Genetic-Epigenetic Interactions in Asthma Revealed by a Genome-Wide Gene-Centric Search

Vladimir Kogan^a Joshua Millstein^a Stephanie J. London^b Carole Ober^c
Steven R. White^d Edward T. Naureckas^d W. James Gauderman^a
Daniel J. Jackson^e Albino Barraza-Villarreal^f Isabelle Romieu^g
Benjamin A. Raby^h Carrie V. Breton^a

^aDepartment of Preventive Medicine, Keck School of Medicine of the University of Southern California, Los Angeles, CA, USA; ^bDivision of Intramural Research, National Institute of Environmental Health Sciences, National Institutes of Health, Department of Health and Human Services, RTP, Research Triangle Park, NC, USA; ^cDepartment of Human Genetics, University of Chicago, Chicago, IL, USA; ^dDepartment of Medicine, University of Chicago, Chicago, IL, USA; ^eUniversity of Wisconsin School of Medicine and Public Health, Madison, IL, USA; ^fDepartment of Environmental Health, Population Health Center, National Institute of Public Health of Mexico, Cuernavaca, Mexico; ^gInternational Agency for Research on Cancer, Section of Nutrition and Metabolism, Lyon, France; ^hDepartment of Medicine, Channing Division of Network Medicine, Brigham and Women's Hospital, Boston, MA, USA

Keywords

DNA methylation · Statistical interactions · Asthma susceptibility · SNPs · Integrative genomics

Abstract

Objectives: There is evidence to suggest that asthma pathogenesis is affected by both genetic and epigenetic variation independently, and there is some evidence to suggest that genetic-epigenetic interactions affect risk of asthma. However, little research has been done to identify such interactions on a genome-wide scale. The aim of this studies was to identify genes with genetic-epigenetic interactions associated with asthma. **Methods:** Using asthma case-control data, we applied a novel nonparametric gene-centric approach to test for interactions between multiple SNPs and CpG sites simultaneously in the vicinities of 18,178 genes across the

genome. **Results:** Twelve genes, *PF4*, *ATF3*, *TPRA1*, *HOPX*, *SCARNA18*, *STC1*, *OR10K1*, *UPK1B*, *LOC101928523*, *LHX6*, *CHMP4B*, and *LANCL1*, exhibited statistically significant SNP-CpG interactions (false discovery rate = 0.05). Of these, three have previously been implicated in asthma risk (*PF4*, *ATF3*, and *TPRA1*). Follow-up analysis revealed statistically significant pairwise SNP-CpG interactions for several of these genes, including *SCARNA18*, *LHX6*, and *LOC101928523* ($p = 1.33\text{E-}04$, $8.21\text{E-}04$, $1.11\text{E-}03$, respectively). **Conclusions:** Joint effects of genetic and epigenetic variation may play an important role in asthma pathogenesis. Statistical methods that simultaneously account for multiple variations across chromosomal regions may be needed to detect these types of effects on a genome-wide scale.

© 2019 S. Karger AG, Basel

Vladimir Kogan and Joshua Millstein contributed equally to this work.

Introduction

Asthma is a chronic inflammatory disorder of the airways, characterized by airway hyperresponsiveness and airflow limitation. Asthma prevalence has increased in recent years [1], afflicting 300 million worldwide, and is highly variable across population centers, ranging from 1 to 18% [2]. In the United States, more than 23 million people were diagnosed with asthma as of August 2015, including over 6 million children [3].

Numerous genome-wide association studies have shown a substantial genomic contribution to the etiology of asthma, with heritability estimates varying between 35 and 95% [4–6]. However, only a fraction of this variation has been explained by specific causal variants, such as those near Chr17q12–21 [7, 8]. Epigenetic mechanisms, another source of disease variation, may explain a portion of this missing heritability [9–11]. Epigenomic alterations, while heritable, may also occur as a response to the endogenous and exogenous environment, and contribute to asthma pathogenesis [9, 12].

There is evidence to suggest that genetic and epigenetic variation interact synergistically to affect gene expression [13]. In fact, interactions between SNPs and DNA methylation in the genomic regions of T-helper 2 pathway genes *IL4R* and others were found to affect asthma risk [14, 15], possibly through expression of these genes.

Methods for detecting statistical interactions in population studies often involve linear or logistic models that include a product term to represent the interaction effect. These pairwise approaches suffer from a multiple testing challenge [16, 17], which is particularly severe if applied on a genome-wide scale. For example, in a study with 1 million SNPs and 450,000 CpG sites, 450 billion tests would be required. In addition, pairwise approaches may not optimally leverage information from dependencies between SNPs and CpG sites across contiguous genomic regions. Using case-control data, we applied a novel gene-centric approach to test for interactions between multiple SNPs and CpG sites simultaneously in the neighborhood of each gene across the whole genome.

Methods

Asthma BRIDGE Data

These data are from the Asthma BioRepository for Integrative Genomic Exploration (Asthma BRIDGE) [18–21], which is publicly accessible and includes 1,542 individuals with asthma and controls with comprehensive phenotype and genomic data. Asth-

ma was defined via a questionnaire by the presence of asthma symptoms, use of an inhaled bronchodilator at least twice per week, or use of a daily asthma medication for the 6 months before the screening interview. Genome-wide SNP and DNA methylation was obtained from whole blood samples from 576 participants. Samples were randomized to avoid confounding by experimental batch. The final data set included 356 asthmatic and 220 nonasthmatic adults, for whom we had complete data.

Genotyping and DNA Methylation

Genotyping methods are described in detail in Torgerson et al. [19]. Briefly, SNPs included all phase 2, release 21 consensus HapMap variants and were obtained from studies participating in the EVE consortium [19, 22]. The genotyping platforms, Illumina (1Mv1, 550k, 610k, 650k) and Affymetrix (500k, 6.0), varied by sample source. Quality control procedures included filtration for call rates (>95%) and tests for agreement with Hardy-Weinberg expectations applied to SNPs oriented to the plus strand. High-density imputation using the 1000 Genomes reference panel was performed with the MACH software [23]. Data were checked for consistency in reference alleles and strand orientation, and SNPs were excluded based on low quality scores and examination of QQ plots comparing the distribution of association *p* values for genotyped and imputed markers. SNPs with minor allele frequencies less than 0.05 were excluded.

A total of 11 plates were run on the Illumina human 450K methylation platform. Data preprocessing steps included the application of norm-exponential background correction on the raw methylation data, plate by plate for all 11 plates. HG19 annotations, including gene symbols, chromosome number, and sequence position, were added via the Bioconductor package, FDb.InfiniumMethylation.hg19. One sample from a nonasthmatic participant was excluded due to having an extreme value of 1.0 for all probes. Methylation of CpG sites were assessed for outliers, and observations exceeding 3 interquartile ranges from the 25th and 75th percentiles, per Tukey's outer fences [24], were removed unless they comprised greater than or equal to 5% of the data for that variable. The 11 norm-exponential corrected plates were combined, and 4 additional samples were excluded as outliers. Dye bias correction was done using one sample as reference for all others, and quantile normalization was applied to the resulting data set.

Canonical Analysis of Set Interactions

Here we describe and apply a new method, canonical analysis of set interactions (CASI), designed to identify statistical interactions between high-order sets of features in a simultaneous approach. In contrast to direct interactions in which DNA sequence variation affects DNA methylation, we are seeking to identify genomic regions underlying genes where there is synergism between the effects of SNPs and methylation on asthma susceptibility.

CASI is described in detail in the Appendix Methods and is freely available to the public in a downloadable R package (<https://github.com/USCbiostats/CASI>). Briefly, it is a hypothesis testing approach to detect differences in correlations between linear combinations of SNPs and CpG sites in cases versus controls. In this study, SNPs and CpG sites were mapped to the nearest gene, thus defining the feature sets, with positions identified using the UCSC genome browser for all RefSeq genes. Overlapping isoforms of the

same gene were combined to form a single full-length version. SNPs and CpG sites residing in flanking regions extending 2 kb on either side were included. A total of 18,178 gene-centric regions were identified across the genome. Using canonical correlation analysis (CCA), coefficients that maximize the correlation, the *canonical correlation*, between linear combinations of the SNPs and CpG sites are estimated in cases. The same coefficients are then used to compute linear combinations in controls. The CASI statistic is based on the scaled difference between Fisher transformed canonical correlation coefficients in cases versus controls. Estimation is performed in cases because it leverages the superior statistical power of the case-only interaction analysis approach [25]; dependencies among predictive variables in cases may indicate multiplicative interactions in their effects on disease [26, 27]. Due to the limitations of CCA in the number of variables relative to the sample size, sparse CCA as described by Witten et al. [28] is applied for sets with dimensions that are too large to be accommodated by conventional CCA. The parametric distribution of this test statistic, the scaled difference between cases and controls, is unknown. Therefore, we use a permutation-based procedure to estimate the null distribution by randomly permuting case status and computing the test statistic. The permutation analysis is conducted repeatedly until a sufficient number of values have been generated under the null. Rather than compute *p* values, a computationally efficient false discovery rate (FDR) approach is used that requires few permutations and yields confidence intervals that account for dependencies among tests as well as the number of permutations conducted [29]. For this study, 100 permutation analyses were conducted, and statistical significance was defined by an FDR threshold of 0.05.

Follow-Up Analyses in Significant Regions

SNPs and CpG sites can be ranked in their contribution to the test statistic by computing correlations, *loadings*, in cases with their corresponding linear combinations, the *canonical variates*. SNP-CpG site pairs that were the greatest contributors to statistically significant tests were identified by selecting all SNP-CpG pairs among those with loadings greater than 0.5 within a gene region. The loading threshold of 0.5 has been proposed as the “operational definition” of a large effect size [30]. Using logistic regression, conventional likelihood ratio tests of multiplicative interactions were conducted, including one-degree of freedom likelihood ratio tests for interaction and three-degree of freedom tests for the combined main and interaction effects. Statistical significance was assessed according to FDR level of 0.05, using a parametric version of Millstein and Volfson [31]. Variables found to be significantly correlated with asthma status (Table 1) were included as adjustment covariates in the logistic models. While the education level was not significant ($p = 0.14$), it was nevertheless included due to prior evidence of association with asthma. Final models were adjusted for gender, ancestry, family history of asthma, age at sample collection, and education level. The ancestry variables were composed of the top two principal components computed by the software Eigensoft [32] using 128 ancestry informative markers [33]. The primary results do not include adjustment for site, because two ABRIDGE sites contributed only cases (Table 1). However, sensitivity analyses were conducted by restricting to sites that included both cases and controls and including site as an adjustment covariate, and the results did not change substantially (Appendix Table A1).

Table 1. Characteristics of participants in the ABRIDGE study

Characteristic	Asthma cases, <i>n</i> (%)	Asthma controls, <i>n</i> (%)
Total number	356 (61.81)	220 (38.19)
Gender		
Male	169 (47.5)	78 (35.5)
Female	187 (52.5)	142 (64.5)
Race/ethnicity		
European	73 (20.5)	41 (18.6)
Hispanic/Latino	174 (48.9)	65 (29.5)
Black or African American	67 (18.8)	92 (41.8)
Other/multiple/uncertain	42 (11.8)	22 (10.0)
Education		
Did not complete high school	41 (11.5)	20 (9.1)
High school or GED degree	96 (27.0)	51 (23.2)
Some college	97 (27.2)	74 (33.6)
College degree	71 (19.9)	44 (20.0)
Post-college coursework	18 (5.1)	15 (6.8)
Graduate or professional degree	18 (5.1)	14 (6.4)
Other	15 (4.2)	2 (0.9)
Site		
GRAAD, JH	39 (11.0)	87 (39.5)
CAG/UAC	32 (9.0)	18 (8.2)
CHS/USC	119 (33.4)	115 (52.3)
CARE; Denver NJH, Tucson/UARC, St. Louis/WUC, Madison/UWM	42 (11.8)	0 (0)
MCCAS	124 (34.8)	0 (0)
Age at methylation measure		
0–10 years	4 (1.1)	0 (0)
10–20 years	134 (37.6)	4 (1.8)
20–30 years	162 (45.5)	143 (65.0)
30–40 years	23 (6.5)	21 (9.5)
40–50 years	16 (4.5)	27 (12.3)
50–60 years	13 (3.7)	21 (9.5)
>60 years	4 (1.1)	4 (1.8)
Family history of asthma		
Yes	173 (48.6)	51 (23.2)
No	183 (51.4)	169 (76.8)
Age at diagnosis		
0–2 years	90 (25.3)	
2–5 years	100 (28.1)	
5–8 years	64 (18.0)	
8–11 years	57 (16.0)	
11–14 years	15 (4.2)	
14–17 years	9 (2.5)	
>17 years	21 (5.9)

Fig. 1. FDR and confidence intervals for 18,178 genomic regions (genes) defined by 2kb up and downstream of each gene. The value of the CASI statistic (x axis) has no direct interpretation other than that larger values are more extreme. 12 significant genomic regions correspond to a threshold of 4.35 (vertical dashed line) and FDR of 0.05, CI = (0.024, 0.104). Integer values shown in green specify the number of genomic regions with CASI statistics at least as extreme as the thresholds specified by the horizontal axis.

Results

Performance of the CASI Approach in Simulation Studies

The simulation results demonstrate that the CASI hypothesis test is sensitive to correlation differences between cases and controls and that sensitivity is associated with magnitude of the differences (Appendix Fig. A1). We also found that CASI is a consistent test with asymptotics depending on effect size and that type I error is very low as the effect size approaches zero (Appendix Fig. A2). CASI performed quite well in the simulation results as compared to other related approaches, especially in the vicinity of conventional levels of FDR that imply statistical significance (Appendix Fig. A3). Applying a conventional pairwise logistic regression approach combined with FDR-evaluated significance to the same simulated data resulted in a complete failure to detect interactions, highlighting the utility of the proposed approach (Appendix Fig. A3).

Demographics of the Asthma BRIDGE Population

Of the 576 participants, 62% had asthma (Table 1). While males and females were evenly distributed among cases, there were more females than males among the controls. Asthma cases had a greater percentage of Hispanics whereas controls had a greater proportion of African Americans. Frequency of asthma prevalence varied by clinic site by design, with some sites recruiting only cases.

CASI Analysis of Asthma BRIDGE Data

Note that p values were not generated in the main analysis even as intermediate statistics because FDR was estimated directly from the observed and permuted CASI statistics. The estimated FDR and corresponding confidence intervals (CIs) for a series of increasingly stringent significance thresholds for the CASI test statistic demonstrate a downward trend with narrow CIs (Fig. 1), indicating that the CASI statistic is informative with respect to distinguishing observed from permuted data. However, this dynamic does not continue beyond about 4.35 where a minimum FDR occurs, implying that more extreme values of the CASI statistic do not correspond to lower rates of false discoveries. The threshold of 4.35 was therefore used to identify the most statistically significant results, yielding 12 genes: *HOPX*, *SCARNA18*, *PF4*, *STC1*, *ATF3*, *OR10K1*, *UPK1B*, *LOC101928523*, *LHX6*, *CHMP4B*, *TPRA1*, and *LANCL1* (FDR = 0.050, CI = 0.024–0.104). Though we highlight these 12 genes due to low FDR, it is clear from the tight CIs bracketing FDR estimates at more permissive significance thresholds that substantially more such interactions are implied by these results.

Pairwise SNP-CpG Interaction Analysis

Pairwise analyses of SNPs and CpG sites from the 12 significant genomic regions with loadings of 0.5 or greater using logistic regression revealed evidence of pairwise interactions for 19 top-loading pairs (Table 2), corresponding to 4 of the 12 genomic regions, i.e., *LOC101928523*, *SCARNA18*, *LHX6*, and *STC1*.

Although the three-degree of freedom tests were significant for most of these pairs, the one-degree of freedom tests tended to be more significant, implying that main effects were not appreciably contributing. Nine individual SNPs and 10 CpG sites in *LHX6*, 3 SNPs and 2 CpG sites in *LOC101928523*, and 3 SNPs and 2 CpG sites in *SCARNA18* met the criteria of having loadings greater than 0.5 (Fig. 2). The SNP-CpG pair with the greatest loadings did not always elicit a significant result for multiplicative interaction (Appendix Table A2); however,

Table 2. Significant interactions for SNPs and CpGs with loadings >0.5

SNP	Methylation	Coefficient SNPxMeth interaction	<i>p</i> value (1-df LRT)	<i>p</i> value (3-df LRT)	Gene
rs67216017(G/GA)	cg14999833	19.79	2.84E-05	1.33E-04	SCARNA18
rs113665237(A/AT)	cg14999833	19.79	2.84E-05	1.33E-04	SCARNA18
rs10061690	cg14999833	19.79	2.84E-05	1.33E-04	SCARNA18
rs10818651	cg01363324	-24.84	1.17E-04	8.21E-04	LHX6
rs13301641	cg07956857	-8.11	1.28E-04	1.11E-03	LOC101928523
rs11792474	cg07956857	-7.86	1.79E-04	1.49E-03	LOC101928523
rs10818651	cg21469772	-6.47	2.49E-04	3.74E-03	LHX6
rs10985567	cg01363324	-23.97	2.99E-04	4.35E-04	LHX6
rs10818651	cg04282082	-4.99	3.59E-04	4.88E-03	LHX6
rs10818651	cg13832372	12.80	3.68E-04	1.44E-03	LHX6
rs10985567	cg21213617	-12.88	6.19E-04	4.51E-03	LHX6
rs10985567	cg13832372	12.61	6.42E-04	4.55E-04	LHX6
rs10985567	cg21469772	-6.19	8.17E-04	4.18E-03	LHX6
rs989798	cg04282082	-4.70	1.17E-03	7.01E-03	LHX6
chr8/BP:23713016 (C/CT)	cg16688533	-3.99	1.20E-03	3.87E-03	STC1
rs989798	cg13832372	12.33	1.42E-03	1.85E-03	LHX6
rs10818651	cg21213617	-11.34	1.54E-03	1.53E-02	LHX6
rs989798	cg21469772	-5.69	1.60E-03	1.24E-02	LHX6
rs10985567	cg04282082	-4.61	1.73E-03	4.51E-03	LHX6

Statistically significant (FDR <0.05) pairwise interactions (1-df LRT) for top loading SNPs and CpGs from the 12 genomic regions identified as significant by CASI. Logistic regression models were adjusted for gender, ancestry/ethnicity, family history of asthma, age at methylation measure, and education level. Very similar *p* values may be indicative of SNPs in high linkage disequilibrium. 1-df LRT, one-degree of freedom likelihood ratio test; 3-df LRT, three-degree of freedom likelihood ratio test.

this is not surprising considering that the CASI statistic is formed from linear combinations of SNPs and CpGs. Multiple genomic regions, *HOPX*, *PF4*, *ATF3*, *UPK1B*, *CHMP4B*, *TPRA1*, and *LANCL1*, did not show evidence of interactions for their top loading SNP-CpG pair (Appendix Table A2). The lack of statistical significance for individual pairs may indicate that it is necessary to evaluate joint interactions between multiple SNPs and CpGs in order to have adequate power to detect the effects.

Additional pairwise interaction analyses were conducted for SNPs and CpGs with loadings greater than 0.5 underlying *LHX6*, *LOC101928523*, and *SCARNA18* (Appendix Tables A3 and A4). Significant interactions were found for *LHX6*, where involved SNPs had negative loadings located at the center haplotype block interacting with CpG sites clustered to the right in a CpG island (Fig. 2; Table 3). Odds ratios (ORs) for the main effects of both SNPs and CpGs in *LHX6* were not statistically significant (Table 3). However, interactions for 7 SNP-CpG pairs were significant, with ORs ranging from 0.53 (0.36–0.77)

to 0.84 (0.74–0.96) (Table 3). High linkage disequilibrium (LD) between SNPs as well as dependencies among CpGs may explain the similarity in interaction effects that is apparent across SNP-CpG pairs. For *LHX6*, minor allele dose was associated with increasingly protective effects of methylation (Table 3). This trend is particularly apparent in the conditional ORs (OR_{meth|SNP}). For the most significant interaction in *LHX6*, individuals with the common GG genotype in rs10818651 had 9% greater odds of asthma for a 5% difference in cg21469772 methylation. Addition of 1 minor allele A reverses the association to a 21% decrease in the odds of asthma for 5% greater methylation, whereas individuals with AA had a 43% decrease. Similarly, for the next most significant interaction (rs10985567 and cg21213617), the odds of asthma for an individual with 5% higher methylation and GG (common homozygote) increases by 24% whereas the odds for an individual with 5% higher methylation and AA decreases by 66%.

Fig. 2. SNP and methylation loadings for genomic regions *LOC101928523*, *LHX6*, and *SCARNA18* with linkage disequilibrium heat maps. Red circles indicate SNPs and blue squares CpG

sites. The horizontal axis indicates base pair (BP) positions in each of the three genomic regions. Heat maps represent dependencies between SNPs and CpGs (alignment is approximate).

Table 3. Tests of multiplicative interactions ($p < 0.05$) within *LHX6* for SNP-CpG pairs

SNP (minor/major alleles) methylation	Minor allele freq.	p value interaction	5% difference in methylation			
			OR _{Meth} SNP (95% CI)	OR _{Interaction} (95% CI)	OR _{SNP} (95% CI)	OR _{Meth.} (95% CI)
rs10818651 (A/G) cg21469772	0.25	2.49E-04	GG: 1.09 (0.97–1.23) AG/GA: 0.79 (0.58–1.07) AA: 0.57 (0.35–0.94)	0.72 (0.60–0.87)	1.00 (0.66–1.53)	0.99 (0.93, 1.07)
rs10818651 (A/G) cg04282082	0.25	3.59E-04	GG: 1.08 (0.99–1.18) AG/GA: 0.84 (0.67–1.07) AA: 0.66 (0.45–0.96)	0.78 (0.67–0.90)	1.00 (0.66–1.53)	1.01 (0.95–1.06)
rs10985567 (A/G) cg21213617	0.23	6.19E-04	GG: 1.24 (0.99–1.55) AG/GA: 0.65 (0.35–1.20) AA: 0.34 (0.13–0.93)	0.53 (0.36–0.77)	1.27 (0.84–1.96)	0.96 (0.83–1.12)
rs10985567 (A/G) cg21469772	0.23	8.17E-04	GG: 1.13 (1.01–1.27) AG/GA: 0.83 (0.61–1.13) AA: 0.61 (0.37–1.00)	0.73 (0.60–0.89)	1.27 (0.84–1.96)	0.99 (0.93–1.07)
rs989798 (T/C) cg04282082	0.20	1.17E-03	CC: 1.08 (1.01–1.17) TC/CT: 0.86 (0.69–1.07) TT: 0.68 (0.47–0.98)	0.79 (0.68–0.92)	1.21 (0.78–1.88)	1.01 (0.95–1.06)
rs10985567 (A/G) cg04282082	0.23	1.73E-03	GG: 1.11 (1.02–1.20) AG/GA: 0.88 (0.70–1.11) AA: 0.70 (0.48–1.03)	0.79 (0.68–0.92)	1.27 (0.84–1.96)	1.01 (0.95–1.06)
rs10985567 (A/G) cg03363289	0.23	4.30E-03	GG: 1.17 (1.06–1.28) AG/GA: 0.98 (0.78–1.23) GG: 0.83 (0.58–1.17)	0.84 (0.74–0.96)	1.27 (0.84–1.96)	1.03 (0.97–1.09)

SNPs (in region chr9:124962877–124992824, GRCh37/hg19) were selected for presentation that reside in the central haplotype block and have negative loadings greater than 0.5 in absolute value. Models were adjusted for gender, ancestry/ethnicity, family history of asthma, age at methylation measure, and education level. CpG sites clustered near the CpG island toward the right of the gene as shown in Figure 2. ORs reflect a 5% change in methylation.

In *LHX6*, the most statistically significant interaction involved the SNP with the greatest loading (rs10818651, loading = -0.86) (Table 2) that resides within the middle LD block (between base pair locations 124972042 and 124982500, indicated by vertical bars in Fig. 2), paired with CpG site cg01363324 (loading = 0.53). The opposite signs of the loadings for this pair (Fig. 2) convey that these variables are negatively correlated among cases (Fig. 3), which is reflected in the interaction pattern. Increasing minor allele dose of the SNP reflects an association between methylation level and asthma that is increasingly negative.

Boxplots of methylation values with jittered points can provide a visual demonstration of relationships underlying the significant interaction effects (Fig. 3). For example, *LHX6* median methylation (cg04282082) is greater in individuals with asthma than controls with the GG genotype (rs10818651), but less within AG individuals. In general, for *LHX6*, increased methylation was protective in

the presence of a minor allele (Table 3). In another example, median methylation of cg16688533 in *STC1* is greater in individuals with asthma versus those without in AA individuals (rs9969426), but those with GG tend to be unmethylated and have asthma. Another clear example can be observed for *LOC101928523*, where median methylation (cg07956857) is greater in CC individuals with asthma (rs13301641), approximately equal in TC individuals but less in TT.

SNPs within the *LOC101928523* genomic region were confined to a single haplotype block (Fig. 2). The highest loading SNPs, rs13301641 (loading = 0.87), rs11792474 (loading = 0.87), and rs75088949 (loading = -0.68) as well as the highest loading CpG site, cg07956857 (loading = -0.89) were the most statistically significant interactions for this genomic region (Appendix Table A3). As with *LHX6*, there was little evidence of main effects; however, unlike *LHX6*, minor allele dose was accompanied by both decreased and increased ORs. The most significant inter-

Fig. 3. Interaction box plots for SNP-methylation pairs with highest loadings in the top 12 genomic regions (Appendix Table A1 for adjusted odds ratios).

action for *LOC101928523*, between rs13301641 and cg07956857, involved loadings with opposite signs (Fig. 2), which reflected negative correlation between methylation level and number of minor alleles among individuals with asthma.

For *SCARNA18*, there were 3 SNPs, rs67216017, rs113665237, and rs10061690, all in high LD, with loadings greater than 0.5. Thus, interaction analysis results are indistinguishable among the 3 (Appendix Table A4). The

loadings for SNPs and CpGs are concordant (Fig. 2), which is consistent with the idea that minor allele dose is associated with increasingly deleterious effects of methylation on asthma susceptibility. The range of methylation for cg14999833 is narrow (Fig. 3), hence it is appropriate to estimate the effect over a 1% change. Deletion of allele “A” near rs67216017 or “T” near rs113665237 appears to have an equivalent effect to presence of minor allele “A” at rs10061690 on the relationship between

methylation at the CpG site cg14999833 and asthma. With respect to cg14999833, a biologically meaningful effect is observed for a 1% difference in methylation in contrast to cg20697188 where a 5% difference is within the range of the observed data.

Discussion

The proposed statistical approach, CASI, was able to overcome the substantial challenges imposed by high dimensional data to identify statistical interactions between methylation variation at CpG sites and DNA variation in SNPs. Part of the explanation for the adequate power that led to this success as well as the superior performance of CASI in simulated data in comparison to other similar methods is the fact that CASI leverages the concept underlying the case-only interaction method, known to be a powerful approach [25, 34]. This characteristic is unique to CASI among related set-based methods designed to detect statistical interactions [35–37]. Another unique characteristic of the CASI approach is the reliance on a computationally efficient permutation-based FDR estimator to define statistical significance. This is a nonparametric approach that is robust to feature distributions and yields CIs that bracket the FDR estimates, allowing uncertainty in those estimates to be quantified. These CIs play an especially important role in studies of weak effects where an FDR level of 0.05 may not be achievable.

A limitation of the CASI procedure is the inability to explicitly adjust for potential confounding covariates. However, this problem was mitigated in follow-up analyses by applying conventional logistic regression with covariates to pairs of features implied by CASI results. Like many conventional approaches, CASI is designed to detect linear associations, which limits the power for non-linear relationships. However, considering that we were able to identify 12 genes with evidence of interactions using a sample size in the hundreds suggests that more interactions may be detectable if the sample size is increased substantially. Though the procedure requires considerable computational resources due to its reliance on permutations, the required number of permutations is small enough to allow genome-wide applications.

Though several previous studies have investigated statistical interactions between genetic and epigenetic variation [13–15, 38], we know of no others that have attempted a genome-wide assessment. CASI identified 12 genomic regions defined by the positions of 12 genes with evidence of statistical interactions between sets of SNPs

and sets of methylated CpGs with respect to asthma risk. Of these 12 genes, 3 have previously been implicated in asthma risk or underlying biological pathways related to pathology of the disease, i.e., *PF4* [39], *ATF3* [40, 41], and *TPRA1* [42]. Three genomic regions not previously implicated in asthma, *LOC101928523*, *SCARNA18*, and *LHX6*, contained the most statistically significant pairwise interactions between the considered individual SNPs and CpGs.

LHX6 (LIM homeobox domain 6) has not been implicated in asthma; however, it is a recognized transcriptional regulator that controls the differentiation and development of lymphoid cells [43]. *LHX6* is known to be regulated epigenetically in lung cancer, where in vitro and in vivo studies found that in normal lung tissue it is readily expressed but downregulated or silenced in lung cancer cells in which the gene is hypermethylated [43]. Other evidence of epigenetic regulation in the vicinity of *LHX6* has been found in head and neck squamous cell carcinomas [44]. Similar to the lung cancer study, hypermethylation of the CpG island in *LHX6* was associated with transcriptional silencing of *LHX6*. These findings suggest that differential methylation near *LHX6* plays a role in lung biology and lends credence to a potential role in asthma.

The 3 genes previously implicated in asthma pathogenesis are *PF4*, *ATF3*, and *TPRA1*. *PF4* (Platelet Factor 4) is a protein-coding gene which functions as an inhibitor of T-cell function [45]. *PF4* activation in the lung is a feature of the late inflammatory response to antigen challenge and may play an important role in allergic inflammation and asthma [46]. *Atf3* (Activating transcription factor 3) is a negative regulator of allergic inflammation in mice challenged with ovalbumin [47] and deficiency in mice leads to the development of significantly increased airway hyperresponsiveness and pulmonary eosinophilia [40]. Significant increases in *ATF3* mRNA have also been observed in patients with mild asthma as compared with nonasthmatic patients [48]. *TPRA1* (transmembrane protein adipocyte associated 1) is an irritant-sensing cation channel expressed in TRPV1-positive, capsaicin-sensitive chemosensory neurons that innervate various organs, including the airways [49]. Various exogenous chemicals have been described to activate *TRPA1*, including agents recognized to trigger and/or worsen asthma such as diisocyanates, cigarette smoke, acrolein, and chlorine [42]. A potential role of *TRPA1* in mediating allergen-induced asthmatic responses has been described in ovalbumin-sensitized mice, in which genetic deletion of *Trpa1* or pretreatment with a selective *Trpa1* antagonist

reduced leukocyte infiltration, decreased cytokine and mucus production, and almost completely abolished airway hyperactivity [49]. Bessac et al. [50] suggested that *TRPA1* may function as an integrator of immunological stimuli modulating inflammation in the airways. Furthermore, chemical irritant-induced activation of *TRPA1* may trigger the release of neuropeptides and chemokines in the airways, thereby exacerbating the cellular and tissue inflammatory response observed in allergic individuals [51].

HOPX, *OR10K1*, *UPK1B*, *CHMP4B*, and *LANCL1* were identified as significantly associated with asthma in the tests for joint interactions but require further investigation into their putative biological functions. There is scant prior evidence of a role for these genes in pulmonary or immune function, lung disease, or asthma, although *HOPX* is involved in the function of regulatory T cells [52].

Conclusions

We demonstrated that simultaneous consideration of genomic and epigenomic variation has the potential to identify genetic risk factors for asthma beyond individual genome-wide association studies or epigenetic screens. These results add to existing evidence suggesting a synergy between genomic and epigenomic variation affecting risk of asthma.

Acknowledgements

We thank the many Asthma Bridge participants who contributed tissue samples and personal data. We also thank the clinicians, and researchers who collected samples and recorded data. This work was supported by NIH grants, NIEHS grant No. R01ES022216, NHLBI grant No. R01HL118455, P01CA196569, and RC2 HL101543.

Appendix

Genetic-Epigenetic Interactions in Asthma Revealed by a Genome-Wide Gene-Centric Search CASI Methodology

To calculate the CASI statistic we consider n observations of SNP and methylation measures with the two types of variables congregated into sets of size p and q , respectively, and assembled into a $(n_Y + n_N) \times (p + q)$ matrix X , where

$$X = \begin{bmatrix} g_Y & m_Y \\ g_N & m_N \end{bmatrix} = \begin{bmatrix} g_{Y1} & \cdots & g_{Yp} & m_{Y1} & \cdots & m_{Yq} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ g_{NY1} & \cdots & g_{Np} & m_{N1} & \cdots & m_{Nq} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ g_{NN1} & \cdots & g_{Np} & m_{N1} & \cdots & m_{Nq} \end{bmatrix}.$$

We designate SNPs by g , methylation probes by m , and asthma case and control by Y and N , respectively. CCA is used to find coefficients $\hat{a}_{kY}$ and $\hat{b}_{kY}$ for g_Y and m_Y , respectively, that maximize the Pearson correlation between their corresponding canonical variates, where each canonical variate is a linear combination of a variable set formed as $\hat{r}_{kY} = g_Y \hat{a}_{kY}$ and $\hat{l}_{kY} = m_Y \hat{b}_{kY}$, where g_Y is a $n_Y \times p$ matrix, m_Y is a $n_Y \times q$ matrix, $\hat{a}_{kY}$ is a $p \times 1$ vector, $\hat{b}_{kY}$ is a $q \times 1$ vector, $\hat{r}_{kY}$ is a $n_Y \times 1$ vector, and $\hat{l}_{kY}$ is a $n_Y \times 1$ vector. The canonical correlations are based on Pearson function,

$$\hat{\rho}_{kY} = \hat{p}_Y \{ \hat{r}_{kY}, \hat{l}_{kY} \} = \hat{a}_{kY}^T S_{12} \hat{b}_{kY} / \left[\sqrt{\hat{a}_{kY}^T S_{11} \hat{a}_{kY}} \sqrt{\hat{b}_{kY}^T S_{22} \hat{b}_{kY}} \right],$$

where

$$S_{(p+q) \times (p+q)} = \begin{bmatrix} S_{11(p \times p)} & S_{12(p \times q)} \\ S_{21(q \times p)} & S_{22(q \times q)} \end{bmatrix}.$$

The first canonical correlation coefficient, $\hat{\rho}_{1Y}$, is the maximum across all possible coefficients, $\hat{a}_{1Y}$, and $\hat{b}_{1Y}$, and subsequent canonical correlations are formed using linear combinations that are uncorrelated with the previous $k-1$ canonical variates. Only the first canonical correlation coefficient is used for the CASI statistic.

$\hat{\rho}_{1Y}$ is calculated from the subset of asthma cases and $\hat{\rho}_{1N}$ is obtained by applying coefficients $\hat{a}_{1Y}$ and $\hat{b}_{1Y}$ to g_N and m_N , control data. Thus, the CASI approach can be thought of as a multivariate analog to computing a case-only interaction estimate and then comparing it to an estimate using controls. It is important to emphasize that variates for the control subset are formed based on maximizing canonical correlation in the case subset; $\hat{r}_{1N} = g_N \hat{a}_{1Y}$ and $\hat{l}_{1N} = m_N \hat{b}_{1Y}$, and correlation calculated by $\hat{\rho}_{1N} = \hat{p}_N \{ \hat{r}_{1N}, \hat{l}_{1N} \} = \text{cor}(\hat{r}_{1N}, \hat{l}_{1N})$. Permuted data sets are generated by permuting case-control phenotype labels. We will use *obs* to denote the observed or nonpermuted data and *perm* to denote SNP and methylation data with permuted case-control labels. Also, we will drop the subscript, 1, since we will be referring to the first canonical coefficient only. For variance stabilization, Fisher transformations,

$$\frac{1}{2} \times \ln \left[\frac{1 + \hat{\rho}}{1 - \hat{\rho}} \right],$$

are applied to the resulting canonical correlation coefficients,

$$Z_{\text{obs},Y} = \frac{1}{2} \times \ln \left[\frac{1 + \hat{\rho}_{\text{obs},Y}}{1 - \hat{\rho}_{\text{obs},Y}} \right]$$

and

$$Z_{obs.N} = \frac{1}{2} \times \ln \left[\frac{1 + \hat{p}_{obs.N}}{1 - \hat{p}_{obs.N}} \right],$$

and analogously for the permuted data,

$$Z_{perm.Y} = \frac{1}{2} \times \ln \left[\frac{1 + \hat{p}_{perm.Y}}{1 - \hat{p}_{perm.Y}} \right]$$

and

$$Z_{perm.N} = \frac{1}{2} \times \ln \left[\frac{1 + \hat{p}_{perm.N}}{1 - \hat{p}_{perm.N}} \right]$$

for $\text{perm.} \in \{1, 2, \dots, B\}$ where B is the number of permutations. The differences from the observed, $Z_{obs.\{Y-N\}} = Z_{obs.Y} - Z_{obs.N}$, and permuted, $Z_{perm.\{Y-N\}} = Z_{perm.Y} - Z_{perm.N}$, Fisher transformed correlations are then scaled by subtracting mean, $\hat{\mu}_{Z_{perm.}} = \text{MEAN}(Z_{perm.\{Y-N\}})$, and dividing by standard deviation, $\hat{\sigma}_{Z_{perm.}} = \text{S.D.}(Z_{perm.\{Y-N\}})$, resulting from permuted data for each gene to obtain the final observed CASI statistic, $(Z_{obs.\{Y-N\}} - \hat{\mu}_{Z_{perm.}}) / \hat{\sigma}_{Z_{perm.}}$, and permuted CASI statistics, $(Z_{perm.\{Y-N\}} - \hat{\mu}_{Z_{perm.}}) / \hat{\sigma}_{Z_{perm.}}$. The rationale behind centering with the permuted sample statistic mean is that the expectation of the canonical correlation coefficient will be different in cases versus controls under the null due to the fact that the maximization is computed in cases. Centering corrects for this difference. Scaling by the permuted sample statistic standard deviation is performed because the standard deviation may depend to some extent on dimensional and distributional characteristics of the sets g and m . Scaling standardizes the statistic for these factors.

It will be assumed here that multiple tests will be conducted in an analysis, such as was performed in this study, i.e., a single test for each gene. Under these circumstances, the CASI statistics are evaluated for statistical significance by computing FDR using permutation-based estimators proposed by Millstein and Volfson [53], which yields confidence intervals that account for the number of permutations conducted as well as dependencies between tests. Significance is assessed directly using CASI statistics, thus no intermediate p values are computed.

Simulation 1 – Tests for Interaction by Examining Equivalence of Set Correlations between Cases and Controls

As the description of the CASI method makes clear, our statistic measures a difference in correlation structure between two class levels of a dichotomous phenotype, case and control. We demonstrate in silico that statistical interactions between linear combinations of variables in two sets will cause a difference in correlations between those linear combinations in cases versus controls, and further that such interactions are detectable by the CASI statistic by measuring the difference in correlation coefficients.

We simulated two groups of genomic variables with correlations simulated in blocks for cases and controls. Representing interactions as the difference in correlation matrices across cases and controls is not common but also not unprecedented, for example, Rajapakse et al. [54] and Peng et al. [55]. This idea is related to the concept described by Millstein et al. [56], that interactions may be implicated if the joint distribution of genomic features depends on disease status.

This prompts the tack we took in how we set up the simulation study. If the off-diagonal blocks of the correlation matrices corresponding to the covariance between two different groupings of

genomic factors differ between cases and controls, we infer interaction. This is also the reason that as constructed for simulation, the correlations within the groupings were the same across case-control status, distributions for groups separately did not vary by disease status.

The simulation study was constructed to show that the CASI method can be used to detect such contrasts in correlation, and therefore interactions. For each disease status, two sets of multivariate normal variables were generated with correlation structures in blocks. Elements of the correlation matrix within blocks were randomly selected but identical within cases and controls. In contrast, the off-diagonal blocks that contains elements for correlations between the two feature sets differ between cases and controls. This is shown in more detail below using matrix symbols. The disease status distinguishing elements are set to 0.1 for controls and range from 0.11 to 0.6 for cases. An equal sample size was generated for cases and controls for each scenario. Statistical significance was assessed using the permutation-based FDR approach from Millstein and Volfson [53], with 100 permutations and FDR estimated over a range of significance thresholds.

Hypothesis Test

To establish the form of the null and alternative hypotheses:

$$\begin{aligned} H_0: p_{\text{case}} &= p_{\text{control}} \\ H_A: p_{\text{case}} &\neq p_{\text{control}} \end{aligned}$$

where p_m ($m \in \{\text{case}, \text{control}\}$) is the correlation structure between two sets of variables within respective phenotype class, m . This can be embodied in a correlation matrix and understood as the intersection of linear column spaces spanned by the two sets. The common space defines statistical variance explained in one set by another, estimated through decomposition of the matrix.

This test signifies that if the correlation patterns are different between cases and controls, we conclude that there is an interaction. A disease-linked interacting factor might be closely related to others. For example, an SNP near more than one genotyped marker in the region. Therefore, our method examines a set of genetic and epigenetic variables in a region jointly and can possibly offer greater statistical power than the classical method of investigating 2-way interactions in univariate logistic regression models.

Simulation Study

We attempt to simulate a simplified version of biological data. In general, we assume that groups of SNPs and CpG sites act in concert based on biological processes. We model this with a 4-block correlation matrix, a block of correlations randomly assigned to pairs from the same gene set ($\text{SNP}_{(\text{gene } i)} \times \text{SNP}_{(\text{gene } i)}$), block of ($\text{CpG}_{(\text{gene } i)} \times \text{CpG}_{(\text{gene } i)}$) correlations assigned likewise, and 2 blocks, one transpose of the other, ($\text{SNP}_{(\text{gene } i)} \times \text{CpG}_{(\text{gene } i)}$), equi-correlated. One way of interpreting this is as a latent factor model, all the SNPs or CPGs from a gene correlated in a block are in turn highly correlated with the same unmeasured latent variables, hypothetically expressed as weighted sums, the canonical variates that underlie the canonical correlation procedure. This interpretation is not necessary to understand the $\text{SNP}_{(\text{gene } i)} \times \text{CpG}_{(\text{gene } i)}$ block as representing correlation structure between two sets of variables. In our simulations each block within the correlation matrix contains 25 elements (5 SNPs $\times$ 5 CpGs for example).

We simulate the SNPs and methylation variables (CpGs) for our healthy controls as jointly Gaussian (multivariate normal) with 0 mean and correlation matrix

$$\Sigma_{\text{control}} = \begin{pmatrix} \mathbf{G}_i \times \mathbf{G}_i & \mathbf{G}_{i(\text{control})} \times \mathbf{M}_{i(\text{control})} \\ \mathbf{M}_{i(\text{control})} \times \mathbf{G}_{i(\text{control})} & \mathbf{M}_i \times \mathbf{M}_i \end{pmatrix},$$

where $\mathbf{G}_i \times \mathbf{G}_i$ and $\mathbf{M}_i \times \mathbf{M}_i$ are 5×5 matrices with 1s along the diagonal and $\rho > 0$ randomly assigned from uniform distribution for all off-diagonal entries. $\mathbf{G}_{i(\text{control})} \times \mathbf{M}_{i(\text{control})}$ and $\mathbf{M}_{i(\text{control})} \times \mathbf{G}_{i(\text{control})}$ are 5×5 matrices, transpose of each other with all elements, $\rho_{\text{control}} > 0$, having the same, fixed value (equi-correlated). For our case patients we again use mean 0 variables, but change our correlation matrix to

$$\Sigma_{\text{case}} = \begin{pmatrix} \mathbf{G}_i \times \mathbf{G}_i & \mathbf{G}_{i(\text{case})} \times \mathbf{M}_{i(\text{case})} \\ \mathbf{M}_{i(\text{case})} \times \mathbf{G}_{i(\text{case})} & \mathbf{M}_i \times \mathbf{M}_i \end{pmatrix},$$

where $\mathbf{G}_i \times \mathbf{G}_i$ and $\mathbf{M}_i \times \mathbf{M}_i$ are 5×5 matrices with the same element values as for controls. $\mathbf{G}_{i(\text{case})} \times \mathbf{M}_{i(\text{case})}$ and $\mathbf{M}_{i(\text{case})} \times \mathbf{G}_{i(\text{case})}$ are 5×5 matrices, transpose of each other with the same element values of $\rho_{\text{case}} > \rho_{\text{control}} > 0$ (equi-correlated).

Effect size is expressed as a measure of the “distance” between the two correlation matrices used for simulating samples. This metric is quantified as “norm” of the difference between two matrices. In particular, we use the spectral norm or 2-norm, which is the largest singular value. If we denote our two matrices as Σ_{case} and Σ_{control} then our distance metric, effect size, is $\|\Sigma_{\text{case}} - \Sigma_{\text{control}}\|$, where $\|\cdot\|$, represents the norm.

Within each class (case and control) and for a range of effect sizes we simulated a sample of size $n \in \{50, 100, 150, 200, 250\}$ and applied CASI method. We then estimated the FDR of this method over 1,000 trials for each. As we can see in Figure A2, despite considerable volatility for very small effect sizes, CASI performed quite well for meaningful differences in correlation. To affirm the ability of our test to detect a difference in correlation structure, for each simulation, we set elements of $\mathbf{G}_{i(\text{case})} \times \mathbf{M}_{i(\text{case})}$ and $\mathbf{G}_{i(\text{control})} \times \mathbf{M}_{i(\text{control})}$ (along with their transposes) blocks in the correlation matrices to different values. For each simulation, we generated a sample of “case” status individuals with $\rho_{\text{case}} \in \{0.11, 0.12, \dots, 0.40\}$ and “control” status subjects using $\rho_{\text{control}} = 0.1$. The respective effect sizes, distances between correlation matrices (spectral norms of difference between two matrices), are $\in \{0.05, 0.075, 0.10, \dots, 1.50\}$.

FDR plots in Figure A1 are snapshots of the entire range of simulations. For each sample size, we garner from our FDR estimates the number of true positive tests that can be identified, hence giving us a measure of power of our test. The results show that for increasing effect and sample sizes our test is capable of detecting non-null hypotheses at near 100% power. The lower right plot of Figure A2 shows consistency for the CASI hypothesis test, the various effect sizes (vertical axis) and the corresponding sample sizes (horizontal axis) at which 100% power is achieved. The jagged power plots for samples 50, 100, 150, 200, and 250 indicate, as expected, that for larger samples we can expect to reach power of 100% at smaller effect sizes. The criteria for identifying power is the proportion of tests identified correctly as true positives at the maximum $FDR \leq 0.05$. From the Figure A2 plots, it is clear that power increases with effect size. We showed this relationship under the assumption of multivariate normal distribution of features, but under conditions CASI will be typically be applied this property is likely to remain valid.

Simulation 2 Study – CASI as a Method for Detecting Multiplicative Interaction

In order to assess type I error rate control, statistical power, and unbiasedness properties of our method to detect multiplicative interaction, we conducted a simulation study in a logistic regression framework. In traditional power analysis, a non-null model is used to generate a sample repeatedly, a statistic is calculated each time and compared to a threshold corresponding to the designated type I error rate. Empirical statistical power is calculated as the proportion of tests that reject the null in favor of alternative hypothesis.

Another, equivalent approach more suitable to use the FDR as a way of conceptualizing type I error rate and evaluating power in hypothesis testing when conducting multiple comparisons is to generate a sample repeatedly from a mixture/amalgam of null and non-null models and calculate empirical statistical power as proportion of instances of the statistic arising from the non-null model that are captured for a given FDR level. Since our method relies on permutation for calculating a statistic we do not have a known reference null distribution to determine significance for calculating power. Rather, we estimate FDR based on the empirical distribution of observed values of our statistic from the multitude of tests by comparing them to realizations from permuted data.

This is akin to generating values from a mixture distribution of two kinds of logistic models. We decided to have 10% of models in our mixture non-null, in deference to 2007 paper by Brad Efron, “Size, Power and False Discovery Rates,” which used the same proportion in its simulation. Their version of FDR is parametric, which contrasts with ours since our method relies on permutation for calculating a statistic. We do not have a known reference null distribution to determine significance for calculating power. Rather, we estimate FDR based on the empirical distribution of observed values of our statistic from the multitude of tests by comparing them to realizations from permuted data.

When conducting simulations, we are interested in examining power for various “effect” sizes. The logistic regression is specified via the coefficients, β_0, β_x ; the first element is the intercept (disease background prevalence) and the second consists of a vector of log OR parameters for the covariates (for example genomic variables such as SNPs). The overall outcome prevalence in the population of interest is fixed in our simulation at 0.05. Modifying any given element of β_x will automatically modify the overall prevalence, unless there is a corresponding change in β_0 . In the simulations we adjust the value of β_0 so that it minimizes the difference between the target outcome prevalence, 0.05, and prevalence induced by the model in conjunction with the assumed marginal exposure and interaction values. In effect, we are generating a population of individuals that hold a gene pair that affects disease through specified logistic model, holding prevalence invariant, and then take a sample from that population.

As stated above, we simulate a collection of genes that is a mixture of null and non-null gene pairs. Non-null gene pair is defined as one in which disease state of an individual is the result of interaction between two sets of SNPs and is generated from a logistic model as follows. Null and non-null disease logistic models are defined above.

Probability of disease in non-null gene pair model:

$$f_0 = \frac{\exp \left(\beta_0 + \sum \beta_{\text{SNP}(\text{gene } i)} \times \text{SNP}(\text{gene } j) + \sum \beta_{\text{SNP}(\text{gene } j)} \text{SNP}(\text{gene } i) + \sum \beta_{\text{intretn}} \right)}{1 + \exp \left(\beta_0 + \sum \beta_{\text{SNP}(\text{gene } i)} \times \text{SNP}(\text{gene } j) + \sum \beta_{\text{SNP}(\text{gene } j)} \text{SNP}(\text{gene } i) + \sum \beta_{\text{intretn}} \right)}$$

Null gene pair is defined as one in which disease state of individual is a result of an additive effect of two sets of SNPs.

Probability of disease in null model:

$$f_1 = \frac{\exp \left(\beta_0 + \sum \beta_{\text{SNP}(\text{gene } i)} \times \text{SNP}(\text{gene } j) + \sum \beta_{\text{SNP}(\text{gene } j)} \text{SNP}(\text{gene } i) \right)}{1 + \exp \left(\beta_0 + \sum \beta_{\text{SNP}(\text{gene } i)} \times \text{SNP}(\text{gene } j) + \sum \beta_{\text{SNP}(\text{gene } j)} \text{SNP}(\text{gene } i) \right)}$$

SNP data is simulated based on minor allele frequency that is randomly chosen from a uniform distribution ranging from 0.05 to 0.49.

For the simulation, 100 non-null and 900 null gene pairs were generated. After application of CASI, the distribution of non-null gene pairs in the resulting ranked set is evaluated as a means of power diagnostic. As we vary interaction OR, we observe the number of non-null genes identified as significant. The larger the proportion of non-null genes (out of a total of 100) detected for low values of FDR the higher the power. In the region bracketing the FDR level of 0.05, CASI performed well in comparison to conventional logistic regression, the approach of Rajapakse et al. [54] and Peng et al. [55] (Fig. A3).

Interpreting CASI Results via Canonical Correlation: Loadings

Having shown that our method is capable of detecting interaction, once CASI is applied to data the next step would be to interpret

results in such a way that conclusion can be drawn about which components of the sets are involved in interaction. This is accomplished through canonical variates. Interpretation of the CASI canonical variates, $\hat{r}_{1Y} = g_Y \hat{a}_{1Y}$ and $\hat{l}_{1Y} = m_Y \hat{b}_{1Y}$, is aided by computing loadings, correlations between canonical variates and the original variables in the sets g_Y and m_Y . The loading of each variable gives a rough gauge of how representative the variate is of the variable and thus the importance of the variable in the test results. There is no specific guidance to how large a loading is considered an important contribution, but we have chosen a threshold of 0.5 because it is regarded as the “operational” definition of a large effect size [57, 58].

CASI Analysis of Asthma BRIDGE Data

Analyses of the top loading SNP-CpG pairs for the 12 significant genes using conventional logistic regression yielded significant interactions for some, e.g., SCARNA8 and LHX6, but not all pairs (Table A1). Interestingly, there was also some suggestion of SNP main effects, e.g., in STC1 and TPRA1. The top loading pair would not necessarily yield a significant interaction in a conventional model, because the effect could more diffuse, distributed over multiple SNPs and CpG sites, implying *joint* interactions that are more detectable by the CASI method. A sensitivity analysis indicated that the results were robust to the distribution of participants with regard to study site. Restricting the analysis to sites that included both cases and controls and adjusting for site as a covariate in the restricted sample yielded similar results, and in some cases increased the statistical significance of the interactions (Table A2).

Table A1. Significant interactions for SNPs and CpGs with loadings >0.5 for the 12 genes identified by CASI, restricted to sites that recruited both asthma cases and controls

SNP	Methylation	Coefficient SNPxMeth interaction	<i>p</i> value (1-df LRT)	<i>p</i> value (3-df LRT)	Gene
rs10818651	cg01363324	-31.72	1.36E-05	1.55E-04	LHX6
rs67216017(G/GA)	cg14999833	21.64	4.11E-05	1.33E-04	SCARNA18
rs113665237(A/AT)	cg14999833	21.64	4.11E-05	1.33E-04	SCARNA18
rs10061690	cg14999833	21.64	4.11E-05	1.33E-04	SCARNA18
rs10985567	cg01363324	-29.40	5.90E-05	1.87E-04	LHX6
rs10818651	cg13832372	16.05	6.43E-05	6.16E-04	LHX6
rs2274748	cg01363324	-26.31	1.26E-04	1.32E-03	LHX6
rs10818651	cg21469772	-7.31	1.82E-04	2.75E-03	LHX6
rs10985567	cg21213617	-15.33	2.14E-04	1.71E-03	LHX6
rs989798	cg21469772	-7.36	2.97E-04	2.07E-03	LHX6
rs989798	cg04282082	-5.88	3.54E-04	1.64E-03	LHX6
rs10818651	cg04282082	-5.54	3.62E-04	4.49E-03	LHX6
rs10985567	cg13832372	14.37	4.12E-04	1.03E-03	LHX6
rs10818651	cg21213617	-13.61	5.12E-04	6.05E-03	LHX6
rs10985567	cg21469772	-6.94	6.22E-04	3.64E-03	LHX6
rs989798	cg01363324	-23.72	6.75E-04	1.73E-03	LHX6
rs10818651	cg13571460	-6.30	7.71E-04	9.61E-03	LHX6
rs13301641	cg07956857	-7.91	8.76E-04	4.64E-03	LOC101928523
rs989798	cg13832372	14.11	8.96E-04	2.14E-03	LHX6
rs989798	cg21213617	-13.27	9.55E-04	6.25E-03	LHX6
rs11792474	cg07956857	-7.61	1.19E-03	6.26E-03	LOC101928523
rs989798	cg13571460	-5.87	1.76E-03	9.81E-03	LHX6
rs10985567	cg04282082	-5.04	1.78E-03	5.20E-03	LHX6
rs2274748	cg13832372	11.79	1.82E-03	1.29E-02	LHX6
rs2274748	cg21469772	-5.10	2.52E-03	2.58E-02	LHX6
rs10985567	cg13571460	-5.70	2.84E-03	1.37E-02	LHX6
rs10760213	cg01363324	-16.89	3.00E-03	2.29E-02	LHX6
rs3793618	cg01363324	-18.14	3.17E-03	2.06E-02	LHX6
rs3793619	cg01363324	-18.14	3.17E-03	2.06E-02	LHX6
rs989798	cg13862711	-4.06	3.32E-03	1.79E-02	LHX6
rs2274748	cg21213617	-9.97	3.80E-03	3.13E-02	LHX6
rs989798	cg03363289	-4.27	3.95E-03	6.35E-03	LHX6
rs1435520	cg13126871	13.49	4.40E-03	2.07E-02	PF4
rs10818651	cg13862711	-3.51	4.45E-03	3.76E-02	LHX6
rs7868147	cg01363324	-16.17	4.48E-03	3.12E-02	LHX6

Pairwise interactions with *p* value <0.05 for top loading SNP and CpG site pairs for the 12 CASI statistically significant genes. Logistic regression models were adjusted for site, gender, ancestry/ethnicity, family history of asthma, age at methylation measure, and education level. Due to multiple SNPs in high LD, some *p* values are similar or identical.

Table A2. Interactions for top loading SNP-CpG pairs for the 12 genes identified by CASI

Gene	Chr.	Position in bp range ^a	SNP (minor/major alleles) methylation	MAF	Sample size	<i>p</i> value Int.	OR _{Cond.} ^b (95% CI)	OR _{Int.} (95% CI)	OR _{SNP} (95% CI)	OR _{Meth.} (95% CI)
HOPX	4q12	57512154–57524688	rs4515223 (A/G) cg19978674	0.48	484	0.15	GG: 1.03 (0.88–1.20) AG/GA: 0.96 (0.75–1.23) AA: 0.90 (0.64–1.26)	0.94 (0.86–1.02)	0.98 (0.72–1.34)	0.99 (0.93–1.06)
SCARNA18	5q14	82358023–82362156	rs10061690 (A/G) cg14999833	0.27	511	2.84E–05	GG: 0.79 (0.56–1.14) AG/GA: 2.14 (0.91–5.01) AA: 5.75 (1.49–22.14)	2.69 (1.64–4.42)	0.83 (0.58–1.18)	1.20 (0.91–1.60)
PF4	4q13	74844542–74849841	rs11574450 (C/A) cg06834998	0.08	531	0.22	AA: 0.89 (0.80–0.99) CA/AC: 1.07 (0.70–1.62) CC: 1.27 (0.62–2.64)	1.20 (0.88–1.63)	1.57 (0.93–2.76)	0.97 (0.89–1.05)
STC1	8p21	23697434–23714320	rs9969426 (G/A) cg16688533	0.13	492	1.62E–02	AA: 1.13 (1.00–1.28) GA/AG: 0.97 (0.74–1.26) GG: 0.83 (0.55–1.25)	0.86 (0.74–0.99)	1.82 (1.15–2.56)	0.95 (0.89–1.02)
ATF3	1q32	212779970–212796119	rs200939984 (A/ACG) cg04323713	0.07021	442	0.48	ACG, ACG: 1.04 (0.77–1.38) A, ACG/ACG, A: 0.93 (0.51–1.68) AA: 0.83 (0.34–2.05)	0.90 (0.66–1.22)	1.01 (0.56–1.82)	0.96 (0.86–1.07)
OR10K1	1q23	158433352–158438293	rs67123344 (A/AT) cg01463139	0.4216	526	3.81E–03	AT, AT: 1.09 (1.02–1.17) A, AT/AT, A: 1.02 (0.90–1.14) AA: 0.95 (0.80–1.12)	0.93 (0.89–0.98)	1.32 (0.98, 1.79)	0.99 (0.95–1.02)
UPK1B	3q13	118890425–118926000	rs56193405 (A/G) cg02741226	0.2031	502	0.53	GG: 0.93 (0.78–1.12) AG/GA: 0.98 (0.70–1.38) AA: 1.03 (0.63–1.71)	1.05 (0.89–1.24)	0.71 (0.50–1.02)	0.93 (0.83–1.05)
LOC101928523	9q31	106759990–106764763	rs13301641 (T/C) cg07956857	0.3212	516	1.28E–04	CC: 1.24 (1.00–1.53) TC/CT: 0.82 (0.54–1.27) TT: 0.55 (0.29–1.05)	0.67 (0.54–0.83)	1.03 (0.76–1.38)	0.92 (0.79–1.06)

Table A2 (continued)

Gene	Chr.	Position in bp range ^a	SNP (minor/major alleles) methylation	MAF	Sample size	<i>p</i> value Int.	OR _{Cond.} ^b (95% CI)	OR _{Int.} (95% CI)	OR _{SNP} (95% CI)	OR _{Meth.} (95% CI)
LHX6	9q33	32397110– 32444173	rs10818651 (A/G) cg04282082	0.2462	448	3.59E–04	GG: 1.08 (0.99–1.18) AG/GA: 0.84 (0.67–1.07) AA: 0.66 (0.45–0.96)	0.78 (0.67– 0.90)	1.00 (0.66– 1.53)	1.01 (0.95– 1.06)
CHMP4B	20q11	127289907– 127311602	rs2378024 (G/C) cg03355213	0.4925	504	0.09	CC: 0.74 (0.53–1.04) GC/CG: 0.91 (0.52–1.59) GG: 1.10 (0.50–2.42)	1.21 (0.97– 1.52)	0.91 (0.68– 1.20)	0.93 (0.79– 1.10)
TPRA1	3q21	211297973– 211343429	rs112458291 (A/G) cg21187194	0.1024	511	0.10	GG: 0.98 (0.80–1.21) AG/GA: 0.72 (0.40–1.31) AA: 0.53 (0.20–1.41)	0.73 (0.50– 1.08)	1.87 (1.16– 3.11)	0.87 (0.72– 1.04)
LANCL1	2q34	211297973– 211343429	rs6727682 (A/C) cg07063745	0.4661	468	0.14	GG: 1.90 (0.89–4.04) AG/GA: 1.41 (0.45–4.48) AA: 1.05 (0.22–4.96)	0.74 (0.50– 1.11)	1.21 (0.89– 1.65)	1.16 (0.95– 1.45)

Model: logit [p(Asthma)] = $\beta_0 + \Sigma \beta_{Covrt} \times Covrt. + \beta_1 SNP + \beta_2 Meth. + \beta_{Int.} (SNP \times Meth)$. Interaction, conditional, and main effect odds ratios are scaled to a 5% change in methylation. Logistic regression models were adjusted for gender, ancestry/ethnicity, family history of asthma, age at methylation measure, and education level. MAF, minor allele frequency. ^a bp position identifies the range for the gene based on the February 2009 human reference sequence (GRCh37/ hg19 assembly) produced by the Genome Reference Consortium. ^b Conditional odds ratio (OR_{Cond.} = OR_{Meth | SNP}) illustrates whether presence of minor alleles increases risk of disease or has a protective effect.

Table A3. Tests of multiplicative interactions ($p < 0.05$) within *LOC101928523* for SNP-CpG pairs

SNP (minor/ major alleles) methylation	MAF	p value Int.	$OR_{Meth SNP}$ (95% CI)	$OR_{Int.}$ (95% CI)	OR_{SNP} (95% CI)	$OR_{Meth.}$ (95% CI)
5% difference in methylation						
rs13301641 (T/C) cg07956857	0.32	1.28E-04	CC: 1.24 (1.00–1.53) TC/CT: 0.82 (0.54–1.27) TT: 0.55 (0.29–1.05)	0.67 (0.54–0.83)	1.03 (0.76–1.38)	0.92 (0.79–1.06)
rs11792474 (G/A) cg07956857	0.32	1.79E-04	AA: 1.23 (0.99–1.52) GA/AG: 0.83 (0.54–1.27) GG: 0.56 (0.30–1.06)	0.68 (0.55–0.84)	1.02 (0.76–1.37)	0.92 (0.79–1.06)
rs75088949 (A/G) cg07956857	0.16	7.52E-03	GG: 0.81 (0.68–0.96) AG/GA: 1.15 (0.74–1.77) AA: 1.62 (0.81–3.24)	1.42 (1.09–1.83)	0.94 (0.63–1.39)	0.92 (0.79–1.06)
1% difference in methylation						
rs13301641 (T/C) cg13488921	0.32	1.35E-02	CC: 0.80 (0.58–1.12) TC/CT: 1.16 (0.62–2.19) TT: 1.68 (0.66–4.27)	1.45 (1.07–1.95)	1.02 (0.76–1.38)	1.11 (0.89–1.37)
rs11792474 (G/A) cg13488921	0.32	1.45E-02	AA: 0.81 (0.58–1.13) GA/AG: 1.16 (0.62–2.19) GG: 1.67 (0.66–4.24)	1.44 (1.07–1.94)	1.02 (0.76–1.37)	1.11 (0.89–1.37)

SNPs are located on chr9:106760350–106764737 with loadings >0.5 in absolute value. ORs reflect a 5 or 1% change in methylation as indicated. Models were adjusted for gender, ancestry/ethnicity, family history of asthma, age at methylation measure, and education level. A 1% difference in methylation is used to estimate effects where measures fall into a narrower range, and a difference of 5% would not be meaningful because it falls outside the available data. MAF, minor allele frequency.

Table A4. Tests of multiplicative interactions ($p < 0.05$) within *SCARNA18* for SNP-CpG pairs

SNP (minor/ major alleles) methylation	MAF	p value Int.	$OR_{Meth SNP}$ (95% CI)	$OR_{Int.}$ (95% CI)	OR_{SNP} (95% CI)	$OR_{Meth.}$ (95% CI)
1% difference in methylation						
rs67216017 (G/GA)	0.27	2.84E-05	GA, GA: 0.96 (0.89–1.03) GA, G/G, GA: 1.16 (0.98–1.38)	1.22 (1.10–1.35)	0.83 (0.58–1.18)	1.04 (0.98–1.10)
cg14999833			GG: 1.42 (1.08–1.86)			
rs113665237 (A/AT)	0.27	2.84E-05	AT, AT: 0.96 (0.89–1.03) AT, A/A, AT: 1.16 (0.98–1.38)	1.22 (1.10–1.35)	0.83 (0.58–1.18)	1.04 (0.98–1.10)
cg14999833			AA: 1.42 (1.08–1.86)			
rs10061690 (A/G)	0.27	2.84E-05	GG: 0.96 (0.89–1.03) AG/GA: 1.16 (0.98–1.38)	1.22 (1.10, 1.35)	0.83 (0.58–1.18)	1.04 (0.98–1.10)
cg14999833			AA: 1.42 (1.08–1.86)			
5% difference in methylation						
rs67216017 (G/GA)	0.27	1.08E-02	GA, GA: 0.82 (0.66–1.02) GA, G/G, GA: 1.14 (0.71, 1.81)	1.38 (1.08–1.77)	0.82 (0.58–1.18)	0.95 (0.80–1.13)
cg20697188			GG: 1.57 (0.77–3.20)			
rs113665237 (A/AT)	0.27	1.08E-02	AT, AT: 0.82 (0.66–1.02) AT, A/A, AT: 1.14 (0.71–1.81)	1.38 (1.08, 1.77)	0.82 (0.58–1.18)	0.95 (0.80–1.13)
cg20697188			AA: 1.57 (0.77–3.20)			
rs10061690 (A/G)	0.27	1.08E-02	GG: 0.82 (0.66–1.02) AG/GA: 1.14 (0.71–1.81)	1.38 (1.08–1.77)	0.82 (0.58–1.18)	0.95 (0.80–1.13)
cg20697188			GG: 1.57 (0.77–3.20)			

SNPs reside in region, chr5:82358023–82362156 with loadings >0.5 in absolute value. ORs reflect a 5 or 1% change in methylation as indicated. Models were adjusted for gender, ancestry/ethnicity, family history of asthma, age at methylation measure, and education level. A 1% difference in methylation is used to estimate effects where measures fall into a narrower range, and a difference of 5% would not be meaningful because it falls outside the available data. MAF, minor allele frequency.

Fig. A1. FDR plots are snapshots from the entire range of simulations. The green values shown horizontally indicate number of CASI tests judged significant for each threshold on the horizontal. For sample size 250, we can glean from a series of FDR estimates the number of true positive tests that can be identified, hence giving us a measure of power of our test. The top plot shows small effect sizes and with the corresponding power quite small given that only approximately 9 or 10 of the tests are identified correctly as significant at a traditional level FDR of 0.05. However, as the effect size increases, as shown in Figures A2 and A3, from the top nearly or exactly 100% of the true positive tests are correctly identified. All tests to which we apply CASI are true positive tests because they are comparing samples from two distinct populations with different correlation structures.

Fig. A2. Power plots show power versus effect size at sample sizes 50, 100, 150, 200, and 250. Sample size designations in the figures refer to the number of individuals in each phenotype class (case and control), with equal amount in each. Since it is a permutation-based approach, we utilize FDR of 0.05 as a significance threshold for calculating power. This is done in lieu of using a threshold that

corresponds to type I error which would typically be employed in a parametric setting. The lower right plot depicts consistency of the CASI method by summarizing power results from the other plots by showing effect sizes and corresponding sample sizes for which 100% power is achieved by the CASI test.

Fig. A3. Comparison of set-based methods and conventional logistic regression with respect to power characteristics. The upper plots show that under these conditions, CASI performs quite well as compared to other approaches in the most important region, where FDR is low, less than approximately 0.2. The lower plots show that under these same conditions, conventional logistic regression essentially fails to make any discoveries at low FDR.

References

- Akinbami LJ, Moorman JE, Bailey C, Zahran HS, King M, Johnson CA, Liu X: Trends in asthma prevalence, health care use, and mortality in the United States, 2001–2010. *NCHS Data Brief* 2012;94:1–8.
- Kroegel C: Global Initiative for Asthma (GINA) guidelines: 15 years of application. *Expert Rev Clin Immunol* 2009;5:239–249.
- Ward BW: Prevalence of multiple chronic conditions among US adults: estimates from the National Health Interview Survey, 2010. *Prev Chronic Dis* 2013;10:E65.
- Palmer LJ, Cookson WO: Genomic approaches to understanding asthma. *Genome Res* 2000;10:1280–1287.
- Zhang Y, Moffatt MF, Cookson WO: Genetic and genomic approaches to asthma: new insights for the origins. *Curr Opin Pulm Med* 2012;18:6–13.
- Ober C, Yao TC: The genetics of asthma and allergic disease: a 21st century perspective. *Immunol Rev* 2011;242:10–30.
- Ferreira MA, McRae AF, Medland SE, Nyholt DR, Gordon SD, Wright MJ, Henders AK, Madden PA, Visscher PM, Wray NR: Association between ORM DL3, IL1RL1 and a deletion on chromosome 17q21 with asthma risk in Australia. *Eur J Hum Genet* 2011;19:458–464.
- Halapi E, Gudbjartsson DF, Jonsdottir GM, Bjornsdottir US, Thorleifsson G, Helgadottir H, Williams C, Koppelman GH, Heinzmann A, Boezen HM: A sequence variant on 17q21 is associated with age at onset and severity of asthma. *Eur J Hum Genet* 2010;18:902–908.
- Reinius LE, Gref A, Saaf A, Acevedo N, Joerink M, Kupczyk M, D'Amato M, Bergstrom A, Melen E, Scheynius A, Dahlen SE, Group BS, Pershagen G, Soderhall C, Kere J: DNA methylation in the Neuropeptide S Receptor 1 (NPSR1) promoter in relation to asthma and environmental factors. *PLoS One* 2013;8:e53877.
- Fu A, Leaderer BP, Gent JF, Leaderer D, Zhu Y: An environmental epigenetic study of ADRB2 5'-UTR methylation and childhood asthma severity. *Clin Exp Allergy* 2012;42:1575–1581.
- Breton CV, Byun HM, Wang X, Salam MT, Siegmund K, Gilliland FD: DNA methylation in the arginase-nitric oxide synthase pathway is associated with exhaled nitric oxide in children with asthma. *Am J Respir Crit Care Med* 2011;184:191–197.
- Ferreira MA, Oates NA, van Vliet J, Zhao ZZ, Ehrich M, Martin NG, Montgomery GW, Whitelaw E, Duffy DL: Characterization of the methylation patterns of MS4A2 in atopic cases and controls. *Allergy* 2010;65:333–337.
- Gutierrez-Arcelus M, Lappalainen T, Montgomery SB, Buil A, Ongen H, Yurovsky A, Bryois J, Giger T, Romano L, Planchon A, Falconnet E, Bielser D, Gagnebin M, Padioleau I, Borel C, Letourneau A, Makrythanasis P, Guipponi M, Gehrig C, Antonarakis SE, Dermizakis ET: Passive and active DNA methylation and the interplay with genetic variation in gene regulation. *eLife* 2013;2:e00523.
- Soto-Ramirez N, Arshad SH, Holloway JW, Zhang H, Schauburger E, Ewart S, Patil V, Karmaus W: The interaction of genetic variants and DNA methylation of the interleukin-4 receptor gene increase the risk of asthma at age 18 years. *Clin Epigenetics* 2013;5:1.
- Zhang H, Tong X, Holloway JW, Rezwani FI, Lockett GA, Patil V, Ray M, Everson TM, Soto-Ramirez N, Arshad SH, Ewart S, Karmaus W: The interplay of DNA methylation over time with Th2 pathway genetic variants on asthma risk and temporal asthma transition. *Clin Epigenetics* 2014;6:8.
- Marchini J, Donnelly P, Cardon LR: Genome-wide strategies for detecting multiple loci that influence complex diseases. *Nat Genet* 2005;37:413–417.
- Cordell HJ: Detecting gene-gene interactions that underlie human diseases. *Nat Rev Genet* 2009;10:392–404.
- Croteau-Chonka DC, Qiu W, Martinez FD, Strunk RC, Lemanske RF Jr, Liu AH, Gilliland FD, Millstein J, Gauderman WJ, Ober C, Krishnan JA, White SR, Naureckas ET, Nicolae DL, Barnes KC, London SJ, Barraza-Villarreal A, Carey VJ, Weiss ST, Raby BA, Asthma BioRepository for Integrative Genomic Exploration Consortium: Gene Expression Profiling in Blood Provides Reproducible Molecular Insights into Asthma Control. *Am J Respir Crit Care Med* 2017;195:179–188.
- Torgerson DG, Ampleford EJ, Chiu GY, Gauderman WJ, Gignoux CR, Graves PE, et al; Mexico City Childhood Asthma Study (MCAAS), Gilliland FD; Children's Health Study (CHS) and HARBORS Study, Burchard EG; Genetics of Asthma in Latino Americans (GALA) Study, Study of Genes-Environment and Admixture in Latino Americans (GALA2) and Study of African Americans, Asthma, Genes & Environments (SAGE), Martinez FD; Childhood Asthma Research and Education (CARE) Network, Weiss ST; Childhood Asthma Management Program (CAMP), Williams LK; Study of Asthma Phenotypes and Pharmacogenomic Interactions by Race-Ethnicity (SAPPHIRE), Barnes KC; Genetic Research on Asthma in African Diaspora (GRAAD) Study, Ober C, Nicolae DL: Meta-analysis of genome-wide association studies of asthma in ethnically diverse North American populations. *Nat Genet* 2011;43:887–892.
- Kothari PH, Qiu W, Croteau-Chonka DC, Martinez FD, Liu AH, Lemanske RF Jr, Ober C, Krishnan JA, Nicolae DL, Barnes KC, London SJ, Barraza-Villarreal A, White SR, Naureckas ET, Millstein J, Gauderman WJ, Gilliland FD, Carey VJ, Weiss ST, Raby BA, Asthma BioRepository for Integrative Genomic Exploration Consortium: The role of local CpG DNA methylation in mediating the 17q21 asthma-susceptibility GSDMB/ORMDL3 expression quantitative trait locus. *J Allergy Clin Immunol* 2018;141:2282–2286.e6.
- Breton CV, Siegmund KD, Joubert BR, Wang XH, Qui WL, Carey V, Nystad W, Haberg SE, Ober C, Nicolae D, Barnes KC, Martinez F, Liu A, Lemanske R, Strunk R, Weiss S, London S, Gilliland F, Raby B, Consortium BRIDGE Consortium: Prenatal tobacco smoke exposure is associated with childhood DNA CpG methylation. *PLoS One* 2014;9:e99716.
- International HapMap Consortium, Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, Gibbs RA, et al: A second generation human haplotype map of over 3.1 million SNPs. *Nature* 2007;449:851–861.
- Liu EY, Li M, Wang W, Li Y: MaCH-admix: genotype imputation for admixed populations. *Genet Epidemiol* 2013;37:25–37.
- Dawson R: How significant is a boxplot outlier. *J Stat Educ* 2011;19:1–12.
- Gauderman WJ: Sample size requirements for association studies of gene-gene interaction. *Am J Epidemiol* 2002;155:478–484.
- Piegorsch WW, Weinberg CR, Taylor JA: Non-hierarchical logistic models and case-only designs for assessing susceptibility in population-based case-control studies. *Stat Med* 1994;13:153–162.
- Yang Q, Khoury MJ, Sun F, Flanders WD: Case-only design to measure gene-gene interaction. *Epidemiology* 1999;10:167–170.
- Witten DM, Tibshirani R, Hastie T: A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics* 2009;10:515–534.
- Millstein J, Volfson D: Computationally efficient permutation-based confidence interval estimation for tail-area FDR. *Front Genet* 2013;4:179.
- Cohen J: A power primer. *Psychol Bull* 1992;112:155–159.
- Millstein J, Chen GK, Breton CV: cit: hypothesis testing software for mediation analysis in genomic applications. *Bioinformatics* 2016;32:2364–2365.
- Patterson N, Price AL, Reich D: Population structure and eigenanalysis. *PLoS Genet* 2006;2:e190.

- 33 Kosoy R, Nassir R, Tian C, White PA, Butler LM, Silva G, Kittles R, Alarcon-Riquelme ME, Gregersen PK, Belmont JW, De La Vega FM, Seldin MF: Ancestry informative marker sets for determining continental origin and admixture proportions in common populations in America. *Hum Mutat* 2009;30:69–78.
- 34 Zhao J, Jin L, Xiong M: Test for interaction between two unlinked loci. *Am J Hum Genet* 2006;79:831–845.
- 35 Peng Q, Zhao J, Xue F: A gene-based method for detecting gene-gene co-association in a case-control association study. *Eur J Hum Genet* 2010;18:582–587.
- 36 Rajapakse I, Perlman MD, Martin PJ, Hansen JA, Kooperberg C: Multivariate detection of gene-gene interactions. *Genet Epidemiol* 2012;36:622–630.
- 37 Yuan Z, Gao Q, He Y, Zhang X, Li F, Zhao J, Xue F: Detection for gene-gene co-association via kernel canonical correlation analysis. *BMC Genet* 2012;13:83.
- 38 Mukherjee N, Lockett GA, Merid SK, Melen E, Pershagen G, Holloway JW, Arshad SH, Ewart S, Zhang H, Karmaus W: DNA methylation and genetic polymorphisms of the Leptin gene interact to influence lung function outcomes and asthma at 18 years of age. *Int J Mol Epidemiol Genet* 2016;7:1–17.
- 39 Tutluoglu B, Gurel CB, Ozdas SB, Musellim B, Erturan S, Anakkaya AN, Kilinc G, Ulutin T: Platelet function and fibrinolytic activity in patients with bronchial asthma. *Clin Appl Thromb Hemost* 2005;11:77–81.
- 40 Gilchrist M, Henderson WR, Clark AE, Simmons RM, Ye X, Smith KD, Aderem A: Activating transcription factor 3 is a negative regulator of allergic pulmonary inflammation. *J Exp Med* 2008;205:2349–2357.
- 41 Marandi Y, Farahi N, Hashjin GS: Asthma: beyond corticosteroid treatment. *Arch Med Sci* 2013;9:521–526.
- 42 Bautista DM, Pellegrino M, Tsunozaki M: TRPA1: A gatekeeper for inflammation. *Annu Rev Physiol* 2013;75:181–200.
- 43 Liu WB, Jiang X, Han F, Li YH, Chen HQ, Liu Y, Cao J, Liu JY: LHX6 acts as a novel potential tumour suppressor with epigenetic inactivation in lung cancer. *Cell Death Dis* 2013;4:e882.
- 44 Estécio M, Youssef E, Rahal P, Fukuyama E, Gois-Filho J, Maniglia J, Goloni-Bertollo E, Issa JJ, Tajara E: LHX6 is a sensitive methylation marker in head and neck carcinomas. *Oncogene* 2006;25:5018–5026.
- 45 Liu CY, Battaglia M, Lee SH, Sun QH, Aster RH, Visentin GP: Platelet factor 4 differentially modulates CD4(+)CD25(+) (regulatory) versus CD4(+)CD25(-) (nonregulatory) T cells. *J Immunol* 2005;174:2680–2686.
- 46 Averill FJ, Hubbard WC, Proud D, Gleich GJ, Liu MC: Platelet activation in the lung after antigen challenge in a model of allergic asthma. *Am Rev Respir Dis* 1992;145:571–576.
- 47 Roussel L, Robins S, Schachter A, Berube J, Hamid Q, Rousseau S: Steroids and extracellular signal-regulated kinase 1/2 activity suppress activating transcription factor 3 expression in patients with severe asthma. *J Allergy Clin Immunol* 2011;127:1632–1634.
- 48 Roussel L, Robins S, Schachter A, Berube J, Hamid Q, Rousseau S: Steroids and extracellular signal-regulated kinase 1/2 activity suppress activating transcription factor 3 expression in patients with severe asthma. *J Allergy Clin Immunol* 2011;127:1632–1634.
- 49 Facchinetti F, Patacchini R: The rising role of TRPA1 in asthma. *Open Drug Discov J* 2010;2:71–80.
- 50 Bessac BF, Jordt SE: Breathtaking TRP channels: TRPA1 and TRPV1 in airway chemosensation and reflex control. *Physiology (Bethesda)* 2008;23:360–370.
- 51 Caceres AI, Brackmann M, Elia MD, Bessac BF, del Camino D, D'Amours M, Witek JS, Fanger CM, Chong JA, Hayward NJ, Homer RJ, Cohn L, Huang X, Moran MM, Jordt SE: A sensory neuronal ion channel essential for airway inflammation and hyperreactivity in asthma. *Proc Natl Acad Sci USA* 2009;106:9099–9104.
- 52 Jones A, Opejin A, Henderson JG, Gross C, Jain R, Epstein JA, Flavell RA, Hawiger D: Peripherally induced tolerance depends on peripheral regulatory T cells that require Hoxp to inhibit intrinsic IL-2 expression. *J Immunol* 2015;195:1489–1497.
- 53 Millstein J, Volfson D: Computationally efficient permutation-based confidence interval estimation for tail-area FDR. *Front Genet* 2013;4:179.
- 54 Rajapakse I, Perlman MD, Martin PJ, Hansen JA, Kooperberg C: Multivariate detection of gene-gene interactions. *Genet Epidemiol* 2012;36:622–630.
- 55 Peng Q, Zhao J, Xue F: A gene-based method for detecting gene-gene co-association in a case-control association study. *Eur J Hum Genet* 2010;18:582–587.
- 56 Millstein J, Conti DV, Gilliland FD, Gauderman WJ: A testing framework for identifying susceptibility genes in the presence of epistasis. *Am J Hum Genet* 2006;78:15–27.
- 57 Faul F, Erdfelder E, Lang AG, Buchner A: G*Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behav Res Methods* 2007;39:175–191.
- 58 Cohen J: A power primer. *Psychol Bull* 1992;112:155–159.