A Functional Genomics Pipeline to Identify High-Value Asthma and Allergy CpGs in the Human Methylome

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

Tel: 773-834-0735

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

Tel: 412-624-8368

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.

Grant Support: 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.

Disclosures: 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.

Abstract

Background: 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.

Objective: To explore this hidden portion of the epigenome, we combined whole-genome bisulfite sequencing (WGBS) with *in silico* evidence of gene regulatory regions 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.

Methods: We identified differentially methylated regions (DMRs) from WGBS in nasal epithelial cell (NEC) DNA from 39 children with and without allergic asthma of both European and African ancestries. We selected CpGs from DMRs, previous allergy or asthma Epigenome-Wide Association Studies (EWAS), or Genome-Wide Association Study (GWAS) loci, and overlapped them with functional annotations for inclusion on a custom Asthma&Allergy array. Using both the Custom and EPIC arrays, we performed EWAS of allergic sensitization (AS) in NEC DNA from children in the URECA birth cohort and using the Custom array in the INSPIRE birth cohort. We assigned each CpG on the arrays to its nearest gene and its promotor capture Hi-C interacting gene and performed expression quantitative trait methylation (eQTM) studies for both sets of genes.

Results: Custom array CpGs were enriched for intermediate methylation (IM) levels compared to EPIC CpGs. IM CpGs were further enriched among those associated with AS and for eQTMs on both arrays.

Conclusions: Our study revealed signature features of high-value CpGs and evidence for epigenetic regulation of genes at AS EWAS loci that are robust to race/ethnicity, ascertainment, age, and geography.

Clinical Implications: These studies identified allergic sensitization-associated differentially methylated CpGs and their target genes in airway epithelium, providing potential epigenetic mechanisms in the development of allergic diseases and suggesting novel drug targets.

Capsule Summary: This study of previously unexplored regions of the airway epithelial methylome revealed novel epigenetic mechanisms regulating genes previously implicated in the pathogenesis of asthma and allergic diseases.

Key Words: EWAS, DNA methylation array, gene expression, airway epithelial cells

Abbreviations: **AS**, allergic sensitization; **DMC**, differentially methylated CpG; **DMR**, differentially methylated region; **eQTM**, expression quantitative trait methylation; **EWAS**, epigenome-wide association study; **FET**, fisher exact test; **GWAS**, genome-wide association study; **IM**, intermediate methylation; **LF**, latent factors; **NEC**, nasal epithelial cells; **PCA**, principal components analysis; **SPT**, skin prick test; **WGBS**, whole-genome bisulfite sequencing.

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. Because the selection of CpGs on the EPIC array was agnostic with respect to disease or tissue types, 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 Asthma&Allergy 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

135 downstream effects on asthma and allergic disease risk[5, 6]. An overview of our study

136 design is shown in **Fig. 1**.

Methods

Cohorts Included in Whole Genome Bisulfite Sequencing (WGBS) Studies or Epigenome-Wide Association Studies (EWAS)

Three longitudinal birth cohorts were included in these studies: the URban Environment and Childhood Asthma (URECA)[7] cohort (n=20) and the Childhood Origins of ASThma (COAST)[8] cohort (n=19) were included in the WGBS studies; URECA (n=280) and the Infant Susceptibility to Pulmonary Infections and Asthma Following RSV Exposure (INSPIRE)[9] (n=474) were included in EWAS of allergic sensitization (AS). 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 characteristics of the 280 URECA and 474 INSPIRE participants in the EWAS are described in **Table I**; detailed descriptions of the three cohorts are included in the Online Repository.

Cohorts included in Cross-Tissue Comparison Studies

In addition to studying nasal epithelial cell DNA with the Asthma&Allergy 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 (n=96). The nasal lavage cell DNA was isolated from URECA subjects at mean age 13.4 years (range = 12.8 - 14.9 years). The buccal, placenta and cord blood cell DNA was collected from

157 participants in the Vitamin C to Decrease the Effects of Smoking in Pregnancy on Infant
158 Lung Function (VCSIP) cohort[10].

159 *Whole Genome Bisulfite Sequencing and DMR Studies*

160 DNA (250 ng) from participants in the URECA and COAST cohorts was transferred to the
161 University of Chicago Genomics Core Facility. Bisulfite conversion and sequencing were
162 conducted using the EZ DNA Methylation-Gold Kit library prep and the Swift Biosciences
163 Accel-NGS Methyl-seq DNA library kit, respectively. Samples were sequenced to a
164 minimum of 330 million reads on the Illumina NovaSEQ6000 (S4 flowcell). Adapters
165 were removed from sequence reads using trimgalore[11] and then mapped using bismark
166 (version 0.18.2)[12] and the hg19 reference assembly. Bismark was also used to remove
167 duplicate reads and to call methylation values. Prior to differential methylated region
168 (DMR) analyses, CpGs were removed if they overlapped with a common single
169 nucleotide polymorphism (SNP) (minor allele frequency [MAF] >0.05) in 1000 genomes
170 CEU or YRI populations[13] or in Blacklisted regions[14]. Variant calling from WGBS was
171 performed using the biscuit algorithm, which also identified sample swaps and
172 contamination. The variant calls from WGBS were compared to either array-based
173 genotypes (COAST) or whole genome-sequence-based genotypes (URECA). All variant
174 calls matched with an accuracy of >95%. One sample in COAST was a duplicate and was
175 excluded from further analysis, leaving 19 samples for analysis (9 allergic asthmatics and
176 10 non-asthma/non-allergic controls). The DMR analyses are described in the Online
177 Repository.

178 *Selection of CpGs and Design of the Custom Array*

179 Our pipeline for selecting high-value CpGs for a custom array was implemented in three
180 steps (**Fig. E1** in this article's Online Repository). Additional details are described in the
181 Online Repository. Our goal was to develop a custom array with 50,000 probes for CpGs
182 that would complement and serve as a "booster" for the EPIC array in studies of asthma
183 and allergic diseases. Briefly, we first identified a set of CpGs based on prior evidence of
184 association with asthma or allergic diseases (atopic dermatitis/eczema, allergic
185 rhinitis/hay fever, and food allergy) from three categories of studies. The first category
186 included the 199,473 CpGs within DMRs from the WGBS study described above, the
187 second included 19,057 CpGs associated with asthma or allergic disease in 17 previous
188 EWAS (**Table E1** in this article's Online Repository), and the third included 570,350 CpGs
189 at 140 GWAS loci (**Table E2** in this article's Online Repository). We then removed CpGs
190 on the EPIC array, in ENCODE blacklisted regions[12], and those that overlapped with a
191 common SNP in 1000 Genomes European (CEU) or African (YRI) populations. In the
192 second step, we further prioritize the remaining 696,225 CpGs, by considering their
193 overlap with six functional annotations: 1) ENCODE[15] TFBSs from all cell types; 2-4)
194 ROADMAP Epigenetics[16] transcriptional start sites, poised enhancers and active
195 enhancers from smooth muscle (E078, E076, E103, E111), epithelial (E055, E056, E059,
196 E061, E058), and blood cells (E062, E034, E045, E044, E043, E039, E041, E042, E040,
197 E037, E048, E038, E047, E029, E050, E032, E046); 5) ATAC-seq in human cultured
198 bronchial epithelial cells exposed to rhinovirus or a vehicle control from asthmatic and

199 non-asthmatic individuals[17], and 6) pcHi-C from *ex vivo* human bronchial epithelial
200 cells[17]. We then required that CpGs within DMRs overlap with at least three annotations
201 or other category of prior evidence, CpGs from prior EWAS overlap with at least one
202 annotation or other category of prior evidence, and CpGs at GWAS loci overlap with at
203 least four annotation categories or other category of prior evidence. After removing
204 duplicates, 92,024 high-value CpGs remained and 53,700 of these CpGs were amenable
205 to the Illumina design algorithm. We selected 44,047 CpGs targeted by 49,999 probes
206 for manufacturing, of which 14% subsequently failed, resulting in a Custom array with
207 38,541 CpGs targeted by 43,605 probes. After removing probes with unknown origins
208 and those without genomic coordinates, 37,863 CpGs passed array QC. The locations of
209 the 37,863 CpGs, their selection among the three primary criteria, and their overlap
210 with the annotation categories are shown in Supplementary Dataset 1. Finally, for our
211 analyses we further removed probes that included a common SNP (MAF≥5%) within 3
212 bp of the CpG interrogation site, probes that failed the minfi QC, and probes on the X
213 chromosome, leaving 37,256 CpGs following both array and processing QC.

214 *Processing Custom and EPIC array methylation data in URECA and INSPIRE*

215 DNA methylation was assessed using the Illumina Allergy&Asthma Custom BeadChip
216 v1.0 or the Illumina Infinium MethylationEPIC BeadChip (Illumina, San Diego, CA)
217 following bisulfite conversion at the University of Chicago Genomics Facility.
218 Methylation data from both arrays were processed using minfi v1.29.1[18]. The
219 annotation and manifest packages for the Allergy&Asthma array v1.0 can be found at

220 https://github.com/hansenlab/IlluminaHumanMethylationAllergymanifest. The details
221 of our protocols for processing the DNA methylation data on each array are described in
222 the Online Repository.

223 *EWAS of AS in the URECA and INSPIRE cohorts*

224 We defined AS as the percent of positive skin prick tests (14 airborne and oral allergens
225 in URECA and 13 airborne allergens in INSPIRE). For each AS EWAS we used the
226 following models:

227 URECA: DNAm ~ proportion positive SPTs + sex + %epithelial cells +
228 $AncPCs_{1-3}$ + LFs_{1-n}

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

231 For the URECA EWAS, we included 14 and 5 latent variables for the Custom[19] and EPIC[20]
232 arrays, respectively, after removing technical variables (see **Table E3** in this article's
233 Online Repository). Cell counts (300 cells/slide) were available on each sample[21].
234 Because we did not have information on cell proportion or genetic ancestry for INSPIRE,
235 we included sex, parent-reported race/ethnicity, DNA concentration, and 21 latent
236 factors as covariates to adjust for cell composition, as well as other unwanted variation
237 in the EWAS. Analyses were performed in R (version 4.1.0) using limma v 3.50.0[22, 23]. To
238 control the false discovery rate, we used a q-value threshold of 0.05. To assess the
239 overlap of AS DMCs with allergic asthma DMCs, we tested the AS DMCs for association

240 with allergic asthma in URECA and INSPIRE. For these studies, we defined allergic
241 asthma cases as children with a diagnosis of asthma at age 10 (URECA) or age 6
242 (INSPIRE) and sensitization to at least one allergen (N=62 and 22, respectively). For
243 these analyses, controls were defined as children without a diagnosis of asthma (ever)
244 and without any positive skin prick tests (n=76 and 271, respectively). We calculated q-
245 values based on the 193 AS DMCs in URECA and the 85 AS DMCs in INSPIRE.

246 *RNA-sequencing and expression Quantitative Trait Methylation (eQTM) studies in URECA*

247 Protocols for processing samples for RNA-seq in nasal epithelial cells from the URECA
248 children have been described[21]. Gene expression data were available in 249 of the
249 children with DNA methylation data. In these samples, 15,643 genes were detected as
250 expressed. To test for cis associations between DNA methylation and gene expression
251 (eQTM studies), we used a linear model that included sex, percent epithelial cells, and
252 three ancestry PCs as covariates (FDR <0.05).

253

254 ## Results

255 ### Identifying DMRs in Whole-Genome Bisulfite Sequences

256 We performed WGBS in airway epithelial cell DNA from 20 African American children
257 (10 with allergic asthma, 10 without asthma or allergies; 11 years old) from the URECA
258 cohort[7] and 19 European American young adults (9 with allergic asthma, 10 without
259 asthma or allergies; 18-20 years old) from the COAST cohort[8]. After quality control (QC)
260 (see **Methods**), analyses for differentially methylated regions (DMRs) between the
261 asthma/allergy cases and non-asthma/non-allergy controls were performed in the
262 African American sample, the European American sample, and the combined sample
263 (see **Methods** for additional details). Overall, we identified 16,611 DMRs that included
264 199,473 CpGs (**Table E4** in this article's Online Repository).

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

266 Our goal was to develop a custom array with 50,000 probes for high-value CpGs that
267 would complement and serve as a "booster" for the EPIC array for studies of asthma
268 and allergic diseases, as described above in **Methods** and in the Online Repository. After
269 manufacture and array QC, 37,863 CpGs targeted by 42,192 probes remained (**Table E5**
270 in this article's Online Repository). The distributions of the 37,256 CpGs on the Custom
271 array and the 789,290 CpGs on the EPIC arrays that passed final QC are shown by
272 annotation category in **Fig. 2A**. Of note, 90% of CpGs on the Custom array overlapped
273 with a transcription factor binding site (TFBS), and 94% overlapped with a predicted

274 enhancer, compared to 56% and 54% of the EPIC array CpGs, respectively (Fisher exact
275 test [FET] $p<10^{-16}$ for both comparisons). Moreover, compared to EPIC CpGs, Custom
276 CpGs were enriched in introns (52.8% vs. 47.2%) and exons (24.0% vs. 22.1%) and
277 depleted in intergenic regions (23.1% vs. 30.7%) and 5'UTRs (6.8% vs. 7.4%) (FET $p<10^{-4}$
278 for all comparisons) **(Fig. 2B**). The CpGs on the Custom array proportionally represented
279 DMR-CpGs from the studies in URECA, COAST and combined analysis in WGBS and CpGs
280 from the three categories of prior evidence (**Fig. E2** in this article's Online Repository).

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

282 To assess the reproducibility of methylation measures of CpGs on the Custom array, we
283 compared methylation levels (measured as β values) in airway epithelial cell DNA in the
284 WGBS to the same CpGs on the Custom and the EPIC arrays for the subset of individuals
285 with all three (19 URECA subjects) or two (17 COAST subjects) measures. We included
286 CpGs with at least 10 WGBS reads that overlapped with an array CpG that passed QC.
287 This resulted in 502,229 CpGs for the WGBS-EPIC comparison in URECA and 24,744 and
288 20,861 CpGs for the WGBS-Custom array comparisons in URECA and COAST,
289 respectively. The β distribution plots were similar and highly correlated between CpGs
290 measured by both WGBS and an array (Spearman's $p<2.2\times10^{-16}$ for each) (**Fig. E3** in this
291 article's Online Repository). Notably, however, compared to the EPIC array, CpGs on the
292 Custom array were depleted for hypermethylated CpGs and enriched for intermediate
293 methylated (IM) CpGs (β values 20-80%).

294 To more broadly assess EPIC and Custom array DNA methylation patterns, we
295 compared the distributions of β values across different tissues using available data for
296 the EPIC array in DNA from airway epithelial cells (this study), airway smooth muscle
297 cells[24], and buccal swabs, placenta cells, and cord blood[10], and for the Custom array in
298 DNA from nasal epithelial cells and nasal lavage cells from URECA children, and buccal
299 cells, cord blood cells, or placenta tissue from infants in the VCSIP cohort[10]. The β
300 distributions were similar across cell sources for the EPIC array (**Fig. 3A**) but showed
301 varying patterns between cell types on the Custom array (**Fig. 3B**). Whereas an average
302 of 68% (range: 60-79%) of CpGs on the EPIC array were either hypomethylated (0-20%)
303 or hypermethylated (80-100%) in all six cell types, most CpGs on the Custom array (66%
304 in nasal epithelial cells and 52% on average across cell types) were IM CpGs with
305 depletions of β values between both 0-10% (24% in nasal epithelial cells and 29% on
306 average) and 90-100% (10% in nasal epithelial cells and 20% on average). In fact, these β
307 value distributions of CpGs on the EPIC array are remarkably similar to those in nine
308 GTEx tissues (**Fig. E4** in this article's Online Repository)[25]. To confirm that differences
309 between the two arrays were a result of the selection criteria used for designing the
310 array, we used the same criteria for the CpGs on the EPIC array. The 26,905 high-value
311 EPIC CpGs also showed an enrichment for IM CpGs and a depletion of both
312 hypomethylated and hypermethylated CpGs (**Fig. E5** in this article's Online Repository).

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

315 Our ultimate goal was to develop a DNA methylation array that could detect asthma- or
316 allergy-associated differential methylation that is missed by the content on the EPIC
317 array. AS is an IgE response to allergens[26] and a crucial step in the development of
318 allergic diseases and common phenotypes of asthma[21, 27, 28]. We conducted an EWAS of
319 AS (see **Methods**) in airway epithelial cell DNA collected from 280 11-year old URECA
320 children using the Custom and the EPIC arrays. The demographic and clinical description
321 of the URECA children is shown in **Table I** and **Fig. E6** in this article's Online Repository.

322 Following QC of the array, we performed two EWAS of AS in the same URECA
323 children using the EPIC and Custom arrays (see **Methods**). Using a q-value threshold of
324 0.05, we identified 1,805 differentially methylated CpGs (DMCs) using the EPIC array and
325 193 DMCs using the Custom array (**Fig. 4A-B**), revealing an enrichment for AS DMCs on
326 the Custom compared to the EPIC array (0.50% vs. 0.23% of CpGs on each array,
327 respectively; FET $p<2.2x10^{-16}$). Among the DMCs on the EPIC array, 295 (16%) were
328 high-value EPIC CpGs, a significant enrichment compared to 3.5% of high-value CpGs
329 among all CpGs on the EPIC array (FET $p<2.2x10^{-16}$). In contrast to the different β
330 distributions of CpGs on each array (**Fig. 3A**), the β distributions of DMCs were similar
331 (**Fig. 4A-B**; middle panels), showing a near complete depletion of both hypomethylated
332 and hypermethylated CpGs and an enrichment of IM CpGs among AS DMCs in airway
333 epithelial cells. The DMCs on both arrays were distributed across the autosomes (**Fig.**
334 **E7A-B** in this article's Online Repository), however, Custom array DMCs showed spikes
335 of association signals compared to the sparse distribution of EPIC array DMCs. In fact,

336 50% of DMCs on the Custom array were within 100bp of the next nearest DMC
337 compared to only 3% of DMCs on the EPIC array (**Fig. 4D**). Regional clustering of DMCs
338 on the Custom array provides greater confidence in and internal validation of EWAS
339 results.

340 Most of the functional annotation categories of DMCs were proportional to all
341 CpGs on each array (**Fig. 2A and 2C**; see **Fig. 2** legend for all statistical test results).
342 However, DMCs from both arrays were depleted at transcription start sites and in areas
343 of open chromatin compared to all CpGs on the arrays. Among the primary criteria
344 (prior evidence), CpGs in DMRs were modestly enriched for DMCs compared to all CpGs
345 on the Custom array and DMCs in prior EWAS studies were modestly enriched on the
346 Custom array but significantly enriched on the EPIC array. In contrast, CpGs at GWAS loci
347 were not enriched among DMCs on the Custom array and only modestly enriched on
348 the EPIC array compared to all CpGs on the array. The genomic locations of the DMCs
349 were proportional to all CpGs on the arrays (**Fig. 2B and 2D**).

350 Because AS is an important step in the development of childhood asthma, we
351 next asked whether AS DMCs were also associated with allergic asthma. Considering
352 only the AS DMCs on the EPIC array, 1,155 (64%) were also associated with allergic
353 asthma at q-value <0.05 (1,805 tests). Among the AS DMCs on the Custom array, 115
354 (60%) were also associated with allergic asthma at q-value <0.05 (193 tests). The effect
355 sizes were significantly correlated between AS DMCs and allergic asthma DMCs (r=0.68).
356 Among the CpGs that were DMCs for both AS and allergic asthma, all showed the same

357 direction of effect; among all AS DMCs, 93% showed the same direction of effect with
358 allergic asthma (r=0.84) (**Fig. 4E**). The sample distributions of AS and allergic asthma are
359 shown in **Fig. E8** in this article's Online Repository.

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

361 The URECA cohort is diverse with respect to ancestry (**Fig. E9** in this article's Online
362 Repository) but includes <1% non-Hispanic White children[7]. Therefore, to both replicate
363 results of the EWAS in URECA children and assess the performance of the Custom array
364 in a primarily non-Hispanic White population, we studied 5- to 7-year-old children from
365 the INSPIRE cohort[9]. An EWAS was performed using the Custom array and nasal
366 epithelial cell DNA from 474 children with measures of AS (see **Methods**). See **Table I** for
367 demographic and clinical characteristics, **Fig. E6** in this article's Online Repository for
368 sensitization rates by allergen, and the Online Repository for details on processing and
369 QC of the array data.

370 At a q-value threshold of <0.05 (37,261 tests), 85 CpGs on the Custom array were
371 DMCs (0.2% of CpGs). The β distribution of the DMCs (**Fig. 4C**) and the patterns of
372 associations (**Fig. E7C** in this article's Online Repository) were similar to those observed
373 in URECA. Among the AS DMCs in INSPIRE, 38% were also allergic asthma DMCs (q-
374 value<0.05 based on 85 tests), all with the same direction of effect. Among all AS DMCs
375 in INSPIRE, 93% had the same direction of effect with allergic asthma (**Fig. 4F**).
376 Moreover, among the AS DMCs in URECA, 25 (13%) were also AS DMCs in INSPIRE,

377 which was a highly significant enrichment ($p<2.2x10^{-16}$), and all AS DMCs had
378 concordant directions of effect. Finally, the effect sizes of the 253 CpGs that were DMCs
379 in either the URECA or INSPIRE AS EWAS were highly correlated ($r=0.61$; $p<2.2x10^{-16}$)
380 (**Fig. 4G**). The high reproducibility of EWAS results in two cohorts demonstrates that the
381 high-value CpGs on the Custom array identified AS DMCs that are robust to ancestry,
382 ascertainment strategies, age at sampling, and geography. The sample distributions of
383 AS and allergic asthma are shown in **Fig. E8** in this article's Online Repository.

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

385 Because nearly all CpGs on the Custom array were selected to overlap with a TFBS or
386 enhancer mark, we expected these CpGs to be correlated with the expression of genes
387 in airway epithelial cells more often than CpGs on the EPIC array[29]. To test this, we used
388 gene expression data[21] in the same cells as those used for DNA methylation studies in
389 249 of the URECA children and defined two sets of genes from among the 15,551 genes
390 detected as expressed in these cells: the nearest gene to each CpG and the promoter
391 capture Hi-C (pcHi-C)-defined target gene in airway epithelial cells (see **Methods**). The
392 latter identifies putative enhancers that physically interact with target gene promoters
393 and account for the 3-dimensional structure of the genome that allows for distal
394 regulatory elements to interact with and regulate the activity of promoters over large
395 distances (e.g., up to 1 Mb or more). Among the nearest expressed genes to DMCs on
396 the Custom (98 genes) and EPIC (1,449 genes) arrays, 63 were the nearest gene to CpGs
397 on both arrays. Among the pcHi-C target genes for DMCs on the Custom (245 genes) or

398 EPIC (1,155 genes) arrays, 95 were target genes of CpGs on both arrays. Thus, although
399 no CpGs overlapped between the arrays, 101 of nearest or pcHi-C target genes were
400 shared on both arrays (**Table E6** in this article's Online Repository). The 318 nearest or
401 pcHi-C target genes on the Custom array that were expressed in airway epithelial cells
402 were enriched in 16 pathways (FDR <0.05), including "Th1 and Th2 cell differentiation,"
403 "JAK-STAT signaling," "Th17 cell differentiation," and "Viral protein interaction with
404 cytokine and cytokine receptor" (**Table E7** in this article's Online Repository and
405 Supplementary Methods in the Online Repository). In contrast, the 2,366 nearest or
406 pcHi-C target genes on the EPIC array that were expressed in airway epithelial cells were
407 not enriched for any pathways (smallest FDR = 0.13). These findings further
408 demonstrate that the CpGs on the Custom array are enriched in pathways relevant to
409 asthma and allergic disease.

410 For each CpG-gene pair, we tested for correlation between DNA methylation and
411 expression levels of their nearest and pcHi-C target gene(s) to identify eQTMs.
412 Significantly more CpGs on the Custom array were eQTMs with their nearest gene (23%)
413 or pcHi-C target gene (16%) compared to CpGs on the EPIC array (11% and 9%
414 respectively; FET $p<2.2x10^{-16}$ in both analyses) (see the Online Repository). The high-
415 value EPIC CpGs were enriched for eQTMs compared to all EPIC CpGs (20% and 12%,
416 respectively; FET $p<2.2x10^{-16}$ in both analyses). Although more AS DMCs were eQTMs
417 compared to non-DMCs on both arrays, significantly more were on the Custom
418 compared to the EPIC array (nearest gene 35% vs. 20% [FET p = 0.0019] and pcHi-C 22%

419 vs 15% [FET p = 0.0082], respectively) (**Table E8** in this article's Online Repository). The β
420 distributions for DMCs on the Custom and EPIC arrays that were eQTMs were further
421 enriched for IM CpGs and depleted of CpGs at the extremes of the distribution (**Fig. 4**
422 and **Fig. E10** in this article's Online Repository).

423 **Further Insights into the Epigenetic Regulation of Gene Expression at EWAS Loci**

424 To examine the DMCs and their associations with gene expression more closely, we
425 generated regional association plots for the 10 most significant DMCs in the URECA
426 EWAS using the EPIC and Custom arrays. The 10 most significant EPIC DMCs were at 10
427 loci (**Fig. E11** in this article's Online Repository). Seven were "solitary" with no other
428 DMCs within 500kb and three had one other DMC within 6.5kb. Among the solitary
429 DMCs, six were high-value EPIC CpGs and all 10 were reported as DMCs in airway
430 epithelial cells in previous EWAS of asthma or allergic phenotypes. Six of the 10 DMCs
431 were within genes and four were intergenic. CpGs from the Custom array were present
432 in four of these regions, but none were AS DMCs. In contrast, the 10 most significant
433 Custom DMCs were at six loci and only two were solitary (**Fig. E12** in this article's Online
434 Repository). One solitary DMC in *ALOX15* was near a high-value and a non-high-value
435 EPIC DMC; the other solitary CpG in the *PDE6A* gene was also a DMC in the INSPIRE
436 EWAS. The other four loci had spikes of association, often with a combination of DMCs
437 from both arrays in URECA and the Custom array in both URECA and INSPIRE. Eight of
438 the 10 most significant DMCs were within genes; two were intergenic.

439 We selected three additional regions as examples of patterns of associations
440 (**Table E9** in this article's Online Repository). The first includes a spike of 11 Custom
441 array DMCs in URECA, 12 in INSPIRE and one high-value EPIC DMC in exons 2 and 3 of
442 the *CISH* (Cytokine Inducible SH2-Containing Protein) gene, a member of the SOCs family
443 of negative regulators of cytokine signaling (**Fig. 5A**). There was one nearby DMC from
444 the EPIC array downstream of *CISH*. The lead Custom and EPIC URECA DMCs were
445 eQTMs for *CISH*, with increased methylation associated with decreased gene expression
446 and fewer allergic sensitizations. The EPIC high-value CpG at this locus was identified in
447 previous asthma/allergy EWAS[30, 31], but this locus has not been associated with asthma
448 or allergic diseases in GWAS. *CISH* expression is induced by IL-13 in bronchial epithelial
449 cells and macrophages[32], and has been implicated in eosinophil physiology[33] and
450 eosinophilic inflammation[32]. Our studies are consistent with these findings and further
451 indicate that increased expression of *CISH* is associated with increased sensitization to
452 allergens and that the regulation of *CISH1* expression may be epigenetically mediated.

453 A second locus included one Custom DMC in URECA and INSPIRE, a second
454 Custom DMC in URECA, and one high-value EPIC DMC in an intron of *SLC22A5* (Solute
455 Carrier Family 22 Member 5) (**Fig. 5B**). The lead DMC was a high-value EPIC CpG, which
456 was also identified in prior EWAS in nasal epithelium[30, 31]. None of the DMCs were
457 eQTMs for *SLC22A5*, but all three were in a region that physically interacted with the
458 promoter of Interferon Response Factor 1 (*IRF1*), 91kb away, and were eQTMs for *IRF1*
459 ($p=2.4x10^{-7}$, beta=0.27 and $p=2.7x10^{-6}$, beta=0.27). This region is at a GWAS locus for

460 adult- and childhood-onset asthma and hay fever[34, 35] (**Table E9** in this article's Online
461 Repository), possibly with sex-specific effects[36]. Genetic variation in the *IRF1* gene has
462 been associated with increased expression of pro-inflammatory genes and IL-13
463 secretion in peripheral blood cells[37]. Our studies further demonstrate long-range
464 epigenetic regulation of *IRF1* expression in airway epithelial cells, with increased
465 methylation levels associated with increased gene expression and decreased numbers
466 of sensitizations.

467 At a third locus, we observed a spike of three Custom array DMCs in URECA and two in
468 INSPIRE upstream of the *HDAC7* (Histone Deacetylase 7) gene (**Fig. 5C**). One Custom
469 DMC (upstream of *HDAC7* and downstream of *VDR*) was an eQTM for *HDAC7*. Histone
470 deacetylases (HDACs) have diverse functions, including regulation of inflammatory
471 genes[38], and impaired barrier function in asthmatics was both induced by HDAC activity
472 and reversed by inhibition of endogenous HDAC[39]. In this study, increased methylation
473 in the *HDAC7* gene was associated with increased expression of *HDAC7* and fewer
474 sensitizations. At this extended locus, two EPIC DMCs were in an intergenic region 42kb
475 upstream of the *VDR* (Vitamin D Receptor) gene. Vitamin D deficiency in childhood has
476 been associated with increased risk for persistent asthma[40] and high dose vitamin D
477 supplementation in pregnancy reduced risk of recurrent wheeze in childhood[41]. Here,
478 increased methylation (EPIC array) near the *VDR* gene was associated with decreased
479 *VDR* expression and sensitization to fewer allergens. The *HDAC7-VDR* region has been
480 identified in GWAS of childhood- and adult-onset asthma[34]. This locus demonstrates the

481 complementarity of the Custom and EPIC arrays in detecting differential methylation

482 relevant to asthma and allergic diseases.

Discussion

The epigenome plays a critical role in regulating gene expression in a context-specific manner, such as in specific cell types or in the presence of environmental exposures or disease states. DNA methylation, for example, is responsive to disease promoting exposures in both *in vitro* cell models[24, 42-44] and *ex vivo* cells from individuals who are healthy and with disease[1, 45-48]. 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 to date include CpGs most relevant to any specific exposure, disease, or cell type. Here, we addressed this question by designing a custom Allergy&Asthma DNAmethylation array with 37,863 high-value CpGs that are not 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 CpGs in the human genome.

Although we did not use methylation levels as a selection criterion, the CpGs selected 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 feature is a signature of high-value, variable and likely 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

504 tissue specific, to vary between individuals, and to play important roles in gene
505 regulation[29, 49]. Our study further revealed that CpGs with intermediate levels of
506 methylation are also more likely to be associated with AS, an important clinical
507 phenotype that reflects both an immune response to past allergen exposures[26] and a
508 risk factor for the development of asthma and allergic diseases[21, 27, 28]. Although the
509 CpGs on the custom array were overall enriched for IM CpGs, AS DMCs were further
510 enriched, with a depletion of very hypomethylated (0-20% methylation) and very
511 hypermethylated (80-100%) CpGs. This important observation was further supported by
512 EWAS results using the EPIC array. Whereas the EPIC array is enriched for
513 hypomethylated and hypermethylated CpGs in all tissues, the AS DMCs from the EPIC
514 were depleted for CpGs at both extremes and enriched for IM CpGs, similar to the DMCs
515 on the custom Allergy&Asthma array. We further showed that the 26,905 high-value
516 EPIC CpGs were also enriched for IM CpGs and enriched among AS DMCs: 16.3% of the
517 DMCs in the AS EWAS were among the high-value EPIC CpGs compared to 3.4% of all
518 EPIC CpGs. These results, combined with previous studies[29, 49], could justify filtering
519 CpGs on the EPIC array to include just IM CpGs (β=20-80%) to enrich for functional CpGs
520 and increase power to detect DMCs associated with exposures or disease outcomes. As
521 proof-of-concept, including only IM EPIC CpGs and using a q-value of 0.05, 2,182 CpGs
522 were AS DMCs (0.70% of all IM CpGs), more than three times the proportion of DMCs
523 among all CpGs (1,805 AS DMCs; 0.23% of all EPIC CpGs). Finally, DMCs that were

524 eQTMs were further enriched for IM CpGs (**Fig. 4** and **Fig. E10** in this article's Online
525 Repository), supporting their role in gene regulation.

526 In addition to the enrichment of IM CpGs on the Custom compared to the EPIC
527 array, the different patterns of associations with AS were notable. Whereas DMCs on
528 the EPIC were more sparsely distributed throughout the genome, the DMCs on the
529 Custom often showed spikes of associations, analogous to GWAS peaks. The spatial
530 organization among DMCs on the Custom array provides confirmatory evidence for their
531 veracity, which is generally not available for EPIC CpGs (**Figure 4D**). Another
532 distinguishing feature of the DMCs on the Custom array was their enrichment in exons
533 and depletions in intergenic regions compared to DMCs on the EPIC array. This is
534 consistent with results of an earlier study that observed enrichments of IM CpGs in
535 exons, where exon expression at those sites correlated with DNA methylation[29].

536 The CpGs included on the custom Allergy&Asthma array were selected based on
537 prior evidence of associations with asthma or allergic diseases using three classes of
538 evidence: within a DMR in our WGBS study, in a previous published EWAS, or at
539 published GWAS loci. Overall, 62% of CpGs on the custom array were selected from
540 DMRs compared to 1.6% of all CpGs on the EPIC array, yet the DMR CpGs were enriched
541 among DMCs on both the custom (77.2%) and EPIC (3.9%) arrays. In contrast, few
542 previous EWAS CpGs were on the Custom array because we excluded all CpGs on the
543 EPIC from the Custom array, yet we still observed a modest enrichment of EWAS CpGs
544 among Custom array DMCs (1.9% overall vs. 4.1% of DMCs) and a highly significant

545 enrichment among DMCs on the EPIC array (4.9% overall vs. 21.7% of DMCs). These
546 represent replications of results from previous EWAS of asthma and allergic phenotypes
547 in the URECA children. Surprisingly, CpGs at asthma and allergy GWAS loci were only
548 modestly enriched among EPIC DMCs and not at all enriched among the Custom DMCs,
549 even though the EWAS in URECA and INSPIRE revealed many examples of AS DMCs at
550 important GWAS loci. The lack of enrichment of Custom array CpGs at GWAS loci among
551 DMCs may reflect the less direct evidence for DNA methylation mechanisms at GWAS
552 loci compared to CpGs from previous EWAS or DMR studies, the large proportion of
553 CpGs at GWAS loci included on the Custom array (44% of all CpGs), and the enrichment
554 of AS DMCs among genic regions whereas GWAS loci are more often in intergenic
555 regions.

556 Our study has limitations. First, due to cost and technical limitations, the Custom
557 array included only 37,863 CpGs out of the 92,024 high-value CpGs that we identified
558 with our pipeline. Therefore, together with the 26,905 high-value EPIC CpGs, we only
559 surveyed 54% of potential high-value CpGs in airway epithelium that are relevant to
560 asthma and allergic diseases. The next generation (v1.1) of the Custom array includes
561 ~8,000 more high-value CpGs, reducing this gap, although alternative array designs or
562 other methods that allow the inclusion of the 118,929 high-value CpGs would be a
563 powerful tool for assessing epigenetic mechanisms in these conditions. Second, we have
564 not yet evaluated the performance of this array in blood cells, a much more easily
565 sampled tissue. Those studies are currently underway. Lastly, we did not functionally

566 validate the correlations between CpGs and gene expression or identify the "causal"
567 CpG(s) driving the correlations, which should be considered in follow-up studies of any
568 of the EWAS regions reported here.

569 In summary, our study revealed high-value CpGs in airway epithelial cells that
570 are not interrogated on the EPIC array. The allergic sensitization EWAS revealed
571 potential epigenetic regulatory mechanisms both for genes previously implicated in
572 asthma or allergy pathogenesis but not identified in GWAS (e.g., *CISH* [32, 33], *NEDD4L* [50, 51],
573 *PRDM10* [52, 53]) as well as for genes at replicated GWAS loci (e.g., *IRF1*, *CLEC16A*,
574 *HDAC7/VDR*, *ALOX15*, *SLC22A5*) (**Table E9** in this article's Online Repository). These
575 associations were robust to race/ethnicity, ascertainment, age, and geography. We
576 suggest that similar pipelines can be used to identify high-value CpGs for other disease
577 groups or specific environmental exposures leading to the discovery of currently
578 invisible portions of the epigenome that are relevant to human health.

References

1. Prunicki M, Cauwenberghs N, Lee J, Zhou X, Movassagh H, Noth E, et al. Air pollution exposure is linked with methylation of immunoregulatory genes, altered immune cell profiles, and increased blood pressure in children. Sci Rep. 2021;11(1):4067.
2. Li Y, Hamilton KJ, Wang T, Coons LA, Jefferson WN, Li R, 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. 2018;115(18):E4189-E98.
3. Tobi EW, Slieker RC, Luijk R, Dekkers KF, Stein AD, Xu KM, et al. DNA methylation as a mediator of the association between prenatal adversity and risk factors for metabolic disease in adulthood. Sci Adv. 2018;4(1):eaao4364.
4. Li M, Zou D, Li Z, Gao R, Sang J, Zhang Y, et al. EWAS Atlas: a curated knowledgebase of epigenome-wide association studies. Nucleic Acids Res. 2019;47(D1):D983-D8.
5. Holgate ST. The sentinel role of the airway epithelium in asthma pathogenesis. Immunol Rev. 2011;242(1):205-19.
6. Akdis CA. The epithelial barrier hypothesis proposes a comprehensive understanding of the origins of allergic and other chronic noncommunicable diseases. J Allergy Clin Immunol. 2022;149(1):41-4.
7. Gern JE, Visness CM, Gergen PJ, Wood RA, Bloomberg GR, O'Connor GT, et al. The Urban Environment and Childhood Asthma (URECA) birth cohort study: design, methods, and study population. BMC Pulm Med. 2009;9:17.
8. Lemanske RF, Jr. The childhood origins of asthma (COAST) study. Pediatr Allergy Immunol. 2002;13(s15):38-43.

601 9. Larkin EK, Gebretsadik T, Moore ML, Anderson LJ, Dupont WD, Chappell JD, et al.
602 Objectives, design and enrollment results from the Infant Susceptibility to Pulmonary Infections
603 and Asthma Following RSV Exposure Study (INSPIRE). BMC Pulm Med. 2015;15:45.
604 10. Shorey-Kendrick LE, McEvoy CT, O'Sullivan SM, Milner K, Vuylsteke B, Tepper RS, et al.
605 Impact of vitamin C supplementation on placental DNA methylation changes related to maternal
606 smoking: association with gene expression and respiratory outcomes. Clin Epigenetics.
607 2021;13(1):177.
608 11. Kechin A, Boyarskikh U, Kel A, Filipenko M. cutPrimers: A New Tool for Accurate Cutting
609 of Primers from Reads of Targeted Next Generation Sequencing. J Comput Biol.
610 2017;24(11):1138-43.
611 12. Krueger F, Andrews SR. Bismark: a flexible aligner and methylation caller for Bisulfite-
612 Seq applications. Bioinformatics. 2011;27(11):1571-2.
613 13. Genomes Project C, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, et al. A
614 global reference for human genetic variation. Nature. 2015;526(7571):68-74.
615 14. Amemiya HM, Kundaje A, Boyle AP. The ENCODE Blacklist: Identification of Problematic
616 Regions of the Genome. Sci Rep. 2019;9(1):9354.
617 15. Consortium EP, Moore JE, Purcaro MJ, Pratt HE, Epstein CB, Shoresh N, et al. Expanded
618 encyclopaedias of DNA elements in the human and mouse genomes. Nature.
619 2020;583(7818):699-710.
620 16. Roadmap Epigenomics C, Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, et al.
621 Integrative analysis of 111 reference human epigenomes. Nature. 2015;518(7539):317-30.

622 17. Helling BA, Sobreira DR, Hansen GT, Sakabe NJ, Luo K, Billstrand C, et al. Altered
623 transcriptional and chromatin responses to rhinovirus in bronchial epithelial cells from adults
624 with asthma. Commun Biol. 2020;3(1):678.
625 18. Aryee MJ, Jaffe AE, Corrada-Bravo H, Ladd-Acosta C, Feinberg AP, Hansen KD, et al.
626 Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA
627 methylation microarrays. Bioinformatics. 2014;30(10):1363-9.
628 19. McKennan C. Factor analysis in high dimensional biological data with dependent
629 observations. arXiv:200911134v1. 2020.
630 20. McKennan C, Nicolae DL. Accounting for unobserved covariates with varying degree of
631 estimability in high dimensional experimental data. Biometrika. 2019;106(4):823-40.
632 21. Altman MC, Calatroni A, Ramratnam S, Jackson DJ, Presnell S, Rosasco MG, et al.
633 Endotype of allergic asthma with airway obstruction in urban children. J Allergy Clin Immunol.
634 2021.
635 22. Phipson B, Lee S, Majewski IJ, Alexander WS, Smyth GK. Robust Hyperparameter
636 Estimation Protects against Hypervariable Genes and Improves Power to Detect Differential
637 Expression. Ann Appl Stat. 2016;10(2):946-63.
638 23. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential
639 expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res.
640 2015;43(7):e47.
641 24. Thompson EE, Dang Q, Mitchell-Handley B, Rajendran K, Ram-Mohan S, Solway J, et al.
642 Cytokine-induced molecular responses in airway smooth muscle cells inform genome-wide
643 association studies of asthma. Genome Med. 2020;12(1):64.

644 25. Oliva M, Demandelis K, Jasnime F, Lu Y, Hahsan H, Kibriya M, et al. Genetic regulation of
645 DNA methylation across tissues reveals thousands of molecular links to complex traits. Research
646 Square. 2021.
647 26. van Ree R, Hummelshoj L, Plantinga M, Poulsen LK, Swindle E. Allergic sensitization:
648 host-immune factors. Clin Transl Allergy. 2014;4(1):12.
649 27. Illi S, von Mutius E, Lau S, Nickel R, Niggemann B, Sommerfeld C, et al. The pattern of
650 atopic sensitization is associated with the development of asthma in childhood. J Allergy Clin
651 Immunol. 2001;108(5):709-14.
652 28. Jackson DJ, Gern JE, Lemanske RF, Jr. The contributions of allergic sensitization and
653 respiratory pathogens to asthma inception. J Allergy Clin Immunol. 2016;137(3):659-65; quiz 66.
654 29. Elliott G, Hong C, Xing X, Zhou X, Li D, Coarfa C, et al. Intermediate DNA methylation is a
655 conserved signature of genome regulation. Nat Commun. 2015;6:6363.
656 30. Cardenas A, Sordillo JE, Rifas-Shiman SL, Chung W, Liang L, Coull BA, et al. The nasal
657 methylome as a biomarker of asthma and airway inflammation in children. Nat Commun.
658 2019;10(1):3095.
659 31. Forno E, Wang T, Qi C, Yan Q, Xu CJ, Boutaoui N, et al. DNA methylation in nasal
660 epithelium, atopy, and atopic asthma in children: a genome-wide study. Lancet Respir Med.
661 2019;7(4):336-46.
662 32. Takeshima H, Horie M, Mikami Y, Makita K, Miyashita N, Matsuzaki H, et al. CISH is a
663 negative regulator of IL-13-induced CCL26 production in lung fibroblasts. Allergol Int.
664 2019;68(1):101-9.

665 33. Burnham ME, Koziol-White CJ, Esnault S, Bates ME, Evans MD, Bertics PJ, et al. Human
666 airway eosinophils exhibit preferential reduction in STAT signaling capacity and increased CISH
667 expression. J Immunol. 2013;191(6):2900-6.
668 34. Pividori M, Schoettler N, Nicolae DL, Ober C, Im HK. Shared and distinct genetic risk
669 factors for childhood-onset and adult-onset asthma: genome-wide and transcriptome-wide
670 studies. Lancet Respir Med. 2019;7(6):509-22.
671 35. Johansson A, Rask-Andersen M, Karlsson T, Ek WE. Genome-wide association analysis of
672 350 000 Caucasians from the UK Biobank identifies novel loci for asthma, hay fever and eczema.
673 Hum Mol Genet. 2019;28(23):4022-41.
674 36. Myers RA, Scott NM, Gauderman WJ, Qiu W, Mathias RA, Romieu I, et al. Genome-wide
675 interaction studies reveal sex-specific asthma risk alleles. Hum Mol Genet. 2014;23(19):5251-9.
676 37. Landgraf-Rauf K, Boeck A, Siemens D, Klucker E, Vogelsang V, Schmidt S, et al. IRF-1
677 SNPs influence the risk for childhood allergic asthma: A critical role for pro-inflammatory
678 immune regulation. Pediatr Allergy Immunol. 2018;29(1):34-41.
679 38. Bhavsar P, Ahmad T, Adcock IM. The role of histone deacetylases in asthma and allergic
680 diseases. J Allergy Clin Immunol. 2008;121(3):580-4.
681 39. Wawrzyniak P, Wawrzyniak M, Wanke K, Sokolowska M, Bendelja K, Ruckert B, et al.
682 Regulation of bronchial epithelial barrier integrity by type 2 cytokines and histone deacetylases
683 in asthmatic patients. J Allergy Clin Immunol. 2017;139(1):93-103.
684 40. Hollams EM, Teo SM, Kusel M, Holt BJ, Holt KE, Inouye M, et al. Vitamin D over the first
685 decade and susceptibility to childhood allergy and asthma. J Allergy Clin Immunol.
686 2017;139(2):472-81 e9.

687 41. Wolsk HM, Chawes BL, Litonjua AA, Hollis BW, Waage J, Stokholm J, et al. Prenatal
688 vitamin D supplementation reduces risk of asthma/recurrent wheeze in early childhood: A
689 combined analysis of two randomized controlled trials. PLoS One. 2017;12(10):e0186657.
690 42. Pacis A, Tailleux L, Morin AM, Lambourne J, MacIsaac JL, Yotova V, et al. Bacterial
691 infection remodels the DNA methylation landscape of human dendritic cells. Genome Res.
692 2015;25(12):1801-11.
693 43. Nicodemus-Johnson J, Naughton KA, Sudi J, Hogarth K, Naurekas ET, Nicolae DL, et al.
694 Genome-wide methylation study identifies an IL-13-induced epigenetic signature in asthmatic
695 airways. Am J Respir Crit Care Med. 2016;193(4):376-85.
696 44. Soliai MM, Kato A, Helling BA, Stanhope CT, Norton JE, Naughton KA, et al. Multi-omics
697 colocalization with genome-wide association studies reveals a context-specific genetic
698 mechanism at a childhood onset asthma risk locus. Genome Med. 2021;13(1):157.
699 45. Cardenas A, Fadadu RP, Van Der Laan L, Ward-Caviness C, Granger L, Diaz-Sanchez D, et
700 al. Controlled human exposures to diesel exhaust: a human epigenome-wide experiment of
701 target bronchial epithelial cells. Environ Epigenet. 2021;7(1):dvab003.
702 46. Wan ES, Qiu W, Baccarelli A, Carey VJ, Bacherman H, Rennard SI, et al. Cigarette
703 smoking behaviors and time since quitting are associated with differential DNA methylation
704 across the human genome. Hum Mol Genet. 2012;21(13):3073-82.
705 47. Bakulski KM, Dou J, Lin N, London SJ, Colacino JA. DNA methylation signature of smoking
706 in lung cancer is enriched for exposure signatures in newborn and adult blood. Sci Rep.
707 2019;9(1):4576.

708 48. Sikdar S, Joehanes R, Joubert BR, Xu CJ, Vives-Usano M, Rezwan FI, et al. Comparison of
709 smoking-related DNA methylation between newborns from prenatal exposure and adults from
710 personal smoking. Epigenomics. 2019;11(13):1487-500.
711 49. Busche S, Shao X, Caron M, Kwan T, Allum F, Cheung WA, et al. Population whole-
712 genome bisulfite sequencing across two tissues highlights the environment as the principal
713 source of human methylome variation. Genome Biol. 2015;16:290.
714 50. Modena BD, Tedrow JR, Milosevic J, Bleecker ER, Meyers DA, Wu W, et al. Gene
715 expression in relation to exhaled nitric oxide identifies novel asthma phenotypes with unique
716 biomolecular pathways. Am J Respir Crit Care Med. 2014;190(12):1363-72.
717 51. Yip KH, Kolesnikoff N, Hauschild N, Biggs L, Lopez AF, Galli SJ, et al. The Nedd4-2/Ndfip1
718 axis is a negative regulator of IgE-mediated mast cell activation. Nat Commun. 2016;7:13198.
719 52. Zhang Y, Huang Y, Chen WX, Xu ZM. Identification of key genes in allergic rhinitis by
720 bioinformatics analysis. J Int Med Res. 2021;49(7):3000605211029521.
721 53. Schlag K, Steinhilber D, Karas M, Sorg BL. Analysis of proximal ALOX5 promoter binding
722 proteins by quantitative proteomics. FEBS J. 2020;287(20):4481-99.
723
724
725
726

Acknowledgments

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.

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).

745 **Table I. Demographic characteristics of URECA and INSPIRE samples.** Also see Fig. E6 in

746 this article's Online Repository for details on allergic sensitizations.

Characteristic	URECA	INSPIRE
Sample size (N)	280	474
Mean Age (range), years	11.1 (10.9-12.4)	6.3 (5.0-7.9)
Female sex	46.8%	47.0%
Ethnicity		
Non-Hispanic White	<1%	65.4%
Non-Hispanic Black	71.8%	18.1%
Hispanic	20.0%	9.5%
Other	7.9%	7.6%
Clinical Definitions		
Asthma	32.1%	9.5%
Allergic asthma	22.1%	4.6%
Allergic sensitization (≥1 +SPT)	63.2%	37.8%
Non-Allergic/Non-Asthma	27.1%	62.2%

747

748

Figure Legends

Fig. 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 **Fig. E1** in this article's Online Repository). 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, 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.

Fig. 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

770 array were significantly enriched ($p<2.2x10^{-16}$) in all categories compared to the EPIC
771 array with the exception of prior EWAS, in which they were depleted on the Custom
772 compared to the EPIC ($p<2.2x10^{-16}$). (**B**) The distributions of CpGs by genomic location.
773 Compared to the EPIC, CpGs on the Custom array were enriched in introns and exons
774 and depleted in intergenic regions and 5'UTRs ($p<10^{-4}$) but did not differ in any other
775 categories. (**C**) Proportions of Custom and EPIC differentially methylated CpGs (DMCs)
776 from each array by three primary criteria and six functional annotation categories.
777 Compared to all CpGs, DMCs from both arrays were depleted at transcription start sites
778 and in areas of open chromatin (Custom $p=1.14x10^{-7}$ and $p=4.11x10^{-9}$; EPIC $p<2.2x10^{-16}$
779 for both, respectively). DMRs were marginally enriched for DMCs compared to all CpGs
780 on the Custom array (p=0.066) but significantly enriched on the EPIC array ($p=1.54x10^{-6}$).
781 DMCs in prior EWAS studies of asthma and allergic diseases were modestly enriched on
782 the Custom array (FET p=0.059) and significantly enriched on the EPIC array ($p<2.2x10^{-}$
783 ^{16}). In contrast, CpGs at GWAS loci for asthma and allergic diseases were not enriched
784 among DMCs on the Custom array (p=0.32) and only modestly enriched on the EPIC
785 array (p=0.042). (**D**) The distribution of DMCs by genomic location. The distribution of
786 DMCs on both arrays did not differ from the distributions of all CpGs.

787 **Fig. 3. Cross-tissue comparisons of methylation levels for CpGs on the EPIC and**
788 **Custom arrays**. Density distribution plots of methylation levels, measured as β values, in
789 96 individuals after quantile (Custom) or quantile + SWAN (EPIC) normalizations (see
790 **Methods**). The x-axis shows the proportion of methylation at CpG sites on the EPIC (**A**)

791 and Custom (**B**) arrays for individuals. All DNA methylation data were processed using
792 the same pipeline. The nasal epithelial cell DNA (EPIC and Custom) and nasal lavage cell
793 DNA (Custom) are shown for the same randomly selected URECA children. The buccal,
794 placenta, and cord blood cells were from infants in the VCSIP Study.

795 **Fig. 4. Allergic sensitization (AS) EWAS results in URECA and INSPIRE children using the**
796 **EPIC and Custom arrays. (A-C)** Upper panel: Volcano plots showing the $\log_2$-fold change
797 in methylation (M-values) by proportion positive skin prick tests (x-axis) and the $-\log_{10}$(P-
798 value) from the EWAS (y-axis). Significant DMCs at a q-value threshold of 0.05 are shown
799 are shown in blue or purple (EPIC and high-value EPIC, respectively) in URECA, and
800 Custom in red (URECA) or turquoise (INSPIRE). The number of AS DMCs that were
801 hypermethylated or hypomethylated are shown as up and down arrows, respectively.
802 **(A-C)** Middle panel: Density plots of the DMC β values in each EWAS. (**A-B**) Lower panel:
803 Density plots of β values for DMCs that are eQTMs for their nearest genes (also see **Fig.**
804 **E10** in this article's Online Repository). **D**) Distributions of the distances of DMCs to the
805 next nearest DMC on the Custom (red) and EPIC (blue) arrays. The distance to nearest
806 DMC is shown on the x-axis; the proportion of DMCs in each distance bin is shown on
807 the y-axis. (**E-F**) Correlation of effect sizes (log fold change) of AS DMCs and allergic
808 asthma in URECA and INSPIRE. The effect size of DMCs for AS (x-axis) and allergic
809 asthma (y-axis) are shown. Red or turquoise dots were associated with allergic asthma
810 at a q-value ≤0.05 in URECA or INSPIRE, respectively; black dots are AS-only DMCs. (**G)**
811 Correlation of effect sizes (log fold change) between DMCs identified in the INSPIRE AS

812 EWAS (x-axis) and the URECA AS EWAS (y-axis). Red or turquoise points are AS DMCs
813 only in URECA or INSPIRE, respectively, at a q-value threshold of 0.05; black points are
814 AS DMCs in both at a q-value ≤0.05. The dashed lines in E-G are the correlation lines.

815 **Fig. 5. DMCs and eQTMs at 3 exemplary loci.** In each panel, locus zoom plots show the
816 DMCs from the three AS EWAS using the EPIC array (URECA; blue points = non-high
817 value CpGs; purple point = high-value [filtered] CpGs) and Custom array (URECA, red
818 points; INSPIRE, turquoise points). The genomic locations and genes are shown on the x-
819 axis; EWAS p-values are shown on the y-axis. The dashed horizontal lines show the 0.05
820 q-value threshold for the Custom array in URECA (red) and for the EPIC array in URECA
821 and Custom array in INSPIRE (blue and purple overlaid). The bar code at the top of the
822 plot shows the location of CpGs at this locus on both arrays. The most significant DMC in
823 the region for a given EWAS (cohort and platform; see legend) is illustrated by a diamond;
824 additional DMCs appear as circles. Correlations between methylation levels (x-axis) of the
825 lead Custom and high-value EPIC CpGs in URECA and expression of relevant genes (y-
826 axis) in epithelial cells from URECA children are shown in the upper right of each panel,
827 and boxplots of the proportion of positive skin prick tests (x-axis) and methylation levels
828 for the lead Custom and high-value EPIC CpGs in URECA (from the EWAS) are shown in
829 the lower right of each panel. **A)** *CISH* locus on chromosome 3. **B)** *SLC22A5-IRF1* locus on
830 chromosome 5. **C)** *HDAC7-VDR* locus on chromosome 12.

831

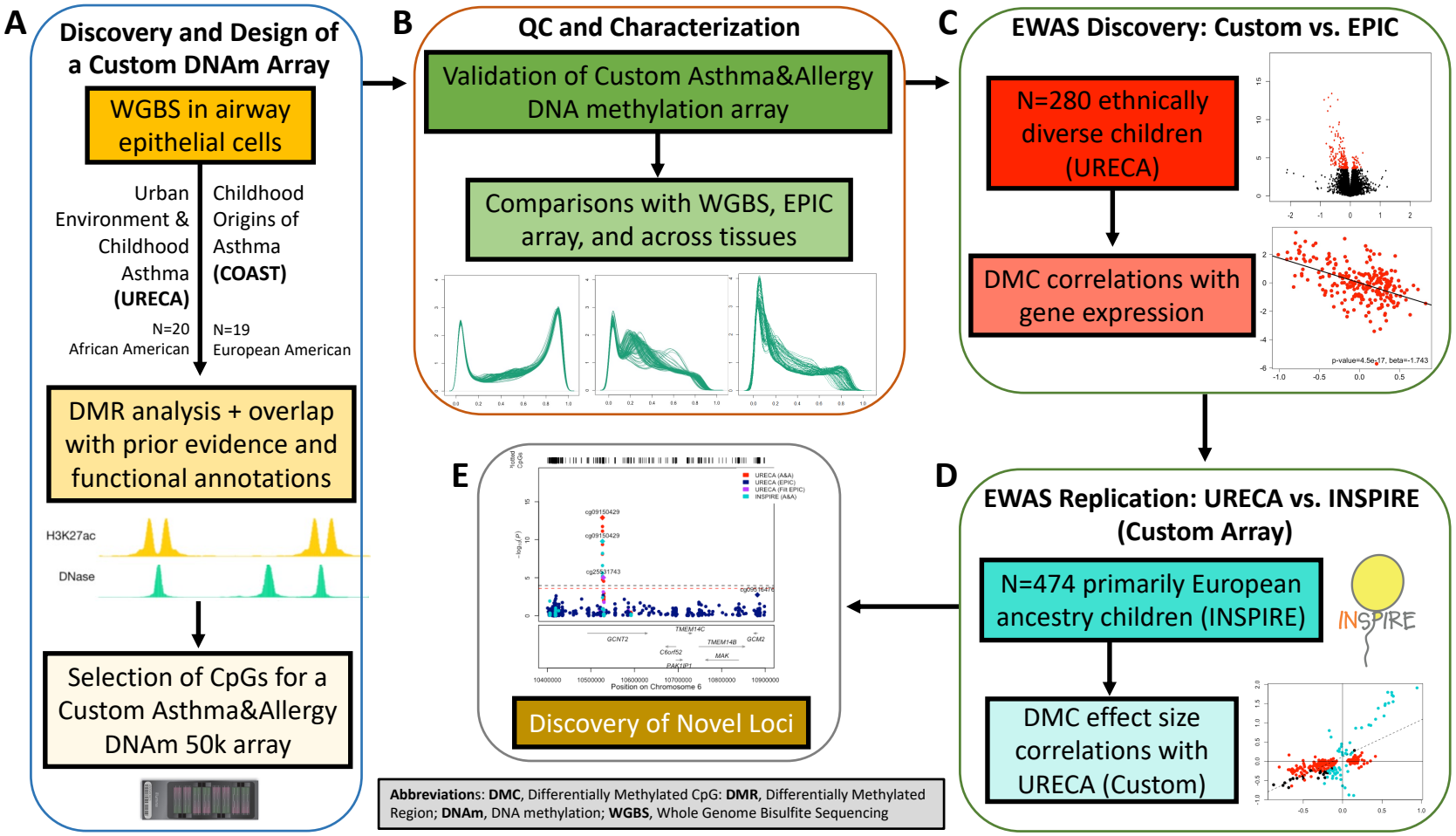
A
Discovery and Design of a Custom DNAm Array
WGBS in airway epithelial cells
Urban Environment & Childhood Asthma (URECA)
Childhood Origins of Asthma (COAST)
N=20 African American
N=19 European American
DMR analysis + overlap with prior evidence and functional annotations
H3K27ac
DNase
Selection of 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
D
EWAS Replication: URECA vs. INSPIRE (Custom Array)
N=474 primarily European ancestry children (INSPIRE)
INSPIRE
DMC effect size correlations with URECA (Custom)
E
Discovery of Novel Loci
Abbreviations: DMC, Differentially Methylated CpG: DMR, Differentially Methylated Region; DNAm, DNA methylation; WGBS, Whole Genome Bisulfite Sequencing

Fig. 1.

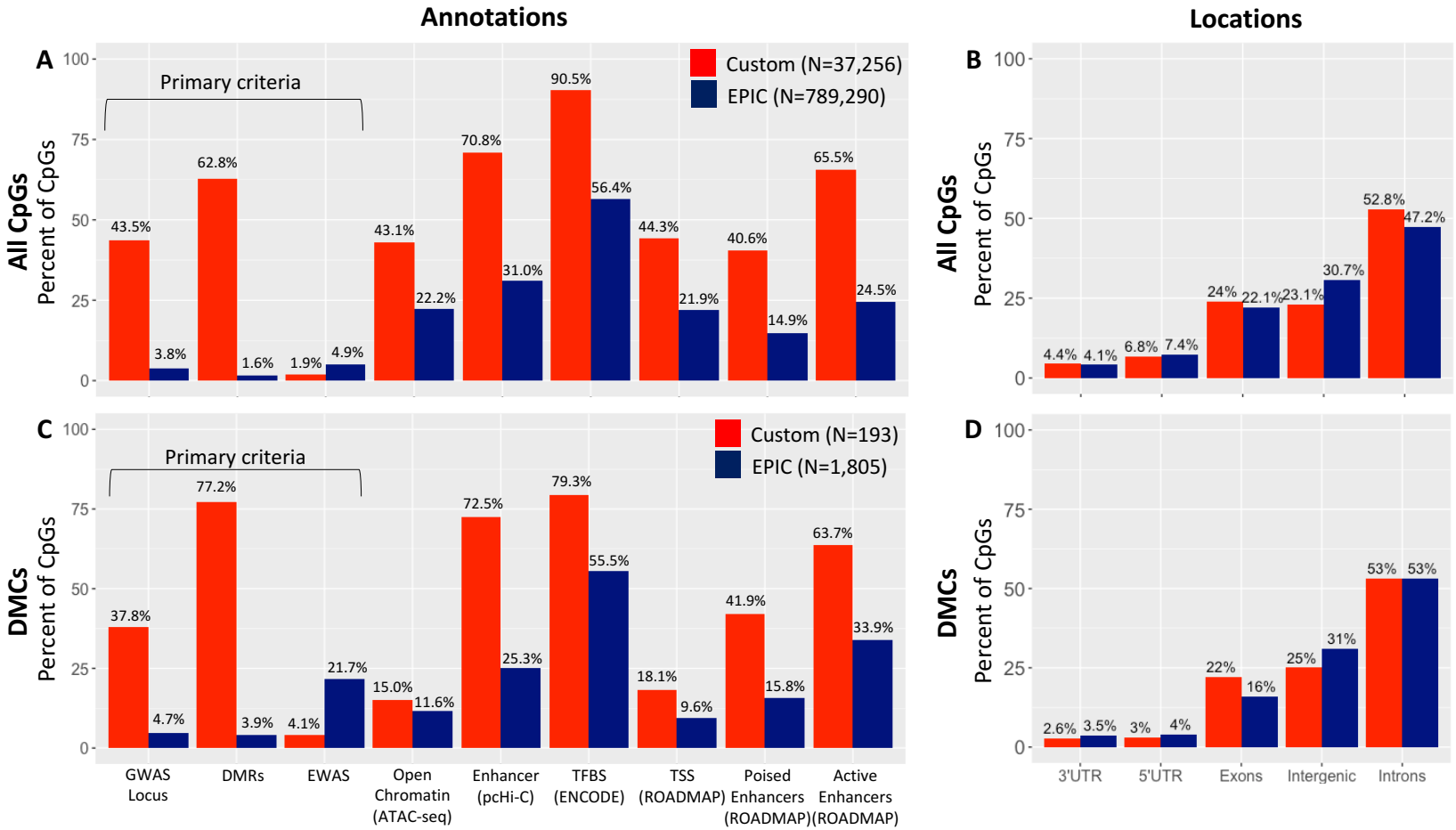
Annotations
Locations
A
B
C
D
All CpGs
DMCs
Percent of CpGs
Primary criteria
Custom (N=37,256)
EPIC (N=789,290)
Custom (N=193)
EPIC (N=1,805)
43.5%
3.8%
62.8%
1.6%
1.9%
4.9%
43.1%
22.2%
70.8%
31.0%
90.5%
56.4%
44.3%
21.9%
40.6%
14.9%
65.5%
24.5%
37.8%
4.7%
77.2%
3.9%
4.1%
21.7%
15.0%
11.6%
72.5%
25.3%
79.3%
55.5%
18.1%
9.6%
41.9%
15.8%
63.7%
33.9%
4.4%
4.1%
6.8%
7.4%
24%
22.1%
23.1%
30.7%
52.8%
47.2%
2.6%
3.5%
3%
4%
22%
16%
25%
31%
53%
53%
GWAS Locus
DMRs
EWAS
Open Chromatin (ATAC-seq)
Enhancer (pcHi-C)
TFBS (ENCODE)
TSS (ROADMAP)
Poised Enhancers (ROADMAP)
Active Enhancers (ROADMAP)
3'UTR
5'UTR
Exons
Intergenic
Introns

Fig. 2.

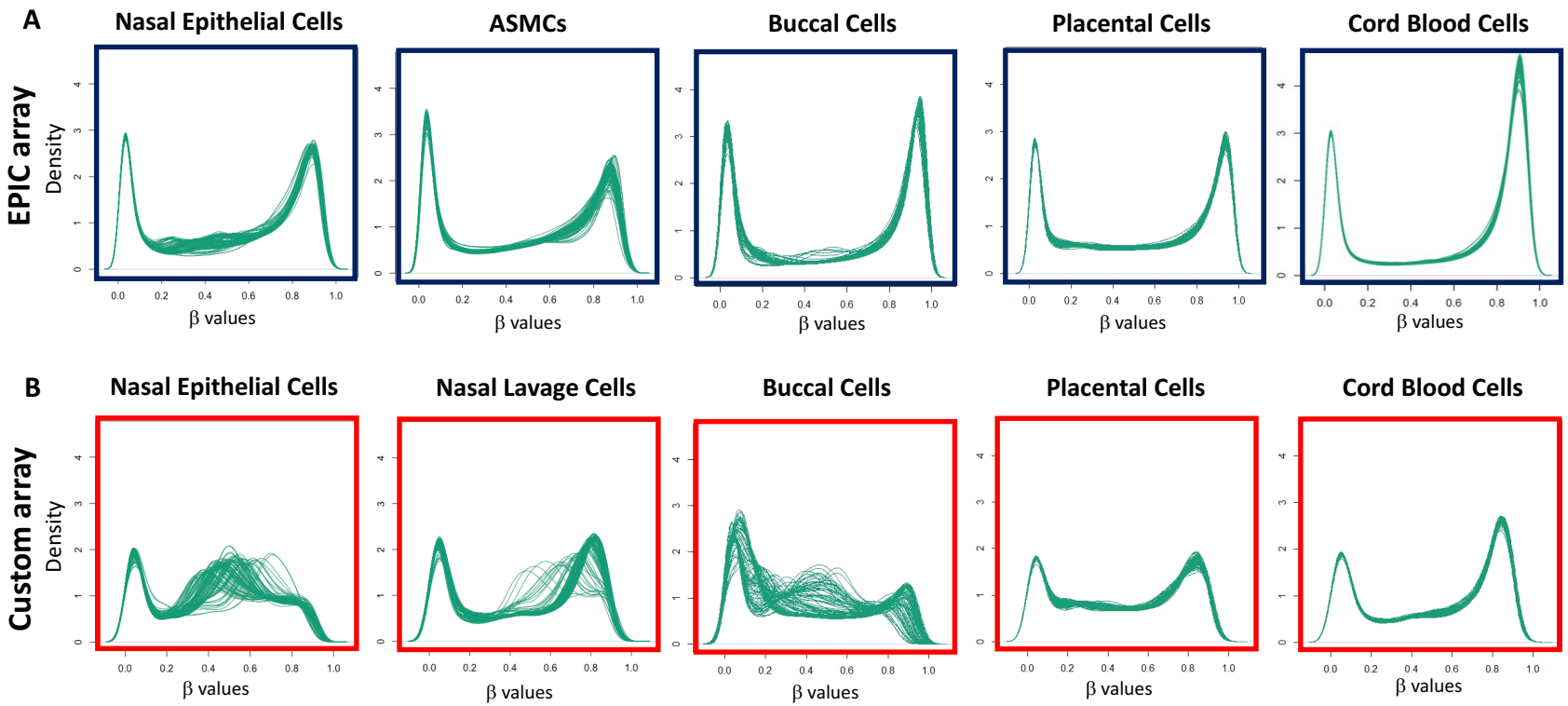
A
Nasal Epithelial Cells
ASMCs
Buccal Cells
Placental Cells
Cord Blood Cells
EPIC array
Density
β values
B
Nasal Epithelial Cells
Nasal Lavage Cells
Buccal Cells
Placental Cells
Cord Blood Cells
Custom array
Density
β values

Fig. 3.

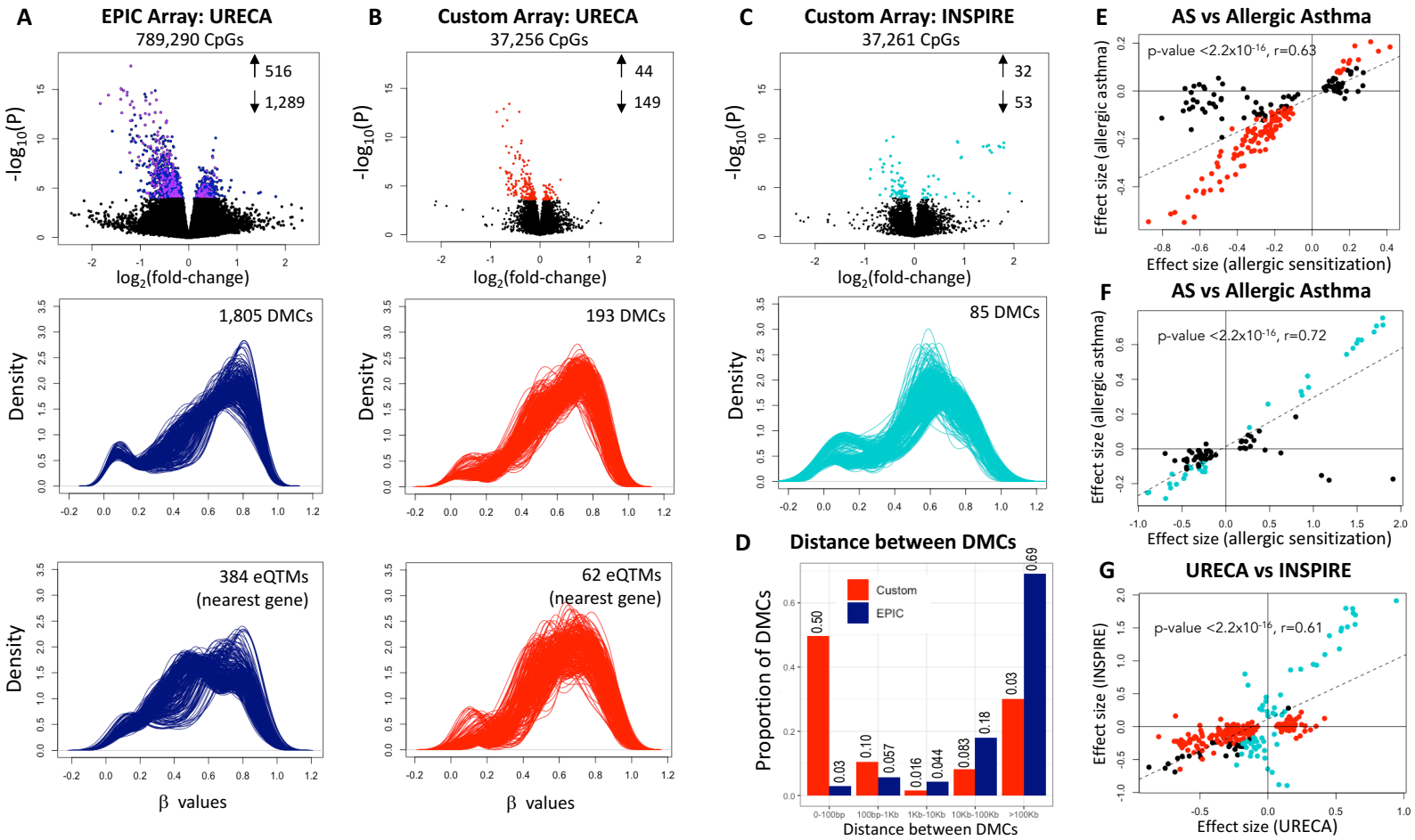
A
EPIC Array: URECA
789,290 CpGs
516
1,289
-log10(P)
log2(fold-change)
1,805 DMCs
Density
384 eQTMs
(nearest gene)
β values
B
Custom Array: URECA
37,256 CpGs
44
149
193 DMCs
62 eQTMs
(nearest gene)
C
Custom Array: INSPIRE
37,261 CpGs
32
53
85 DMCs
D
Distance between DMCs
Custom
EPIC
Proportion of DMCs
0.50
0.03
0.10
0.057
0.016
0.044
0.083
0.18
0.03
0.69
0-100bp
100bp-1Kb
1Kb-10Kb
10Kb-100Kb
>100Kb
Distance between DMCs
E
AS vs Allergic Asthma
p-value <2.2x10-16, r=0.63
Effect size (allergic asthma)
Effect size (allergic sensitization)
F
AS vs Allergic Asthma
p-value <2.2x10-16, r=0.72
Effect size (allergic asthma)
Effect size (allergic sensitization)
G
URECA vs INSPIRE
p-value <2.2x10-16, r=0.61
Effect size (INSPIRE)
Effect size (URECA)

Fig. 4.

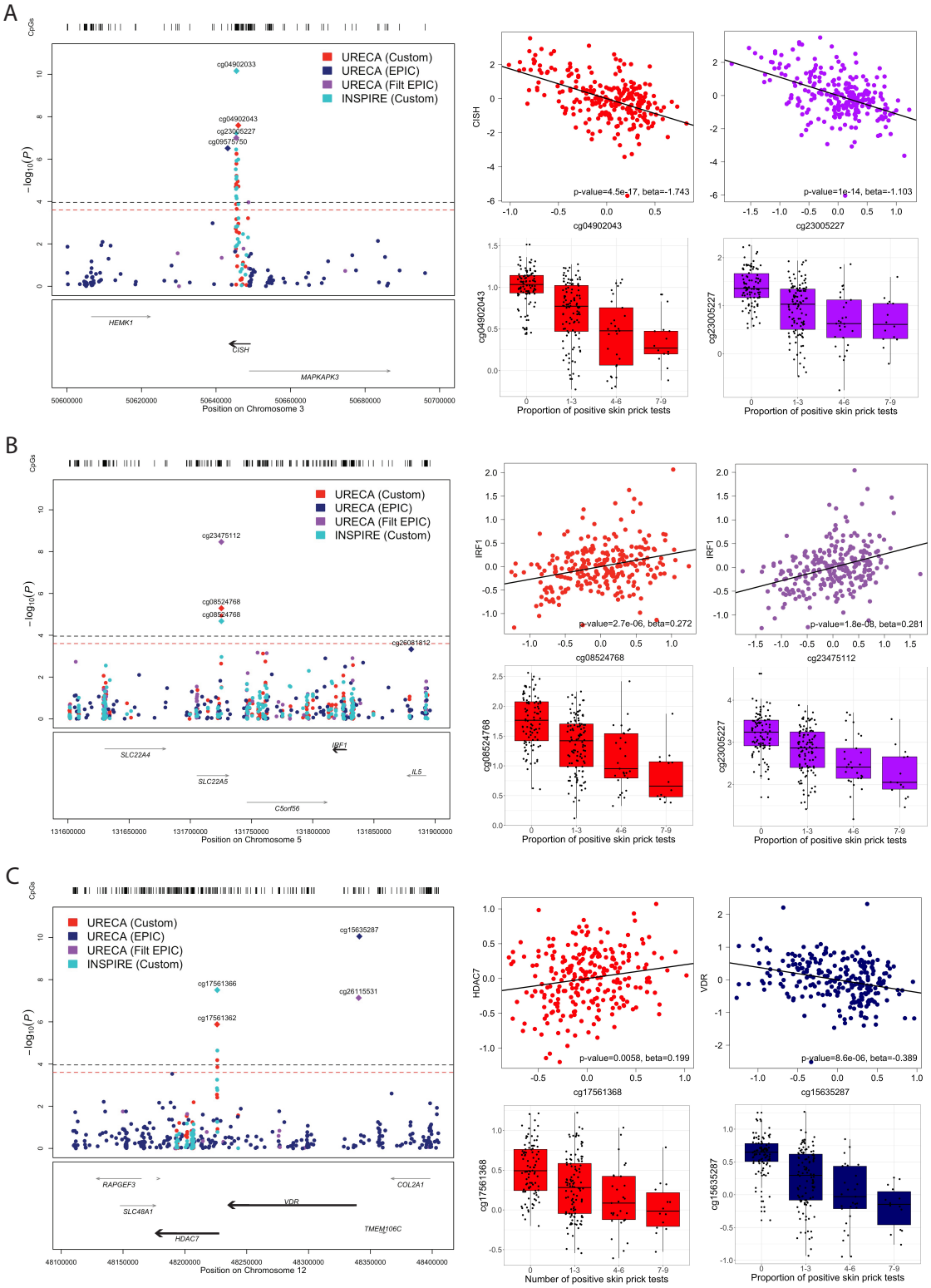
A
CpGs
URECA (Custom)
URECA (EPIC)
URECA (Filt EPIC)
INSPIRE (Custom)
cg04902033
cg04902043
cg23005227
cg09575750
−log10(P)
HEMK1
CISH
MAPKAPK3
Position on Chromosome 3
CISH
p-value=4.5e-17, beta=-1.743
cg04902043
p-value=1e-14, beta=-1.103
cg23005227
Proportion of positive skin prick tests
B
cg23475112
cg08524768
cg08524768
cg26081812
SLC22A4
SLC22A5
IRF1
IL5
C5orf56
Position on Chromosome 5
IRF1
p-value=2.7e-06, beta=0.272
cg08524768
p-value=1.8e-08, beta=0.281
cg23475112
C
cg15635287
cg17561366
cg26115531
cg17561362
RAPGEF3
SLC48A1
HDAC7
VDR
COL2A1
TMEM106C
Position on Chromosome 12
HDAC7
p-value=0.0058, beta=0.199
cg17561368
VDR
p-value=8.6e-06, beta=-0.389
cg15635287
Number of positive skin prick tests
Proportion of positive skin prick tests

Fig. 5