DNA methylation and allergic sensitizations, a genome-scale longitudinal study during adolescence

Supplemental Material

Hongmei Zhang[1*], Akhilesh Kaushal[2], Simon Kebede Merid[3], Erik Melen[3,4], Göran Pershagen[3], Faisal I. Rezwan,[5] Luhang Han[6], Susan Ewart[7], S. Hasan Arshad[4,8], Wilfried Karmaus[1], John W. Holloway[4,9]

1 Division of Epidemiology, Biostatistics, and Environmental Health Sciences, School of Public Health, University of Memphis, Memphis, TN, USA. **2** Center for Precision Environmental Health, Baylor College of Medicine, Houston, TX, USA. **3** Institute of Environmental Medicine, Karolinska Institutet, Sweden. **4** Sachs' Children's Hospital, Stockholm, Sweden. **5** Clinical and Experimental Sciences, Faculty of Medicine, University of Southampton, Southampton, UK. **6** Department of Mathematical Sciences, University of Memphis, TN, USA. **7** College of Veterinary Medicine, Michigan State University, East Lansing, MI, USA. **8** David Hide Asthma and Allergy Research Centre, Isle of Wight, UK. **9** Human Development and Health, Faculty of Medicine, University of Southampton, Southampton, UK.

* Corresponding author, hzhang6@memphis.edu

Methods

Screening

The method in *ttScreening* utilized training and testing data in robust linear regressions. Surrogate variables to explain unknown factors were estimated and included in the regressions to adjust for unknown effects. The method has a better control on types I and II errors than Bonferroni- or false discovery- based methods. [1, 2] We applied the package to assess the association of genome-scale DNAm with the status of sensitization (yes/no) and with the status of poly-sensitization status (poly- vs. mono-sensitization). No other known covariates were included in the screening process. Details of the method have been discussed elsewhere. [1, 2] Briefly, two thirds of the data were randomly selected and used for training and the remaining for testing. In total, 100 pairs of randomly selected training and testing data sets were used. For each training data set, a robust linear regression was fitted to assess the association. If this was statistically significant, it was further evaluated in the testing data using the same model. A CpG was selected as an informative site if it showed statistical significance in at least 50% of training and testing data pairs. Other CpGs that did not meet this criterion were treated as non-informative CpGs potentially not associated with allergic sensitization status or poly-sensitization status and excluded from subsequent analyses.

Agreement with specific IgEs

The CpGs associated with sensitization defined by skin-prick test identified in the IoW cohort and replicated in the BAMSE cohort were further assessed using screening-specific IgEs on a mixture of food allergens and a mixture of inhalant allergens in the IoW cohort. Subjects showing positive results for either FX5 or Padiatop tests were treated as allergic sensitization positive. We examined the agreement

between SPT and screening-specific IgE in the determination of allergic sensitization to one more allergens using Kappa statistic. To assess the association of DNAm at the identified CpGs with allergic sensitization determined by screening specific IgEs, the same statistical analyses were utilized as those applied to SPT-determined allergic sensitizations. The statistical significance level was set at 0.05.

Correlation of DNA methylation with gene expression

We quantitatively examined biological relevance of methylation at the identified CpGs by assessing the association with the transcript expression level of the corresponding gene in the BAMSE cohort. For intergenic CpGs, we assessed the association with transcripts levels of genes within a +/-250 kb region surrounding the CpG site. Paired DNAm and gene expression in peripheral white blood cells of participants aged 16 years in BAMSE were used in this assessment. DNAm and mRNA expression were first adjusted for age, sex, and cell type proportions.[3] The residuals of DNAm and expression after adjustment were then analysed using linear regression models to evaluate correlation. The statistical significance level was set at 0.05.

Results

Agreement with specific IgEs on main effects: The 35 CpGs associated with allergic sensitization status based on SPT were further assessed to test whether the associations remained consistent when sensitization was defined using specific IgE. At all the 35 CpGs (identified via SPT), the associations between DNAm and allergic sensitization status based on specific IgEs were statistically significant at the 0.05 level. Furthermore, the directions of associations were all consistent with those identified in IoW with similar coefficients (Table S1).

The calculated Kappa statistics was 0.72 with a 95% confidence interval of (0.66, 0.79), indicating a strong agreement between the two approaches in the determination of sensitization status to any allergen.

Table S1. Effects of any allergic sensitization (determined based on specific screening IgE) on DNAm at the identified 35 CpGs via linear mixed models in IoW.

Name	Relation to CpG islands	Gene Name	Location	Chr.	Est.	Raw p-value
cg01888561	Island	*SEC14L1*	TSS1500	17	-0.059	**5.9×10^{-3}**
cg02245534	N_Shore	*METRNL*	Body	17	-0.049	**3.8×10^{-3}**
cg02475695	S_Shore	*NHLRC4*	TSS1500	16	-0.086	**3.4×10^{-4}**
cg02803925	—	*PCYT1A*	Body	3	-0.202	**4.4×10^{-4}**
cg03493123	N_Shore	*B4GALT7*	Body	5	-0.065	**2.7×10^{-4}**
cg05072552	N_Shore	*CFL1*	Body	11	-0.045	**3.5×10^{-3}**
cg05390183	N_Shelf	*Intergenic*		1	-0.052	**2.5×10^{-3}**
cg06070625	—	*MITF*	Body; TSS200	3	-0.083	**3.9×10^{-4}**
cg06099697	Island	*ZFPM1*	Body	16	-0.068	**4.8×10^{-4}**
cg07124719	—	*GDEP*	TSS200	4	-0.085	**2.0×10^{-4}**
cg07721901	—	*FAM81B*	TSS1500	5	-0.065	**1.5×10^{-3}**
cg09249800	Island	*ACOT7*	Body	1	-0.183	**4.2×10^{-4}**
cg09705784	—	*DNAH17*	Body	17	-0.105	**9.2×10^{-4}**
cg10159529	—	*IL5RA*	TSS1500	3	-0.084	**8.8×10^{-4}**
cg11699125	Island	*ACOT7*	Body	1	-0.220	**2.5×10^{-4}**
cg11988722	S_Shelf	*Intergenic*		20	-0.068	**8.1×10^{-4}**
cg12077460	—	*MFHAS1*	Body	8	-0.060	**2.9×10^{-3}**
cg12105691	—	*C3orf50*	Body	3	-0.111	**1.3×10^{-4}**
cg14011077	—	*Intergenic*		9	-0.215	**8.3×10^{-5}**
cg14436861	—	*WEE1*	3'UTR	11	-0.102	**2.3×10^{-4}**
cg15482717	S_Shelf	*FADD*	3'UTR	11	-0.074	**2.7×10^{-4}**
cg15710961	—	*DST*	Body	6	-0.095	**1.4×10^{-3}**
cg17203290	—	*C8orf47*	3'UTR	8	-0.083	**1.2×10^{-3}**
cg17429587	S_Shelf	*NCOR2*	Body	12	-0.063	**6.9×10^{-3}**
cg17933300	S_Shelf	*SCOC*	TSS200	4	-0.055	**0.013**
cg17971251	—	*SEC16B*	Body	1	-0.076	**2.0×10^{-3}**
cg18666454	N_Shore	*KCNH2*	Body	7	-0.099	**2.9×10^{-3}**
cg19210306	Island	*C13orf27*	5'UTR	13	0.155	**0.012**
cg20315954	—	*PMP22*	Body	17	-0.105	**1.1×10^{-4}**
cg21220721	Island	*ACOT7*	Body	1	-0.232	**4.3×10^{-4}**
cg25087851	S_Shelf	*GPR44*	TSS1500	11	-0.076	**4.9×10^{-4}**

cg25479097	S_Shelf	*C13orf35*	5'UTR	13	-0.094	**7.0×10^{-4}**
cg25854298	—	*ASCC1*	Body	10	-0.095	**1.8×10^{-3}**
cg26508444	—	*FAM53B*	Body	10	-0.057	**6.4×10^{-4}**
cg27469152	—	*EPX*	3'UTR	17	-0.078	**1.1×10^{-4}**

Note: 1) Chr..=Chromosome, Est.=Estimation. 2) Results statistically significant at 0.05 are in bold. 3) The "—" in the second column represents CpGs not identified as being related to CpG islands.

Correlation of DNA methylation with gene expression: For the 33 CpGs with differential DNAm replicated in BAMSE, we assessed the association of methylation with expression of the corresponding genes listed in Table 2 in the text. Out of the 33 CpGs, expression of genes corresponding to 4 CpGs were low quality and excluded from the analyses (cg05390183, cg12105691, cg19210306, cg25087851). In addition, two CpGs were intergenic, cg11988722 and cg14011077. For these two CpGs, we assessed the association with transcripts levels of genes within +/- 250 kb of the CpG. This resulted in total 29 CpGs available in BAMSE (27 are non-intergenic and 2 intergenic). At 4 of the 27 non-intergenic CpGs, the associations of DNAm and gene expression were statistically significant at the 0.05 significance level (Table S2a). The most significant association in terms of effect size and statistical significance was observed for methylation of cg10159529 with expression of *IL5RA* (coefficient=-12.58, p-value=6.76×10^{-20}). The CpG cg10159529 was located in the TSS1500 of the *IL5RA* gene. For the two CpGs located in the intergenic region, it was found that DNAm at both CpGs was associated with expression of nearby genes (p-value<0.05. Table S2b).

Table S2a. Biological relevance assessment in BAMSE for the identified 35 CpGs on their association with gene expressions.

			BAMSE	
Name	**Gene Name**	**Location**	**Reg. Coeff.**	**Raw p-value**
cg01888561	SEC14L1	TSS1500	0.11	0.94
cg02245534	METRNL	Body	0.70	0.087
cg02475695	NHLRC4	TSS1500	0.39	0.19
cg02803925	PCYT1A	Body	0.45	0.18
cg03493123	B4GALT7	Body	0.37	0.38
cg05072552	CFL1	Body	NA	NA
cg05390183	Intergenic		NA	NA

cg06070625	MITF	Body; TSS200	-0.064	0.80
cg06099697	ZFPM1	Body	-0.53	0.17
cg07124719	GDEP	TSS200	0.40	0.52
cg07721901	FAM81B	TSS1500	-0.67	0.18
cg09249800	ACOT7	Body	NA	NA
cg09705784	DNAH17	Body	0.040	0.78
cg10159529	IL5RA	TSS1500	**-12.58**	**6.76×10^{-20}**
cg11699125	ACOT7	Body	0.21	0.40
cg11988722	Intergenic		*	*
cg12077460	MFHAS1	Body	**-1.12**	**0.0068**
cg12105691	C3orf50	Body	NA	NA
cg14011077	Intergenic		*	*
cg14436861	WEE1	3'UTR	**-1.25**	**0.0046**
cg15482717	FADD	3'UTR	0.42	0.37
cg15710961	DST	Body	-0.38	0.17
cg17203290	C8orf47	3'UTR	0.65	0.20
cg17429587	NCOR2	Body	0.32	0.27
cg17933300	SCOC	TSS200	-0.34	0.39
cg17971251	SEC16B	Body	0.057	0.84
cg18666454	KCNH2	Body	0.052	0.86
cg19210306	C13orf27	5'UTR	NA	NA
cg20315954	PMP22	Body	**-1.94**	**2.35×10^{-5}**
cg21220721	ACOT7	Body	0.22	0.18
cg25087851	GPR44	TSS1500	NA	NA
cg25479097	C13orf35	5'UTR	1.26	0.056
cg25854298	ASCC1	Body	-0.28	0.35
cg26508444	FAM53B	Body	0.42	0.35
cg27469152	EPX	3'UTR	0.30	0.46

Note: 1) Reg. Coeff.=Regression Coefficient, 2) Results statistically significant at 0.05 are in bold. 3) *For the two CpGs that are intergenic, we assessed the association with transcripts levels of genes within a 500 kb region of the CpG sites (250 kb upstream and 250 kb downstream) (Table S2b).

Table S2b. Biological relevance assessment in BAMSE for the identified 2 intergenic CpGs on their association with gene expressions within a 500 kb region of the CpGs (250 kb upstream [-] and 250 kb downstream [+]).

		BAMSE		
Name	**Gene Name**	**Reg. Coeff.**	**Raw p-value**	**Distance from a gene to the CpG** (bp)
cg11988722	TP53INP2	**1.26**	**2.07×10^{-4}**	124,866
	ACSS2	**0.39**	**0.016**	-47,457
	PIGU	-0.37	0.22	-152,071
	GGT7	0.26	0.28	43,703
	NCOA6	0.36	0.54	-7,518
	GSS	-0.11	0.60	126,660
	NCOA6	-0.06	0.86	-7,518
	MYH7B	-0.01	0.97	-146,246
cg14011077	PPP1R26	**-0.65**	**0.029**	-9,320
	SOHLH1	-0.28	0.44	229,046
	MRPS2	-0.13	0.61	-29,502
	LCN9	0.21	0.64	-192,840
	LCN1	0.15	0.69	-50,956
	PAEP	0.10	0.75	-91,276
	GLT6D1	-0.14	0.76	169,058
	KCNT1	-0.01	0.97	-231,710
	OBP2A	-0.01	0.99	-75,673

Note: 1) Reg. Coeff.=Regression Coefficient. 2) Results statistically significant at 0.05 are in bold.

Further assessment on the quality of DNA methylation at 13 CpGs. In total, 13 of the 33 CpGs were on a list of 190,672 spurious probes on the Illumina Infinium HumanMethylation450 BeadChip, suggested by Naeem et al.[4] Of these 13 CpGs, 5 were located in regions containing SNPs, 6 in regions containing repeat sequences, one in regions where insertions or deletions (INDELs) were found, only one showed a difference of more than 0.3 between whole-genome bisulfite sequencing and Illumina 450K bead arrays, and 12 having at least one known SNP(s) and INDEL(s) overlapping the probe region (Table S3). The numbers did not add up to 13 because some CpG sites were with multiple concerns (e.g., cg05390183 located in regions containing SNPs and also in regions with insertions or deletions). These

sites may be more likely to contain outlier values that influence results. To examine the possibility of this concern, we tested for multimodality using the R package diptests[5] and visually inspected the density plots of methylation beta values. However, none of the multimodality tests were statistically significant at the 0.05 statistical significance level (the last column of Table S3), and the sample density plots of DNAm in beta values at all the 13 CpGs did not support multimodality either (Figure S1).

Table S3. The 13 CpGs on a list of 190,672 CpGs suggested by Naeem et al. as possibly spurious sites, the category of concerns and p-values of multimodality tests.

CpG sites	SNP+INDEL_count	Multi Map	INDELs	SNP-at-CpG	Repeat	WGBS_HM450K_GT_0.3	p-values for multmodality
cg05390183	2	0	1	1	0	0	0.42
cg06099697	1	0	0	1	0	0	1.00
cg07124719	1	0	0	1	1	0	1.00
cg11699125	2	0	0	0	0	0	0.99
cg11988722	1	0	0	0	1	0	1.00
cg14011077	2	0	0	0	0	0	1.00
cg15482717	1	0	0	0	1	0	0.75
cg17971251	1	0	0	0	1	0	0.74
cg20315954	4	0	0	0	1	0	1.00
cg21220721	0	0	0	0	0	1	0.77
cg25087851	1	0	0	1	0	0	1.00
cg25479097	1	0	0	1	0	0	0.98
cg25854298	1	0	0	0	1	0	1.00

Note: 1) SNP+INDEL_count: the total number of known SNPs (dbSNP ver135) and INDELs overlapping the probe region. 2) “1” indicates that the corresponding probe hybridizes to multiple genomic loci and “0” represents that the probe maps to unique genomic loci. 3) INDELs: the total number of INDELs which reside in the region hybridized by the probe. 4) SNP-at-CpG: “1” indicates that the given probe mapped to sequence with the SNP at the interrogated CpG, and "0" represents that the probe does not contain any SNP at the interrogated CpG site. 5) Repeat: “1” illustrates that the probe spans a region in the genome containing repeat sequence elements, and “0” represents that the probe hybridizes to non-repeat regions. 6) WGBS_HM450K_GT_0.3: “1” represents that, for a given probe, the absolute beta difference between whole-genome bisulfite sequencing (WGBS) and Illumina HumanMethylation450 (HM450K) bead array is greater than 0.3, and “0” shows that the probe holds absolute beta difference (WGBS-HM450K) less than or equal to 0.3.

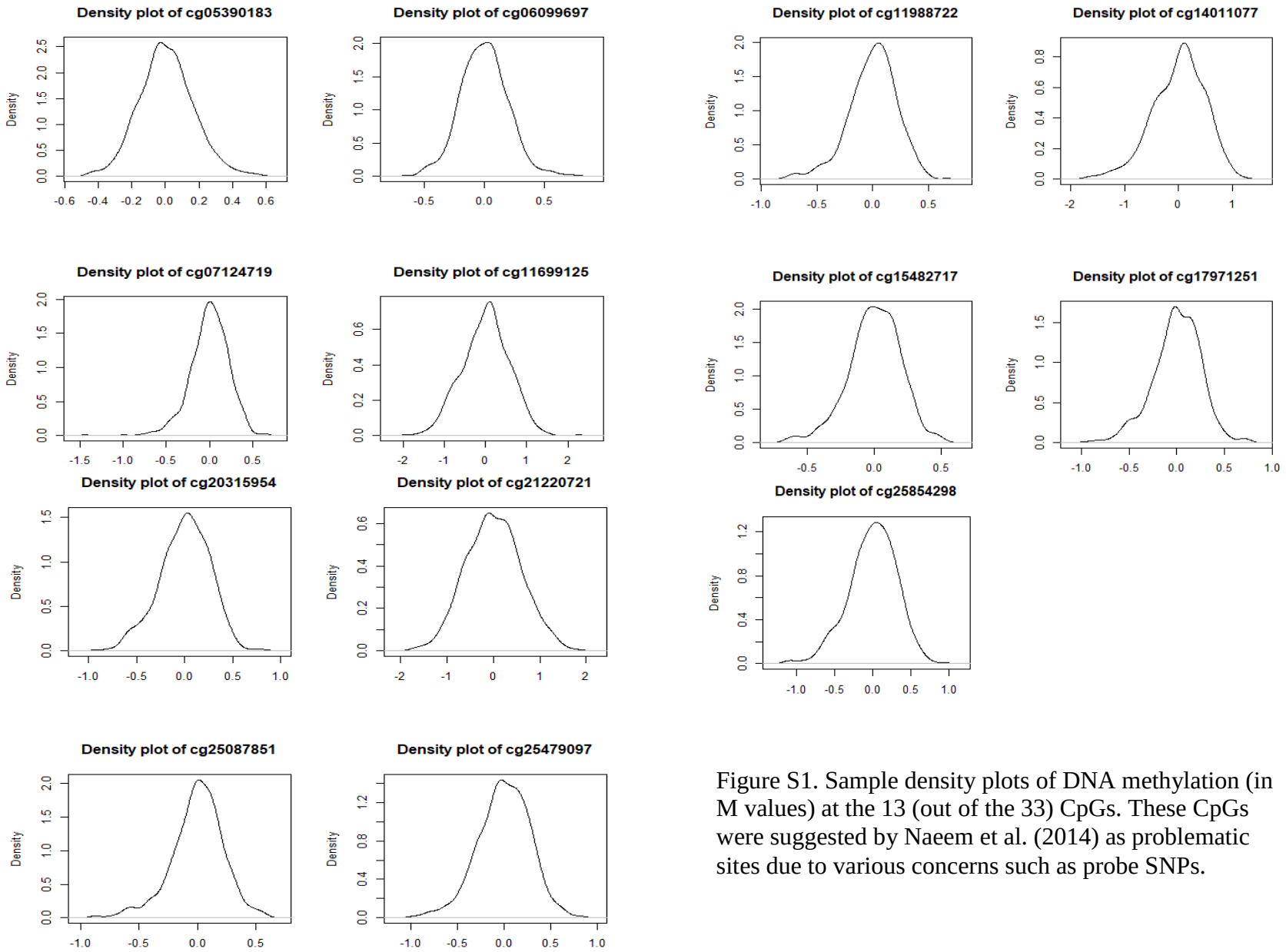

Figure S1. Sample density plots of DNA methylation (in M values) at the 13 (out of the 33) CpGs. These CpGs were suggested by Naeem et al. (2014) as problematic sites due to various concerns such as probe SNPs.

Agreement with specific IgEs on interaction effects effects (**sIgE×Age**): For age-specific sensitization effects on DNAm, consistent associations at the three CpGs (cg14121142 on *DDN*, cg23842695 on *PRKD2*, and cg26496795 on *ANKRD20A8P* identified via SPT) were also observed (Table S4).

Table S4. Results from testing the identified 3 CpGs (Table 4a in the text) by using screening specific IgE to determine the sensitization status to at least one allergen.

				Main effects (sIgE)		**Interaction effects (sIgE×Age)**	
CpG Name	**Gene Name**	**Location**	**Chr.**	**Est.**	**Raw p-value**	**Est.**	**Raw p-value**
cg14121142	*DDN*	Body	12	-0.81	0.003	0.75	**0.004**
cg23842695	*PRKD2*	Body	19	0.18	0.015	-0.17	**0.009**
cg26496795	*ANKRD20B*	Body	2	0.025	0.44	-0.07	**0.012**

Note: 1) Chr..=Chromosome, Est.=Estimation. 2) Results statistically significant at 0.05 are in bold.

The 48 CpGs passed screening based on comparison in DNAm between the poly- and mono-sensitizations:

Table S5. Results from ttScreening for candidate CpGs with DNAm potentially associated with poly-sensitization status.

CpG Sites	Frequency of Selection	Intercept	Effects of poly-sensitization	Raw P-value
cg00561338	64	1.951	0.381	5.27×10^{-5}
cg00612107	55	-3.791	0.163	1.76×10^{-5}
cg00788688	59	4.269	-1.078	3.4×10^{-4}
cg00958815	66	3.228	0.363	4.70×10^{-6}
cg01381613	68	1.993	0.380	2.71×10^{-5}
cg01818634	56	3.339	-0.378	1.11×10^{-5}
cg02002217	62	-0.667	1.213	4.98×10^{-5}
cg04003990	67	-2.012	1.513	5.20×10^{-6}
cg04441857	57	-4.153	0.436	2.07×10^{-5}
cg06031422	60	3.352	-0.344	3.62×10^{-6}
cg06418871	56	-2.344	-0.810	1.71×10^{-4}
cg07499182	60	0.345	-0.680	5.93×10^{-5}
cg07542871	59	0.727	-1.312	3.03×10^{-4}
cg08234308	54	-4.891	-0.459	1.62×10^{-4}
cg08356028	66	3.518	-0.514	2.09×10^{-5}
cg08471713	55	-3.124	1.333	2.89×10^{-4}
cg08553156	58	-4.208	-0.306	6.09×10^{-4}
cg08762603	50	2.010	-0.693	1.21×10^{-4}
cg08767938	53	-1.359	-0.312	4.81×10^{-5}
cg09379340	51	-2.507	-0.631	3.54×10^{-4}
cg09451235	54	-4.272	0.475	9.73×10^{-5}
cg09547119	55	-2.150	0.590	3.24×10^{-5}
cg09869015	52	7.050	-0.801	2.32×10^{-4}
cg10020897	51	-4.805	0.280	8.90×10^{-5}
cg10188668	57	4.813	0.510	5.59×10^{-5}
cg10360139	51	3.850	-0.574	7.31×10^{-4}
cg11371879	57	6.351	-0.475	2.16×10^{-4}
cg11772527	50	3.082	-0.208	4.26×10^{-6}
cg12645858	50	3.260	0.235	9.08×10^{-5}
cg13446199	52	-1.410	0.475	1.17×10^{-4}
cg13692739	57	-2.089	0.680	2.03×10^{-5}
cg13828227	54	4.135	-0.200	7.35×10^{-6}
cg16049707	62	-4.958	-1.221	8.05×10^{-6}
cg17133762	53	-4.720	-0.391	3.81×10^{-5}

cg18192325	52	0.553	0.550	4.96×10^{-5}
cg19879537	58	3.321	-0.221	8.30×10^{-5}
cg20241256	50	-3.611	-0.570	3.21×10^{-5}
cg20980408	51	2.624	0.354	2.73×10^{-5}
cg21117808	69	0.482	0.891	4.82×10^{-5}
cg22752533	59	-3.509	0.607	1.84×10^{-4}
cg23109891	50	0.990	0.514	1.80×10^{-5}
cg23584176	58	-0.812	0.920	4.37×10^{-4}
cg24996339	51	3.547	0.270	1.57×10^{-4}
cg25152909	57	3.823	-0.284	4.04×10^{-5}
cg25426203	54	4.088	-0.338	4.08×10^{-5}
cg25591377	62	1.940	0.606	1.25×10^{-4}
cg26333594	65	-4.057	-0.401	1.59×10^{-5}
cg26386044	52	2.458	-0.370	8.04×10^{-5}

Note: Frequency of selection is a statistic outputted from the ttScreening package, which is used to quantify the strength of being informative of a candidate CpG site with respect to its association with a variable of interest (in this case, it was the poly-sensitization status). The larger the frequency, the higher potential of being informative.

1. Ray, M., X. Tong, and H. Zhang, *ttScreening: Genome-wide DNA methylation sites screening by use of training and testing samples (R Package)*, 2014.
2. Ray, M.A., et al., *An Efficient Approach to Screening Epigenome-Wide Data.* Biomed Res Int, 2016. **2016**: p. 16.
3. Gref, A., et al., *Genome-Wide Interaction Analysis of Air Pollution Exposure and Childhood Asthma with Functional Follow-up.* Am J Respir Crit Care Med, 2017. **195**(10): p. 1373-1383.
4. Naeem, H., et al., *Reducing the risk of false discovery enabling identification of biologically significant genome-wide methylation status using the HumanMethylation450 array.* BMC Genomics, 2014. **15**: p. 51.
5. Maechler, M., *Diptest: Hartigan's dip test statistic for unimodality—corrected.* R package version 0.75-7. See https://CRAN. R-project. org/package= diptest, 2015.