國立交通大學

統計學研究所

碩士論文

在全基因關聯分析中
利用總體的基因表現量作為穩定表現型

Using Global Gene Expression as Endophenotypes in a Genome-wide Association Study

研 究 生：蔡佩芳

指導教授：黃冠華　博士

中華民國九十七年七月

在全基因關聯分析中
利用總體的基因表現量作為穩定表現型

Using Global Gene Expression as Endophenotypes in a Genome-wide Association Study

研 究 生：蔡佩芳　　Student: Pei-Fang Tsai

指導教授：黃冠華　　Advisor: Dr. Guan-Hua Huang

國立交通大學

統計學研究所

碩士論文

A Thesis
Submitted to institute of Statistics
College of Science
National Chiao Tung University
in Partial Fulfillment of the Requirements
for the Degree of
Master
in
Statistics
July 2008

Hsinchu, Taiwan, Republic of China

中華民國九十七年七月

在全基因關聯分析中
利用總體的基因表現量作為穩定表現型

研究生：蔡佩芳　　　　指導教授：黃冠華　博士

國立交通大學統計學研究所

摘要

在生物學上，穩定表現型(endophenotype)和疾病有著相同的遺傳路徑，但穩定表現型卻比診斷上的表現型(phenotype)更為接近其相關的基因，這也顯示穩定表現型在複雜疾病上基因研究的重要性。穩定表現型為主的基因遺傳分析比表現型為主的基因遺傳分析更容易找到致病基因。由穩定表現型所發展的指標(PHE)，穩定表現型的基因遺傳性所佔比例，用在判別出有可能的穩定表現型。

在這篇報告裡，氣喘是一種基因間作用和環境因素複雜的疾病，我們利用基因表現量當作穩定表現型找尋可能的氣喘基因。我們利用指標(PHE)判斷哪些探針組的基因表現量是穩定表現型，接著作指標(PHE)的檢定。針對多重檢定的問題，我們利用 q-值來控制錯誤發現率和調整。我們也對每一個基因表現量做全基因關聯分析，比較基因表現量指標(PHE)中有顯著和沒有顯著之間基因遺傳性的變化。最後，我們論文中有(1) 評估利用基因表現量當作穩定表現型找尋跟疾病有相關基因的適當性，(2) 檢驗從基因表現量為主的分析，辨別出的基因和文獻中提過疾病相關的基因重複的多寡，(3) 從這些基因表現量判定為穩定表現型中，評估它們基因遺傳的特徵。

關鍵字: *順式作用 (cis effect); 穩定表現型 ; 表現量位置 ; 全基因關聯分析 ; 基因表現量 ; 遺傳率 ; 反式作用 (trans effect)*

Using Global Gene Expression as Endophenotypes in a Genome-wide Association Study

Student: Pei-Fang Tsai   Advisor: Dr. Guan-Hua Huang
Institute of Statistics
National Chiao Tung University

ABSTRACT

Endophenotype, which involve the same biological pathways as diseases but presumably are closer to the relevant gene action than diagnostic phenotypes, have emerged as an important concept in the genetic studies of complex diseases. Endophenotype-based genetic analysis is more likely to succeed than phenotype-based one in terms of search for the susceptibility genes. The index, proportion of heritability explained (PHE), has been proven useful in identifying potential endophenotypes.

In this report, we use global gene expression as endophenotypes in search for the susceptibility genes underlying asthma, which is a disease caused by complex interactions of genetic and environmental factors. We judge which gene expressions of probe sets are endophenotypes by using the index PHE and do hypothesis test of PHE. For the problem of multiple testing, we utilize the q-value to control for the false discovery rate (FDR) for significance judgment. We also perform genome-wide association tests for each gene expression and compare various genetic properties between gene expressions with and without significant PHE values. At the end, this thesis has (1) evaluated the appropriateness of using global gene expressing as endophenotypes in searching possible phenotype-related genes, (2) examined the overlap between genes identified by the gene-expression-based analysis and genes already identified in the literature, and (3) assessed genetic characteristics of gene expressions that are identified as the endophenotypes.

Key words: *Cis effect ; Endophenotype ; Expression quantitative trait loci ; Genome-wide association study ; Gene expression ; Heritability ; Trans effect*

誌 謝

這兩年來，首先感謝老師　黃冠華的指導，常常因為一點小問題就跑去找老師，老師都很熱心地幫忙解決問題和疑惑。最後還讓老師非常頭痛地改我的論文，對老師感到非常不好意思。也感謝老師一年來的訓練，學到很多看書和抓重點的技巧，也學習到面對一個全新領域一開始應有的應對方式，這一年來獲益良多。最後對老師在說一聲 謝謝。也感謝其他老師的教導，讓我對統計更有了解。

接下來要感謝我的同學們，像是同老師的同學們 重耕、彥銘和仲竹，彼此間會互相幫忙指導和提醒對方事情。另外也感謝夙吟、姿蒨、郃嵐和瑜達聽我練習和給我意見。也非常感謝其他的同學們，因為有你們讓我的研究所生活很精彩。

最後感謝我的家人，常常給我鼓勵和指導一些論文的事項，謝謝你們!

在此，謹以此篇論文，獻給我親愛的家人和同學，還有陪伴我的好朋友們!

蔡佩芳　2008.07.02

Contents

Figures and Tables Content

1 Introduction

In diseases with classic or Mendelian genetics as their distal causes, genotypes are usually indicative of phenotypes. However, this degree of genetic certainty does not exist for complex disease [1]. These “complex” diseases are influenced by multiple genes, environmental factors and their interactions on phenotypes. As a result, the direct relationship between phenotype and genotype is disrupted, so that the same genotype may result in different phenotypes, or different genotypes may result in the same phenotype. To facilitate the identification of influential genetic markers of complex disease, endophenotype approach has been advocated.

Endophenotypes are useful for theorizing about clinical phenotypes and mark the path between genotype and phenotype. The endophenotype is closer to the underlying gene than the phenotype in the course of disease’s natural history and can increase the chance of identifying genotype (Figure 1). Huang et al. [2] defined an endophenotype to be “a trait for which a test of null hypothesis of no genetic heritability implies the corresponding null hypothesis based on the phenotype of interest” and develop a formal statistical methodology for accessing the utility of endophenotypes, motivated by the conditioning strategy used for identifying surrogate endpoints in clinical research. The methodology is especially useful for the situation where underlying genotype is unknown and researchers use endophenotypes to increase opportunities of finding susceptible disease genes. Similar to validating surrogate endpoints, various indices can be used to validate endophenotypes. One of the indices is the proportion of heritability explained (PHE) by the endophenotype, similar to PTE introduced by Freedman et al. [3] in the surrogate endpoint study. The greater the PHE value, the more likely the intermediate variable is an endophenotype. Hsieh et al. [4] utilized the delta method to evaluate the variance of PHE.

There is a problem about signal-SNP association test between SNP genotypes and case-control status: it cannot discover SNPs that are weakly related to the disease by itself, but can have great impacts on the disease variability after combining with other SNPs. A large PHE value represents that endophenotype and phenotype share many genes. Because endophenotype is closer to genotype, there will be some genes that are significant in endophenotype, but are not significant in phenotype. We can utilize the PHE to find SNPs that may be weakly associated to the disease phenotype. Hopefully, these additional SNPs can increase our chances in searching possible phenotype-related genes.

Gene expression is a measurement of mRNA and mRNA is transcribed from a DNA template and carries coding information to the sites of protein synthesis. Because mRNA is closer to DNA genotype, we use global gene expression as endophenotypes in search for the susceptibility genes underlying asthma, which is a disease caused by complex interactions of genetic and environmental factor. In this thesis, we first get global gene expressions of probe sets in Epstein-Barr virus Iymphoblastoid cell lines (EBVL) measured with Affymetrix HG-U133 Plus 2.0 chip [5-6]. Gene expression values are preprocessed by the robust multi-array averaging (RMA). We judge which gene expressions of probe sets are endophenotypes by using the index PHE and do hypothesis test of PHE. For the problem of multiple testing, we utilize the q-value to control for the false discovery rate (FDR) for significance judgment. We also perform genome-wide association tests for each gene expression and compare various genetic properties between gene expressions with and without significant PHE values.

Therefore, this thesis aims at

(1) Evaluating the appropriateness of using global gene expressing as endophenotypes in searching possible phenotype-related genes,

(2) Examining the overlap between genes identified by the gene-expression-based analysis and genes already identified in the literature, and

(3) Assessing genetic characteristics of gene expressions that are identified as the endophenotypes.

2 Literature Review

2.1 Statistical validation of surrogate endpoints

Surrogate endpoints have been frequently utilized in most clinical research, when the primary endpoint is too difficult or costly or time-consuming to obtain. Clinically meaningful biomarkers of the disease projected as a surrogate endpoint in a clinical trial is expected to ultimately demonstrate treatment effect on the primary endpoint if a treatment effect shown on the markers. Surrogate endpoints have been of clinical interest for decades, but it was not until Prentice published a seminal paper in 1989 that formal statistical investigation started. Prentice defined a surrogate endpoint to be "a response variable for which a test of null hypothesis of no relationship to the treatment groups under comparison is also a valid test of the corresponding null hypothesis based on the true (clinical) endpoint".

Prentice's definition can be written as

$$f(S \mid X) = f(S) \Leftrightarrow f(T \mid X) = f(T)$$

Where T denotes the status of a primary endpoint, S denotes the status of a surrogate end-point, X is the treatment variable, f(S) is the distribution of S, and f(S|X) is the conditional distribution of S given X. Validation of Prentice's definition involves the following two criteria:

$$f(T \mid S) \neq f(T) \quad \text{and} \quad f(T \mid S, X) = f(T \mid S)$$

[3, 7-8]. The first criterion states that the surrogate endpoint must be correlated with the primary clinical endpoint, and the second criterion is that the surrogate endpoint should fully capture the treatment effect on the treatment effect on the primary endpoint.

The surrogate endpoint described by Prentice mediates all of the effect of treatment on the primary endpoint, that is

$$X \rightarrow S \rightarrow T$$

A more complex, but more likely, situation arises when treatment has a direct effect on the primary endpoint that is not mediated through the surrogate [9]:

$$\begin{tikzcd} X \arrow[r] \arrow[rr, bend right] & S \arrow[r] & T \end{tikzcd}$$

Freedman et al. [3] proposed to focus on the proportion of the treatment effect mediated through the surrogate. A good surrogate is one that explains large proportion of that effect. The proposal can be made in the content of generalized linear models [10]. The net effect of X on T can be assessed through the regression coefficient β_T in the generalized linear model

$$g[E(T)] = \alpha_T + \beta_T X$$

Where g(.) is the link function connecting the mean response and covariates, and the effect of X on T after inclusion of S is the regression coefficient β_{TS} in the following generalized linear model

$$g[E(T)] = \alpha_{TS} + \beta_{TS} X + \gamma_{TS} S$$

The proportion of the treatment effect (on the primary endpoint) explain (PTE) by the surrogate is given by

$$PTE = 1 - \frac{\beta_{TS}}{\beta_T}$$

The $100(1-\alpha)\%$ confidence limits of PTE can be calculated using the delta method.

2.2 Statistical validation of endophenotype

Notation:

$i = 1, \ldots, I$ **:** representing the different family

$j = 1, \ldots, n_i$: representing the jth member of this family

P_{ij} : The observed phenotype in the jth member of the ith family

x_{ij} : A vector of observed covariates

σ_A^2: The variances arising from polygenic additive effects,

σ_D^2: The variance arising from polygenic dominance effects

σ_C^2: The variance arising from the shared environmental effects

2.2.1 Model

Endophenotypes are useful for theorizing about clinical phenotypes and can mark the path between the genotype and the phenotype. Verification of existence of the pathway genotype-endophenotype-phenotype is the key of validating endophenotypes. Analogous to Prentice's definition [7] that surrogate endpoint to be "a response variable for which a test of null hypothesis of no relationship to the treatment groups under comparison is also a valid test of the corresponding null hypothesis based on the true (clinical) endpoint", Huang et al. [2] define an endophenotype to be "a trait for which a test of null hypothesis of no genetic heritability implies the corresponding null hypothesis based on the phenotype of interest". More specifically, suppose P is the phenotype of interest, E is the selected endophenotype, and G represents an underlying genetic structure that fulfills the specified assumptions in calculating heritability, then the proposed definition is:

$$f(E\,|\,G) = f(E) \Rightarrow f(P\,|\,G) = f(P) \qquad (1)$$

The endophenotype's definition has two important features [2]. First, "imply" replaces ‘’if and only if’’ statement in Prentice's definition of surrogate endpoints in avoidance of a problematic implication arisen in Begg and Leung [11]. This change places endophenotype in higher upstream of the pathway from genotype to phenotype. Second, genetic heritability represents the proportion of variability attributable to genetic factors and can be obtained in a variance component approach [12-13]. This is a perfect fit to our situation since it does not require knowledge of specific culprit genes and allows the likelihood of multiple gene influences.

Suppose the condition

$$f(P \mid E, G) = f(P \mid E) \qquad (2)$$

Huang et al. [2] takes (2) in a variance component model as the operational criterion for proposed endophenotype definition. It then requires heritability of phenotype becomes null, conditioning on candidate endophenotype, and implies genetic heritability of phenotype is captured by endophenotype.

Given an observed phenotype, significance of (2) can be judged through the following variance component analysis for discrete trait [2, 14]:

$$\begin{aligned} &P_{ij} = \alpha_H + \gamma_H E_{ij} + \tau_H x_{ij} + G_{ij} + \varepsilon_{ij}, \\ &\varepsilon_{ij} \sim Normal(0, \sigma_R^2) \\ &G_{ij} \sim Normal(0, [\sigma_A^2 + \sigma_D^2 + \sigma_C^2]) \\ &\operatorname{cov}(G_{ij}, G_{ik}) = 2\phi_{ij,ik}\sigma_A^2 + \Delta_{ij,ik}\sigma_D^2 + \lambda_{ij,ik}\sigma_C^2,\ j \neq k \end{aligned} \qquad (3)$$

(4) Where α_H is a baseline mean, E_{ij} is his/her corresponding specified endophenotype. The term ε_{ij} is the residual error term representing the effect of non-family factors. The term G_{ij} is the random effect for the underlying genetic structure. The term $\phi_{ij,ik}$ denotes the kinship coefficient between individual ij and ik: the probability of randomly drawing a single allele in individual ij that is identical by descent (ibd) to a single allele at the same locus randomly drawn from individual ik. The term $\Delta_{ij,ik}$ is the probability that both alleles at a locus are shared ibd by individuals ij and ik. the elements, $\lambda_{ij,ik}$, is

simply binary indicator denoting whether two individuals live together ($\lambda_{ij,ik}$ =1) or apart ($\lambda_{ij,ik}$ =0).

The (broad sense) heritability of P_{ij} , conditional on E_{ij} is

$$h = \frac{\sigma_A^2 + \sigma_D^2}{\sigma_A^2 + \sigma_D^2 + \sigma_C^2 + \sigma_R^2} \tag{4}$$

The significance of rejecting the hypothesis h=0 indicates the fulfillment of (2).

Table 1 details the term $\phi_{ij,ik}$ and $\Delta_{ij,ik}$ values for selected relative pairs and the total genetic variances that these imply [15].

Table 1 Genetic components of variance assuming random mating.

Relationship	ϕ	Δ	Genetic covariance
same person	$1/2$	1	$\sigma_A^2 + \sigma_D^2$
Parent-child	$1/4$	0	$1/2\sigma_A^2$
Full sibling	$1/4$	$1/4$	$1/2\sigma_A^2 + 1/4\sigma_D^2$
Half sibling	$1/8$	0	$1/4\sigma_A^2$
Monozygous twins	$1/2$	1	$\sigma_A^2 + \sigma_D^2$
Grandparent-grandchild	$1/8$	0	$1/4\sigma_A^2$
Uncle/aunt-nephew/niece	$1/8$	0	$1/4\sigma_A^2$
First cousins	$1/16$	0	$1/8\sigma_A^2$
Double first cousins	$1/8$	$1/16$	$1/4\sigma_A^2 + 1/16\sigma_D^2$
Spouses	0	0	0

For a discrete phenotype of ordinal scale, the liability threshold model can be used in the preceding variance component setting [16-17]. The model postulates the existence of an unobserved continuous trait (i.e., liability L_{ij}), and a set of thresholds $t_1, t_2, ..., t_{K-1}$ that partition the liability distribution into intervals corresponding to distinct phenotypic states:

$$P_{ij} = \begin{cases} 1, & if\ L_{ij} < t_1 \\ 2, & if\ t_1 < L_{ij} < t_2 \\ \vdots & \quad\vdots \\ K, & if\ t_{K-1} < L_{ij} \end{cases}$$

The liability L_{ij} is then assumed to follow the same distribution as the P_{ij} in model (3) and heritability can be obtained based on the liability.

Huang et al. [2] have provided the index to evaluate the validation of endophenotypes that is the proportion of heritability explained (PHE) by the endophenotype defined as

$$PHE = 1 - \frac{h}{h_{NE}} \tag{5}$$

Where h_{NE} is the heritability calculated from the variance component analysis (3) without including the endophenotype E_{ij} with any other covariates. A good endophenotype is one that explains a large proportion of heritability, thus, the greater PHE value, the more likely E_{ij} an endophenotype.

2.2.2 Variance of PHE

Hsieh et al. [4] redefined

$$
\begin{aligned}
h_1^{(t)} &= \frac{\sigma_A^2}{\sigma_A^2+\sigma_D^2+\sigma_C^2+\sigma_R^2} \\
h_2^{(t)} &= \frac{\sigma_D^2}{\sigma_A^2+\sigma_D^2+\sigma_C^2+\sigma_R^2} \\
h_3^{(t)} &= \frac{\sigma_D^2}{\sigma_A^2+\sigma_D^2+\sigma_C^2+\sigma_R^2} \\
h_4^{(t)} &= \sigma_A^2+\sigma_D^2+\sigma_C^2+\sigma_R^2
\end{aligned}
$$

Where t is representing the different models.

So the broad-sense heritability $h \equiv h_1^{(1)} + h_2^{(1)}$. Similarly, $h_{NE} \equiv h_1^{(2)} + h_2^{(2)}$.

Hsieh et al. [4] use the delta method [4, 18] to evaluate the variance of PHE. The first-order Taylor approximations give

$$
\begin{aligned}
Var(1-\frac{h}{h_{NE}}) &= Var(\frac{h}{h_{NE}}) \\
&\approx \frac{1}{\mu_{h_{NE}}^2} Var(h) + \frac{\mu_h^2}{\mu_{h_{NE}}^4} Var(h_{NE}) - 2\frac{\mu_h}{\mu_{h_{NE}}^3} Cov(h, h_{NE}) \\
&\approx \frac{1}{\mu_{h_{NE}}^2}\{Var(h_1^{(1)}) + Var(h_2^{(1)}) + 2Cov(h_1^{(1)}, h_2^{(1)})\} \\
&\quad + \frac{\mu_h^2}{\mu_{h_{NE}}^4}\{Var(h_1^{(2)}) + Var(h_2^{(2)}) + 2Cov(h_1^{(2)}, h_2^{(2)})\} \\
&\quad - 2\frac{\mu_h}{\mu_{h_{NE}}^3}\{Cov(h_1^{(1)}, h_1^{(2)}) + Cov(h_1^{(1)}, h_2^{(2)}) \\
&\quad + Cov(h_2^{(1)}, h_1^{(2)}) + Cov(h_2^{(1)}, h_2^{(2)})\}
\end{aligned}
$$

Use $\hat{h}_1^{(1)} + \hat{h}_2^{(1)}$ to estimate μ_h and use $\hat{h}_1^{(2)} + \hat{h}_2^{(2)}$ to estimate $\mu_{h_{NE}}$. The estimators for $\hat{h}_1^{(1)}, \hat{h}_2^{(1)}, \hat{h}_1^{(2)}$, and $\hat{h}_2^{(2)}$ are obtained from the SOLAR computer package.

The remaining terms, such as

$$Var(\hat{h}_1^{(1)}),Var(\hat{h}_2^{(1)}),Var(\hat{h}_1^{(2)}),Var(\hat{h}_2^{(2)}),Cov(\hat{h}_1^{(1)},\hat{h}_2^{(1)}),Cov(\hat{h}_1^{(2)},\hat{h}_2^{(2)}),$$
$$Cov(\hat{h}_1^{(1)},\hat{h}_2^{(2)}),Cov(\hat{h}_1^{(1)},\hat{h}_2^{(2)}),Cov(\hat{h}_2^{(1)},\hat{h}_1^{(2)}),Cov(\hat{h}_2^{(1)},\hat{h}_2^{(2)})$$

will be solved in **2.2.3 [4]**.

2.2.3 The covariance of $\hat{h}_q^{(t)}$ and $\hat{h}_{q^*}^{(t^*)}$

Suppose two models are

$$P_{ij} = x_{ij}^{'(1)}\beta^{(1)} + G_{ij}^{(1)} + \varepsilon_{ij}^{(1)}$$

and

$$P_{ij} = x_{ij}^{'(2)}\beta^{(2)} + G_{ij}^{(2)} + \varepsilon_{ij}^{(2)}$$

Where $\varepsilon_{ij}^{(t)} \sim N(0,(\sigma_R^2)^{(t)}) \equiv N(0,(1-h_1^{(t)}-h_2^{(t)}-h_3^{(t)})h_4^{(t)})$

$G_{ij}^{(t)} \sim N(0,(\sigma_A^2+\sigma_D^2+\sigma_C^2)^{(t)}) \equiv N(0,h_1^{(t)}h_4^{(t)}+h_2^{(t)}h_4^{(t)}+h_3^{(t)}h_4^{(t)})$,

and $Cov(G_{ij},G_{ik})\,[j \neq k] = (2\phi_{ij,ik}\sigma_A^2 + \Delta_{ij,ik}\sigma_D^2 + \lambda_{ij,ik}\sigma_C^2)^t \equiv$

$2\phi_{ij,ik}h_1^{(t)}h_4^{(t)} + \Delta_{ij,ik}h_2^{(t)}h_4^{(t)} + \lambda_{ij,ik}h_3^{(t)}h_4^{(t)}$, and G_{ij} us the random effect for the underlying genetic structure. Assumed I is the total number of family and there are n_i members in the ith family. Let $h^{(t)} = (h_1^{(t)},h_2^{(t)},h_3^{(t)},h_4^{(t)})$, then we have

$$Cov(\hat{h}_q^{(t)},\hat{h}_{q^*}^{(t^*)})$$
$$\approx\left[\sum_{r=1}^{R}\left\{\left(\frac{\partial V_r^{(t)}}{\partial h_q^{(t)}}\right)'\left(W^{-1(t)}\frac{\partial W^{(t)}}{\partial h_q^{(t)}}W^{-1(t)}\right)\left(\hat{S}_r^{(t)}-\hat{V}_r^{(t)}\right)+\left(\frac{\partial V_r^{(t)}}{\partial h_q^{(t)}}\right)'W^{-1(t)}\left(\frac{\partial V_r^{(t)}}{\partial h_q^{(t)}}\right)\right\}\right]^{-1}$$
$$\times\left[\sum_{r=1}^{R}\left\{\left(\frac{\partial V_r^{(t)}}{\partial h_q^{(t)}}\right)'W^{-1(t)}\left(\hat{S}_r^{(t)}-\hat{V}_r^{(t)}\right)\left(\hat{S}_r^{(t)}-\hat{V}_r^{(t)}\right)'W^{-1(t^*)}\left(\frac{\partial V_r^{(t^*)}}{\partial h_q^{(t^*)}}\right)'\right\}\right]$$

$$\times\left[\sum_{r=1}^{R}\left\{\left(\frac{\partial V_r^{(t^*)}}{\partial h_{q^*}^{(t^*)}}\right)'\left(W^{-1(t^*)}\frac{\partial W^{(t^*)}}{\partial h_{q^*}^{(t^*)}}W^{-1(t^*)}\right)\left(\hat{S}_r^{(t^*)}-\hat{V}_r^{(t^*)}\right)+\left(\frac{\partial V_r^{(t^*)}}{\partial h_{q^*}^{(t^*)}}\right)'W^{-1(t^*)}\left(\frac{\partial V_r^{(t^*)}}{\partial h_{q^*}^{(t^*)}}\right)\right\}\right]^{-1}$$

$q=1,2,3,4 \quad q^*=1,2,3,4 \quad t=1,2 \quad t^*=1,2$

Where

$S_r^{(t)}=(r_{r1}^{(t)}r_{r1}^{(t)},r_{r1}^{(t)}r_{r2}^{(t)},\ldots,r_{r1}^{(t)}r_{rn_r}^{(t)},\ldots,r_{rn_r}^{(t)}r_{rn_r}^{(t)})'$,

$r_{rj}^{(t)}=P_{rj}-x_{rj}^{(t)}\beta^{(t)}$,

$V_r^{(t)}=E(S_r^{(t)};\beta^{(t)},h^{(t)})$ as given by Covariance after transformation in table 1,

$$W_{r\times r}^{(t)}=\begin{cases}2\sigma_{ij}^{(t)2} & \textit{for the i, jth pairs}\\ \sigma_{il}^{(t)}\sigma_{jm}^{(t)}+\sigma_{im}^{(t)}\sigma_{jl}^{(t)} & \textit{for the i, jth and l, mth pairs}\end{cases}$$

and $$\frac{\partial W^{(t)}}{\partial h^{(t)}}=\begin{cases}4\sigma_{ij}\dfrac{\partial\sigma_{ij}}{\partial h} & \textit{for the i, jth pairs}\\ \dfrac{\partial\sigma_{il}}{\partial h}\sigma_{jm}+\sigma_{il}\dfrac{\partial\sigma_{jm}}{\partial h}+\dfrac{\partial\sigma_{im}}{\partial h}\sigma_{jl}+\sigma_{im}\dfrac{\partial\sigma_{jl}}{\partial h} & \textit{for the i, jth and l, mth pairs}\end{cases}$$

The table 2 shows the interested derivative of covariance components, related $\hat{h}_1$ and $\hat{h}_2$ for relative pairs.

Table 2 The derivative of covariance components for relative pairs.

Relationship	$\frac{\partial V}{\partial h_1}$	$\frac{\partial V}{\partial h_2}$	$\frac{\partial \tilde{V}}{\partial h_1}$	$\frac{\partial \tilde{V}}{\partial h_2}$
Same person	0	0	0	0
Parent-child	$1/2h_4$	0	$1/2$	0
Full sibling	$1/2h_4$	$1/4h_4$	$1/2$	$1/4$
Half sibling	$1/4h_4$	0	$1/4$	0
Monozygous twins	h_4	h_4	1	1
Grandparent-grandchild	$1/4h_4$	0	$1/4$	0
Uncle/aunt-nephew/niece	$1/4h_4$	0	$1/4$	0

First cousins	$1/8h_4$	0	$1/8$	0
Double first cousins	$1/4h_4$	$1/16h_4$	$1/4$	$1/16$
Spoused	0	0	0	0

Based on Table 2, Heish et al. [4] express the result of the above-mentioned as follow:

$$
\begin{aligned}
&Cov(\hat{h}_q^{(t)}, \hat{h}_{q^*}^{(t^*)}) \\
&\approx \left[\sum_{r=1}^{R} \left\{ \hat{h}_4^{(t)} \left(\frac{\partial \tilde{V}_r^{(t)}}{\partial h_q^{(t)}} \right)' \left(W^{-1(t)} \frac{\partial W^{(t)}}{\partial h_q^{(t)}} W^{-1(t)} \right) \left(\hat{S}_r^{(t)} - \hat{V}_r^{(t)} \right) \right. \right. \\
&\left. \left. + \hat{h}_4^{(t)} \left(\frac{\partial \tilde{V}_r^{(t)}}{\partial h_q^{(t)}} \right)' W^{-1(t)} \left(\frac{\partial \tilde{V}_r^{(t)}}{\partial h_q^{(t)}} \right) \hat{h}_4^{(t)} \right\} \right]^{-1} \\
&\times \left[\sum_{r=1}^{R} \left\{ \hat{h}_4^{(t)} \left(\frac{\partial \tilde{V}_r^{(t)}}{\partial h_q^{(t)}} \right) W^{-1(t)} \left(\hat{S}_r^{(t)} - \hat{V}_r^{(t)} \right) \left(\hat{S}_r^{(t^*)} - \hat{V}_r^{(t^*)} \right)' W^{-1(t^*)} \left(\frac{\partial \tilde{V}_r^{(t^*)}}{\partial h_{q^*}^{(t^*)}} \right) \hat{h}_4^{(t^*)} \right\} \right] \\
&\times \left[\sum_{r=1}^{R} \left\{ \hat{h}_4^{(t^*)} \left(\frac{\partial \tilde{V}_r^{(t^*)}}{\partial h_q^{(t^*)}} \right)' \left(W^{-1(t^*)} \frac{\partial W^{(t^*)}}{\partial h_q^{(t^*)}} W^{-1(t^*)} \right) \left(\hat{S}_r^{(t^*)} - \hat{V}_r^{(t^*)} \right) \right. \right. \\
&\left. \left. + \hat{h}_4^{(t^*)} \left(\frac{\partial \tilde{V}_r^{(t^*)}}{\partial h_q^{(t^*)}} \right)' W^{-1(t^*)} \left(\frac{\partial \tilde{V}_r^{(t^*)}}{\partial h_q^{(t^*)}} \right) \hat{h}_4^{(t^*)} \right\} \right]^{-1} \\
&q = 1,2,3,4 \quad q^* = 1,2,3,4 \quad t = 1,2 \quad t^* = 1,2
\end{aligned}
$$

2.2.4 Hypothesis test

For having more statistical meaning of PHE, we utilize the confidence interval to get more information about PHE. We hope to find a value that it means that there exist a useful endophenotype when PHE value is larger than the value. That is, do one-sided confidence interval.

The hypothesis is

$$\begin{cases} H_0 : PHE = a \\ H_1 : PHE > a \end{cases}$$

Under null hypothesis and significance level of α, we reject H_0 if the lower bound of one-sided confidence interval of PHE, $\widehat{PHE} - Z_{1-\alpha} \times s.e.(\widehat{PHE})$, is larger than a.

2.3 The whole-genome association test for quantitative outcomes

For each of the genotyped SNP markers, we are interested in testing whether observed genotypes and quantitative phenotypes are associated. We let G_{ijm} denote the observed genotype at maker m for individual j in family i.

We label two allele “A” and “a” and define a genotype score, g_{ijm}

$$g_{ijm} = \begin{cases} 0 & \text{if } G_{ijm} = a/a \\ 1 & \text{if } G_{ijm} = A/a \\ 2 & \text{if } G_{ijm} = A/A \end{cases}$$

First we built the model for each SNP

$$E(P_{ij}) = \mu + \beta_g g_{ij} + \beta_x x_{ij}$$

Here μ is the population mean, β_g is the additive effect for each SNP, and β_x is a vector of covariate effects [19].

Chen et al. [19] extend the model with

$$E(P_{ij}) = \mu + \beta_g \overline{g}_{ij} + \beta_x x_{ij}$$

where $\bar{g}_{ij}$ is the expected genotype score and define as

$$\begin{aligned}\bar{g}_{ijm} &= E(g_{ijm} \mid G_i, \theta, F) \\ &= 2P(G_{ijm} = A/A \mid G_i, \theta, F) + P(G_{ijm} = A/a \mid G_i, \theta, F)\end{aligned}$$

where θ is a vector of intermarker recombination fractions and F is a vector of allele frequencies for each marker. To allow for correlation between different observed phenotypes within each family, we define the variance-covariance matrix Ω_i for family i as

$$\Omega_{ijk} = \begin{cases} \sigma_A^2 + \sigma_D^2 + \sigma_C^2 & if \;\; j = k \\ 2\phi_{ij,ik}\sigma_A^2 + \Delta_{ij,ik}\sigma_D^2 + \lambda_{ij,ik}\sigma_C^2 & if \;\; j \neq k \end{cases}$$

Chen et al. [19] provide an approach is to first fit a simple variance-components model to the data (with parameters $\mu, \beta_x, \sigma_A^2, \sigma_R^2$ but without parameters β_g, σ_D^2). This model provides a vector of fitted values for each family, which they denote $E(P_i)^{(base)}$, and an estimate of the variance-covariance matrix for each family, which they denote $\Omega_i^{(base)}$. Using these two quantities, we define the score statistic

$$T^{SCORE} = \frac{\left\{\sum_i [\bar{g}_i - E(\bar{g}_i)]' \left[\Omega_i^{(base)}\right]^{-1} \left[y_i - E(y_i)^{(base)}\right]\right\}^2}{\sum_i [\bar{g}_i - E(\bar{g}_i)]' \left[\Omega_i^{(base)}\right]^{-1} [\bar{g}_i - E(\bar{g}_i)]}$$

Where $\bar{g}_i$ is a vector with expected genotype scores for each individual in the ith family, calculated conditional on the available marker data, and $E(\bar{g}_i)$ is a vector with identical elements that give the unconditional expectation of each genotype score. T^{SCORE} is approximately distributed as χ^2 with 1 df. As usual, LOD scores were defined as

$$LOD \text{ score} \equiv \chi^2 / 2\ln(10)$$

2.4 The preprocessing method for gene expression levels

We interpret the three main steps of data preprocessing.

2.4.1 Background adjustment

Because partial measured probe intensities maybe caused by non-specific hybridization or the noise in the optical detection system, background adjustment is essential to rid of these intensities not exactly expressed from genes. Observed probe intensities need to be adjusted to give the accurate expression levels of specific hybridization [20]. Some methods make use of MM probes to adjust, but some are not.

2.4.2 Normalization

During the process of carrying out the microarray experiment involving multiple arrays, there are many obscuring sources of variation involved, such as physical problems with the arrays, laboratory conditions, hybridization reactions, labeling, and scanner difference. In order to compare measurements from different arrays, implying different tissue, some proper normalization is necessary.

2.4.3 Summarization

Due to Affymetrix platform designing multiple probes to represent a gene, summarization is needed to combine these probe intensities to a single value. For each gene, the background adjusted and normalized intensities are used to be summarized into one measurement that estimates the expression level.

2.4.4 RMA

$i = 1, ..., I : \textit{representing the different array (sample)}$

$j = 1, ..., J : \textit{representing the probe pair in the gene}$

$g = 1, ..., G : \textit{representing the probe set (gene)}$

RMA [21- 22], Robust Multi-array Analysis, is an expression measure consisting of three particular preprocessing steps: convolution background correction, quantile normalization, and a summarization based on a multi-array model fit robustly using the median polish algorithm. These RMA authors proposed a procedure ignoring the MM intensities and using only the PM intensities.

The RMA convolution background correction method is motivated by looking at the distribution of probe intensities. The model observed PM as the sum of a background intensity bg_{ijg} caused by optical and nonspecific binding, and signal intensity s_{ijg} .

$$PM_{ijg} = bg_{ijg} + s_{ijg} \ , i = 1,\ldots,I \ , \ j = 1,\ldots,J \ , \ g = 1,\ldots,G$$

Under the model above, the background corrected probe intensities will be given by $B(PM_{ijg})$, where $B(PM_{ijg}) \equiv E(s_{ijg} \mid PM_{ijg})$. To obtain a computationally feasible $B(\cdot)$ we consider the closed-form transformation obtained when assuming that s_{ijg} is distributed exponential and bg_{ijg} is distributed normal, and the results obtained using $B(\cdot)$ work well in practice [21-22].

Next, perform the quantile normalization, which is to make the distribution of probe intensities for each array the same [22, 23]. In order to summarize the probe intensities, RMA introduced a log scale linear additive model. The model is:

$$T(PM_{ij}) = e_i + a_j + \varepsilon_{ij} \ ,$$

where PM_{ijg} represents the PM intensity of array $i = 1,\ldots,I$ and probe pair $j = 1,\ldots,J$, for any given probe set g. $T(\cdot)$ represents the transformation that background corrects, normalizes, and logs the PM intensities, e_i represents the log2 scale expression value found on arrays i, a_j represents the log scale affinity effects for probes j , and ε_{ij} represents error [22, 24]. To protect against outlier probes, they use a robust procedure, such as median polish, to estimate model parameters [21-22].

The estimate of e_i as the log scale measure of expression refers to as robust multi-array average (RMA).

2.5 The differential expression methods

2.5.1 Fold-change

Fold-change analysis is used to identify genes with expression ratios or differences between a treatment and a control that are outside of a given cutoff or threshold. Intensity values may be compared using ratio, $\log_2(ratios)$, or difference. Biologist favors fold-change equal to 2 as the threshold of differential expression.

2.5.2 Two sample t-test

The simplest statistic method for comparing means between two groups is two sample t-test. The variances of the two samples may be assumed to be equal or unequal. The approach of unequal variance assumption is also called Welch's t-test. For any given gene g, suppose that the number of samples in sample 1 and in sample 2 are M and N respectively. Here we describe the two tests briefly.

Two sample t-test for equal variance:

$$sample\ 1: X_{g1},\ldots., X_{gM} \sim N(\mu_1,\sigma^2)$$

$$sample\ 2: Y_{g1},\ldots.., Y_{gN} \sim N(\mu_2,\sigma^2)$$

$$\mathrm{H}_0 : \mu_1 = \mu_2 \qquad versus \qquad H_1 : \mu_1 \neq \mu_2$$

$$test\ statistic: \frac{\bar{X}-\bar{Y}}{\sqrt{\frac{1}{M}+\frac{1}{N}}\, S_p} \sim T_{M+N-2},$$

$$where\ S_p^{\ 2} = \frac{\sum_{i=1}^{M}(X_i-\bar{X})^2 + \sum_{i=1}^{N}(Y_i-\bar{Y})^2}{M+N-2}.$$

Two sample t-test for unequal variance (Welch’s t-test):

$$sample\ \ 1: X_{g1},\ldots., X_{gM} \sim N(\mu_1, \sigma_1{}^2)$$

$$sample\ \ 2: Y_{g1},\ldots.., Y_{gN} \sim N(\mu_2, \sigma_2{}^2)$$

$$\mathrm{H}_0\text{:}\mu_1 = \mu_2 \qquad versus \qquad H_1 : \mu_1 \neq \mu_2$$

$$test\ statistic: \frac{\bar{X} - \bar{Y}}{\sqrt{(\frac{S_X^2}{M} + \frac{S_Y^2}{N})}} \sim T_v\ (approximately),$$

$$where\ S_X^2 = \frac{1}{M-1}\sum_{i=1}^{M}(X_i - \bar{X})^2 \ \ , \ \ S_Y^2 = \frac{1}{N-1}\sum_{i=1}^{N}(Y_i - \bar{Y})^2\ and$$

$$v = \frac{(\frac{S_X^2}{M} + \frac{S_Y^2}{N})^2}{\frac{S_X^4}{M^2(M-1)} + \frac{S_Y^4}{N^2(N-1)}}.$$

After performing the test and the conclusion leads to reject H_0, we consider that this gene is a differentially expressed gene.

2.5.3 SAM (Significance Analysis of Microarrays)

It was proposed by Tusher, Tibshirani and Chu (2001) [22, 25]. The method is based on a modified version of the standard t-statistic to adjust the high variance probably caused by a low expression level.

The “relative difference” d_g, $g = 1, \cdots, p$ genes:

$$d_g = \frac{r_g}{s_g + s_0}$$

Here r_g is a score, s_g is a standard deviation, and s_0 is an exchangeability factor in the denominator to ensure that the variance of d_g is independent of gene expression level. In two-sample t-test for equal variance,

$$r_g = \bar{x}_{g2} - \bar{x}_{g1}$$

$$s_g = \sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right) \times \frac{\sum_{i \in group1} (x_{gi} - \bar{x}_{g1})^2 + \sum_{i \in group2} (x_{gi} - \bar{x}_{g2})^2}{n_1 + n_2 - 2}},$$

where $\bar{x}_{g1}$ and $\bar{x}_{g2}$ are defined as the average levels of expression for gene g in group 1 and group 2, and x_{gi} is defined as the expression level for gene g and sample i. Group 1 and 2 have n_1 and n_2 genes, respectively. Then rank all genes by the observed relative difference d_g and denote the new arrangements as $d_{(g)}$. B sets of permutations of the samples are taken to obtain the expected relative difference $\bar{d}_{(g)}^{\,*}$ by a similar way (For more details, see [22, 25-26]). A scatter plot of $d_{(g)}$ vs. $\bar{d}_{(g)}^{\,*}$ is used and the genes apart from the $d_{(g)} = \bar{d}_{(g)}^{*}$ line by a distance greater than the threshold Δ are regarded as differentially expressed genes.

2.6 Multiple testing procedures

If thousands of hypotheses are tested simultaneously, the probability of false positives by chance increases. We use an example to understand the question: when a two-sample t-test is performed on a gene, the probability by which the gene's expression level will be considered significantly different between two groups of samples is expressed by the p-value. The p-value is the probability that a gene's expression levels are different between the two groups due to chance. A p-value of 0.05 signifies a 5% probability that the gene's mean expression value in one condition is different than the mean in the other condition by chance alone. If 10,000 genes are tested, 5% or 500 genes might be called significant by chance alone.

Table 3 probability of calling 1 or more false positives by chance.

Number of genes tested (N)	False positives incidence	Probability of calling 1 or more false positives by chance ($100(1-0.95^N)$)
1	1/20	5%
2	1/10	10%
20	1	64%
100	5	99.40%

In microarray data analysis, false positives are genes that are found to be statistically different between conditions, but are not in reality. We need to adjust p-values derived from multiple statistical tests to correct for occurrence of false positives.

2.6.1 Type I error rates

The null hypotheses in M tests, $H_0(1),\ldots,H_0(M)$, we conduct 2×2 table to interpret the different number of tests in different conditions.

Table 4 Type I and type II errors in multiple hypothesis testing.

	Number of not rejecting H_0	Number of rejecting H_0	
Number of true Non-differential genes (H_0 is true)	U	V (false positive) (type I error)	M_0
Number of true Differential genes (H_0 is not true)	T (false negative) (type II error)	S	M_1
	$M-R$	R	M

Type I error rates is defined as a parameters, $\theta = \theta(F_{V,R})$, of the joint distribution $F_{V,R}$ of the numbers of Type I errors V and the number of rejecting hypotheses R. Such a general representation covers the following commonly-used Type I error rates [20].

(1) Generalized family-wise error rate (gFWER), or probability of at least (k+1) Type I errors.

$$gFWER(k) \equiv \Pr(V > k).$$

When k=0, the gFWER is the usual family-wise error rate (FWER), or probability of at least one Type I error. $FWER(k) \equiv \Pr(V > 0).$

(2) Tail probabilities for the pro portion of false positives (TPPFP) among the rejected hypotheses,

$$TPPFP(q) \equiv \Pr(V / R > q), \quad q \in (0,1).$$

(3) False discovery rate (FDR), or expected value of the proportion of false positives among the rejected hypotheses (Benjamini and Hochberg, 1995),

$$FDR \equiv E\left[V / R\right].$$

The convention that $V / R \equiv 0$ if $R = 0$ is used.

2.6.2 Adjusted p-values

Given M null hypotheses being tested $H_0(1),\ldots,H_0(M)$, the adjusted p-value $\tilde{P}_{0a}(m)$, for null hypothesis $H_0(m)$, is defined as the smallest Type I error level α at which one would reject $H_0(m)$, that is,

$$\tilde{P}_{0a}(m) \equiv \inf\left\{\alpha \in \left[0,1\right] : T(m) \in C(m)\right\}, \quad m = 1,\ldots,M$$

Where $T(m)$ is the test statistic and $C(m)$ is the rejection region for test m .It need to have null distribution of $T(m)$ [22].

The smaller adjusted p-value, the evidence against the corresponding null hypothesis is stronger. The difference between adjusted p-value and unadjusted p-values that the unadjusted p-value, $p_0(m)$, for null hypothesis $H_0(m)$ is the smallest type I error rate of the single hypothesis testing procedure at which one would reject $H_0(m)$.

2.6.3 The q-value

Storey et al. [27] define a new false discovery rate, *pFDR*

$$pFDR = E(\frac{V}{R} \mid R > 0) = \Pr(H_0 \text{ is true} \mid T \in C)$$

Where T is the test statistic and C is the rejection region. The term ’positive’ has been added to reflect the fact that we are conditioning on the event that positive findings have occurred.

As a natural extension to $pFDR$, the q-value has the following definition [27].

Definition 1 For an observed statistic $T = t$, the q-value of t is defined to be

$$q(t) = \inf_{\{C: t \in C\}} \{pFDR(C)\}$$

Definition 2 For a set of hypothesis tests conducted with independent p-values, the q-value of the observed p-value p is

$$q(p) = \inf_{\gamma \geq p} \{pFDR(\gamma)\}$$

Where the nested set of rejection regions take the form $[0, \gamma]$.

The q-value was discussed, which is the pFDR analogue of the p-value. Whereas it can be inconvenient to have to fix the rejection region or the error rate beforehand, the q-value requires us to do neither.

2.7 Datasets

The dataset we used is from the Gene Expression Omnibus (GEO) in NCBI. The Gene Expression Omnibus (GEO) is a public repository that archives and freely distributes microarray and other forms of high-throughput data submitted by the scientific community. In addition to data storage, a collection of web-based interfaces and applications are available to help users query and download the experiments and gene expression patterns stored in GEO.

In 2007, Moffatt et al. [5] brought up this paper about asthma and offered expression file free to download. Moffatt et al. [5] mentioned that the study subjects were recruited from family (MRC-A) and case-control panels (MAGICS and UK-C). The family panel included a 207 predominantly (99 %) nuclear families (MRC-A). these were recruited through a proband with severe (step 3) childhood onset asthma and contained 295 sib pairs, 11 half-sib pairs and 3 singletons (counting all possible sibs). The study included siblings regardless of asthma status. Lymphoblastoid cell lines (LCLs) were derived from peripheral blood Iymphocytes on probands and siblings. Cells were harvested at log phase from roller cultures in the first growth after transformation. Global gene expression in Epstein-Barr virus Iymphoblastoid cell lines (EBVL) was measured with the Affymetrix HG-U133 Plus 2.0 chip in family panel. The other one, in case-control panel, included 437 non-asthmatic Caucasian UK controls (UK-C) children, 728 asthmatic children of German in the Multicentre Asthmatic Genetics in Childhood Study (MAGICS) study with physician-diagnosed asthma for comparison with 694 reference children recruited in the cross sectional International Study of Asthma and Allergies in Childhood (ISAAC) study. The study genotyped all children in the primary association study with the Illumina Sentrix HumanHap300 BeadChip. Additional, the study typed the parents and children in the MRC-A panel with the Illumina Sentrix Human-I Genotyping BeadChip.

This paper had an own database, “mRNA by SNP Browser”, that provides graphical overviews of whole-genome association studies of datasets with very rich phenotypic information, such as global surveys of gene expression. The software incorporates a generic eQTL database and provides a graphic interface for browsing association between 54,675 transcript levels and 406,912 SNPs. For each transcript, the browser can tabulate and plot association test statistics (p-value<0.001), estimates of effect size and allele information across the genome. The browser automatically links results to the UCSC genome browser where users can examine each transcript in its genomic context. In addition to browsing the results by transcript or by position, results can be searched for information on specific SNPs. LD and tag information are provided for SNPs not in the database [6].

3 Materials and Methods

DNA transcribes to mRNA, mRNA translates to protein which may affect the disease. Because mRNA is closer to DNA and gene expression value is measured in mRNA, we use global gene expression of probe sets as endophenotypes in search for the susceptibility genes underlying asthma. Gene expression values are preprocessed by the robust multi-array averaging (RMA) to adjust for background noises, normalize expression levels and summarize multiple probs. We judge which gene expressions of probe sets are endophenotype by using the index PHE and do hypothesis test of PHE=0 : Endophenotype and phenotype share no genes. For the problem of multiple testing, we utilize the q-value to control for the false discovery rate (FDR) for significance judgment. We also perform genome-wide association tests for each gene expression of probe set.

We then derive SNPs under four conditions; they are SNPs (1) significantly associated with gene expressions of all probes sets, (2) significantly associated with gene expressions of probe sets with PHEs greater than zero, (3) significantly associated with gene expressions of probe sets with PHEs significantly different from zero at unadjusted p-values smaller than 0.05, and (4) significantly associated with gene expressions of probe sets with PHEs significantly different from zero at q-values smaller than 0.05.

Through systematic literature reviews (Hoffjan et al. [28], Ober et al. [29] and Zhang et al. [30]), we identify 144 genes that have been reported to be associated with asthma or atopy phenotypes. Among them, there are 25 genes that have been repeatedly reported in six or more populations [29]. We further identify additional 125 genes that are related to these 25 genes. We then identify the overlap between these 269 (=144+125) genes and genes identified by the above gene-expression-based analysis.

Finally, we are interests in realizing the genetic properties of gene expressions that are identified as the endophenotypes. We use various plots to aid us observing some phenomenon.

3.1 Datasets

In 2007, Moffatt et al. [5] brought up this paper about asthma. The subjects of this study can be divided into two master parts: the case-control panel and the family panel. We are interested in the family panel. There are 404 children from the family panel. The family panel included 207 predominantly (99 %) nuclear families. These were recruited through a proband with severe childhood onset asthma. The study included siblings regardless of asthma status. Global gene expression in Epstein-Barr virus Iymphoblastoid cell lines (EBVL) was measured with the Affymetrix HG-U133 Plus 2.0 chip in the family panel. The study genotyped 830 offspring and parents with the Illumina Sentrix HumanHap-I BeadChip and 378 offspring with the Illumina Sentrix HumanHap300 BeadChip in family panel.

We downloaded the files of Gene expression levels in 404 children from the family panel. This dataset is available at http://www.ncbi.nlm.nih.gov/ego/, the GEO accession: GSE8052 [6].

3.2 Statistical method of endophenotype

We use global gene expressions of probe sets as endophenotypes. Epstein-Barr virus Iymphoblastoid cell lines (EBVL) were derived from participants and global gene expression was measured with Affymetrix HG-U133 Plus 2.0 chip. One of the indices determining the gene expression of probe set an endophenotype is the proportion of heritability explained (PHE) by the endophenotype [2].

Given a phenotype of case-control status, the variance component analysis for discrete trait [2, 14]:

$$
\begin{aligned}
&P_{ij} = \alpha_H + \gamma_H E_{ij} + \tau_H Z_{ij} + G_{ij} + \varepsilon_{ij},\\
&\varepsilon_{ij} \sim Normal(0, \sigma_R^2)\\
&G_{ij} \sim Normal(0, [\sigma_A^2 + \sigma_D^2 + \sigma_C^2])\\
&\mathrm{cov}(G_{ij}, G_{ik}) = 2\phi_{ij,ik}\sigma_A^2 + \Delta_{ij,ik}\sigma_D^2 + \lambda_{ij,ik}\sigma_C^2 \text{ , } j \neq k
\end{aligned}
$$

Where P_{ij} is the observed phenotype in the jth member of the ith family, E_{ij} is his/her corresponding specified endophenotype, Z_{ij} is his/her other covariates. ε_{ij} is the residual error term representing the effect of non-family factors. G_{ij} is the random effect for the underlying genetic structure. The components σ_A^2 , σ_D^2 and σ_C^2 represent the variance arising from polygenic additive effects, polygenic dominance effects and shared environmental effects, respectively.

The (broad sense) heritability of P_{ij} , conditional on E_{ij} is

$$h = \frac{\sigma_A^2 + \sigma_D^2}{\sigma_A^2 + \sigma_D^2 + \sigma_C^2 + \sigma_R^2}$$

Then we calculate PHEs of all probe sets.

$$PHE = 1 - \frac{h}{h_{NE}}$$

Where the term h is the heritability calculated from the variance component analysis with disease status as response variable and covariates including the endophenotype and other covariates. The term h_{NE} is the heritability calculated from the variance component analysis with disease status as response variable and covariates without endophenotype [2]. The greater of the PHE value, the more likely he intermediate variable an endophenotype. Hsieh et al. [4] utilize the delta method to evaluate the variance of PHE. For having more statistical meanings of PHE, We do hypothesis test of PHE to get more information about PHE.

The hypothesis is

$$\begin{cases} H_0 : PHE = 0 \\ H_1 : PHE > 0 \end{cases}$$

Under $H_0 : PHE = 0$, we suppose the phenotype and endophenotype share no genes and significance level of α =0.05. We reject H_0 if the lower bound of one-sided confidence interval of PHE, $\widehat{PHE} - Z_{1-\alpha} \times s.e.(\widehat{PHE})$, is larger than 0 and plot the confidence of interval of PHEs over all chromosomes.

If multiple hypotheses are tested simultaneously, the probability of false positives by chance increases. We utilize the q-value which is a natural pFDR analogue of the p-value to correct for occurrence of false positives. These q-values were calculated by applying the QVALUE (http://faculty.washington.edu/~jstorey/qvalue/) package.

Where h and h_{NE} were obtained from the results by performing variance component analysis using the SOLAR computer package [14], so $\widehat{PHE}$ is estimated. The estimator of $s.e.(\widehat{PHE})$ used delta method [4].

Then we classify PHEs into four conditions, there are total PHEs, PHEs higher than zero, significant PHEs with unadjusted p-value smaller than the significant level of 5%, and significant PHEs with q-value smaller than the significant level of 5%.

3.3 The preprocessing and differential method

All analyses of expression value used quantile normalization, after performing robust multi-array averaging (RMA), to remove non-biological variation, enforce normality, deal with outlier and summary the intensity values [21-22, 24].

We want to know if a disease may be caused by large expression of particular genes resulting in variation between diseased and normal tissues. The differential method used to detect the genes expressed differentially between case-control samples is significance analysis of Microarrays (SAM) [22, 25-26]. The term q-value in SAM output smaller than 0.05, it represents these probe sets may be different between diseased and normal subset.

3.4 Genome-wide association tests for gene expressions

A database, “mRNA by SNP Browser”, that provides the results of Genome-wide association test for each gene expression of probe set [31]. Association analysis was applied with Merlin (FASTASSOC option), after probabilistically inferring missing genotypes. This database is downloaded at http://www.sph.umich.edu/csg/liang/asthma/.

We find significant SNPs from results of associate test between expression value of each probe set and all SNPs genotypes with LOD score$\geq$6 (about equivalent to the false discovery rate (FDR) = 0.056) [6]. We define these significant SNPs as “eSNPs”.

3.5 Asthma genes and overlap rate

Asthma is a disease of chronic airway inflammation that affects nearly 155 million individuals worldwide. Like other atopic diseases, asthma is a complex disorder caused by interactions between multiple genes of small to modest effect and equally important environmental factors. Asthma has an important genetic component but no clear pattern of inheritance and heritability estimates of asthma vary between 36-79%.

To this day, there are over 100 genes that have been reported to be associated with asthma or related phenotypes. We look for three review papers: (1) “Association studies for asthma and atopic diseases: a comprehensive review of the literature” [28], there are 64 genes in 251 papers about asthma or atopy from 1982 to 2002, (2) “Asthma genetics 2006: the long and winding road to gene discovery” [29], there are

118 genes in nearly 500 papers about asthma or atopy from 1982 to 2005, and (3) “Recent advances in asthma genetics” [30], there are 34 genes in 87 papers about asthma or atopy from 2005 to 2007. These review papers identified the genes by searching the public database using the keywords “association” or “case-control” together with each of the following terms: “asthma”, “bronchial hyperresponsiveness”, “BHR”, “atopy”, ”SPT”, “atopic dermatitis”, “IgE”, or “drug response”. This identified a group of genes with at least one significant association reported and then they searched for all other studies of those genes. We collect and make up the 144 genes that have been associated with asthma or atopy phenotypes in at least one study, although many of these studies are methodologically limited and need replication (Appendix I).

Ober’s paper in 2006 also identified 25 genes that have been associated with asthma or atopy phenotype in six or more populations. We use these 25 genes as the basis to identify more related genes. The PubGene (http://www.pubgene.org/) was used for identifying related genes. The PubGene not only catalogues individual genes but gene pairs. It uses co-citation to create networks of gene identifiers, allowing the possibility for the discovery of relationships between two genes via an intermediary gene. Co-citation suggests biological relationship between the implicated genes. There are 125 genes that are related with these 25 genes (Appendix II).

We calculate overlap rate by comparing significant SNPs (LOD>6) from probe sets with significant PHEs with 269(=144+125) genes mentioned in the front statement. The significant PHEs can be either with unadjusted p-value smaller than 0.05, or with q-value smaller than 0.05.

3.6 Plots

We use plot to assess genetic characteristics of gene expressions that are identified as the endophenotypes.

3.6.1 Density plot of PHEs

We plot density plots and observe the variations of PHEs' distributions under four conditions of PHEs, there are total PHEs, PHEs >0, PHEs with unadjusted p-value < 0.05 and PHEs with q-value < 0.05.

3.6.2 The scatter plot of heritability of gene expressions versus PHEs

We calculate the heritability of gene expression in each probe set. Then we plot the scatter plot under four conditions: (1) total heritability of gene expressions for all probe sets versus total PHEs, (2) heritability of gene expressions for probe sets with PHEs >0 versus PHEs >0, (3) heritability of gene expressions for probe sets with significant PHEs with unadjusted p-values < 0.05 versus significant PHEs with unadjusted p-values < 0.05, (4) heritability of gene expressions for probe sets with significant PHEs with q-values < 0.05 versus significant PHEs with q-values < 0.05. Under four conditions, we observe the relationship between heritability of gene expressions and PHEs.

3.6.3 The bar-plot of proportions of probe sets with max significant SNPs' LOD >6 versus PHEs

We derive four conditions of significant SNP's LOD >6, there are the results of genome-wide association tests for (1) gene expressions for all probe sets, (2) probe sets with PHEs >0, (3) gene expressions for probe sets with PHEs with unadjusted p-values < 0.05, (4) gene expressions for probe sets with PHEs with q-value < 0.05.

Under four conditions, we plot the bar-plot of proportions of probe sets with max significant SNP's LOD >6 versus PHEs. The purpose of these plot is to observe variable numbers of underlying genes by observing proportions of probe sets with max significant SNPs’ LOD >6.

The steps of plot:

Step1. No matter which conditions, we divide the PHEs into 5 groups according to 5 quantiles.

Step2. To evaluate the number of probe sets with eSNPs in each quantile of PHEs.

Step3. Each quantile of PHEs, to evaluate the proportion

$$=\frac{\#\text{ of probe sets with significant SNPs (max LOD >6)}}{\#\text{ of probe sets}}.$$

Step4. The bar-plot: Proportions versus quantiles of PHEs.

3.6.4 The bar-plot of number of cis eSNPs <100 kb or cis eSNPs >100 kb or trans versus PHEs

A cis-regulatory element or cis-element is a region of DNA or RNA that regulates the expression of genes located on that same strand. These cis-regulatory elements are often binding sites of one or more trans-acting factors. In contrast, trans-regulatory elements are species which may modify the expression of genes distant from the gene that was originally transcribed to create them. To demonstrate the concept (this is not a specific example), a transcription factor which regulates a gene on chromosome 6 might itself have been transcribed from a gene on chromosome 11.

To summarize, cis-elements are present on the same strand as the gene they regulate whereas trans-elements can regulate genes distant from the gene from which they were transcribed (http://en.wikipedia.org/wiki/Cis-regulatory_element).

So we defined the cis eSNPs is that the strongest cis effect for a given expression values of probe sets was then mapped by testing SNPs located at the location of this probe set in the same chromosome. If cis eSNPs located within 100 kb window centered at the location of this probe set in the same chromosome, we define them as “cis eSNPs <100 kb”. If cis eSNPs located outside 100 kb window centered at the location of this probe set in the same chromosome, we define them as “cis eSNPs >100 kb”. Trans is that eSNPs are in other chromosomes and not the same chromosomes of this probe set.

We derive four conditions of significant SNP’s LOD >6 (eSNPs), there are the results of genome-wide association tests for (1) gene expressions for all probe sets, (2) probe sets with PHEs >0, (3) for gene expressions for probe sets with PHEs with unadjusted p-values < 0.05, (4) for gene expressions for probe sets with PHEs with q-value < 0.05.

Under four conditions**,** we plot

a. The bar-plot of number of cis eSNPs <100 kb versus PHEs.

b. The bar-plot of number of cis eSNPs >100 kb versus PHEs.

c. The bar-plot of number of trans versus PHEs.

The purpose of these plots is to observe the various numbers of cis eSNPs < 100 kb, cis eSNPs > 100 kb or trans and infer the various number of underlying genes.

The steps of plot:

Step1. No matter which conditions, we divide PHEs into 5 groups according to 5 quantiles.

Step2. To evaluate the number of cis eSNPs <100 kb, cis eSNPs >100 kb and trans separately in each quantile.

Step3. The bar-plots: Number (cis eSNPs <100kb, cis eSNPs >100kb and trans) versus quantiles of PHEs.

3.6.5 The bar-plot of number of cis eSNPs <100 kb or cis eSNPs >100 kb or trans versus heritability of gene expressions and differential expressions.

Heritability of gene expression is the proportion of gene expressional variation that is attributable to genetic variation. We calculate the heritability of gene expression value in each probe set and use the q-values from output of SAM as differential expression value.

We calculate heritability of gene expressions (differential expressions) under four conditions: (1) gene expressions for all probe sets, (2) gene expressions for probe sets with PHEs >0, (3) gene expressions for probe sets with PHEs with unadjusted p-values < 0.05, (4) gene expressions for probe sets with PHEs with q-value < 0.05.

Under four conditions, we plot

a. The bar-plot: numbers of cis eSNPs <100 kb versus heritability of gene expressions and differential expressions.

b. The bar-plot: numbers of cis eSNPs >100 kb versus heritability of gene expressions and differential expressions.

c. The bar-plot: numbers of trans versus heritability of gene expressions and differential expressions.

The purpose of these plots is to observe that the various numbers of cis eSNPs < 100 kb, cis eSNPs > 100 kb or trans corresponding to heritability of gene expressions and differential expression values.

The steps of plot:

Step1. No matter which conditions, we divide heritability of gene expressions and differential expressions into 5 groups according to 5 quantiles.

Step2. To evaluate the numbers of cis eSNPs <100 kb, cis eSNPs >100 kb and trans separately in each quantile of heritability of gene expressions and differential expressions.

Step3. The bar-plots: Numbers (cis eSNPs <100kb, cis eSNPs >100kb and trans) versus quantiles of heritability of gene expressions and differential expressions.

3.6.6 The density plot of LOD score for cis eSNPs <100 kb, cis eSNPs >100 kb and trans

We derive four conditions of SNP’s LOD scores, there are the result of genome-wide association tests for (1) gene expressions for all probe sets, (2) for gene expressions for probe sets with PHEs >0, (3) gene expressions for probe sets with PHEs with unadjusted p-values < 0.05, (4) gene expressions for probe sets with PHEs with q-value < 0.05.

Under four conditions, we plot the density plot of LOD scores for cis eSNPs <100 kb, cis eSNPs >100 kb and trans and observe these distributions of cis eSNPs <100 kb, cis eSNPs > 100 kb and trans.

4 result

4.1 Test of PHEs and the distribution of different conditions of PHEs

There are 54675 gene expressions of probe sets and calculate 54675 PHEs. Among total PHEs, the minimum is -0.3708, the maximum is 0.315897, the mean is -0.01053 and the median is -0.00679. The results of hypothesis test: PHE=0 and solving the problem of multiple test by utilizing the q-value which is a natural pFDR analogue of the p-value to correct for occurrence of false positives: there are 19876 probe sets with PHEs greater than 0, 522 probe sets with PHEs with unadjusted p-value smaller than 0.05 and 38 probe sets with q-value smaller than 0.05 (Table 5). In Figure 2-a, most the low bounds of PHEs are smaller than zero. In Figure 2-b, the only low bounds of PHEs are higher than zero. Most PHEs are between 0.1 and 0.2, 25 PHEs are greater than 0.2 and one PHE is greater than 0.3 (Table 6).

In Density plot of four conditions of PHEs, total PHEs are mostly distributed between -0.15 and 0.15(Figure 3-a). Significant PHEs with unadjusted p-values < 0.05 are mostly distributed between 0.05 and 0.2 (Figure 3-c). Significant PHEs with q-value < 0.05 are most distributed between 0.15 and 0.3 (Figure 3-d). Under four conditions of PHEs, the distribution is from (-0.15, 0.15) to (0.15, 0.25). After testing and adjusting the problem of multiple tests, these significant PHEs have stronger evidences that asthma and endophenotype share more genes to provide more genetic information.

4.2 The scatter plot of heritability of gene expression versus PHEs

The scatter plots are under four conditions: (1) total heritability of gene expressions for all probe sets versus total PHEs, (2) heritability of gene expressions for probe sets with PHEs >0 versus PHEs >0, (3) heritability of gene expressions for probe sets with significant PHEs with unadjusted p-values < 0.05 versus significant PHEs with unadjusted p-values < 0.05, (4) heritability of gene expressions for probe sets with significant PHEs with q-values < 0.05 versus significant PHEs with q-values < 0.05. Then we fit two lines in scatter plot, one is regression line and the other is smooth line of loss function.

Under (1) condition, two lines in plots are increasing with greater PHEs. There is a positive correlation between heritability of gene expression and PHEs (Figure 4).

Under (2) condition, the smooth line is slowly increasing (Figure 5-a) and the other is slowly decreasing in plot (Figure 5-b). Two lines are almost horizontals. We infer that heritability of gene expression have very weak relationship with PHEs.

Under (3) condition, most of the smooth line is decreasing in plot. But this smooth line is increasing with few greater PHEs (Figure 6-a). The other in plot is decreasing. Besides greater PHEs, there may be a negative correlation between heritability of gene expression and PHEs (Figure 6-b).

Under (4) condition, the smooth line is irregular (Figure 7-a) and in Figure 7-b, the other is almost horizontal (Figure 7-b). There is no relationship between heritability of gene expression and PHEs.

4.3 The overlap rate

The overlap rate by comparing significant SNPs (LOD>6) from probe sets with significant PHEs with 269 genes (144 genes: asthma or atopy genes, 125 genes: genes associated with these 25 asthma or atopy genes in six or more populations).The significant PHEs can be (1) with unadjusted p-value smaller than 0.05, (2) with q-value smaller than 0.05.

Under the (1) condition of significant PHEs, there are 767 significant SNPs (LOD > 6) from genome-wide association test. There are 14 significant SNPs (LOD > 6) overlap with 144 asthma or atopy genes. There is 0 significant SNPs (LOD > 6) overlap with 125 genes associated with these 25 asthma or atopy genes in six or more populations (Table 7).

The overlap rate =

$$\frac{\#\ \text{significant SNPs (LOD >6) overlap with asthma or atopy genes plus genes associated with an asthma or atopy phenotype in six or more populations.}}{\#\ \text{significant SNPs (LOD >6)}}$$

$$= \frac{14}{767} \approx 0.015$$

Under the (2) condition of significant PHEs, there are 12 significant SNPs (LOD > 6) (Table 9, 10). There is 0 significant SNPs (LOD > 6) overlap with 269 genes. The overlap rate is zero.

These overlap rates are not high. We infer that some significant SNPs may be weakly associated with asthma or atopy. Because we lack the genome-wide association data, the validation of this assumption does not analyze.

4.4 The result of association test under four conditions of PHEs

We utilize genome-wide association tests for gene expression of probe set and take the SNPs' LOD score greater than 6 as significant SNPs. We define these significant SNPs as "eSNPs" and show various plots of eSNPs to observe the genetic properties.

4.4.1 The bar-plot of proportion of probe sets with max significant SNPs' LOD >6 versus PHEs

Four conditions of significant SNP's LOD >6 (eSNPs), there are the results of genome-wide association tests for (1) gene expressions for all probe sets, (2) probe sets with PHEs >0, (3) gene expressions for probe sets with PHEs with unadjusted p-values < 0.05, (4) gene expressions for probe sets with PHEs with q-value < 0.05.

Under (1) and (2) conditions, the proportions are almost similar and slowly increasing when PHEs are larger (Figure 8-a, b). The underlying genes may be more because of greater proportion of probe sets with max significant SNPs' LOD > 6.

Under (3) condition, the proportions are decreasing (Figure 8-c). When PHEs are larger, proportions of probe sets with max significant SNPs' LOD > 6 are smaller and the underlying genes may be less. Although most underlying genes are not significant in greater PHEs, some remaining underlying genes may be very significant to provide more genetic information.

4.4.2 The bar-plot of the number of cis eSNPs <100 kb or cis eSNPs >100 kb or trans versus PHEs

Four conditions of significant SNP's LOD >6 (eSNPs), there are the results of genome-wide association tests for (1) gene expressions for all probe sets, (2) probe sets with PHEs >0, (3) gene expressions for probe sets with PHEs with unadjusted p-values < 0.05, (4) gene expressions for probe sets with PHEs with q-value < 0.05.

We do the bar-plot with the number of cis eSNPs <100 kb at first. Under (1) condition, the numbers will slowly increasing when PHEs are larger (Figure 9-a). When PHEs are greater, the number of underlying genes may be more. Under (2) condition, the numbers are almost similar (Figure 9-b). Under (3) condition, the numbers are irregularly up-and-down (Figure 9-c).

To compare with Figure 9-c and Figure 8-c. The smallest proportion is in 5th quantile of PHEs (Figure 9-c), but the number of cis eSNPs < 100 kb in 5th quantile of PHEs is not least. This may be because some probe sets does not significantly associated to underlying genes at all, but some probe sets significantly associated to many underlying genes.

Then we do the bar-plot with the number of cis eSNPs > 100 kb. Under (1) and (2) conditions, the pattern of Figure 10-a and 10-b are similar with Figure 9-a and 9-b to result in the same conclusions. Under (3) condition, the numbers are almost less, besides 2nd quantile of PHEs (Figure 10-c). It represents more number of underlying genes in 2nd quantile.

Finally we do the bar-plot with the number of trans. Under (1) condition, the numbers are slowly increasing when PHEs are larger (Figure 11-a). When PHEs are greater, the number of underlying genes may be more. Under (2) condition, the numbers are increasing, besides 1st quantile of PHEs (Figure 11-b). Under (3) condition, the numbers are almost less, besides 2nd and 3rd quantile of PHEs (Figure 11-c). It represents more number of underlying genes in 2nd and 3rd quantile.

Under the condition of (4), it only has 12 eSNPs (Table 6).

4.4.3 The bar-plot of number of cis eSNPs <100 kb or cis eSNPs >100 kb or trans versus heritability of gene expressions and differential expressions.

There are four conditions of heritability of gene expressions (differential expression): (1) gene expressions for all probe sets, (2) gene expressions for probe sets with PHEs >0, (3) gene expressions for probe sets with PHEs with unadjusted p-values < 0.05, (4) gene expressions for probe sets with PHEs with q-value < 0.05.

We do the bar-plot with the numbers of cis eSNPs <100 kb, cis eSNPs > 100 kb and trans, the numbers increase with greater heritability of gene expressions under (1) and (2) conditions (Figure 12-a,b; Figure 13-a,b; Figure 14-a,b). The higher heritability represents the more proportion of genetic component to result in more numbers of eSNPs < 100 kb. The numbers are irregular with heritability of gene expressions under (3) condition (Figure 12-c, Figure 13-c, Figure 14-c).The greater heritability of gene expressions not sure to result more number of cis eSNPs < 100 kb. There is no relationship between PHEs with unadjusted p-value < 0.05 and heritability of gene expressions.

Under (1) and (2) conditions, the numbers are similar with differential expression values (Figure 12-a, b; Figure 13-a, b; Figure 14-a, b). Under (3) condition, the numbers are irregular with differential expression value (Figure 12-c; Figure 13-c; Figure 14-c). These differential expressions are significant with the differences of case and control status and not necessary associated with significant SNPs (LOD > 6).

4.4.4 The density plot of LOD score for cis eSNPs <100 kb, cis eSNPs >100 kb and trans

Four conditions of SNP’s LOD scores, there are the result of genome-wide association tests for (1) gene expressions for all probe sets, (2) for gene expressions for probe sets with PHEs >0, (3) gene expressions for probe sets with PHEs with unadjusted p-values < 0.05, (4) gene expressions for probe sets with PHEs with q-value < 0.05.

No matter which conditions, if the LOD scores increase, the effect in cis eSNPs <100 kb is stronger than the other two. The effects in trans were weaker than in cis eSNPs <100kb or cis eSNPs >100kb and distributed at lower LOD score (about LOD<5). Most LOD greater than 6 were in cis eSNPs <100 kb (Figure 15-21).

5Conclusions and Discussion

5.1 Conclusions

We start with 54675 gene expressions of probe sets, preprocessed by the robust multi-array averaging (RMA). These preprocessed gene expressions are used as endophenotypes in search for the susceptibility genes underlying asthma. We judge which gene expressions are endophenotypes by the index PHE. The greater the PHE value, the more likely the intermediate variable being an endophenotype and sharing more genes with the phenotype. We do hypothesis test (H_0: PHE=0: the probe set and phenotype share no genes versus H_1: PHE>0) and adjusted for multiple testing by utilizing the q-value to control for the false discovery rate (FDR). Then we perform genome-wide association tests for each gene expression. We are interested in assessing genetic characteristics of gene expressions that are identified as the endophenotypes.

The conclusion of these plots: For all probe sets, the greater the PHEs, the more underlying genetic components the gene expressions have. For probe sets with PHEs greater than zero, the genetic properties are not obvious, not like for all probe sets. For probe sets with PHEs with unadjusted p-value < 0.05, the proportion of probe sets with max significant SNPs' LOD > 6 decreases when the PHE value increases. However, there is an increasing numbers of underlying genes when the PHE value increases. This may be because some probe sets does not significantly associated to underlying genes at all, but some probe sets significantly associated to many underlying genes. The larger heritability of gene expressions is not sure to result more number of cis eSNPs < 100 kb. So there is no relationship between PHEs with unadjusted p-value < 0.05 and heritability of gene expressions. To sum up, these PHEs are greater, the genetic components are less or irregular. For probe sets with

PHEs with q-value smaller than 0.05, the irregular pattern make us hardly observing the genetic properties.

The genetic effect in cis eSNPs <100 kb is stronger than the other two in greater LOD scores (LOD>6). Trans effects were weaker than the other two and distributed at lower LOD score (LOD<5).

5.2 Discussion

If All SNPs put into gene-gene interaction test, the number of test is too many. In signal-SNP association test between SNP genotypes and case-control status, it only discoveries significant SNPs with disease and neglect Some SNPs are weakly related with disease, but affected disease after combining with other SNPs. So we use the properties of endophenotype to find these significant SNPs weakly associated with disease from association test between gene expression of each probe set and all SNP genotypes. We collect these significant SNPs from single-SNP association test of case-control and continuous outcome (gene expression) and then do gene-gene interaction to add more opportunity for searching possible underlying disease genes.

But there is a problem, what criteria of PHE judging the expression value of probe sets as an endophenotype is appropriate. In the future, we will keep on overcoming this problem and utilize the real SNP data that provided by Moffatt et al [5] to confirm the assumption.

Reference

1. Gottesman II, Gould TD. The endophenotype concept in psychiatry: etymology and strategic intentions. *Am J Psychiatry* **160**, 636-645 (2003).
2. Huang GH, Chen CH, Chen WJ. Statistical Validation of Endophenotypes Using a Surrogate Endpoint Analytic Analogue with Application to Schizophrenia (2005).
3. Freedman LS, Graubard BI, Schatzkin A. Statistical validation of intermediate endpoints for chronic diseases. *Stat Med* **11**, 167-178 (1992).
4. Hsieh CC, Confidence Interval and Simulation Studies for the Proportion of Heritability Explained by Endophenotypes (2006).
5. Moffatt MF*, Kabesch M*, Liang L*, Dixon AL, Strachan D, Heath S, Depner M, von Berg A, Bufe A, Rietschel E, Heinzmann A, Simma B, Frischer T, Willis-Owen SAG, Wong KCC, Illig T, Vogelberg C, Weiland SK, von Mutius E, Abecasis GR3, Farrall M, Gut IG, Lathrop GM, Cookson WOC. Genetic variants regulating ORMDL3 expression are determinants of susceptibility to childhood asthma. *Nature* **448**, 470-473 (2007).
6. Dixon AL*, Liang L*, Moffatt MF*, Chen W, Heath S, Wong KCC, Taylor J, Gut I, Farrall M, Lathrop GM, Abecasis GR, Cookson WOC. A genome-wide association study of global gene expression. *Nat Genet* **39**, 1202-1207 (2007).
7. Prentice RL. Surrogate endpoints in clinical trials: definitions and operational criteria. *Stat Med* **8**, 431-440 (1989).
8. Buyse M, Molenberghs G. Criteria for validation of surrogate endpoints in randomized experiments. *Biomrtrics* **54**, 1014-1029 (1998).
9. De Gruttola VG, Glax P, DeMets DL, Downing GJ, Ellenberg SS,Friedman L, Gail MH, Prentice R, Wittes J, Zeger SL. Considerations in the evaluation of surrogate endpoints in clinical trials: Summary of a National Institutes of Health workshop. *Control Clin Trials* **22**, 485-502 (2001).

10. McCullagh P, Nelder JA. Generalized Linear Models, 2nd edition. Chapman and Hall, London (1989).

11. Begg C, Leung DHY. On the use of surrogate end point in randomized trials (with comments). *JRSS A* **163**, 15-28 (2000).

12. Hopper JL. In Biostatistical Genetic Epidemiology, Elston, Olson and Palmer eds. Wiley, Chichester, 371-372 (2002).

13. Hopper JL, In Biostatistical Genetics and Genetic Epodemiology, Elston, Olson and Palmer eds. Wiley, Chichester, 778-788 (2002).

14. Almasy L, Blanero J. Multipoint quantitative-trait linkage analysis in general pedigree. *Am J Hum Genet* **62**, 1198-1121 (1998).

15. Burton P.R., Tobin M.D. Handbook of Statistical Genetics, 2nd edition, Balding D.J., Bishop M.and Cannings C eds. John Wiley & Sons, Ltd, p. 855-879 (2003).

16. Duggirala R, Williams JT, Williams-Blangero S, Blangero J. A variance component approach to dichotomous trait linkage analysis using a threshold model. *Genet Epidemiol* **14**, p. 987-992 (1997).

17. Falconer DS. Introduction to Quantitative Genetics, Third edn. John WIley &Sons, New York (1989).

18. Casella G, Berger RL. Statistical Inference, 2nd edition. p. 240-245 (2001).

19. Chen WM and Abecasis GR. Family-Based association Tests for Genome-wide Association Scans. *Am. J. Hum. Genet* **81**, 913-926 (2007).

20. Gentleman R, Carey VJ, Huber W, Irizarry RA, Dudoit S. Bioinformatics and Computational Biology Solutions Using R and Bioconductor. Springer, p. 250-271 (2005).

21. Irizarry,R.A., Hobbs,B., Collin,F., Beazer-Barclay,Y.D., Antonellis,K.J., Scherf,U. and Speed,T.P. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* **4**, 249-264 (2003).

22. Wang YY. Validity and Reloability of Combinations of Preprocessing and Differential Expression Methods for Affymetrix Genehip Micoarrays (2007).

23. Bolstad,B.M., Irizarry,R.A., Astrand,M. and Speed,T.P. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* **19**, 185-193 (2003).

24. Irizarry, R.A., Bolstad,B.M., Collin,F., Cope,L.M., Hobbs,B. and Speed,T.P. Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Research.* p. 31, e15 (2003).

25. Tusher,V.G., Tibshirani,R. and Chu,G. Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Science* **98**, 5116-5121 (2001).

26. Chu,G., Narasimhan,B., Tibshirani,R. and Tusher,V. SAM. Significance Analysis of Microarrays–Users guide and technical document. Technical Report, Stanford University. *http://www-stat.stanford.edu/~tibs/SAM/sam.pdf.*

27. John D. Storey. A direct approach to false discovery rates. *J. R. Statist. Soc. B* **64**, 479-498 (2001).

28. Sabine Hoffjan, Dan Nicolae, Carole Ober. Association studies for asthma and atopic diseases: a comprehensive review of the literature. *Respir Res* **4**, p. 14 (2003).

29. C Ober, S Hoffjan. Asthma genetics 2006: the long and winding road to gene discovery. *Genes and Immunity* **7**, 95-100 (2006).

30. Jian Zhang, Peter D Pare, Andrew J Sandford. Recent advances in asthma genetics. *Respir Res* **9**, p.4 (2008).

31. Wei-Min Chen and Goncalo R. Abecasis. Fammily-Based Association Tests for Genome-wide Association Scans. *Am. J. Hum. Genet*. 81, 913-926 (2007)

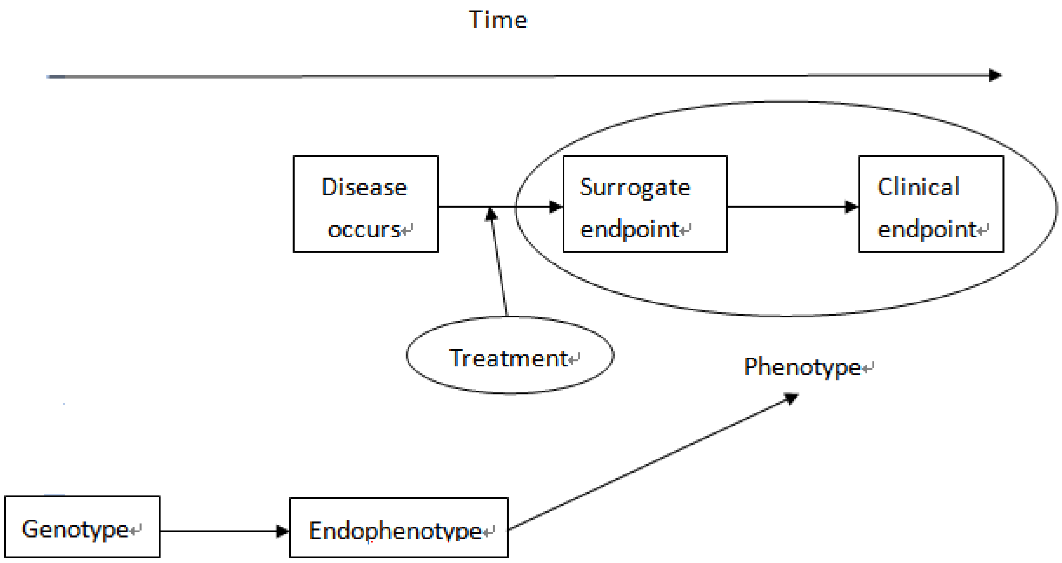

Figure 1　A surrogate endpoint versus an endophenotype in the disease process.

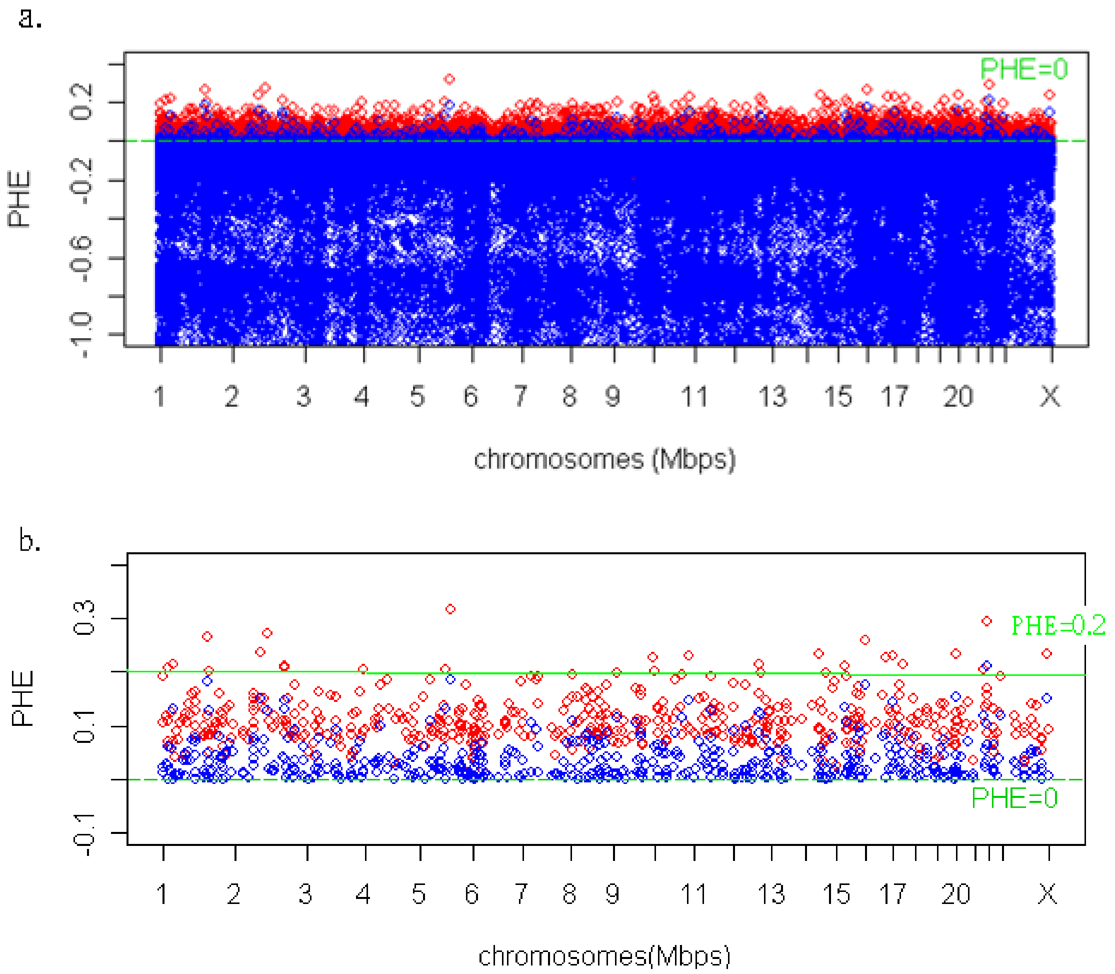

Figure 2 The confidence interval of PHEs on all chromosomes (red: PHE, blue: the low bound of PHE). | a. the confidence interval of total PHEs, b. the confidence interval of the low bound of PHEs >0

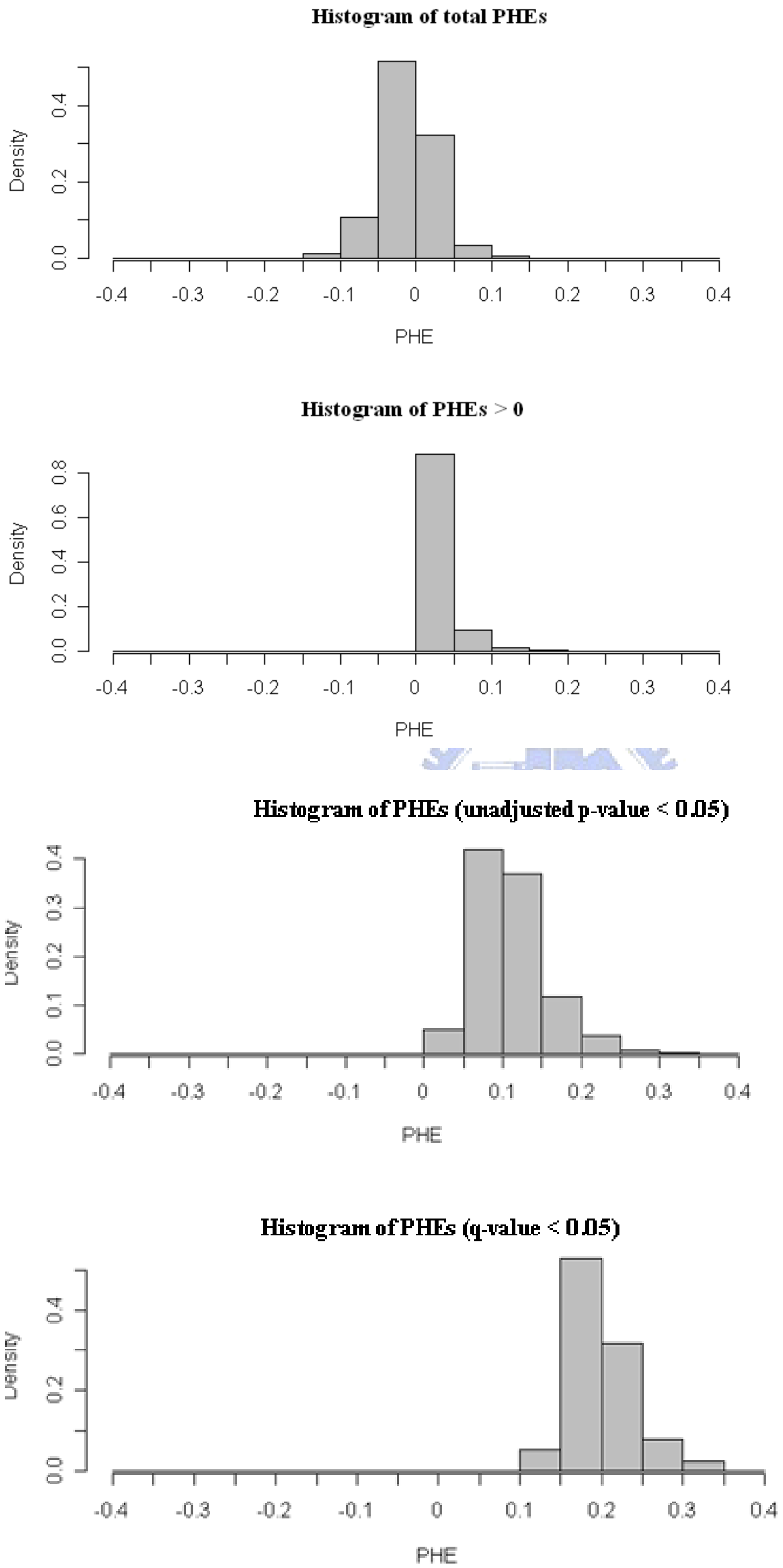

Figure 3 The density plot of PHEs. | a. Density plot of total PHEs. b. Density plot of PHEs >0. c. Density plot of PHEs with unadjusted p-value <0.05. d. Density plot with q-value<0.05.

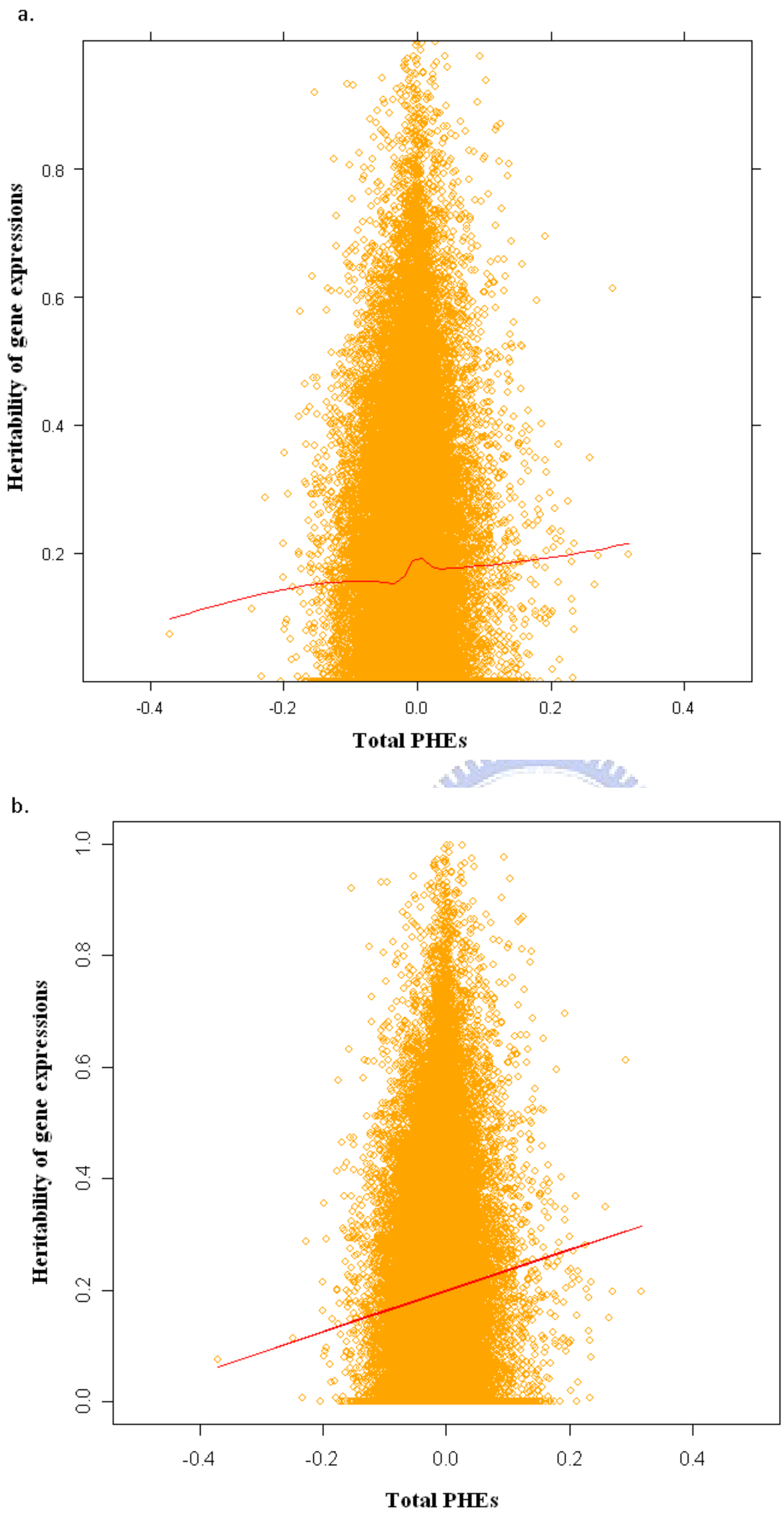

Figure 4 The scatter plot of heritability versus total PHEs.

a. Scatter plot with a smooth line of loss function.

b. Scatter plot with a regression line.

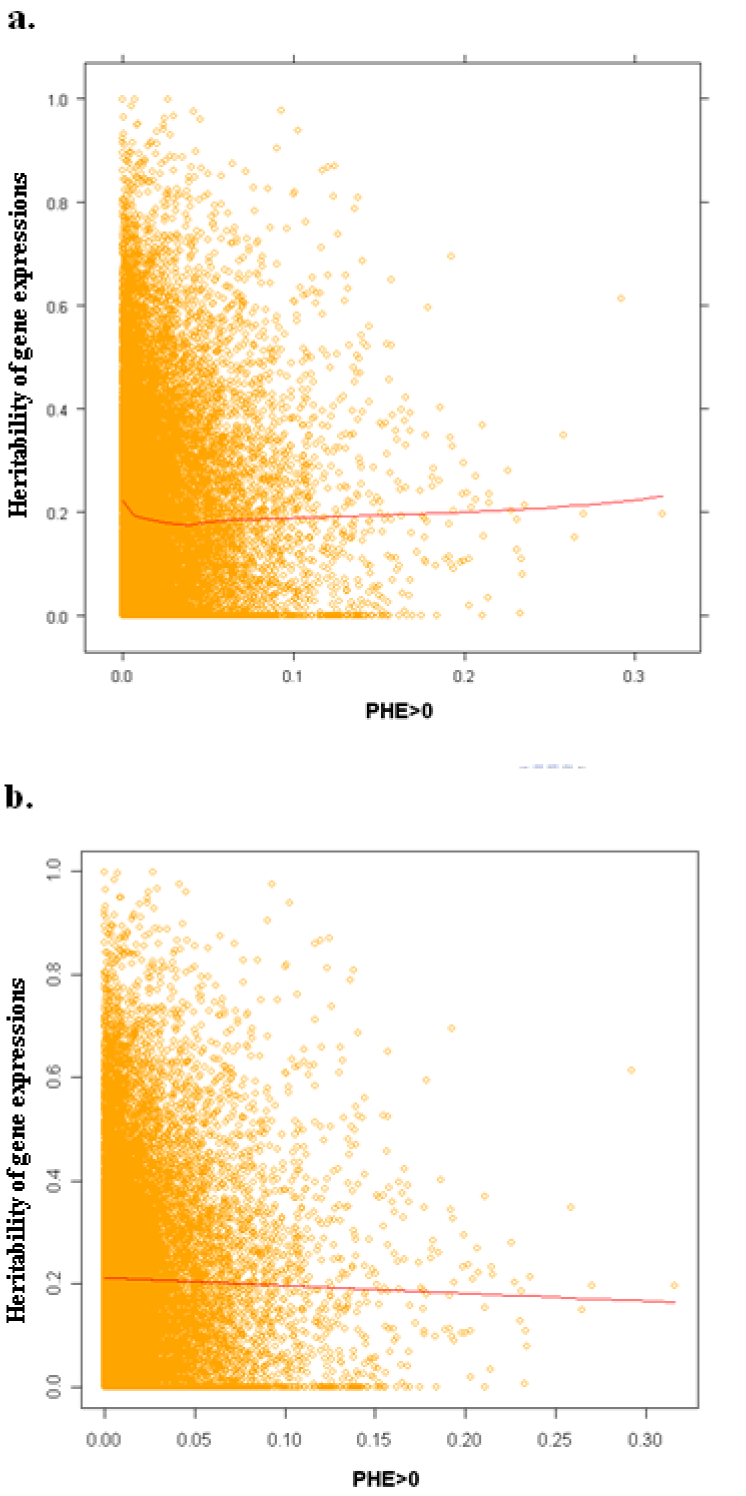

Figure 5 The scatter plot of heritability versus PHEs >0.

a. Scatter plot with a smooth line of loss function.

b. Scatter plot with a regression line.

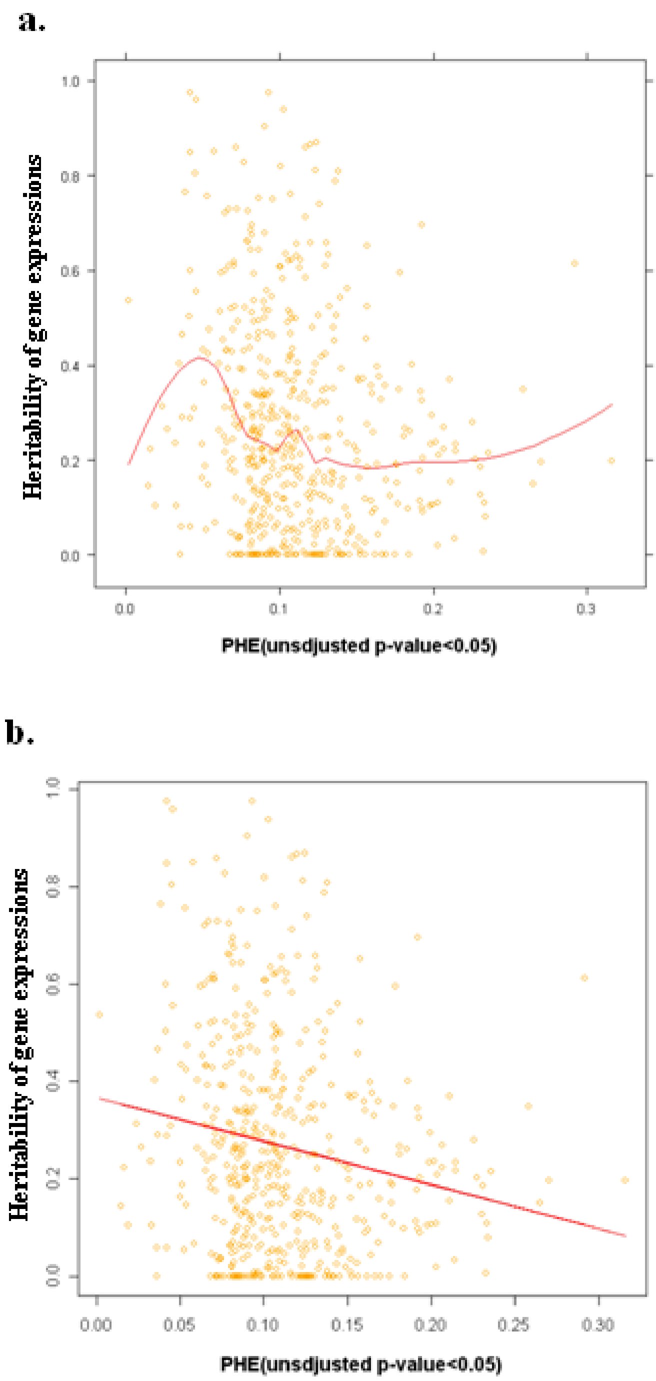

Figure 6 The scatter plot of heritability versus PHEs with unadjusted p-value <0.05.

a. Scatter plot with a smooth line of loss function.

b. Scatter plot with a regression line.

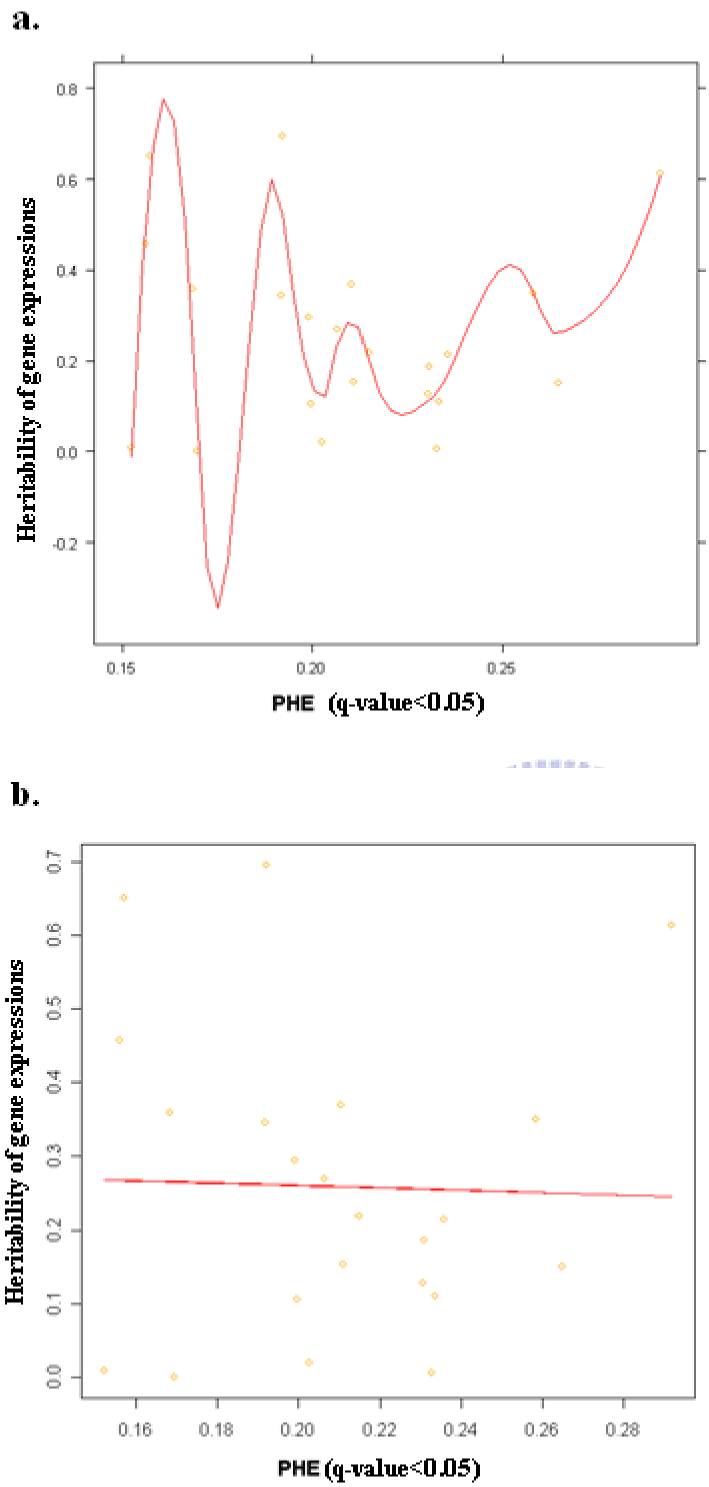

Figure 7 The scatter plot of heritability versus PHEs with q-value <0.05.

a. Scatter plot with a smooth line of loss function.

b. Scatter plot with a regression line.

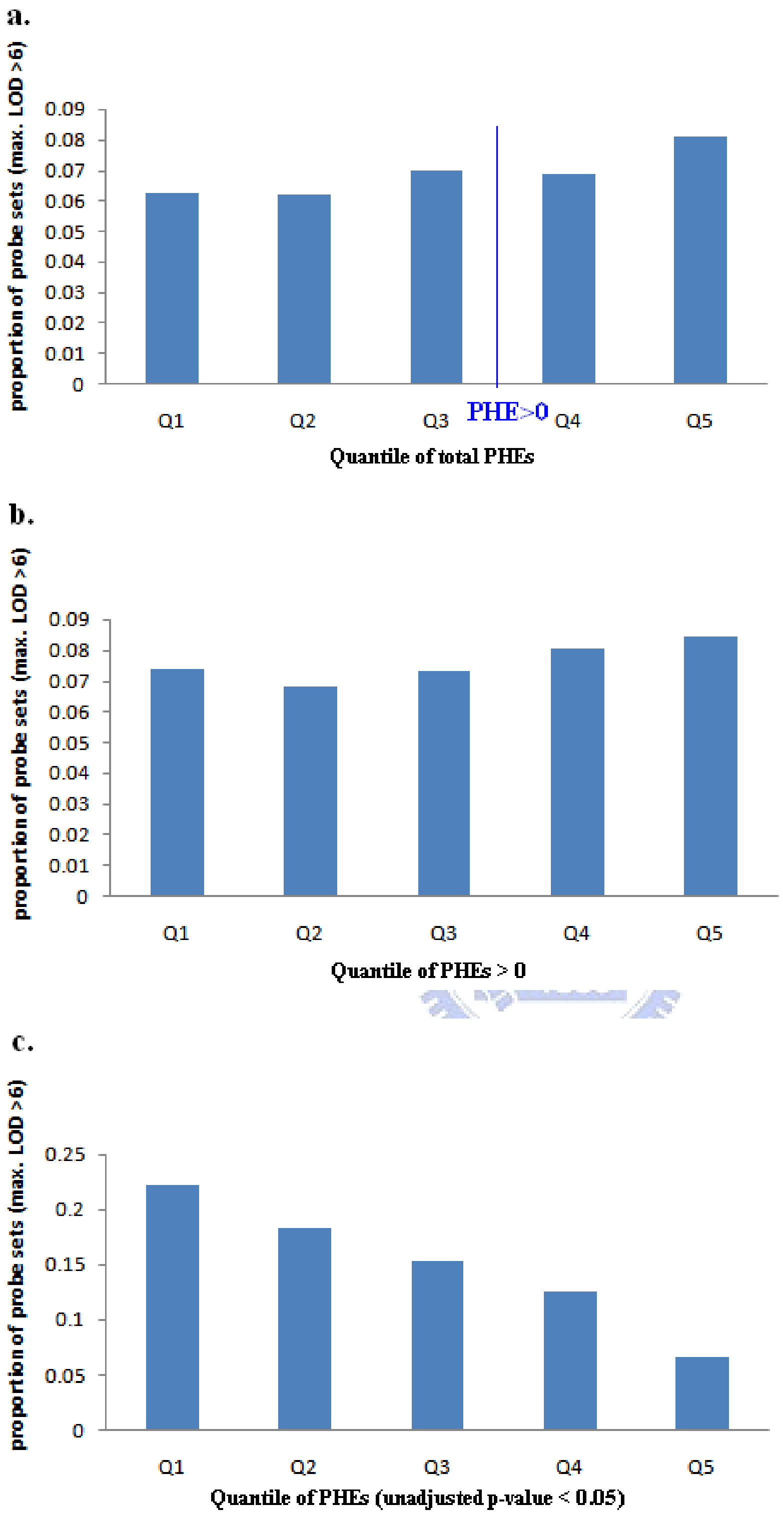

Figure 8 The bar-plot of proportion of probe sets with max significant SNPs' LOD >6 versus PHEs.

a. The bar-plot of proportion of probe sets with max significant SNPs' LOD >6 versus total PHEs.

b. The bar-plot of proportion of probe sets with max significant SNPs' LOD >6 versus PHEs >0.

c. The bar-plot of proportion of probe sets with max significant SNPs' LOD >6 versus PHEs with unadjusted p-value <0.05.

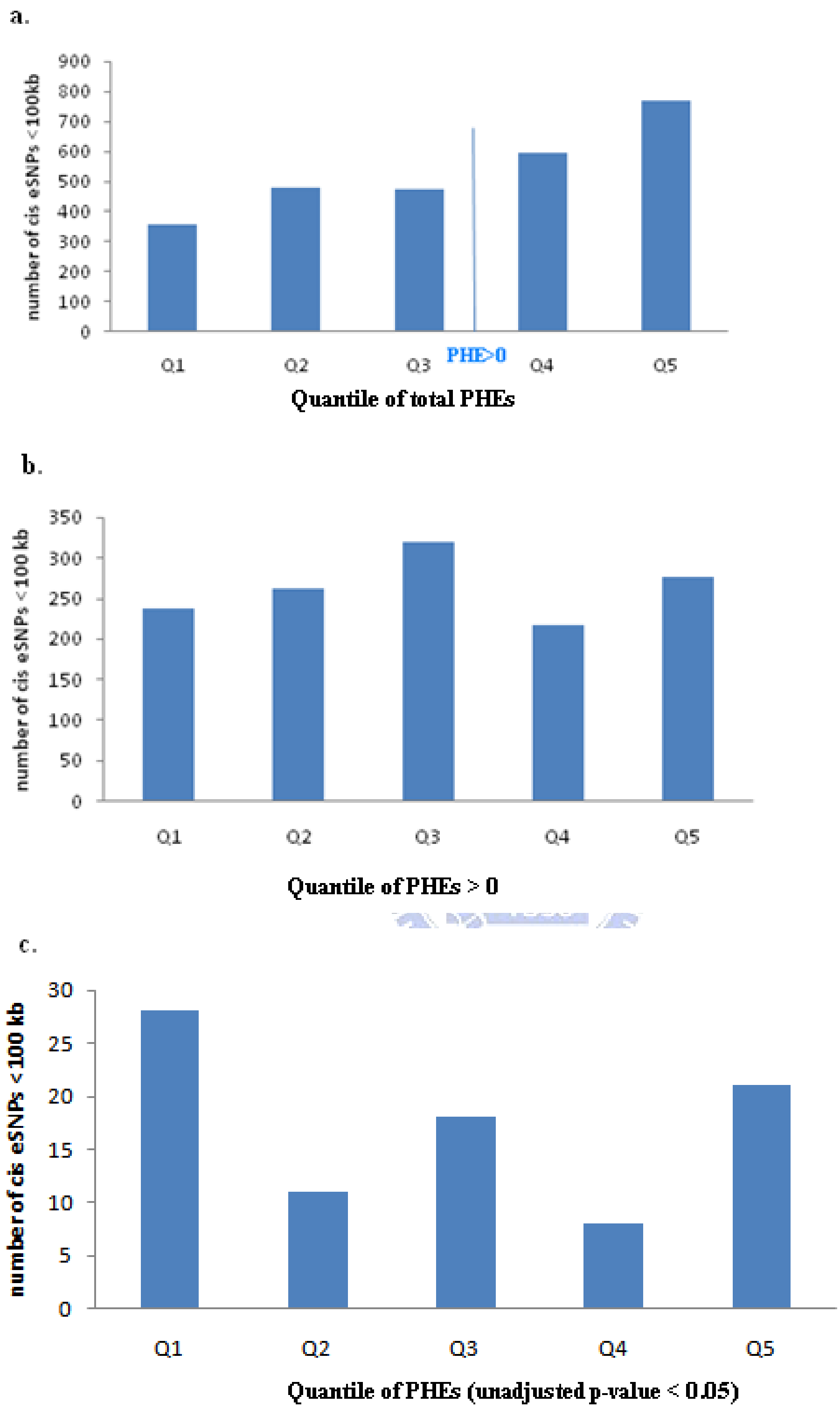

Figure 9　The bar-plot of the number (cis eSNPs <100kb) versus PHEs.

a.　The bar-plot of the number versus total PHEs.

b.　The bar-plot of the number versus PHEs>0.

c.　The bar-plot of the number versus PHEs with unadjusted p-value <0.05.

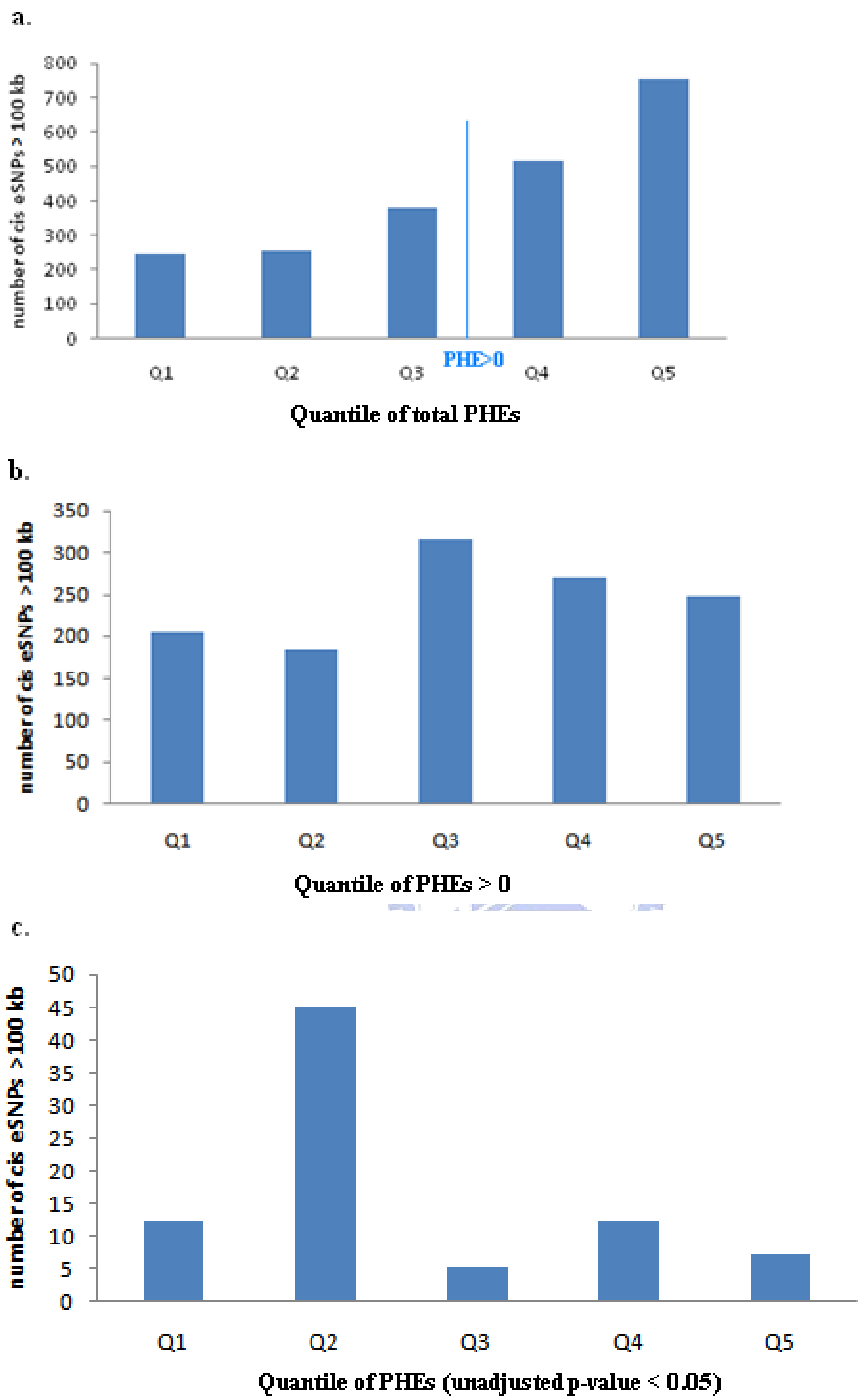

Figure 10 The bar-plot of the number (cis eSNPs >100kb) versus PHEs.

a. The bar-plot of the number versus total PHEs.

b. The bar-plot of the number versus PHEs>0.

c. The bar-plot of the number versus PHEs with unadjusted p-value <0.05.

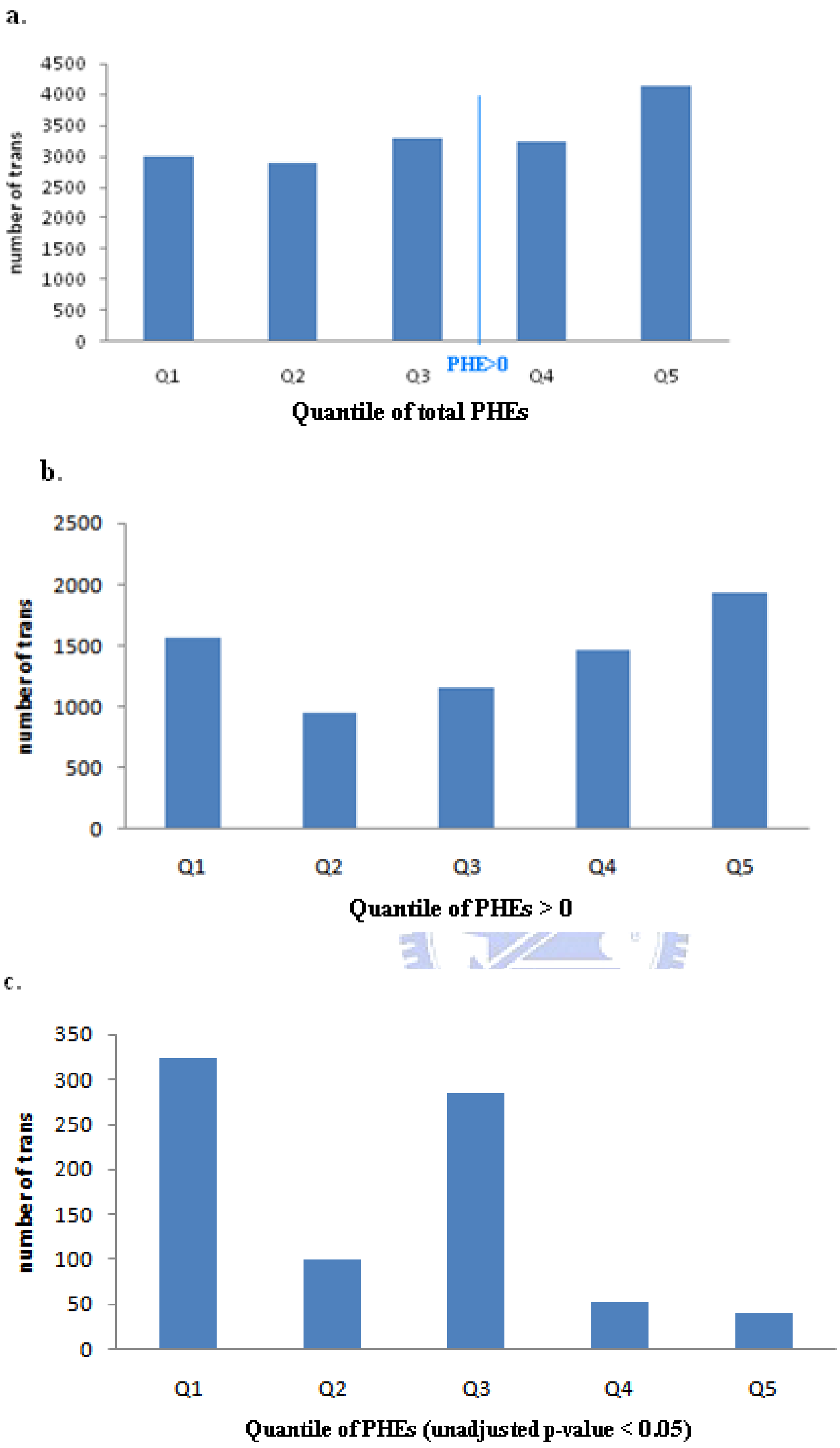

Figure 11 The bar-plot of the number (trans) versus PHEs.

a. The bar-plot of the number versus total PHEs.

b. The bar-plot of the number versus PHEs>0.

c. The bar-plot of the number versus PHEs with unadjusted p-value <0.05.

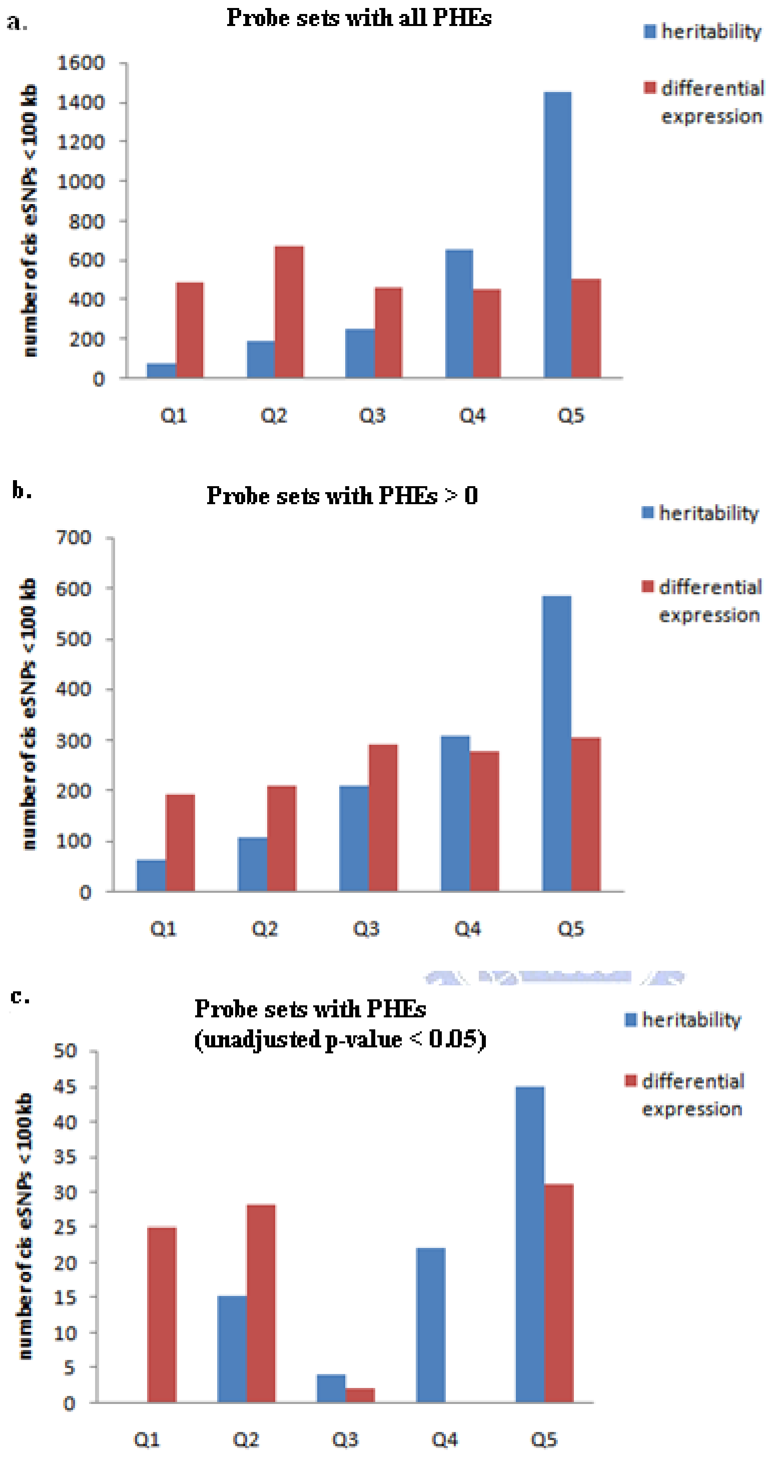

Figure 12 The bar-plot of the number (cis eSNPs <100 kb) versus heritability and differential expression.

a. The bar-plot of the number versus heritability and differential expression for probe sets with total PHEs.

b. The bar-plot of the number versus heritability and differential expression for probe sets with PHEs>0.

c. The bar-plot of the number versus heritability and differential expression for probe sets with PHEs with unadjusted p-value <0.05.

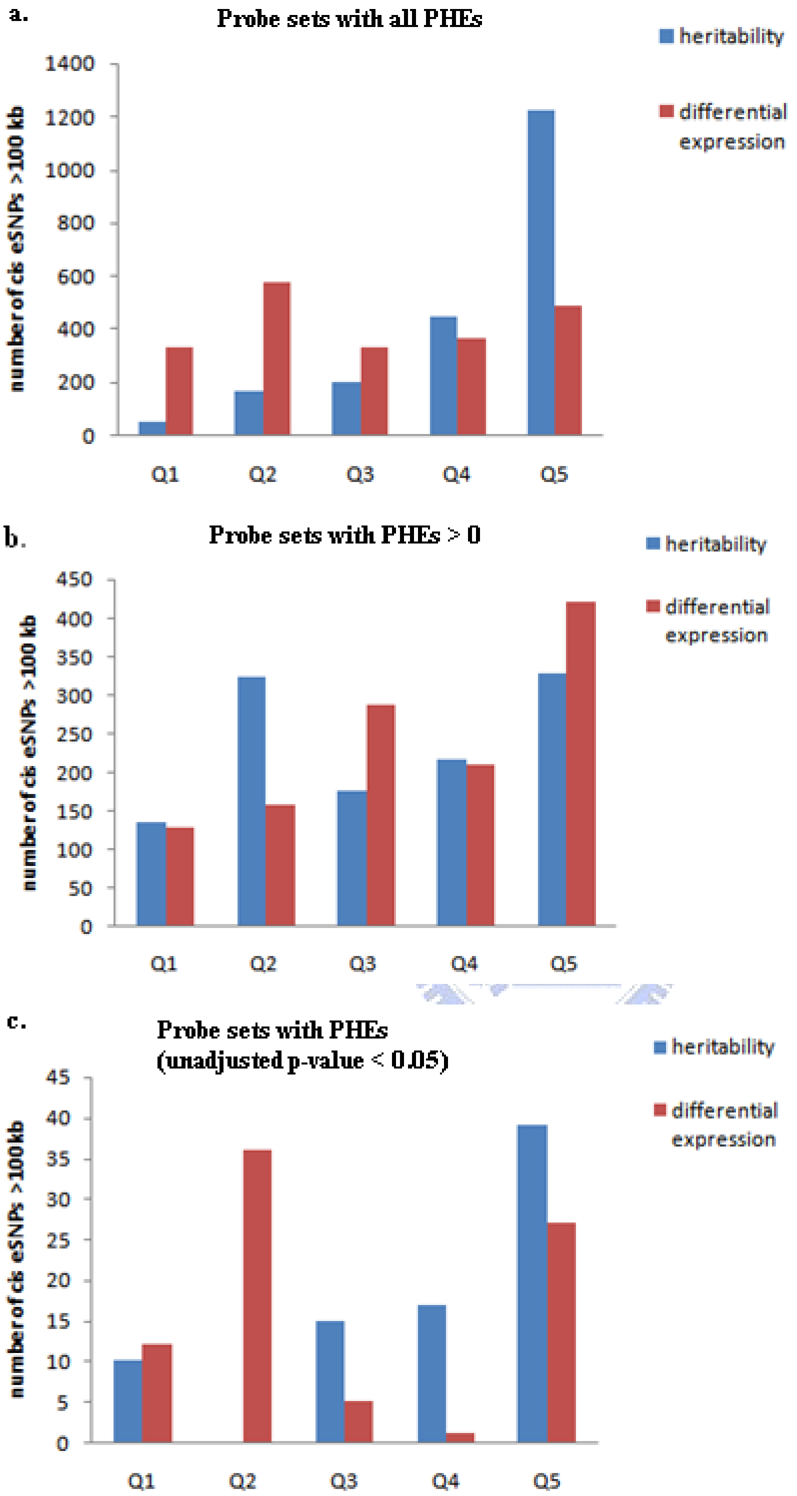

Figure 13 The bar-plot of the number (cis eSNPs >100 kb) versus heritability and differential expression.

a. The bar-plot of the number versus heritability and differential expression for probe sets with total PHEs.

b. The bar-plot of the number versus heritability and differential expression for probe sets with PHEs>0.

c. The bar-plot of the number versus heritability and differential expression for probe sets with PHEs with unadjusted p-value <0.05.

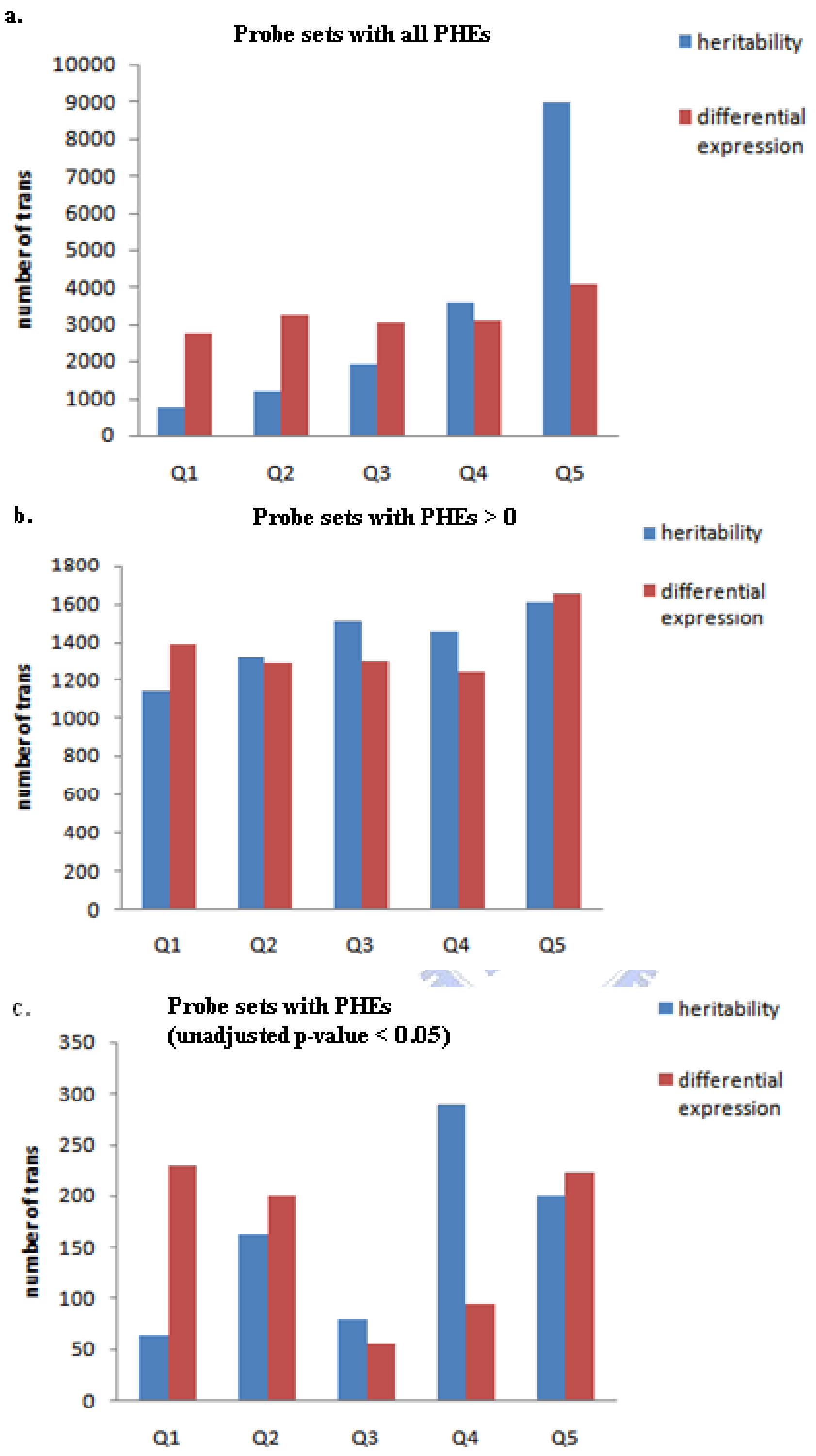

Figure 14 The bar-plot of the number (trans) versus heritability and differential expression.

a. The bar-plot of the number versus heritability and differential expression for probe sets with total PHEs.

b. The bar-plot of the number versus heritability and differential expression for probe sets with PHEs>0.

c. The bar-plot of the number versus heritability and differential expression for probe sets with PHEs with unadjusted p-value <0.05.

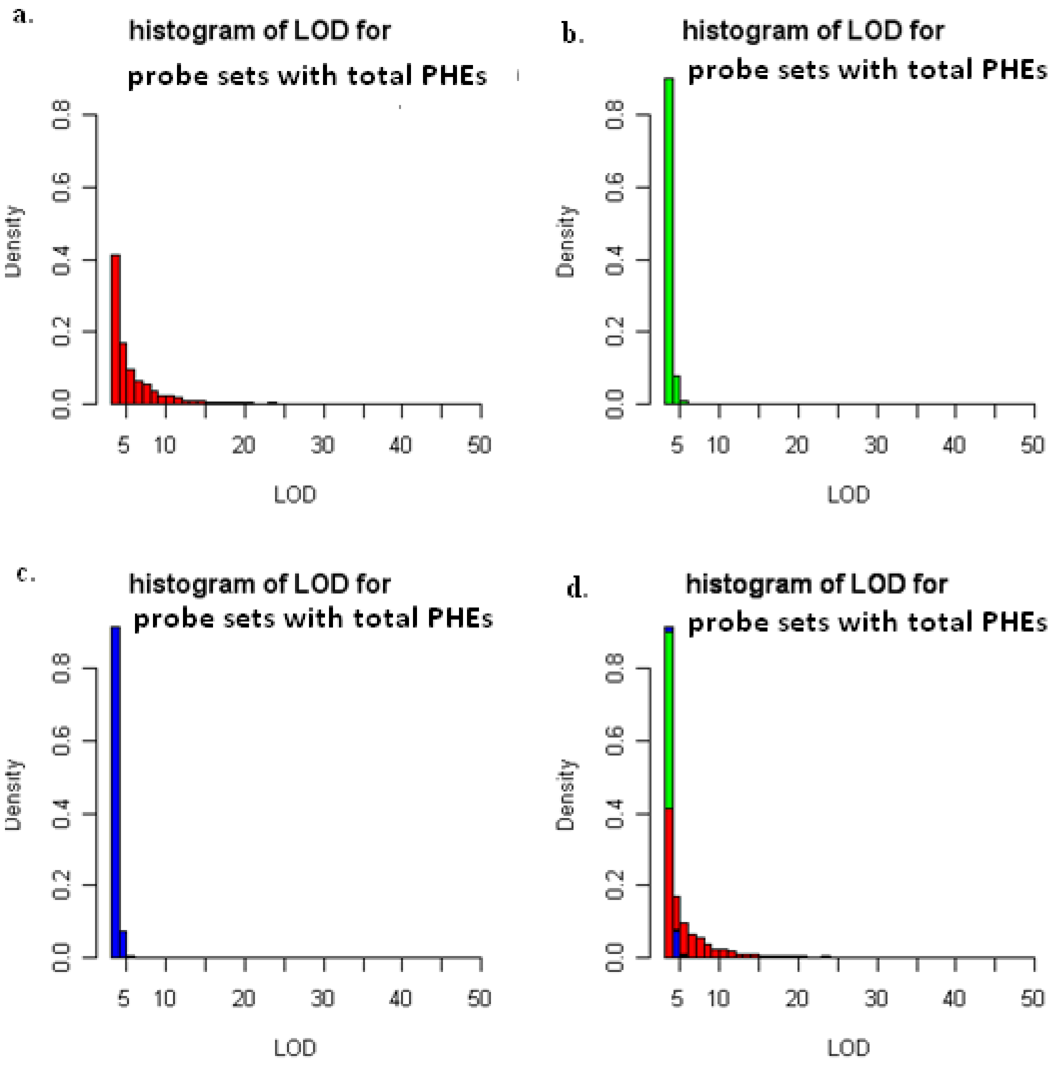

Figure 15 The density plot of LOD score for cis eSNPs <100 kb (red), cis eSNPs >100 kb (green) and trans (trans) between X-limit (3, 50) and Y-limit (0, 0.8) for probe sets with total PHEs.

a. The density plot of LOD score for cis eSNPs <100 kb.

b. The density plot of LOD score for cis eSNPs >100 kb.

c. The density plot of LOD score for trans.

d. The density plot of LOD score for cis eSNPs <100 kb, cis eSNPs >100 kb and trans.

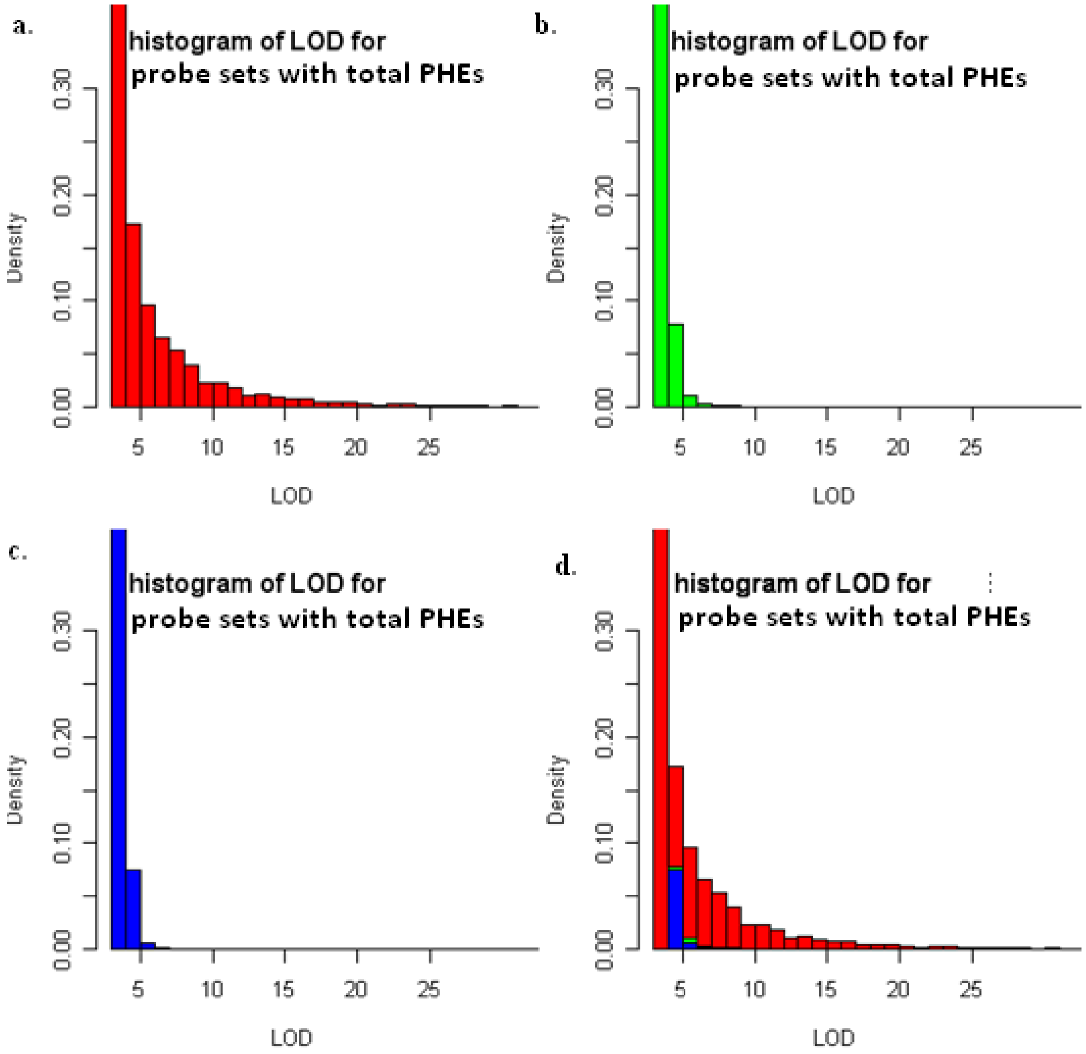

Figure 16 The density plot of LOD score for cis eSNPs <100 kb (red), cis eSNPs >100 kb (green) and trans (blue) between X-limit (3, 25) and Y-limit (0, 0.3) for probe sets with total PHEs.

a. The density plot of LOD score for cis eSNPs <100 kb.

b. The density plot of LOD score for cis eSNPs >100 kb.

c. The density plot of LOD score for trans.

d. The density plot of LOD score for cis eSNPs <100 kb, cis eSNPs >100 kb and trans.

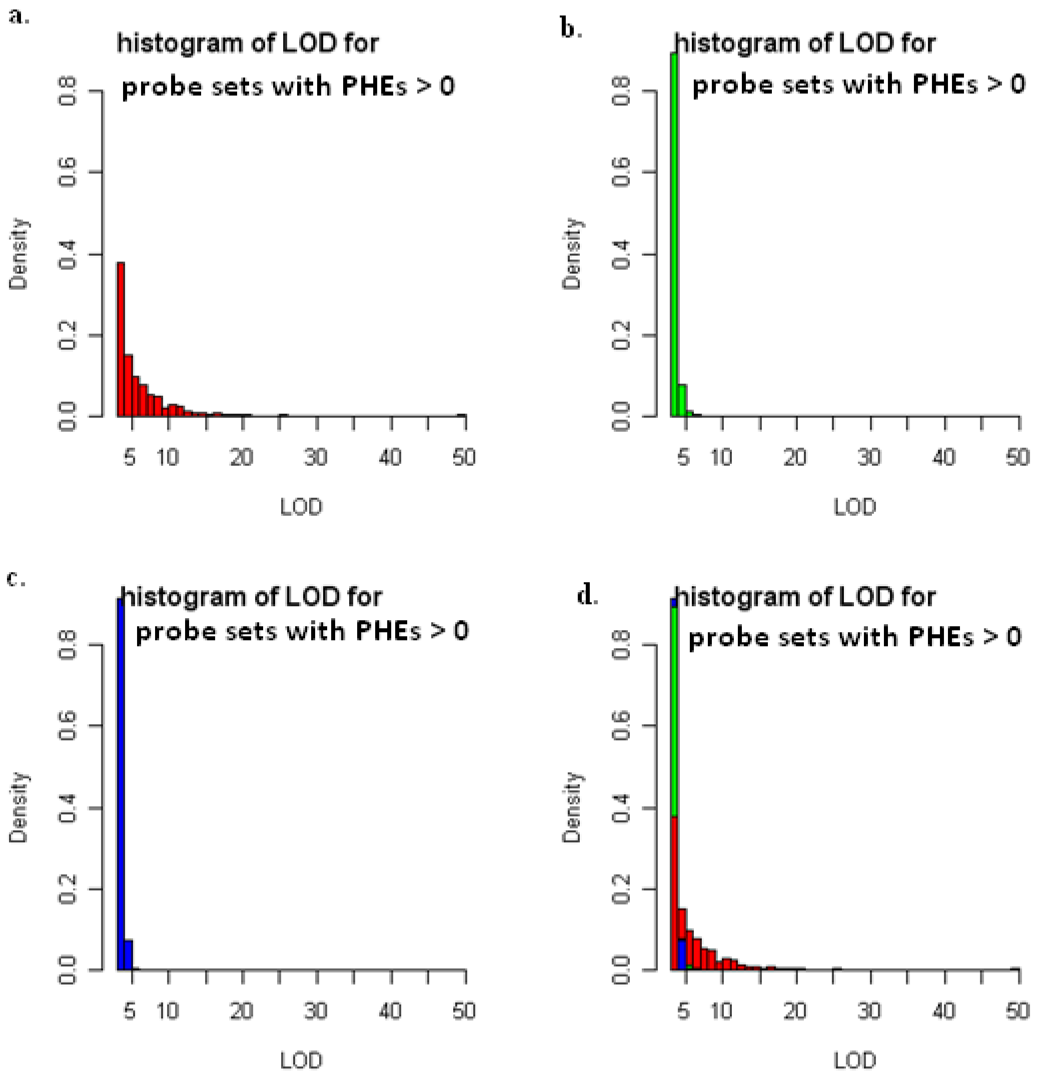

Figure 17 The density plot of LOD score for cis eSNPs <100 kb (red), cis eSNPs >100 kb (green) and trans (blue) between X-limit (3, 50) and Y-limit (0, 0.8) for probe sets with PHEs>0.

a. The density plot of LOD score for cis eSNPs <100 kb.
b. The density plot of LOD score for cis eSNPs >100 kb.
c. The density plot of LOD score for trans.
d. The density plot of LOD score for cis eSNPs <100 kb, cis eSNPs >100 kb and trans.

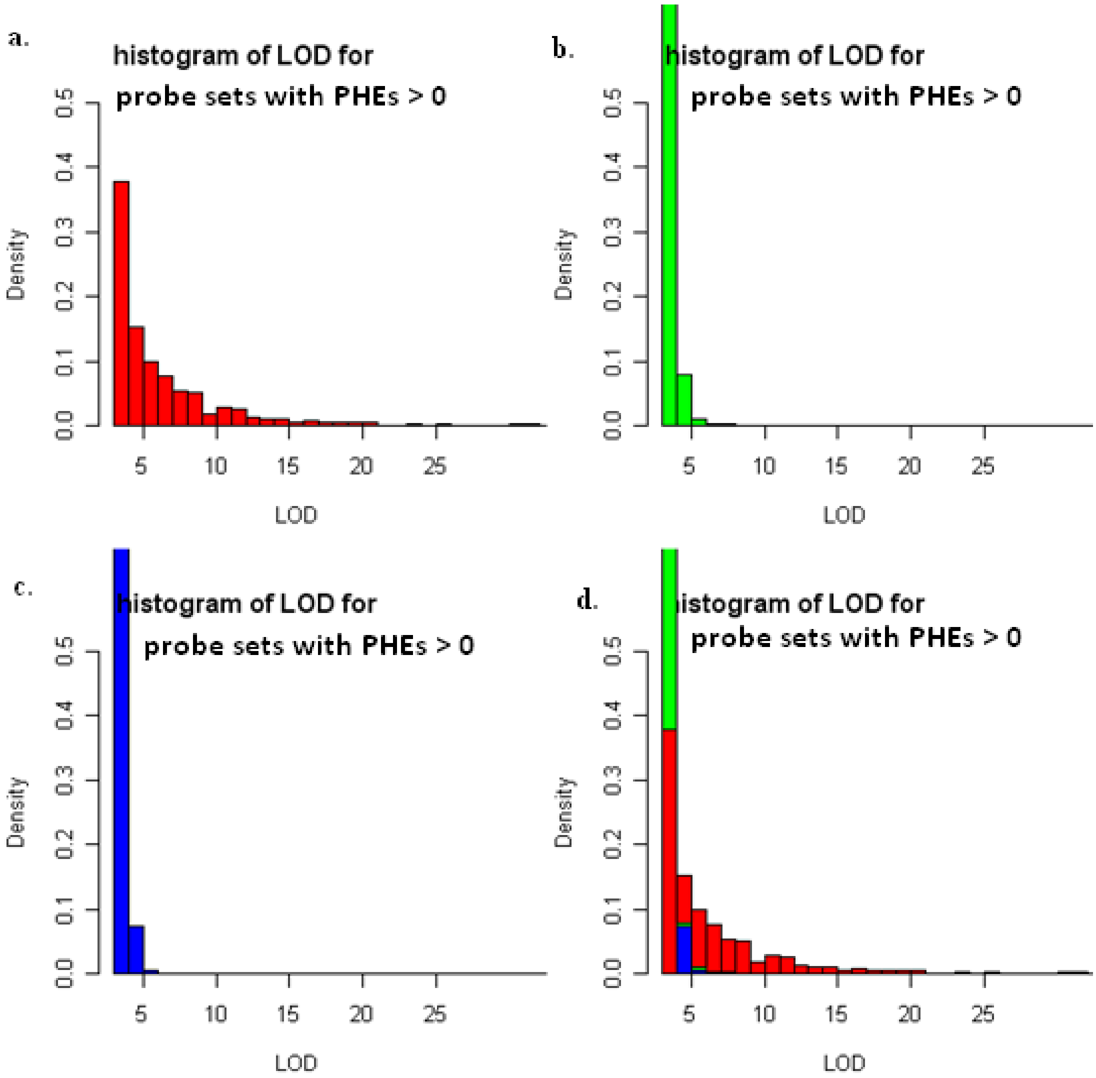

Figure 18 The density plot of LOD score for cis eSNPs <100 kb (red), cis eSNPs >100 kb (blue) and trans (blue) between X-limit (3, 25) and Y-limit (0, 0.5) for probe sets with PHEs>0.

a. The density plot of LOD score for cis eSNPs <100 kb.
b. The density plot of LOD score for cis eSNPs >100 kb.
c. The density plot of LOD score for trans.
d. The density plot of LOD score for cis eSNPs <100 kb, cis eSNPs >100 kb and trans.

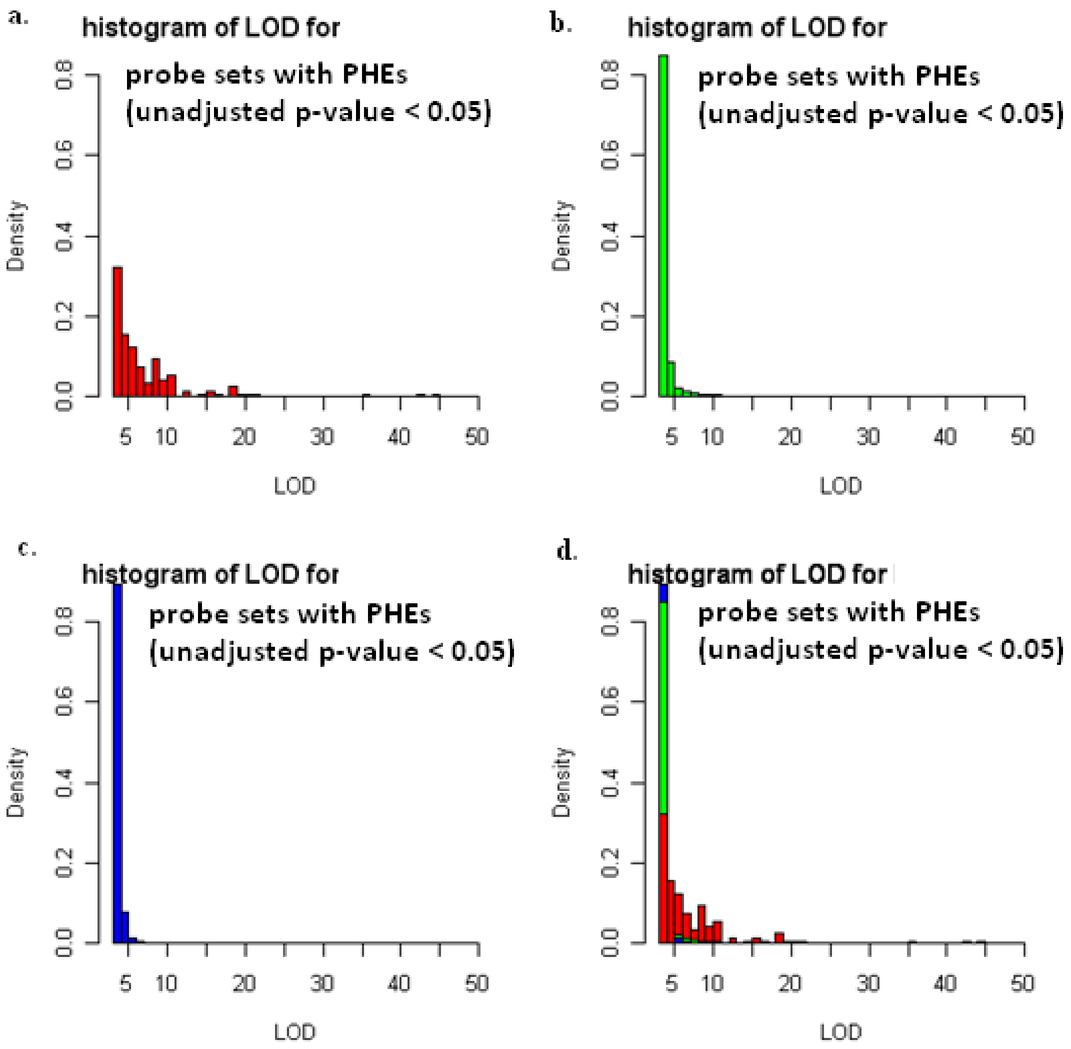

Figure 19 The density plot of LOD score for cis eSNPs <100 kb (red), cis eSNPs >100 kb (green) and trans (blue) between X-limit (3, 50) and Y-limit (0, 0.8) for probe sets with PHEs with unadjusted p-value <0.05.

a. The density plot of LOD score for cis eSNPs <100 kb.
b. The density plot of LOD score for cis eSNPs >100 kb.
c. The density plot of LOD score for trans.
d. The density plot of LOD score for cis eSNPs <100 kb, cis eSNPs >100 kb and trans.

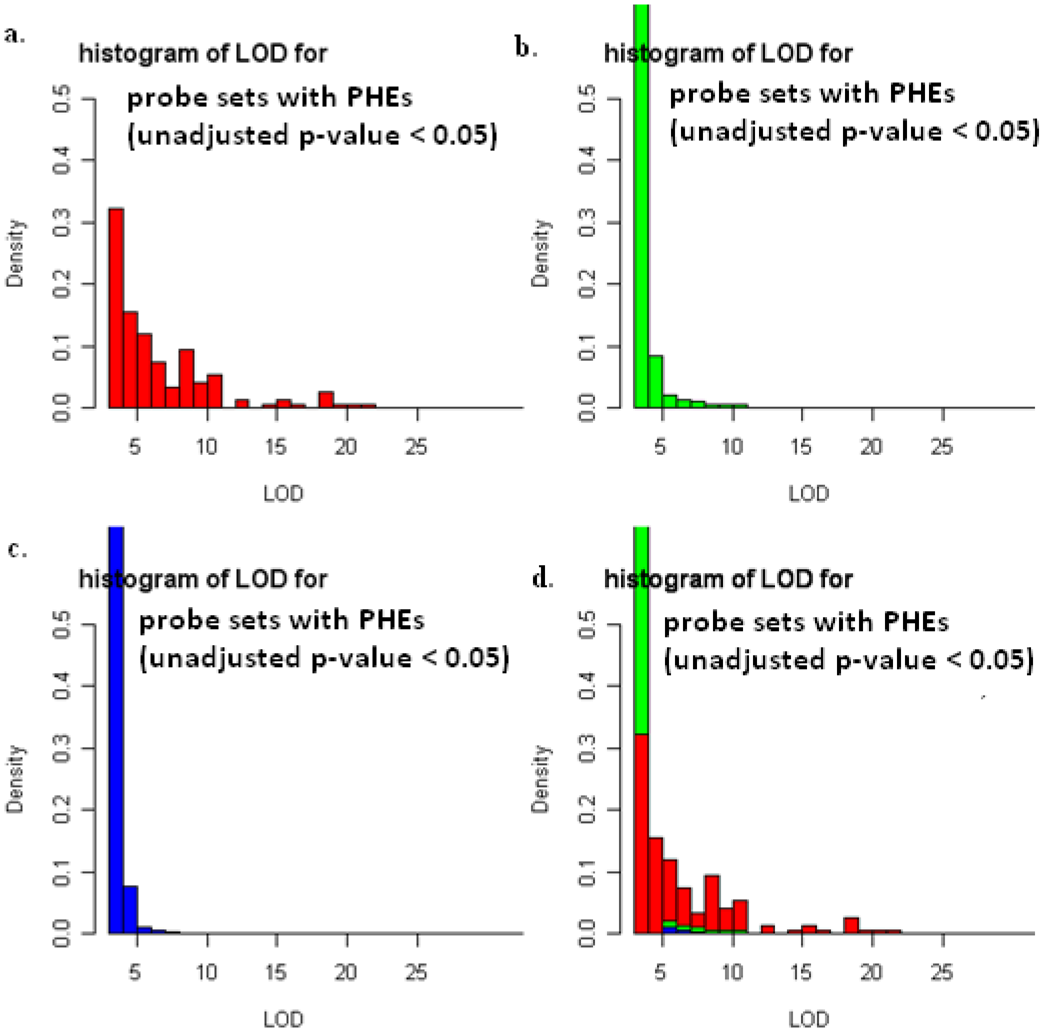

Figure 20 The density plot of LOD score for cis eSNPs <100 kb (red), cis eSNPs >100 kb (green) and trans (blue) between X-limit (3, 25) and Y-limit (0, 0.5) for probe sets with PHEs with unadjusted p-value <0.05.

a. The density plot of LOD score for cis eSNPs <100 kb.
b. The density plot of LOD score for cis eSNPs >100 kb.
c. The density plot of LOD score for trans.
d. The density plot of LOD score for cis eSNPs <100 kb, cis eSNPs >100 kb and trans.

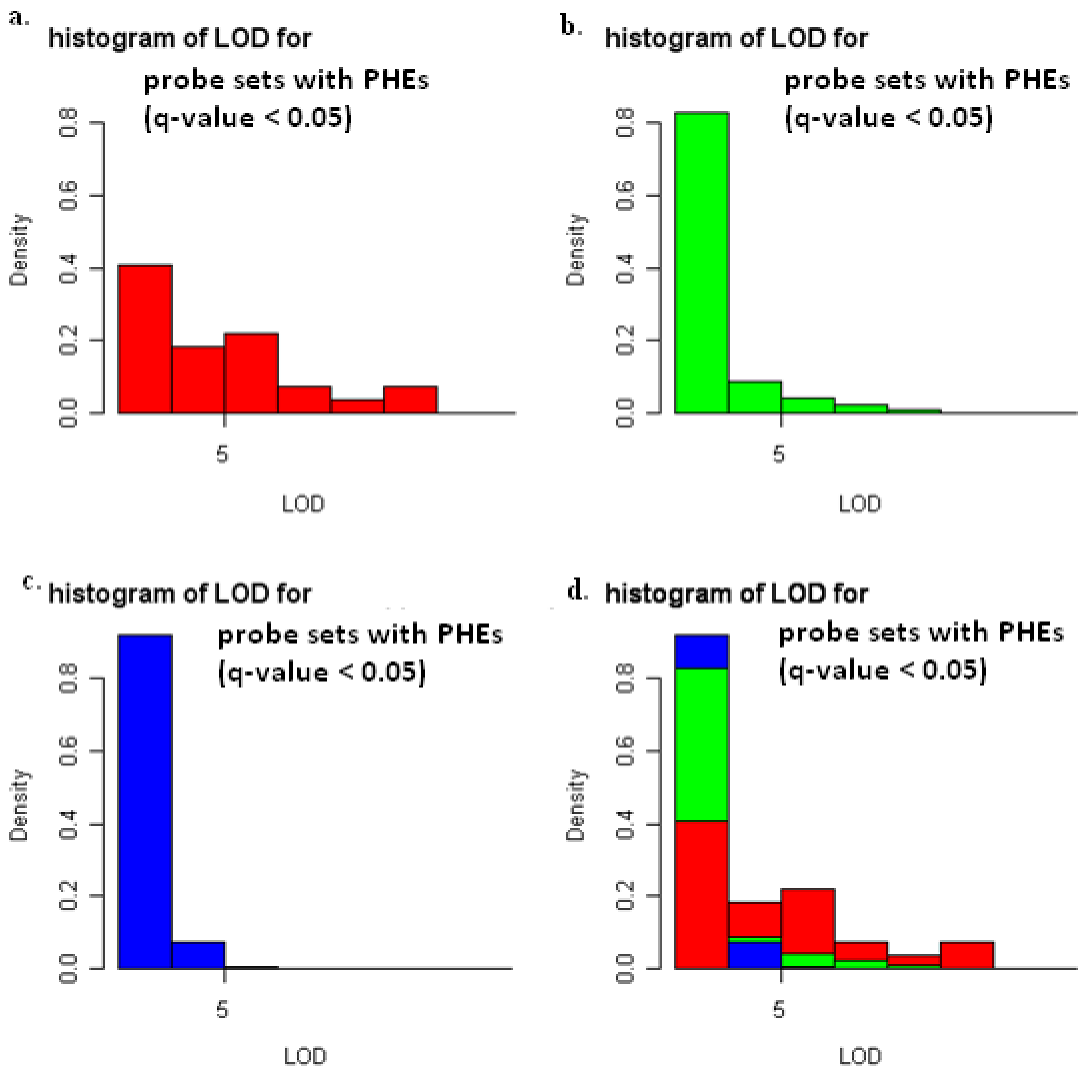

Figure 21 The density plot of LOD score for cis eSNPs <100 kb (red), cis eSNPs >100 kb (green) and trans (blue) between X-limit (3, 10) and Y-limit (0, 0.8) for probe sets with PHEs with q-value <0.05.

a. The density plot of LOD score for cis eSNPs <100 kb.
b. The density plot of LOD score for cis eSNPs >100 kb.
c. The density plot of LOD score for trans.
d. The density plot of LOD score for cis eSNPs <100 kb, cis eSNPs >100 kb and trans.

Table 5 The PHEs of probe sets with q-value < 0.05.

ProbeID	low	PHE	chr	Start	End	q-value
223949_at	0.210989	0.291867	21	42675053	42689269	1.54E-05
223952_x_at	0.14619	0.210583	2	1.7E+08	1.7E+08	0.000156
238076_at	0.18335	0.264808	1	1.51E+08	1.51E+08	0.000156
203627_at	0.176716	0.258253	15	97010284	97320636	0.000248
226841_at	0.127579	0.192336	11	58732558	58734705	0.001084
235835_at	0.153614	0.232668	19	63631810	63651932	0.00113
239938_x_at	0.131783	0.202743	5	88066154	88066571	0.001813
209772_s_at	0.101687	0.157232	24	19542359	19542777	0.001813
220177_s_at	0.100664	0.156129	21	42665067	42689269	0.001813
209071_s_at	0.128678	0.19985	1	1.6E+08	1.6E+08	0.001813
214254_at	0.150264	0.23343	23	1.51E+08	1.51E+08	0.001813
65472_at	0.151357	0.235664	2	85744034	85745751	0.001813
218390_s_at	0.148063	0.230837	10	1.2E+08	1.2E+08	0.001813
219799_s_at	0.130186	0.206607	2	1.7E+08	1.7E+08	0.003115
239199_at	0.145202	0.230571	17	579571	580102	0.003115
1570156_s_at	0.132495	0.211029	15	30849385	30850684	0.003243
202615_at	0.10535	0.169551	9	77562838	77875925	0.004321
222842_at	0.132372	0.214778	1	35942865	35991660	0.005079
1553296_at	0.093781	0.152233	3	1.02E+08	1.02E+08	0.005079
1552514_at	0.11781	0.192029	22	40719270	40748979	0.005473
1562098_at	0.120392	0.199272	12	98922496	98923786	0.008119
223948_s_at	0.101249	0.168402	21	42675053	42689269	0.00885
1570087_at	0.186288	0.315897	22	41759088	41772869	0.013429
239904_at	0.113997	0.193411	6	83136598	83137111	0.013429
239786_at	0.125965	0.214028	3	1.39E+08	1.39E+08	0.013429
226997_at	0.108961	0.18607	2	30775651	30778742	0.014572
242477_at	0.116073	0.198662	9	44414427	44415310	0.014812
238676_at	0.117222	0.203013	1	52266829	52267579	0.018643
226818_at	0.075209	0.130817	18	26899939	26901242	0.019778
266_s_at	0.063488	0.110892	24	19540661	19542776	0.020873
244783_at	0.098588	0.175553	9	1.01E+08	1.01E+08	0.029632
235667_at	0.098162	0.175046	10	15595956	15596395	0.029632
204837_at	0.107754	0.193471	9	6522466	6635650	0.03267
230282_at	0.095278	0.172182	16	79644602	79645066	0.035667
231237_x_at	-0.90096	-0.08877	6	29748238	29753058	0.03723

242142_at	0.105077	0.190967	6	1.14E+08	1.14E+08	0.03723
219067_s_at	-0.17093	-0.05114	15	73122679	73129132	0.040256
240089_at	0.104381	0.192788	4	1.21E+08	1.21E+08	0.04628

Table 6 The probe sets with PHEs > 0.2.

ProbeID	PHE	var	low	chr
1570087_at	0.3159	0.0062	0.1863	5
223949_at	0.2919	0.0024	0.2110	21
1559680_at	0.2701	0.0074	0.1288	2
238076_at	0.2648	0.0025	0.1834	1
203627_at	0.2583	0.0025	0.1767	15
65472_at	0.2357	0.0026	0.1514	2
224459_at	0.2341	0.0148	0.0342	14
214254_at	0.2334	0.0026	0.1503	23
235835_at	0.2327	0.0023	0.1536	19
218390_s_at	0.2308	0.0025	0.1481	10
239199_at	0.2306	0.0027	0.1452	17
216407_at	0.2275	0.0078	0.0823	16
230717_at	0.2255	0.0138	0.0326	9
222842_at	0.2148	0.0025	0.1324	1
220978_at	0.2146	0.0101	0.0491	17
239786_at	0.2140	0.0029	0.1260	12
1570156_s_at	0.2110	0.0023	0.1325	15
223952_x_at	0.2106	0.0015	0.1462	2
1563113_at	0.2069	0.0077	0.0623	1
219799_s_at	0.2066	0.0022	0.1302	2
1556538_at	0.2038	0.0134	0.0131	3
238676_at	0.2030	0.0027	0.1172	21
239938_x_at	0.2027	0.0019	0.1318	5
1557548_at	0.2018	0.0073	0.0614	10

Table 7 Genes of significant SNPs (LOD>6) overlap with asthma or atopy genes

SNP	chr	LOD	Gene
rs2844484	6	6.146	LTA
rs2239704	6	6.403	LTA
rs1041981	6	7.141	LTA
rs10776482	4	7.124	TLR10
rs4129009	4	13.141	TLR10
rs10776483	4	7.144	TLR10
rs11096955	4	6.049	TLR10
rs11096956	4	9.563	TLR10
rs3024498	1	8.219	IL10
rs3024496	1	7.823	IL10
rs1518111	1	9.300	IL10
rs3024490	1	7.265	IL10
rs1800872	1	8.131	IL10
rs1800896	1	7.473	IL10

Table 8 The probe set with max eSNPs' LOD >6 (PHEs with q-value < 0.05).

probe set ID	PHE	hertibility	q-value(SAM)	chr	cis eSNPs <100kb	cis eSNPs >100kb	trans
1558102_at	0.156884	0.275246	0.515981	15	5	7	0

Table 9 Cis eSNPs < 100 kb (PHEs with q-value < 0.05).

SNP	chr	LOD
rs7172665	15	8.28
rs7177599	15	8.03
rs8023669	15	7.83
rs8035001	15	6.26
rs11259964	15	6.40

Table 10 Cis eSNPs > 100 kb (PHEs with q-value <0.05).

SNP	chr	LOD
rs1877240	15	6.630
rs8042254	15	6.430
rs6603044	15	7.910
rs8038619	15	12.53
rs8040998	15	7.720
rs4386103	15	6.230
rs8033380	15	6.750

Appendix I

Asthma genes (review papers from 2003, 2006, 2008)											
AACT(SERPINA3)	CCL11	CMA1	DCNP1	GATA3	HLA	IL15	IL5RA	MRP1	PTGER3	TBXA2R	VDR
ACE	CCL2	COX2	DEFB1	GCLM	HNMT	IL16	IL8	MUC7	PTGIR	TCRA/D	
ACP1	CCL24	CRHR1	DPP10	GPRPA	ICOS	IL17F	IL8RA	NAT2	RANTES	TGFB1	
ADAM33	CCL26	CRTH2	ECP	GSTM1	IFNG	IL18	IRF1	NOD1	SCCE	TGFB2	
ADRB2	CCL5	CSF2	EDN1	GSTP1	IFNGR1	IL1A	IRF2	NOS1	SDF1	TIMP1	
AGT	CCR3	CSTA	EDNRA	GSTT1	IFNGR2	IL1B	ITGB3	NOS2A	SELP	TLR10	
AICDA	CCR5	CTLA4	EOTAXIN1	HAVCR1	IGHG	IL1RL1	KCNS3	NOS3	SOCS1	TLR2	
ALOX5	CD14	CXCL12	EOTAXIN2	HAVCR2	IKAP	IL1RN	LTA	ORMDL3	SPINK5	TLR4	
CHIA	CD40	CXCR3	EP2	HLADPB1	IL10	IL27	LTA	PAFAH	STAT3	TLR6	
C3	CD86	CYFIP2	FCER1B	HLA-DQA1	IL12B	IL3	LTC4S	PGDS	STAT4	TLR9	
C3AR1	CFTR	CYSLTR1	FCER2	HLA-DQB1	IL12RB1	IL4	MCP1	PHF11	STAT6	TNF	
C5	CHRM3	CYSLTR2	FLAP	HLADRB1	IL13	IL4RA	MIF	PTGDR	TAP1	UGRP1	
CC16/CC10	CLCA1	DAP3	FLG	HLA-G	IL13RA1	IL5	MMP9	PTGER2	TBX21	VCAM1	

Appendix II

25 genes that have been associated with asthma or atopy phenotype in >6 populations

Gene	Related Gene								
HLA-DRB1	TNF	IFNG	RA	MYLK	MLCK	HLA-DPB1	DRB1	HLA-A	LOC642072
	HLA-B	HLA-C	INFRSF10A	INFRSF10B					
IL13	IL4	IL5	IFNG	TNF	IL10				
IL4	IL13	IFNG	IL5	IL10	TNF				
IL10	IL4	CD4	IL5	CD8A	IFNG	TNF			
IL4RA	IL5	IL2	IL10	IL4	CD4	ISG20	IL13	IFNG	IL2RA
LTA	IL4	IL1B	IFNG	IL1A	TNF	IL10			
LTC4S	ALOX5	LTA4H	STK32C	GSTA1	MGST2	YWHAZ	CYSLTR1	GSTA2	CYSLTR2
	SEC23IP	PGCP							

NOS1	CALM3	CALM1	NOS3	NOS2A	NANOS1	CALM2	TNF		
CCL5	EEF1A1	CCR5	CCL2	CCL4L1	CCR2	CCL4L2	CCL3	CXCL9	IFNG
	TRIM24	CXCL10	CCL4	TNF					
SPINK5	NETS	KLK6	PLG	DSG1	KLK1	KLK5	SPINK5L3	DSP	SERPINA13
	KLK7	DSG3							
STAT6	JUN	STK32C	IL5	IL10	IL4	CD4	IL13	CD8A	IFNG
	FAM48A	TNF							
TBXA2R	PTGER4	PTGER3	STK32C	SAC	YWHAZ	PTGIR	PTGER1	PTGS2	PTGDS
	ADCY7	INS	SEC23IP						
TGFB1	PLAT	PLG	NUDT6	IL10	C20orf181	IL4	FGF13	IFNG	TNF
	FGF2								
TNF	IL1A	MAPK14	IL10	IL4	IL1B	IFNG	MAPK1	AHSA1	
CC16/CC10	PLA2G4A	TFF1	STK32C	YWHAZ	TFF2	SFTPB	SFTPA2	SEC23IP	TFF3
	SFTPC								
CD14	ITGB2	TLR4	IL10	CD4	ITGAM	LTBR	NDUFA2	CD8A	IFNG
	TLR2	TNF							
CTLA4	CELIAC3	CD40	CD80	CD4	CD28	CD8A	CD86	IFNG	
FCER1B	DNAH8	IGER	IL10	PTPRC	PHF11	GCG	IL4	PPY	CANT1
	IFNG	INS							
NOD1	CCK	TLR4	ACUG	IL10	GAST	NOD2	RIPK2	IFNG	TLR2
	TNF								
HLA-DQB1	HLA-DQA1	DRB1	IDDM2	IDDM1	HLA-DRB1	INS	HLA-DPB1	RA	
GRPA(AAA1)	PSORS#	DNAH8	KRT5	HLA-DRB4	SLC7A10	SLC3A2	ATP5E	SLC36A1	KRT19
	PSORS1	KRT14	HLA-C	GJA8	CALM2				
GSTM1	SLC45A2	GSTP1	GSTM2	GSR	GSPT2	GSPT1	CYP1A2	CYP1A1	
GSTP1	CYP1A2	CYP1A1	G6PD	XDH	GSTA1	GSR	EPHX1	CAT	GSTA2
	SLC45A2								
ADAM33	MS4A2	DPP10	IL5	IGER	IL10	PHF11	IL4	NPS	IL13
	IFNG	NPSR1							
ADRB2	INSR	ADRB1	PDE4B	STK32C	SAC	GCG	ADCY7	ADRB3	INS
	SEC23IP								