related DNA methylation between newborns
682 from prenatal exposure and adults from personal smoking. *Epigenomics* **11**, 1487-1500
683 (2019).
684 28. S. Busche *et al.*, Population whole-genome bisulfite sequencing across two tissues
685 highlights the environment as the principal source of human methylome variation.
686 *Genome Biol* **16**, 290 (2015).
687 29. S. Basnet *et al.*, CDHR3 Asthma-Risk Genotype Affects Susceptibility of Airway
688 Epithelium to Rhinovirus C Infections. *Am J Respir Cell Mol Biol* **61**, 450-458 (2019).
689 30. A. Kechin, U. Boyarskikh, A. Kel, M. Filipenko, cutPrimers: A New Tool for Accurate
690 Cutting of Primers from Reads of Targeted Next Generation Sequencing. *J Comput Biol*
691 **24**, 1138-1143 (2017).
692 31. 1000 Genomes Project Consortium *et al.*, A global reference for human genetic
693 variation. *Nature* **526**, 68-74 (2015).
694 32. H. M. Amemiya, A. Kundaje, A. P. Boyle, The ENCODE Blacklist: Identification of
695 Problematic Regions of the Genome. *Sci Rep* **9**, 9354 (2019).
696 33. K. D. Hansen, B. Langmead, R. A. Irizarry, BSmooth: from whole genome bisulfite
697 sequencing reads to differentially methylated regions. *Genome Biol* **13**, R83 (2012).
698 34. L. Paparo *et al.*, Epigenetic features of FoxP3 in children with cow's milk allergy. *Clin*
699 *Epigenetics* **8**, 86 (2016).
700 35. R. Berni Canani *et al.*, Differences in DNA methylation profile of Th1 and Th2 cytokine
701 genes are associated with tolerance acquisition in children with IgE-mediated cow's milk
702 allergy. *Clin Epigenetics* **7**, 38 (2015).
703 36. M. Pividori, N. Schoettler, D. L. Nicolae, C. Ober, H. K. Im, Shared and distinct genetic risk
704 factors for childhood-onset and adult-onset asthma: genome-wide and transcriptome-
705 wide studies. *Lancet Respir Med* **7**, 509-522 (2019).

706 37. A. Johansson, M. Rask-Andersen, T. Karlsson, W. E. Ek, Genome-wide association
707 analysis of 350 000 Caucasians from the UK Biobank identifies novel loci for asthma, hay
708 fever and eczema. *Hum Mol Genet* **28**, 4022-4041 (2019).
709 38. C. Bycroft *et al.*, The UK Biobank resource with deep phenotyping and genomic data.
710 *Nature* **562**, 203-209 (2018).
711 39. A. Winters *et al.*, The MALT1 locus and peanut avoidance in the risk for peanut allergy. *J*
712 *Allergy Clin Immunol* **143**, 2326-2329 (2019).
713 40. Encode Project Consortium *et al.*, Expanded encyclopaedias of DNA elements in the
714 human and mouse genomes. *Nature* **583**, 699-710 (2020).
715 41. Roadmap Epigenomics Consortium *et al.*, Integrative analysis of 111 reference human
716 epigenomes. *Nature* **518**, 317-330 (2015).
717 42. B. A. Helling *et al.*, Altered transcriptional and chromatin responses to rhinovirus in
718 bronchial epithelial cells from adults with asthma. *Commun Biol* **3**, 678 (2020).
719 43. M. P. Conomos, M. B. Miller, T. A. Thornton, Robust inference of population structure
720 for ancestry prediction and correction of stratification in the presence of relatedness.
721 *Genet Epidemiol* **39**, 276-293 (2015).
722 44. M. J. Aryee *et al.*, Minfi: a flexible and comprehensive Bioconductor package for the
723 analysis of Infinium DNA methylation microarrays. *Bioinformatics* **30**, 1363-1369 (2014).
724 45. J. Maksimovic, L. Gordon, A. Oshlack, SWAN: Subset-quantile within array normalization
725 for illumina infinium HumanMethylation450 BeadChips. *Genome Biol* **13**, R44 (2012).
726 46. J. T. Leek *et al.*, Tackling the widespread and critical impact of batch effects in high-
727 throughput data. *Nat Rev Genet* **11**, 733-739 (2010).
728 47. W. E. Johnson, C. Li, A. Rabinovic, Adjusting batch effects in microarray expression data
729 using empirical Bayes methods. *Biostatistics* **8**, 118-127 (2007).
730 48. B. Phipson, S. Lee, I. J. Majewski, W. S. Alexander, G. K. Smyth, Robust Hyperparameter
731 Estimation Protects against Hypervariable Genes and Improves Power to Detect
732 Differential Expression. *Ann Appl Stat* **10**, 946-963 (2016).
733 49. M. E. Ritchie *et al.*, limma powers differential expression analyses for RNA-sequencing
734 and microarray studies. *Nucleic Acids Res* **43**, e47 (2015).
735
736
737

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:

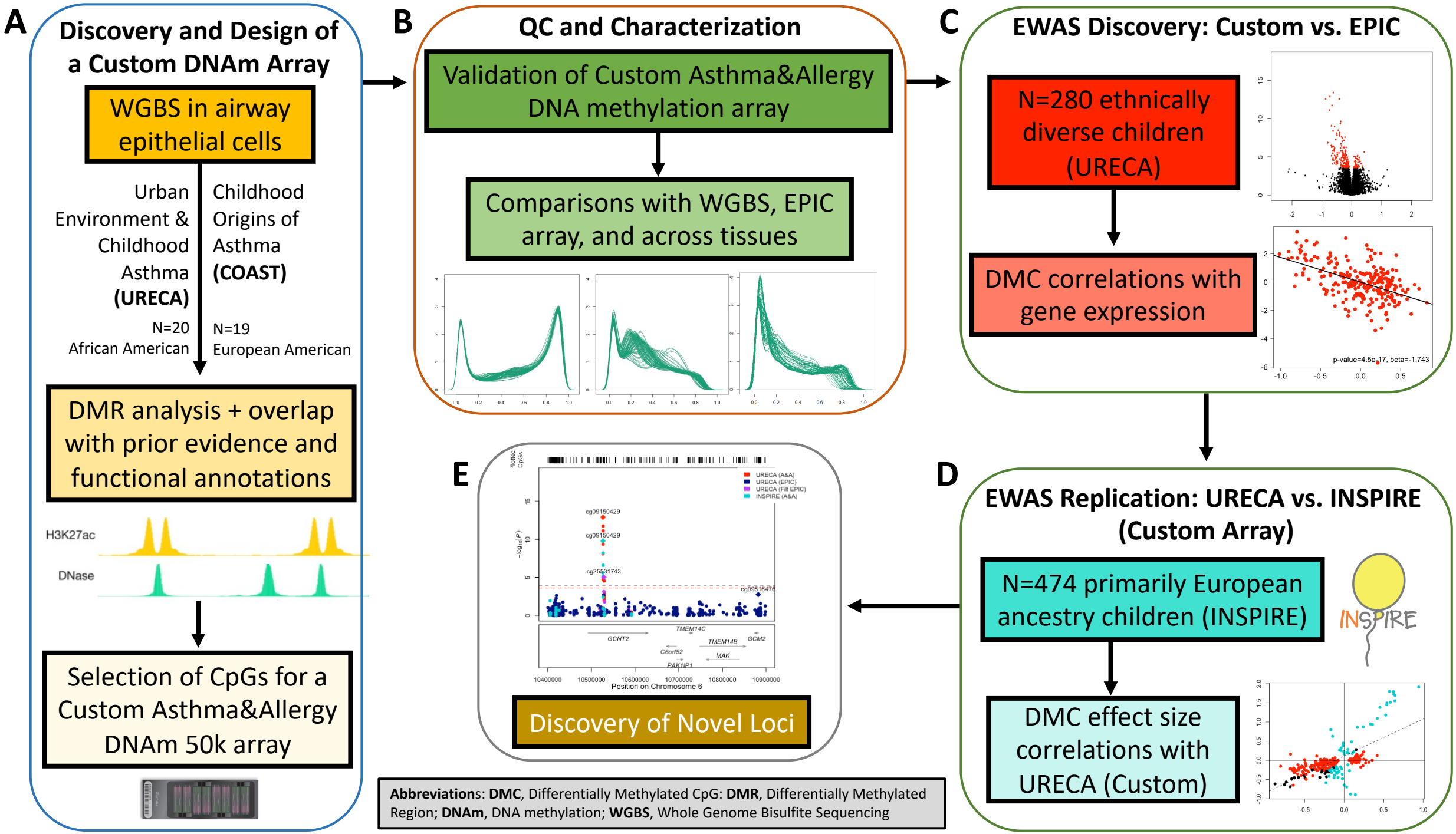
ECHO Components – Coordinating Center: Duke Clinical Research Institute, Durham, NC: Smith PB, Newby KL.

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.

Data Sharing Plans

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


A
Discovery and Design of a Custom DNAm Array
WGBS in airway epithelial cells
Urban Environment & Childhood Asthma (URECA)
N=20 African American
Childhood Origins of Asthma (COAST)
N=19 European American
DMR analysis + overlap with prior evidence and functional annotations
H3K27ac
DNase
Selection of CpGs for a Custom Asthma&Allergy DNAm 50k array
B
QC and Characterization
Validation of Custom Asthma&Allergy DNA methylation array
Comparisons with WGBS, EPIC array, and across tissues
C
EWAS Discovery: Custom vs. EPIC
N=280 ethnically diverse children (URECA)
DMC correlations with gene expression
p-value=4.5e-17, beta=-1.743
D
EWAS Replication: URECA vs. INSPIRE (Custom Array)
N=474 primarily European ancestry children (INSPIRE)
INSPIRE
DMC effect size correlations with URECA (Custom)
E
URECA (A&A)
URECA (EPIC)
URECA (Filt EPIC)
INSPIRE (A&A)
cg09150429
cg25531743
TMEM14C
GCNT2
TMEM14B
GCM2
C6orf52
MAK
PAK1IP1
Position on Chromosome 6
Discovery of Novel Loci
Abbreviations: DMC, Differentially Methylated CpG: DMR, Differentially Methylated Region; DNAm, DNA methylation; WGBS, Whole Genome Bisulfite Sequencing

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 Fig. 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.

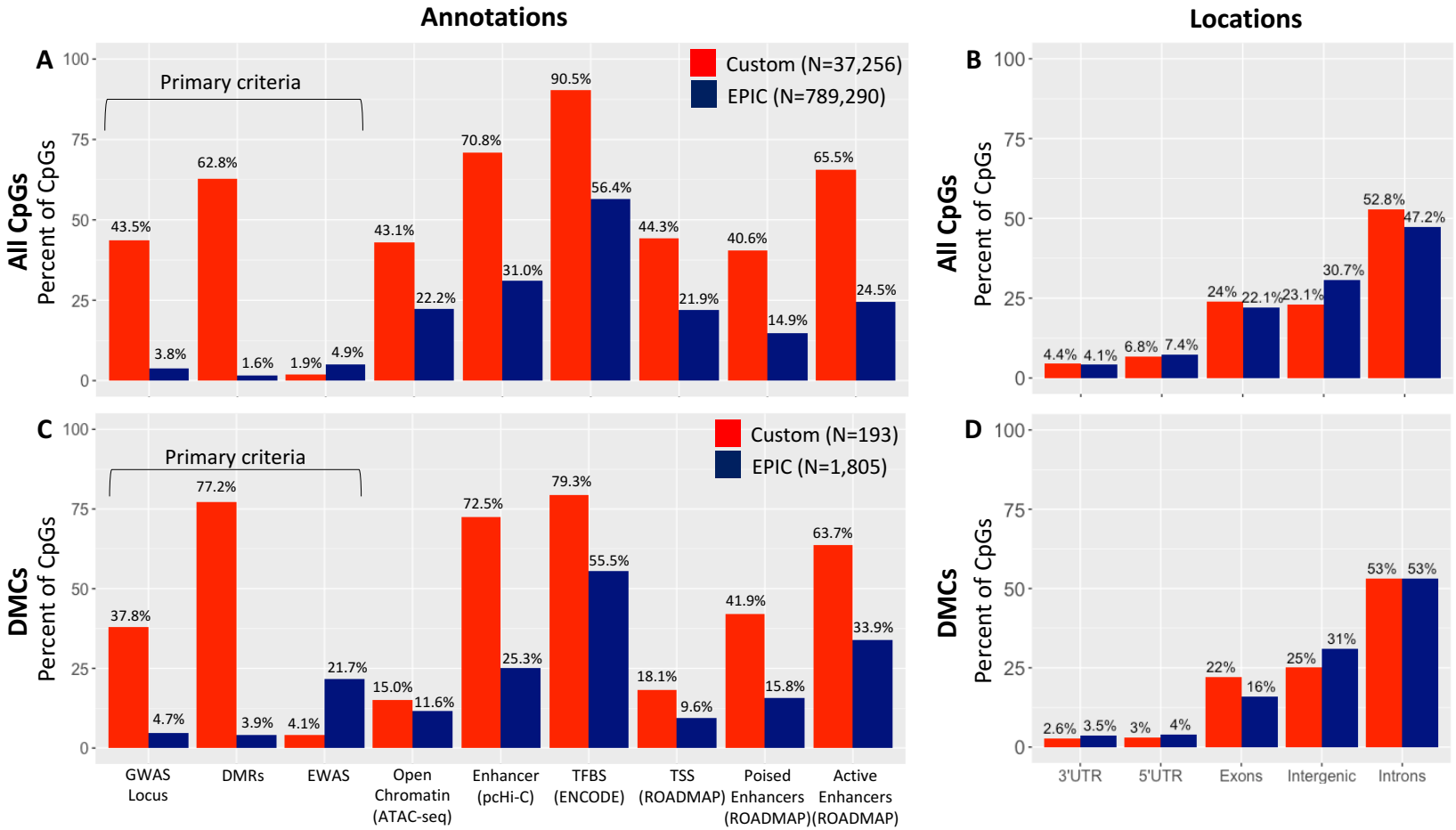
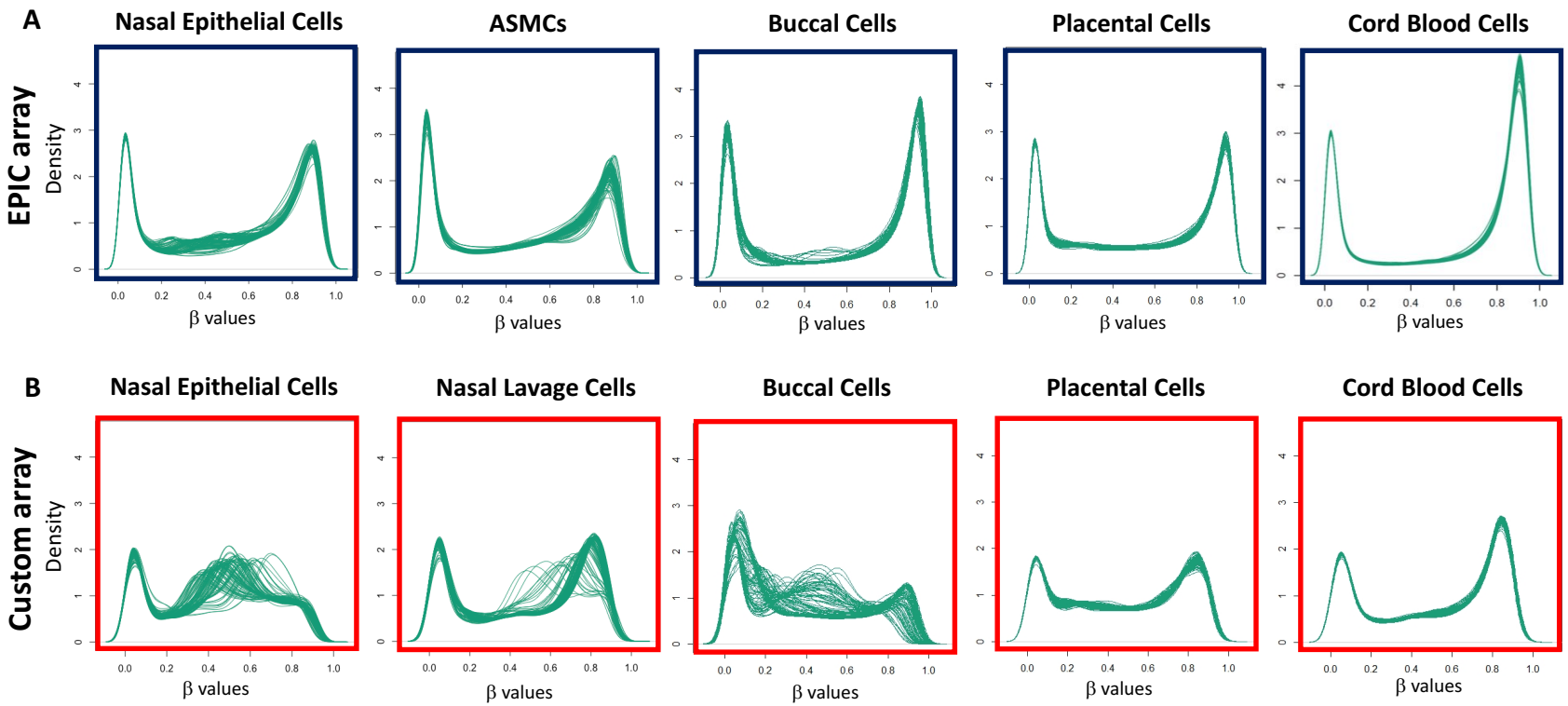


Figure 2. Proportions of Custom and EPIC CpGs by primary criteria, functional annotation category, and genomic location. All comparisons were performed with a Fisher's Exact Test. **(A)** Proportions of Custom CpGs that passed processing QC in URECA on each array by three primary

796 criteria and six functional annotation categories. CpGs on the Custom array were significantly enriched ($p<2.2x10^{-16}$) in all categories compared
797 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.
798 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
799 did not differ in any other categories. (**C**) Proportions of Custom and EPIC DMCs from each array by three primary criteria and six functional
800 annotation categories. Compared to all CpGs, DMCs from both arrays were depleted at transcription start sites and in areas of open chromatin
801 (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
802 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
803 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
804 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
805 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
806 distributions of all CpGs.



807

808 **Figure 3. Cross-tissue comparisons of methylation levels on the EPIC and Custom arrays**. Density distribution plots of methylation levels,

809 measured as β values, in 96 individuals after quantile (Custom) or quantile + SWAN (EPIC) normalizations (see **Methods**). The x-axis shows the

810 proportion of methylation at CpG sites on the EPIC (**A**) and Custom (**B**) arrays for each individual. All DNA methylation data were processed using

811 the same pipeline. The nasal epithelial cell DNA (EPIC and Custom) and nasal lavage cell DNA (Custom) were from the same randomly selected

812 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

813 the VCSIP Study (see **Methods**).

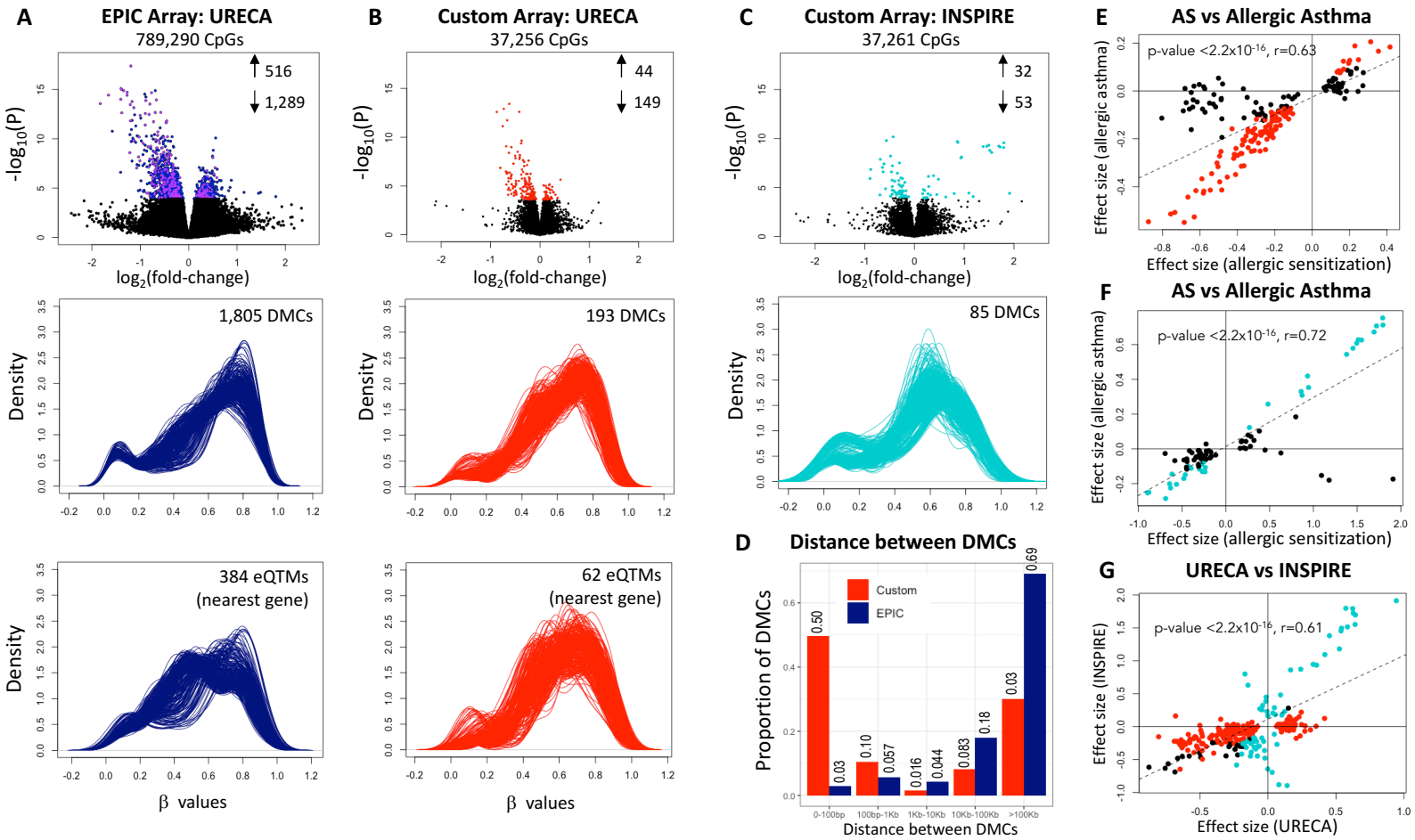


Figure 4. Asthma EWAS results in URECA and INSPIRE children using the EPIC and Custom arrays. (A-C) Upper panel: Volcano plots showing the $\log_2$-fold change in methylation proportion (M-values) with and without allergic sensitization (x-axis) and the $-\log_{10}$(P-value) from the EWAS (y-

817 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
818 red (Custom) in URECA or turquoise (Custom) in INSPIRE. The number of DMCs that were hypermethylated or hypomethylated in the children
819 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
820 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
821 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**)
822 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
823 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
824 INSPIRE, respectively; black dots are AS-only DMCs. (**G)** Correlation of effect sizes (log fold change) between DMCs identified in the INSPIRE AS
825 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
826 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.

827