

BRIEF COMMUNICATION

# ***GSTP1* is a hub gene for gene–air pollution interactions on childhood asthma**

M. W. Su<sup>1</sup>, C. H. Tsai<sup>1</sup>, K. Y. Tung<sup>1</sup>, B. F. Hwang<sup>2</sup>, P. H. Liang<sup>3</sup>, B. L. Chiang<sup>4</sup>, Y. H. Yang<sup>4</sup> & Y. L. Lee<sup>1,5</sup>

<sup>1</sup>Institute of Epidemiology and Preventive Medicine, College of Public Health, National Taiwan University, Taipei; <sup>2</sup>Department of Occupational Safety and Health and Graduate Program, College of Public Health, China Medical University, Taichung; <sup>3</sup>School of Pharmacy, College of Medicine, National Taiwan University, Taipei; <sup>4</sup>Department of Pediatrics, National Taiwan University Hospital, Taipei; <sup>5</sup>Institute of Biomedical Sciences, Academia Sinica, Taipei, Taiwan

**To cite this article:** Su MW, Tsai CH, Tung KY, Hwang BF, Liang PH, Chiang BL, Yang YH, Lee YL. *GSTP1* is a hub gene for gene–air pollution interactions on childhood asthma. *Allergy* 2013; DOI: 10.1111/all.12298.

## **Keywords**

asthma; gene–air pollution interactions; multifactor dimensionality reduction; ordered subset information gain analysis.

## **Correspondence**

Yungling Leo Lee, MD, PhD, Institute of Epidemiology and Preventive Medicine, College of Public Health, National Taiwan University, Taipei 100, Taiwan.  
Tel.: 886-2-3366-8016  
Fax: 886-2-2351-1955  
E-mail: leolee@ntu.edu.tw

Accepted for publication 5 September 2013

DOI:10.1111/all.12298

Edited by: Stephan Weidinger

## **Abstract**

There is growing evidence that multiple genes and air pollutants are associated with asthma. By identifying the effect of air pollution on the general population, the effects of air pollution on childhood asthma can be better understood. We conducted the Taiwan Children Health Study (TCHS) to investigate the influence of gene–air pollution interactions on childhood asthma. Complete monitoring data for the ambient air pollutants were collected from Taiwan Environmental Protection Agency air monitoring stations. Our results show a significant two-way gene–air pollution interaction between glutathione S-transferase P (*GSTP1*) and PM<sub>10</sub> on the risk of childhood asthma. Interactions between *GSTP1* and different types of air pollutants have a higher information gain than other gene–air pollutant combinations. Our study suggests that interaction between *GSTP1* and PM<sub>10</sub> is the most influential gene–air pollution interaction model on childhood asthma. The different types of air pollution combined with the *GSTP1* gene may alter the susceptibility to childhood asthma. It implies that *GSTP1* is an important hub gene in the anti-oxidative pathway that buffers the harmful effects of air pollution.

In recent years, there is growing evidence that gene–air pollution interaction is related to asthma (1, 2). Previous studies have generated conflicting results for gene–pollution interaction in terms of the presence and direction of interactions (3). As the underlying mechanism of air pollution effects on asthma is not fully understood, this study provide a system biology approach to comprehensively elucidate the role of

genes in gene–air pollution interaction that may help us to understand the response to exposure to various air pollutants (4).

## **Methods**

In TCHS, we recruited 4134 seventh-grade children from 14 Taiwanese communities, of which 3812 children provided samples of their oral mucosa for genotyping in 2007 (5–7). Communities were selected with the aim of maximizing the variability and minimizing the correlations of exposures to outdoor pollutants based on historic routine air monitoring data. Information pertaining to the child's health status was collected with a questionnaire and completed pulmonary function tests (8). Children were confirmed to have asthma if there was an affirmative answer to the question 'Has a doctor ever diagnosed this child as having asthma?' The final study population comprised of 295 children with asthma and

## **Abbreviations**

*ADRB2*,  $\beta$ 2-adrenergic receptor; CO, carbon monoxide; FEV<sub>1</sub>, forced expiratory volume in 1 sec; FVC, forced vital capacity; *GSTM1*, glutathione S-transferase M1; *GSTP1*, glutathione S-transferase P; *GSTT1*, glutathione S-transferase T1; HWE, Hardy–Weinberg equilibrium; MDR, multifactor dimensionality reduction; MMEF, maximal mid-expiratory flow; NO<sub>2</sub>, nitrogen dioxide; *NQO1*, quinone oxidoreductase; O<sub>3</sub>, ozone; OSIG, ordered subset information gain; PM<sub>10</sub>, particles; ROS, reactive oxygen species; SO<sub>2</sub>, sulfur dioxide; TCHS, Taiwan Children Health Study; *TNFA*, tumor necrosis factor alpha.

3517 controls without asthma (Table S1). Based on historic routine air monitoring data from Taiwan Environmental Protection Administration monitoring station in each community (6, 9), the 3-year average (2005–2007) concentrations of 24-h NO<sub>2</sub>, CO, SO<sub>2</sub>, PM<sub>10</sub> and 10:00 am to 6:00 pm 8-h O<sub>3</sub> were calculated in each community. To avoid bias of differential changes in air pollutants during 2005–2007, we monitored respective fluctuations during this period. As no changes were observed, a 3-year average of air pollution exposure was applied in this study. In this study, we tried to identify the air pollution effects on asthma and to divide participants into high exposure and low exposure groups. However, the association between asthma and PM<sub>2.5</sub> was complicated and could not be simply explained by high and low exposure of PM<sub>2.5</sub>. We therefore decided to exclude PM<sub>2.5</sub> in further analyses.

To investigate the influence of gene–air pollution interactions on childhood asthma, we examined the most commonly studied genes involved in the oxidative stress pathway which are suggested to be associated with asthma for genotyping (10). SNPs were selected from previous reports, by functional relevance and haplotype-tagging capacity. A total of 10 SNPs in *TNFα*, *ADRB2*, *EPXH1*, *GSTs* and *NQO1* genes in anti-oxidative pathway were assessed by real-time polymerase chain reaction (PCR). The ordered subset information gain (OSIG) analysis was used to objectively define the asthma-related exposure level of ambient air pollutants (Note S1). The OSIG analysis was conducted using Microsoft Excel. Gene–air pollution interactions were detected by a model-free and nonparametrical open-source MDR software package (11). Finally, to visualize and interpret the MDR results, we performed interaction graph analyses using the Orange software package (12) to represent the information gain, or percentage of entropy removed between two factors.

## Results

All SNPs were under the Hardy–Weinberg equilibrium (HWE) ( $P > 10^{-4}$ ) (Table S2). The genotyping call rate for each SNP was over 98%. The OSIG analysis splits each air pollutant with the maximal information gain to assign participants with high-level or low-level exposure (Figure S1). A daily average ambient PM<sub>10</sub> level of 51.55 µg/m<sup>3</sup> showed the maximal information gain among all air pollutants. More than 41.7% of children in our study lived in high-PM<sub>10</sub> communities (ambient PM<sub>10</sub> levels > 51.55 µg/m<sup>3</sup>).

The interaction between *GSTP1* and PM<sub>10</sub> possessed the highest testing balanced accuracy at 56.34%, with a cross-validation consistency of 10/10 and statistically significant based on 1000 permutation tests (Table 1). Based on 1000 permutation tests, we evaluated the statistical significance of top ten two-way interaction models (Table S3). Children with a *GSTP1* Ile/Ile genotype living in low PM<sub>10</sub> communities were more likely to suffer from asthma than those with a *GSTP1* Ile/Val or Val/Val genotype (Figure S2). In high PM<sub>10</sub> communities, *GSTP1* Ile/Val or Val/Val genotype was the high-risk genotype of childhood asthma. Chi-square tests also revealed significant differences between high-risk and low-risk classification on childhood asthma (Figure S2) ( $\chi^2 = 19.2348$ ,  $P < 0.0001$ ).

In each PM<sub>10</sub> exposure cutoff level, the testing balanced accuracy of *GSTP1* and PM<sub>10</sub> was estimated by MDR (Figure S3). The interaction between the *GSTP1* and PM<sub>10</sub> at 51.55 µg/m<sup>3</sup> showed the highest testing balanced accuracy, consistent with the PM<sub>10</sub> results in Figure S1.

Table 2 shows the result of information gain analysis on each gene–air pollution combination. The positive information gain value indicates the synergistic effect of gene–air

**Table 1** Summary of gene–air pollution interaction for childhood asthma from MDR models

| Model                                                                                                                  | Training bal. acc. (%) | Testing bal. acc. (%) | Cross-validation consistency |
|------------------------------------------------------------------------------------------------------------------------|------------------------|-----------------------|------------------------------|
| PM <sub>10</sub>                                                                                                       | 53.74                  | 49.68                 | 6/10                         |
| <i>GSTP1</i> (rs1695), PM <sub>10</sub>                                                                                | <b>56.62</b>           | <b>56.34</b>          | <b>10/10</b>                 |
| <i>GSTP1</i> (rs1695), <i>EPXH1</i> (rs2234922), PM <sub>10</sub>                                                      | 58.09                  | 54.27                 | 5/10                         |
| <i>EPXH1</i> (rs1051740), <i>NQO1</i> (rs1800566), <i>ADRB2</i> (rs1042713), O <sub>3</sub>                            | 61.18                  | 50.31                 | 4/10                         |
| <i>EPXH1</i> (rs1051740), <i>EPXH1</i> (rs2234922), <i>NQO1</i> (rs1800566), <i>ADRB2</i> (rs1042713), NO <sub>2</sub> | 65.53                  | 48.59                 | 3/10                         |

The maximum cross-validation consistency and maximum balanced accuracy are indicated in bold type.

**Table 2** Information gain analysis of each gene–air pollution combination

|                  | <i>GSTP1</i><br>(rs1695) | <i>GSTM1</i> | <i>GSTT1</i> | <i>EPXH1</i><br>(rs1051740) | <i>EPXH1</i><br>(rs2234922) | <i>ADRB2</i><br>(rs1042714) | <i>ADRB2</i><br>(rs1042713) | <i>NQO1</i><br>(rs1800566) | <i>NQO1</i><br>(rs2917666) | <i>TNFα</i><br>(rs1800629) |
|------------------|--------------------------|--------------|--------------|-----------------------------|-----------------------------|-----------------------------|-----------------------------|----------------------------|----------------------------|----------------------------|
| PM <sub>10</sub> | <b>0.738</b>             | −0.029       | −0.011       | 0.060                       | 0.013                       | 0.034                       | 0.044                       | 0.079                      | 0.203                      | 0.117                      |
| SO <sub>2</sub>  | <b>0.222</b>             | −0.021       | −0.026       | 0.118                       | 0.022                       | 0.098                       | 0.138                       | 0.095                      | 0.198                      | 0.078                      |
| O <sub>3</sub>   | <b>0.304</b>             | 0.001        | −0.037       | 0.146                       | 0.050                       | 0.034                       | 0.103                       | 0.132                      | 0.191                      | 0.090                      |
| NO <sub>2</sub>  | 0.296                    | −0.013       | −0.015       | 0.021                       | <b>0.438</b>                | 0.033                       | 0.024                       | 0.077                      | 0.034                      | 0.082                      |
| CO               | <b>0.199</b>             | −0.007       | −0.055       | 0.184                       | 0.159                       | 0.077                       | 0.035                       | 0.080                      | 0.038                      | 0.098                      |

Values with maximum entropy removed % by gene–air pollutant combination for each pollutant are indicated in bold type.

pollutant combinations. An interaction graph was used to visualize and interpret the gene–air pollution network (Figure S4). The percentage of entropy explained by the interaction between *GSTP1* and  $PM_{10}$  (0.78%) was larger than the percentage of entropy explained by *GSTP1* (0.01%), or  $PM_{10}$  (0.33%) independently. In addition, a gene–air pollution interaction between *EPHX1* (rs2234922) and  $NO_2$  is indicated. The negative percentage of entropy explained by the interaction between each of the two air pollutants indicates the dependence of air pollutants.

## Discussion

In many studies, the effect of genetic or environmental factors associated with the case–control status of participants (13) is difficult to replicate. Differences in exposure level definition may hamper replication of their findings between different studies. In this study, we developed OSIG analysis based on the environmental risk factor and disease case–control status to objectively define the disease-related air pollution exposure standard in the general population. Furthermore, we systematically analyzed anti-oxidative genes with common air pollutants and elucidated the gene–air pollution interaction network on childhood asthma. Using the MDR approach, the two-way gene–air pollution interaction between *GSTP1* and  $PM_{10}$  is elucidated. It is suggested that the *GSTP1* enzyme activity may be influenced by the genotype of *GSTP1* (14) and associate with the risk for allergic diseases through an impaired defense against oxidant injury (15). Different *GSTP1* genotypes with different enzyme activity are thought to be the major reason for the nonuniform effect between *GSTP1* and  $PM_{10}$  (Figure S2). The hub genes of interaction networks confer environmental resilience and phenotypic stability. Information gain values of *GSTP1*–air pollutant combinations were higher than all other combinations for certain air pollutants, indicating that *GSTP1* is the most important hub gene in the anti-oxidative pathway to buffer the stochastic changes of multiple air pollutants on childhood asthma. Only the combination of *EPHX1* (rs2234922)– $NO_2$  showed slightly higher information gain value than the *GSTP1*– $NO_2$  combination. There are some potential limitations of this study. The definition of asthma was based on parental reporting, but questionnaires are widely used to define childhood respiratory outcomes in many epidemiologic studies (7, 9). Another potential limitation is the exposure assessment of air pollution, which relied upon data obtained from monitoring stations not located at schools. However, the schools chosen were all located in the vicinity of the monitoring stations and almost all Taiwanese children attend schools within 1 km of their homes (9). Therefore, our data represent exposure both in the school and in the home, and so it is reasonable to use the community level exposure assessment to represent individual level exposure. Numerous studies over the years have linked  $PM_{2.5}$  to asthma. Multiple peaks in our results of  $PM_{2.5}$  OSIG analysis imply that the association between asthma and  $PM_{2.5}$  is very complex and cannot be simply explained by high and low exposure of  $PM_{2.5}$ . We also found maximal

information gain in  $PM_{2.5}$  was 0.00053, much lower than the maximal information gain in  $PM_{10}$  (0.001). Therefore, our results in OSIG analysis constitute the evidence that  $PM_{10}$  is the most important air pollutant on childhood asthma. This study provided for the first time a comprehensive candidate gene approach for gene–air pollution interaction analyses identifying *GSTP1* as a hub gene. With genome-wide association data available, hypothesis-free analyses will further extend our current knowledge in identifying novel gene–environment interaction pathways.

## Acknowledgments

The authors thank all the field workers who supported the data collection, the school administrators and teachers, and especially the parents and children who participated in this study.

## Author contributions

MWS, KYT, PHL, CHT, BFH, BLC, YHY and YLL conceived and designed the experiments; MWS and KYT performed the experiments; MWS analyzed the data; MWS, KYT and CHT contributed reagents/materials/analysis tools; MWS and YLL wrote the paper.

## Funding

This study was supported by grants #98-2314-B-002-138-MY3, #96-2314-B-006-053 and #95-2314-B-006-103 from Taiwan National Science Council. The funders had no role in study design, data collection, analysis, decision to publish or manuscript preparation.

## Conflicts of interest

We declare that we have no conflicts of interest.

## Supporting Information

Additional Supporting Information may be found in the online version of this article:

**Figure S1.** Results of ordered-subset information gain (OSIG) analysis for childhood asthma with air pollutants.

**Figure S2.** The best two-way interaction model between *GSTP1* and  $PM_{10}$  for childhood asthma.

**Figure S3.** Testing balanced accuracy of *GSTP1*– $PM_{10}$  interaction models by ambient  $PM_{10}$  level.

**Figure S4.** The percentage of entropy removed by each factor and by each pairwise combination of factors.

**Table S1.** Demographic characteristics and pulmonary function indices for study participants in TCHS.

**Table S2.** Genotyping characteristics of each single nucleotide polymorphism.

**Table S3.** Top ten models for two-way gene–air pollution interaction for childhood asthma.

**Note S1.** Ordered-subset information gain (OSIG) analysis.

## References

1. Romieu I, Moreno-Macias H, London SJ. Gene by environment interaction and ambient air pollution. *Proc Am Thorac Soc* 2010;**7**:116–122.
2. Kim BJ, Seo JH, Jung YH, Kim HY, Kwon JW, Kim HB et al. Air pollution interacts with past episodes of bronchiolitis in the development of asthma. *Allergy* 2013;**68**: 517–523.
3. Minelli C, Wei I, Sagoo G, Jarvis D, Shaheen S, Burney P. Interactive effects of antioxidant genes and air pollution on respiratory function and airway disease: a HuGE review. *Am J Epidemiol* 2011;**173**: 603–620.
4. London SJ. Gene-air pollution interactions in asthma. *Proc Am Thorac Soc* 2007;**4**:217–220.
5. Tsai CH, Tung KY, Chen CH, Lee YL. Tumour necrosis factor G-308A polymorphism modifies the effect of home dampness on childhood asthma. *Occup Environ Med* 2011;**68**:771–776.
6. Tung KY, Tsai CH, Lee YL. Microsomal epoxide hydroxylase genotypes/diplotypes, traffic air pollution, and childhood asthma. *Chest* 2011;**139**:839–848.
7. Tsai CH, Huang JH, Hwang BF, Lee YL. Household environmental tobacco smoke and risks of asthma, wheeze and bronchitic symptoms among children in Taiwan. *Respir Res* 2010;**11**:11.
8. Lee YL, Hwang BF, Chen YA, Chen JM, Wu YF. Pulmonary function and incident bronchitis and asthma in children: a community-based prospective cohort study. *PLoS One* 2012;**7**:e32477.
9. Hwang BF, Lee YL. Air pollution and prevalence of bronchitic symptoms among children in Taiwan. *Chest* 2010;**138**:956–964.
10. Castro-Giner F, Kunzli N, Jacquemin B, Forsberg B, de Cid R, Sunyer J et al. Traffic-related air pollution, oxidative stress genes, and asthma (ECHRS). *Environ Health Perspect* 2009;**117**:1919–1924.
11. Su MW, Tung KY, Liang PH, Tsai CH, Kuo NW, Lee YL. Gene-gene and gene-environmental interactions of childhood asthma: a multifactor dimension reduction approach. *PLoS ONE* 2012;**7**:e30694.
12. Demsar J, Zupan B, Leban G, Curk T. Orange: from experimental machine learning to interactive data mining. Knowledge Discovery in Databases: Pkdd 2004, Proceedings 2004;**3202**:537–539.
13. Renz H, Conrad M, Brand S, Teich R, Garn H, Pfefferle PI. Allergic diseases, gene-environment interactions. *Allergy* 2011;**66**(Suppl 95):10–12.
14. Zimniak P, Nanduri B, Pikula S, Bando-rowicz-Pikula J, Singhal SS, Srivastava SK et al. Naturally occurring human glutathione S-transferase GSTP1-1 isoforms with isoleucine and valine in position 104 differ in enzymic properties. *Eur J Biochem* 1994;**224**:893–899.
15. Tischer CG, Gref A, Standl M, Bauer M, Bergstrom A, Brauer M et al. Glutathione-S-transferase P1, early exposure to mould in relation to respiratory and allergic health outcomes in children from six birth cohorts. A meta-analysis. *Allergy* 2013;**68**:339–346.

## Supplementary Table

**Table S1. Demographic characteristics and pulmonary function indices for study participants in TCHS.**

| Characteristic             | Asthmatics<br>(N = 295) | Controls<br>(N = 3517) |
|----------------------------|-------------------------|------------------------|
| Gender                     |                         |                        |
| Boys                       | 164 (55.6%)             | 1708 (48.6%)           |
| Girls                      | 131 (44.4%)             | 1809 (51.4%)           |
| Age, yr                    | 12.7 ± 0.9              | 12.7 ± 0.6             |
| Pulmonary function indices |                         |                        |
| FVC (% predicted)          | 99.8 ± 13.3             | 99.9 ± 13.7            |
| FEV1 (% predicted)         | 97.6 ± 14.1             | 100.0 ± 14.1           |
| MMEF (% predicted)         | 93.5 ± 23.9             | 100.4 ± 21.9           |
| FEV1/FVC (%)               | 88.7 ± 6.2              | 91.2 ± 5.3             |

Values were presented by n (%) or mean ± SD.

FEV1, forced expiratory volume in one second.

MMEF, maximal mid-expiratory flow.

FVC, functional vital capacity.

**Table S2. Genotyping characteristics of each single nucleotide polymorphism.**

| Gene                                      | Chromosome position | Location     | Polymorphism | MAF   | HWE   | P-value* |
|-------------------------------------------|---------------------|--------------|--------------|-------|-------|----------|
| <i>TNF<math>\alpha</math></i> , rs1800629 | 6p21.3              | 5' near gene | A/G          | 0.105 | 0.025 | 0.875    |
| <i>ADRB2</i> , rs1042713                  | 5q31-q32            | Missense     | A/G          | 0.434 | 6.957 | 0.008    |
| <i>ADRB2</i> , rs1042714                  | 5q31-q32            | Missense     | C/G          | 0.094 | 0.285 | 0.593    |
| <i>EPXH1</i> exon 4, rs2234922            | 1q42.1              | Exon         | A/G          | 0.129 | 2.622 | 0.105    |
| <i>EPHX1</i> exon 3, rs1051740            | 1q42.1              | Exon         | C/T          | 0.476 | 1.934 | 0.164    |
| <i>GSTP1</i> , rs1695                     | 11q13               | Exon         | A/G          | 0.193 | 1.281 | 0.257    |
| <i>NQO1</i> , rs1800566                   | 16q22               | Exon         | C/T          | 0.465 | 6.967 | 0.008    |
| <i>NQO1</i> , rs2917666                   | 16q22               | Exon         | C/G          | 0.159 | 1.822 | 0.177    |
| <i>GSTT1</i>                              | 22q11.23            |              |              |       |       |          |
| <i>GSTM1</i>                              | 1p13.3              |              |              |       |       |          |

MAF, minor allele frequency; HWE, Hardy Weinberg equilibrium.

\*p-value for Hardy-Weinberg Equilibrium.

**Table S3. Top ten models for two-way gene-air pollution interaction for childhood asthma.**

| Rank* | Models                                      | Training Bal.<br>Acc. (%) | Testing Bal.<br>Acc. (%) <sup>#</sup> | <i>p</i> -value <sup>&amp;</sup> |
|-------|---------------------------------------------|---------------------------|---------------------------------------|----------------------------------|
| 1     | <i>GSTP1</i> (rs1695), PM <sub>10</sub>     | <b>56.62</b>              | <b>56.33</b>                          | <b>0.005</b>                     |
| 2     | <i>NQO1</i> (rs2917666), PM <sub>10</sub>   | 54.41                     | 53.95                                 | 0.066                            |
| 3     | <i>EPHX1</i> ( rs2234922), CO               | 54.14                     | 53.85                                 | 0.074                            |
| 4     | <i>ADRB2</i> (rs1042714), SO <sub>2</sub>   | 54.02                     | 53.39                                 | 0.097                            |
| 5     | <i>GSTP1</i> (rs1695), O <sub>3</sub>       | 54.79                     | 53.35                                 | 0.099                            |
| 6     | <i>EPHX1</i> ( rs1051740), O <sub>3</sub>   | 54.11                     | 52.30                                 | 0.174                            |
| 7     | <i>EPHX1</i> ( rs2234922), PM <sub>10</sub> | 54.08                     | 52.25                                 | 0.180                            |
| 8     | <i>NQO1</i> (rs1800566), O <sub>3</sub>     | 54.07                     | 51.71                                 | 0.235                            |
| 9     | <i>GSTP1</i> (rs1695), SO <sub>2</sub>      | 54.03                     | 51.67                                 | 0.243                            |
| 10    | <i>NQO1</i> (rs1800566), PM <sub>10</sub>   | 54.03                     | 51.58                                 | 0.250                            |

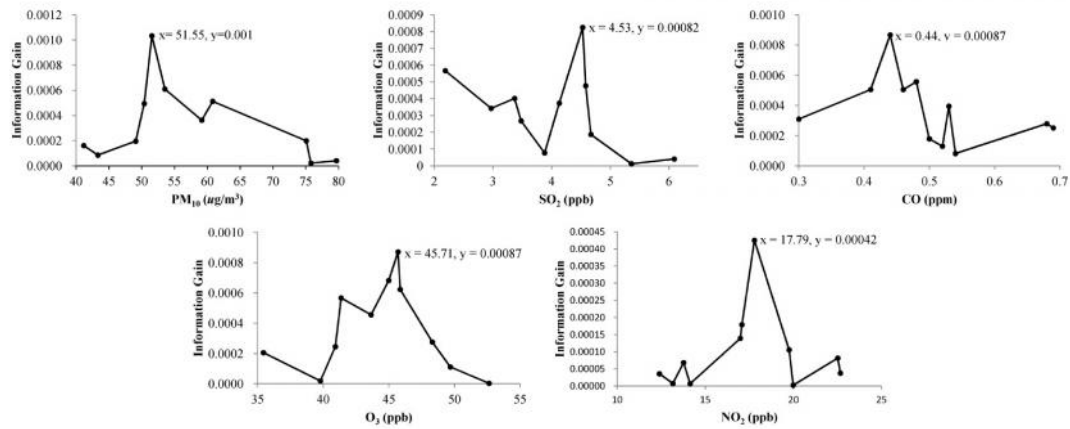
\*Ranking is determined by testing balance accuracy.

<sup>&</sup>Two-way interactions were validated based on 1000 permutations. Significant two-way interaction models are in bold type.

<sup>#</sup> Testing balanced-accuracy is obtained from the sum of true positive plus true negative divided by the total number of samples in the testing dataset.

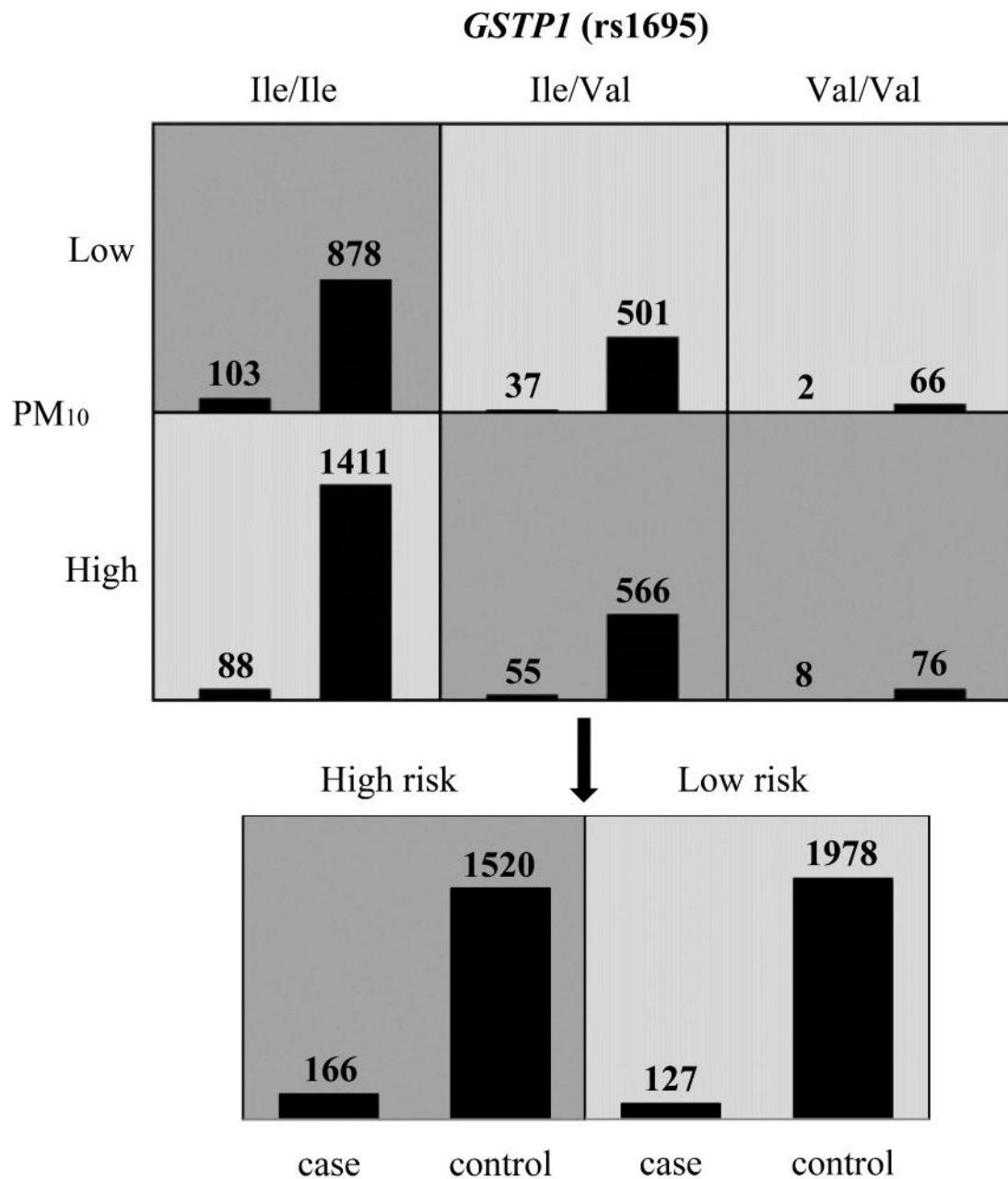
## Supplementary Figure

**Figure S1. Results of ordered-subset information gain (OSIG) analysis for childhood asthma with air pollutants.**



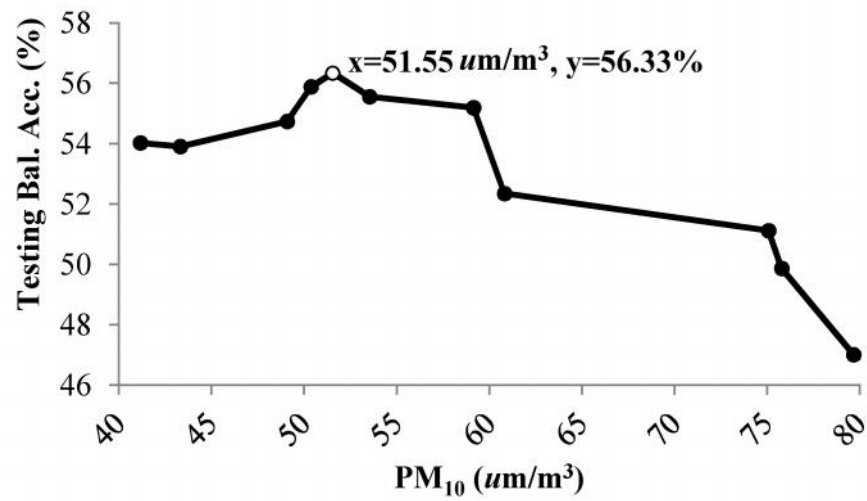
OSIG analysis, performed for five common air pollutants, showed that the maximal information gain of different air pollution exposure level. At least 2 communities in each subset, and therefore 11 air pollution exposure level in each air pollutant. (a) Particulate matter with 50% cut-off aerodynamic diameter of 10 μm (PM<sub>10</sub>). (b) Sulfur dioxide (SO<sub>2</sub>). (c) Carbon monoxide (CO). (d) 10:00 am to 6:00 pm 8-h 3 Ozone (O<sub>3</sub>). (e) Nitrogen dioxides (NO<sub>2</sub>).

**Figure S2. The best two-way interaction model between *GSTP1* and PM<sub>10</sub> for childhood asthma.**



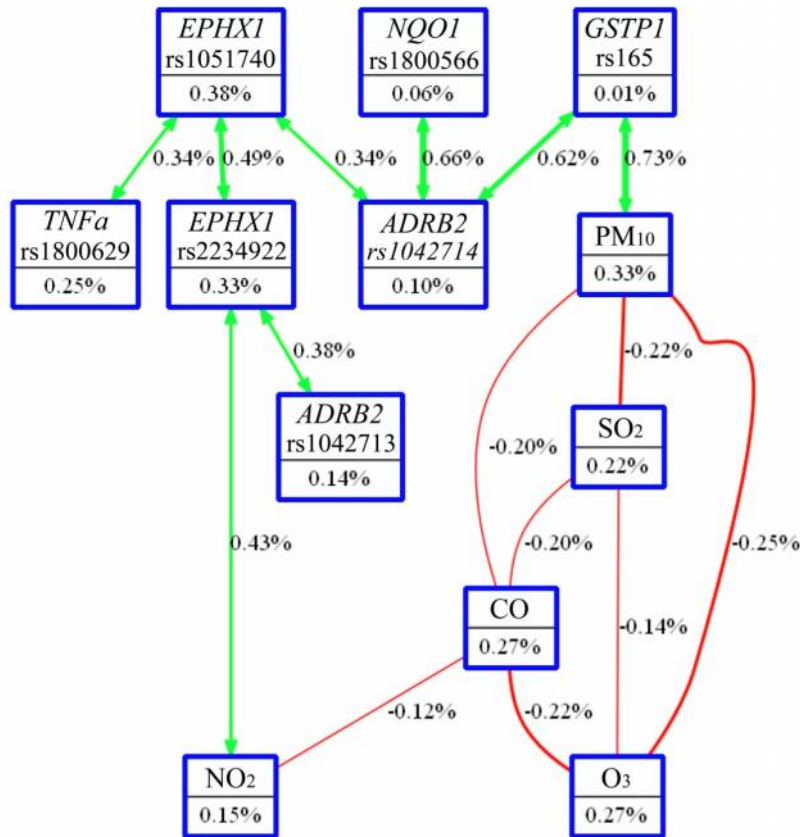
The different *GSTP1* genotypes and PM<sub>10</sub> exposure level combinations displayed by MDR. High risk combinations are in grey and low risk genotype combinations in white. The left bar represent the asthma group and the right bar represent the controls. The validation of high and low risk classification by chi-square test (19.2348,  $p < 0.0001$ ).

**Figure S3. Testing balanced accuracy of *GSTP1*-PM<sub>10</sub> interaction models by ambient PM<sub>10</sub> level.**



Optimal ambient PM<sub>10</sub> cut-off point was validated by the two-way interaction models between *GSTP1* and PM<sub>10</sub>.

**Figure S4. The percentage of entropy removed by each factor and by each pairwise combination of factors.**



The entropy removed in case-control status of each gene or air pollutant is shown in percentage below the label in a box. The arrow accompanied by a percent of entropy removed described synergistic interactions between factors. The negative percent of entropy removed depicted as a line indicated the factor redundancy. A positive percentage of entropy removed by two factors (plotted in green) indicates a synergistic interaction, whilst a negative percentage (plotted in red) indicates redundancy.

## Supplementary Note

### Ordered-subset information gain (OSIG) analysis

Study participants were ordered by air pollution exposure level of 14 communities. For each air pollutant, study participants were classified into low and high exposure-specific subsets. Due to the population size, at least two communities were included in the exposure-specific subset, so there are 11 cut-off points in each air pollutant. Participants living in an area where ambient levels of a certain air pollutant level were higher than the cut-off point were classified as high-exposure, and vice versa. Information gain was calculated for each cut-off point to determine the optimal cut-off point.

Information gain derived by entropy is a very important concept in data mining (8). Entropy is a nonlinear function used to measure the uncertainty associated with a random variable. If entropy is zero, all participants in the dataset belong to the same disease case or control status. If entropy is one, there are equal numbers of cases and controls in the dataset. Information gain of certain air pollution cut-off point  $X$  and the disease case-control status variable  $Y$  is defined by

$$I(X; Y) = H(Y) - H(Y|X) \quad (1)$$

where  $H(Y)$  is the Shannon entropy of  $Y$  (9), and  $H(Y|X)$  is the conditional entropy of  $Y$  when given a certain air-pollution cut-off point  $X$ . Information gain was used to compare the difference of case-control impurity and to evaluate how well the given cut-off point of a certain air pollutant separates the study participants. Entropy can be defined by

$$H(Y) = \sum_y p(y) \log \frac{1}{p(y)} \quad (2)$$

where  $p(y)$  is the frequency of disease (case) or healthy (control) participants respectively. The conditional entropy is defined by

$$H(Y|X) = \sum_{x,y} p(x,y) \log \frac{1}{p(y|x)} \quad (3)$$

For each cut-off point  $X$ , we can classify study participants into two exposure-specific subsets  $x$ .  $p(x, y)$  is the frequency of study participants in either the case or the control group that belong to a certain exposure-specific subset.  $p(y|x) = p(x, y)/p(x)$ , where  $p(x)$  is the frequency of participants that belongs to the exposure-specific subsets  $x$ .

For each air pollutant, we selected a cut-off point with the maximal information gain to distinguish the high- and low exposure participants according to asthma case-control status.