Published in final edited form as:
Lect Notes Comput Sci. 2009 ; 5483: 1–12. doi:10.1007/978-3-642-01184-9_1.

Association Study between Gene Expression and Multiple Relevant Phenotypes with Cluster Analysis

Zhenyu Jia[*],
Department of Pathology and Laboratory Medicine, University of California, Irvine, CA 92697, USA, zjia@uci.edu

Yipeng Wang,
The Sidney Kimmel Cancer Center, San Diego, CA 92121, USA, ywang@skcc.org

Kai Ye,
European Bioinformatics Institute, Hinxton, Cambridge, CB10 1SD, United Kingdom, kye@ebi.ac.uk

Qilan Li,
Division of Medicinal Chemistry, Leiden University, 2300 RA Leiden, The Netherlands, q.li@lacdr.leidenuniv.nl

Sha Tang,
Department of Pediatrics, University of California, Irvine, CA 92697, USA, shat@uci.edu

Shizhong Xu, and
Department of Botany and Plant Sciences, University of California, Riverside, CA 92521, USA, shxu@ucr.edu

Dan Mercola
Department of Pathology and Laboratory Medicine, University of California, Irvine, CA 92697, USA, dmercola@uci.edu

Abstract

A complex disease is usually characterized by a few relevant disease phenotypes which are dictated by complex genetical factors through different biological pathways. These pathways are very likely to overlap and interact with one another leading to a more intricate network. Identification of genes that are associated with these phenotypes will help understand the mechanism of the disease development in a comprehensive manner. However, no analytical model has been reported to deal with multiple phenotypes simultaneously in gene-phenotype association study. Typically, a phenotype is inquired at one time. The conclusion is then made simply by fusing the results from individual analysis based on single phenotype. We believe that the certain information among phenotypes may be lost by not analyzing the phenotypes jointly. In current study, we proposed to investigate the associations between expressed genes and multiple phenotypes with a single statistics model. The relationship between gene expression level and phenotypes is described by a multiple linear regression equation. Each regression coefficient, representing gene-phenotype(s) association strength, is assumed to be sampled from a mixture of two normal distributions. The two normal components are used to model the behaviors of phenotype(s)-relevant genes and phenotype(s)-irrelevant genes, respectively. The conclusive classification of coefficients determines the association

*To whom all correspondence should be addressed.

status between genes and phenotypes. The new method is demonstrated by simulated study as well as a real data analysis.

Keywords

Association; classification; gene expression; multiple phenotypes

1 Introduction

Intensive efforts have been tried to identify genes that are associated with disease phenotype of interest [1,2]. Sorting out the phenotype associated genes will help understand the mechanism of the disease and develop efficient treatment to the disease. However, a complex disease is usually characterized by more than one phenotypes reflecting different facets of the disease. For example, Mini-Mental State Examination (MMSE) and Neurofibrillary Tangle count (NFT) are often used for quantifying the severity of the Alzheimer's Disease. Another example is that obesity is commonly defined as a body mass index (BMI, weight divided by height squared) of $30 kg/m^2$ or higher. However, other indices, such as waist circumference, waist hip ratio and body fat percent, can also be applied as obesity indicators. Some of the disease phenotypes may be well correlated to one another but are not necessarily controlled by the same biological pathway. The gene networks involved in the development of the phenotypes may overlap and interact via the common genes shared by those networks. It is very useful to uncover genes representing the involved pathways as well as the cross-talk between those pathways.

The typical approach is to find the genes that are associated with individual phenotype first, and then combine the findings based on separate analysis. Genes shared by different phenotypes can be regarded as the 'nodes' between the participating pathways. Several methods can be used for the single-trait association analysis. The most simple way is to calculate the Pearson's correlation for each gene and the phenotype [1,3]. Genes are then sorted on the basis of the magnitude of the correlation coefficients. The genes on the top of the list are claimed to be associated with the phenotype. [4] developed a simple linear regression model to detect genes that are related to the phenotype. Subsequently, [5] and [6] enhanced the prototype to take into account the non-linear association between the gene expression levels and the phenotypic values. In those model-based methods, the gene expression level is described as a linear function of phenotypic value. The significance level of the regression coefficient is suggestive of the strength of the association between gene and phenotype. Classification approach is then utilized to cluster genes based on the magnitude of the regression coefficients. With the linear-model-based method, the association study has been turned into a classification problem. To our best knowledge, no analytical model has been reported to deal with multiple phenotypes simultaneously in such association study. We consider that, to some extent, information among phenotypes may be Phenotype Association Study of Genomic Data 3 lost in single-trait method since phenotypes are analyzed separately, and such missing information is important for inferring the interaction between relevant phenotypes.

In current study, we developed an advanced model which analyzes the association between genes and multiple phenotypes jointly. The expression level of a gene is described as a linear regression function of the phenotypes and their interactions. The gene specific regression coefficient, either main effect or interaction, is assumed to be sampled from a mixture of two normal distributions. One normal component has mean zero and a very tiny variance; while the other normal component also has mean zero but a much larger variance. If the coefficient is from the first normal component, then its size is close to zero and there is no indication of association; otherwise, if the coefficient is from the second normal component, the size of the

coefficient is nontrivial suggesting the association between the gene and the phenotype(s). Such mixture model setting originated from the spike and slab distribution proposed by [7]. We compared the new method with other single-trait approaches by synthesized data as well as real data generated from microarray experiment. The new method appeared to be more desirable for gene-phenotype association study.

2 Methods

2.1 Linear Regression Setting and Bayesian Modelling

Let Z_{js} be the standardized phenotypic value for trait j and subject s, where $j = 1, \ldots, p$ and $s = 1, \ldots, n$. Here all the phenotypes are normalized under the same scale, *i.e.*, [−1, 1]. For example, suppose $\Psi_j = [\Psi_{j1}, \ldots, \Psi_{jn}]$ is the original measurements for trait j, and if Ψ_{js} is a continuous variable, then

$$Z_j=[Z_{j1},\ldots,Z_{jn}]=2\times\frac{\Psi_j-\mathrm{Min}(\Psi_j)}{\mathrm{Max}(\Psi_j)-\mathrm{Min}(\Psi_j)}-1.$$

If Ψ_{js} is a categorical phenotype, we use $Z_{js} = \{-1, 1\}$ for binary variable or $Z_{js} = \{-1, 0, 1\}$ for trinary variable. In current study, only binary or trinary variable are considered for simplicity. Let $Y_i = [Y_{i1}, \ldots, Y_{in}]$ be the expression levels of gene i, where $i = 1, \ldots, m$. We may use following linear model to describe the relationship between the gene expression and phenotypes,

$$Y_i=b_{i0}+\sum_{j=1}^{p}Z_jb_{\mathrm{ij}}+e_i, \tag{1}$$

where b_{i0} is the intercept, b_{ij} is the regression coefficient for Z_j, and $e_i = [e_{i1}, \ldots, e_{in}]$ is random error with multivariate normal distribution $N(\mathbf{0}, I\sigma^2)$. Note that the number of genes is usually much greater than that of phenotypes. So, it is plausible to regress gene expression on phenotypes to reduce the dimension of parameters. If the interactions between the phenotypes are also considered, model 1 can be rewritten as

$$Y_i=b_{i0}+\sum_{j=1}^{p}Z_jb_{\mathrm{ij}}+\sum_{j=1}^{p-1}\sum_{k>j}^{p}Z_jZ_kb_{\mathrm{ijk}}+e_i. \tag{2}$$

For simplicity, we only consider the 2-way interaction in current study. If we treat the interaction terms as ordinary regression terms, there will be a total of $(p^2 + p)/2$ regression terms in the linear model. We can further rewrite model 2 as

$$Y_i=\beta_{i0}+\sum_{l=1}^{(p^2+p)/2}X_l\beta_{\mathrm{il}}+e_i, \tag{3}$$

where β's represent b series, and X's represent Z series. Note that β_0 is irrelevant to the association study. It can be simply removed from model via a linear contrasting scheme, *i.e.*,

$$y_i = \sum_{l=1}^{(p^2+p)/2} x_l \beta_{il} + \varepsilon_i, \tag{4}$$

where $\sum_{s=1}^{n} y_{is} = 0$ and $\sum_{s=1}^{n} x_{js} = 0$. After linear contrasting transformation, the n pieces of information for each gene are no longer independent to one another because the constraint of "sum-to-zero" has been imposed to y_i and x_l. Therefore, the last element of vector y_i should be removed and y_i becomes an $N \times 1$ vector for $N = n - 1$. Accordingly, x_l becomes an $N \times 1$ vector, and ε_i is now $N(\mathbf{0}, R\sigma^2)$ distributed with a known $N \times N$ positive definite matrix R (see [5] for more details).

To simplify the presentation, we use the following notation to express different probability densities throughout the current study:

$$p(\text{variable} \mid \text{parameter list}) = \text{DensityName}(\text{variable}; \text{parameter list}).$$

We assume that each regression coefficient β_{ij} in model 4 is sampled from the following two-normal-components-mixture distribution:

$$p(\beta_{il} \mid \eta_{il}, \sigma_1^2) = (1 - \eta_{il})\text{Normal}(\beta_{il}; 0, \delta) + \eta_{il}\text{Normal}(\beta_{il}; 0, \sigma_1^2), \tag{5}$$

where the two normal distributions are both centered at zero but have different levels of spread. Such mixture setting was originally suggested by [8]. We use the first normal distribution to model trivial effects by enforcing a fixed tiny variance $\delta = 10^{-2}$; while, for the second normal component, a larger variance σ_l^2 is utilized to govern nonzero effects. The unobserved variable $\eta_{il} = \{0, 1\}$ is used for indicating whether β_{il} is sampled from $N(0, \delta)$ or $N(0, \sigma_l^2)$. Our goal is to infer the posterior distribution of η_{il} and estimate the likelihood of association. We further describe η_{il} by a Bernoulii distribution with probability ρ_l, denoted by

$$p(\eta_{il} \mid \rho_l) = \text{Bernoulii}(\eta_{il}; \rho_l). \tag{6}$$

The parameter ρ_l and its counterpart $1 - \rho_l$, which can be viewed as mixing proportion of the two-component mixture model, regulate the number of genes that are connected with the lth regression term. This parameter ρ_l is *per se* governed by a Dirichlet distribution, denoted by Dirichlet(ρ_l; 1, 1). The variance components of the hierarchical model are assigned with scaled inverse chi-square distributions, denoted by $\text{Inv} - \chi^2(\sigma_l^2; d_0, \omega_0)$. We chose $d_0 = 5$ and $\omega_0 = 50$ for σ_l^2 , and chose $d_0 = 0$ and $\omega_0 = 0$ for σ^2. The hyper-parameters were selected based on our previous experience of similar analyses.

2.2 Markov Chain Monte Carlo

The inference of the parameters of interest is accomplished by using MCMC sampling, which is a algorithm for sampling from probability distributions based on constructing a Markov chain that has the desired distribution as its equilibrium distribution. We initialize all the parameters at very first step. Then the probability distribution of any parameter given the other parameters is derived. The value of the parameter is then replaced by a sample drawn from the obtained distribution. Such process is repeated sequentially until all the parameters have been

updated. We call it a sweep if we finish updating all the parameters for one time. A Markov chain is constructed by sweeping for many thousands of times. The state of the chain after a large number of sweeps is then used as a sample from the desired distribution. The detailed sampling scheme for each variable is described as follows.

1. Variable η_{il} is drawn from Bernoulli(η_{il} ; π_{il}), where

$$\pi_{\mathrm{il}}=\frac{\rho_l \mathrm{N}(\beta_{\mathrm{il}};0,\sigma_l^2)}{\rho_l \mathrm{N}(\beta_{\mathrm{il}};0,\sigma_l^2)+(1-\rho_l)\mathrm{N}(\beta_{\mathrm{il}};0,\delta)} \tag{7}$$

2. Variable β_{il} is drawn from $\mathrm{N}(\beta_{\mathrm{il}};\mu_\beta,\sigma_\beta^2)$, where

$$\mu_\beta=\left[x_l^T R^{-1} x_l+\frac{\sigma^2}{\eta_{\mathrm{il}}\sigma_l^2+(1-\eta_{\mathrm{il}})\delta}\right]^{-1} x_l^T R^{-1}\Delta y_i, \tag{8}$$

$$\sigma_\beta^2=\left[x_l^T R^{-1} x_l+\frac{\sigma^2}{\eta_{\mathrm{il}}\sigma_l^2+(1-\eta_{\mathrm{il}})\delta}\right]^{-1}\sigma^2 \tag{9}$$

and

$$\Delta y_i=y_i-\sum_{l\prime\neq l}^{(p^2+p)/2} x_{l\prime}\beta_{\mathrm{il}\prime}. \tag{10}$$

3. Sample σ_l^2 from

$$\mathrm{Inv}-\chi^2\left(\sigma_l^2;\sum_{i=1}^m \eta_{\mathrm{il}}+5,\sum_{i=1}^m \eta_{\mathrm{il}}\beta_{\mathrm{il}}^2+50\right).$$

4. Sample σ^2 from

$$\mathrm{Inv}-\chi^2\left(\sigma^2;\mathrm{mn},\sum_{i=1}^m (y_i-\sum_{l=1}^{(p^2+p)/2} x_l\beta_{\mathrm{il}})^T R^{-1}(y_i-\sum_{l=1}^{(p^2+p)/2} x_l\beta_{\mathrm{il}})\right).$$

5. Sample ρ_l from

$$\mathrm{Dirichlet}\left(\rho_l;\sum_{i=1}^m \eta_{\mathrm{il}}+1,m-\sum_{i=1}^m \eta_{\mathrm{il}}+1\right).$$

Usually it is not hard to construct a Markov Chain with the desired properties. The more difficult problem is to determine how many sweeps are needed to converge to the stationary distribution within an acceptable error. Diagnostic tool, such as the R package “coda” [9], can be used to estimate the required length of the chain for convergency. Once the chain has converged, the burn-in iterations from the beginning of the chain are discard; while for the remaining portion

of the chain, one sample in every 10 sweeps is saved to form a posterior sample for statistical inference.

3 Experimental Analysis

3.1 Simulated Study

In simulated study, we generated three phenotypes (I, II and III) for 100 individuals. Of the three phenotypes, the first two are correlated while the third is relatively irrelevant to the other two. The correlations between these pseudo-phenotypes are given in Figure 1(a). Such setting was meant to mimic the situation we met with in real data analysis (see next subsection). We also generated 1000 genes for each individual based on the phenotypic values we simulated. Of the 1000 genes, 20 are associated with phenotype I, 20 are associated with phenotype II, and 20 are associated with phenotype III. Note that the genes associated with different phenotypes are not mutually exclusive. The relationship between the genes associated with the three phenotypes are shown in Figure 1(a). For simplicity, all main effects and interaction effects are set to 1, and the residual variance is set to 0.01.

We first used the new method (denoted by method 1) to analyze the simulated data. All of the 44 true associated genes were identified and placed in the right positions of Venn Diagram (Figure 1(b)). We then used single-trait methods of [4] (denoted by method 2) and [6] (denoted by method 3) to analyze the same data set. The results are shown in Figure 1(c) and Figure 1 (d), respectively. Only 41 true associated genes were identified by single-trait methods 2 and 3. Besides, many spurious associations have been claimed by methods 2 and 3. For example, no simulated genes are associated with all three phenotypes; however, methods 2 and 3 declared more than 10 genes shared by the three phenotypes. We define a linkage as a true gene-phenotype association, and a non-linkage if gene is not related to phenotype. Therefore, a total of 60 linkages and 2940 ($1000 \times 3 - 60$) non-linkages have been generated for the simulated study. Numerically, the empirical Type I error is defined as

$$\alpha = \frac{\text{Number of undetected linkages}}{\text{Total number of linkages}},$$

and empirical Type II error is defined as

$$\beta = \frac{\text{Number of declared non-linkages}}{\text{Total number of non-linkages}}.$$

The type I and type II error rates for three methods are listed in Table 1. The new method, which considers multiple traits jointly, appeared superior to the single-trait methods 2 and 3. For single-trait analysis, we have one more time proved that the model based on non-linear association outperforms the model of simple linear association. (The comparison between methods 2 and 3 was originally presented in [6]). We additionally noted that, for single-trait methods 2 and 3, many genes are in common for phenotypes I and II because these two phenotypes are highly correlated (cor = 0.35 and p value = 0.00038). Genes that are authentically linked to a phenotype are likely to be claimed being associated with another correlated phenotype in single-trait analysis. This reasoning can also be supported by the phenomena that the number of genes shared by phenotypes II and III tended to be larger than the number shared by phenotypes I and III and the correlation between phenotypes II and III (cor = −0.14) is slightly larger than that between phenotypes I and III (cor = −0.11).

We simulated 19 more data sets with the same setting and repeated the analysis by 19 times. The results are almost identical to the presented one (data not shown). We then varied the residual variance from 0.01 to 0.81, and used the new method to analyze the additional simulated data sets. The analytical results are given in Table 2. It seems that the type II error (β) inflates faster than type I error (α) as residual variance increases.

3.2 Analysis of Mice Data

To demonstrate the new method, we used a mice data generated for obesity study [2]. In this study, Affymetrix GeneChip Mouse Expression Array was used to survey the expression levels of more than 40,000 transcript for 60 ob/ob mice. A total of 25 obesity related phenotypes were collected for the experimental units. The data is publicly available at gene expression omnibus (GEO) with accession no. GSE3330. We first removed invariant genes to lessen the computational burden. About 5000 transcripts with variance greater than 0.05 were saved for further analysis. Similar pre-screening scheme has been used for array data analyses [10,11]. For the sake of convenience, we only considered three phenotypes that have been often examined in related researches. These three phenotypes were plasma glucose level (mg/dl), plasma insulin level (ng/ml) and body weight. The phenotypic data were collected at eight weeks of age. In this subsection, we only compared the new multiple-traits method (method 1) to single-trait method 3, since method 3 has been repeatedly verified being more efficient than other single-trait approaches, such as method 2. We used methods 1 and 3 to analyze this mice data and the results are shown in Figure 1(e) and Figure 1(f), respectively. Note that glucose level (Glu) is correlated with insulin level (Ins); while there is no obvious correlation between body weight (Wt) and these two traits. The three phenotypes for simulated study (see previous subsection) were generated in concordance to this observation. Comparing Figure 1 (e) with Figure 1(f), we found that a lot more genes were declared to be associated with both Glu and Ins by method 3, which is no wonder since these two phenotypes are well correlated. Similar false linkages due to correlation between phenotypes have already been noticed in simulated study. We compared the gene identity of genes identified by both methods (Table 3). The results from two analyses are essentially consistent.

4 Discussion

Gene-phenotype association analysis is a pilot study for disclosing the biological process of disease development. Typically, data on more than one trait are collected in disease-related studies. More often than not, these disease phenotypes are correlated indicating commonality of the pathways responsible for those phenotypes. However, the existing analytical tools solely deal with a phenotype at one time; thus, potential information shared by multiple phenotypes are bound to be lost. We consider that such missing information is important for inferring the interaction between correlated phenotypes and it can be picked up only through a joint analysis of multiple traits. For the first time, we proposed to analyze the associations between gene expressions and multiple phenotypes in a single statistical model. In simulated study, the new multiple-traits method appeared superior to the single-trait methods.

In real data analysis, more genes have been detected by method 3 [6] than the new method. This is because method 3 considers the gene-phenotype association beyond linearity; while, for simplicity in current study, only main effects (first order or linear association) and 2-way interaction are taken into account for the new method. However, the new method can easily extend to include higher order of main effects and interactions. The enhanced model should be more powerful for discovering relevant genes. We contrasted our analytical findings (by methods 1 and 3) with the obesity related genes obtained from NCBI. Only a small portion of identified genes have been previously reported, leaving a large number of newly detected genes worthy of a closer look in the future researches.

Mapping quantitative trait loci (QTL) has been widely used for inferring genomic loci that dictate phenotypes of interest [12,13,14,15,16,17,18,19]. However, QTL analysis are different from gene-phenotype association analysis discussed in the current study. Data for QTL analysis consist of phenotypic data and genotypic data (markers and their genotypes); while gene-phenotype association analysis involve gene expression data and phenotypic data. For quite a long time, QTL mapping also has been limited to single-trait scheme. Recently, the first multiple-traits QTL method was developed by using multivariate linear regression models [20]. Nevertheless, this multiple-traits QTL method can not be directly adopted by gene-phenotype association analysis. Usually, the dimension of gene expression data is dramatically larger than that of genotypic data. It seems impractical to regress phenotypes (at most of tens) on expression of many thousands of genes. Such model, even if it worked for a simple regression setting, would be very difficult to include higher order effects. It is natural to reverse the roles of phenotypes and genes by regressing gene expression on phenotypes. With the workable dimension of predictive variables, interactions between phenotypes or underlying pathways can be easily appreciated by adding in cross-product terms. On the other hand, clustering along gene coordinate allows effective information sharing between correlated genes.

If only gene expression data and genotypic data have been collected from biological experiment, one may carry out expression quantitative trait loci (eQTL) mapping to find likely loci that account for the variations of expression traits or gene expression [21,22,23]. Single-trait QTL approaches are typically utilized for eQTL analysis since the numbers of expression traits is far beyond the capacity of current multiple-traits analysis. Obviously, potential gene-gene correlations have been overlooked by analyzing each gene separately. To solve this problem, [24] and [10] developed mixture models to restore the lost information by clustering along gene coordinate. If phenotypic data, gene expression data and genotypic data are all available, we may conduct genetical mapping for expression traits and regular phenotypes jointly. Association between phenotype and expressed gene is determined if they are mapped to the same locus on the genome. In addition, gene-gene relationship, either *cis*- or *trans*-, will be clear at the same time.

Acknowledgment

This research was supported by the National Institute of Health SPECS Consortium grant CA 114810-02, and the Faculty Career Development Awards of UCI to ZJ.

References

1. Blalock EM, Geddes JW, Chen KC, Porter NM, Markesbery WR, Land-field PW. Incipient Alzheimer's disease: Microarray correlation analyses reveal major transcriptional and tumor suppressor responses. Proceedings of the National Academy of Sciences of the United States of America 2004;101:2173–2178. [PubMed: 14769913]
2. Lan H, Chen M, Flowers JB, Yandell BS, Stapleton DS, Mata CM, Mui ET, Flowers MT, Schueler KL, Manly KF, Williams RW, Kendziorski C, Attie AD. Combined expression trait correlations and expression quantitative trait locus mapping. Plos Genetics 2006;2:e6. [PubMed: 16424919]
3. van Bakel H, Strengman E, Wijmenga C, Holstege FCP. Gene expression profiling and phenotype analyses of s. cerevisiae in response to changing copper reveals six genes with new roles in copper and iron metabolism. Physiol. Genomics 2005;22:356–367. [PubMed: 15886332]
4. Jia Z, Xu S. Clustering expressed genes on the basis of their association with a quantitative phenotype. Genetical Research 2005;86:193–207. [PubMed: 16454859]
5. Qu Y, Xu SH. Quantitative trait associated microarray gene expression data analysis. Molecular Biology and Evolution 2006;23:1558–1573. [PubMed: 16731570]

6. Jia, Z.; Tang, S.; Mercola, D.; Xu, S. Detection of quantitative trait associated genes using cluster analysis. In: Marchiori, E.; Moore, JH., editors. EvoBIO 2008. LNCS. Vol. vol. 4973. Heidelberg: Springer; 2008. p. 83-94.
7. Mitchell TJ, Beauchamp JJ. Bayesian variable selection in linear regression. Journal of the American Statistical Association 1988;83:1023–1036.
8. George EI, Mcculloch RE. Variable selection via gibbs sampling. Journal of the American Statistical Association 1993;88:881–889.
9. Raftery AE, Lewis SM. One long run with diagnostics: Implementation strategies for markov chain monte carlo. Statistical Science 1992;7:493–497.
10. Jia Z, Xu S. Mapping quantitative trait loci for expression abundance. Genetics 2007;176:611–623. [PubMed: 17339210]
11. Ghazalpour A, Doss S, Zhang B, Wang S, Plaisier C, Castellanos R, Brozell A, Schadt EE, Drake TA, Lusis AJ, Horvath S. Integrating genetic and network analysis to characterize genes related to mouse weight. Plos Genetics 2006;2:1182–1192.
12. Lander ES, Botstein D. Mapping mendelian factors underlying quantitative traits using rflp linkage maps. Genetics 1989;121:185–199. [PubMed: 2563713]
13. Jiang CJ, Zeng ZB. Mapping quantitative trait loci with dominant and missing markers in various crosses from two inbred lines. Genetica 1997;101:47–58. [PubMed: 9465409]
14. Xu SZ, Yi NJ. Mixed model analysis of quantitative trait loci. Proceedings of the National Academy of Sciences of the United States of America 2000;97:14542–14547. [PubMed: 11114174]
15. Broman KW, Speed TR. A model selection approach for the identification of quantitative trait loci in experimental crosses. Journal of the Royal Statistical Society Series B-Statistical Methodology 2002;64:641–656.
16. Yi NJ, Xu SZ. Mapping quantitative trait loci with epistatic effects. Genetical Research 2002;79:185–198. [PubMed: 12073556]
17. Yi NJ, Xu SZ, Allison DB. Bayesian model choice and search strategies for mapping interacting quantitative trait loci. Genetics 2003;165:867–883. [PubMed: 14573494]
18. Rzhetsky A, Wajngurt D, Park N, Zheng T. Probing genetic overlap among complex human phenotypes. Proceedings of the National Academy of Sciences of the United States of America 2007;104:11694–11699. [PubMed: 17609372]
19. Oti M, Huynen MA, Brunner HG. Phenome connections. Trends in Genetics 2008;24:103–106. [PubMed: 18243400]
20. Banerjee S, Yandell BS, Yi N. Bayesian quantitative trait loci mapping for multiple traits. Genetics 2008;179:2275–2289. [PubMed: 18689903]
21. Schadt EE, Monks SA, Drake TA, Lusis AJ, Che N, Colinayo V, Ruff TG, Milligan SB, Lamb JR, Cavet G, Linsley PS, Mao M, Stoughton RB, Friend SH. Genetics of gene expression surveyed in maize, mouse and man. Nature 2003;422:297–302. [PubMed: 12646919]
22. Hubner N, Wallace CA, Zimdahl H, Petretto E, Schulz H, Maciver F, Mueller M, Hummel O, Monti J, Zidek V, Musilova A, Kren V, Causton H, Game L, Born G, Schmidt S, Muller A, Cook SA, Kurtz TW, Whittaker J, Pravenec M, Aitman TJ. Integrated transcriptional profiling and linkage analysis for identification of genes underlying disease. Nature Genetics 2005;37:243–253. [PubMed: 15711544]
23. Jansen RC, Nap JP. Genetical genomics: the added value from segregation. Trends in Genetics 2001;17:388–391. [PubMed: 11418218]
24. Kendziorski CM, Chen M, Yuan M, Lan H, Attie AD. Statistical methods for expression quantitative trait loci (eqtl) mapping. Biometrics 2006;62(1):19–27. [PubMed: 16542225]

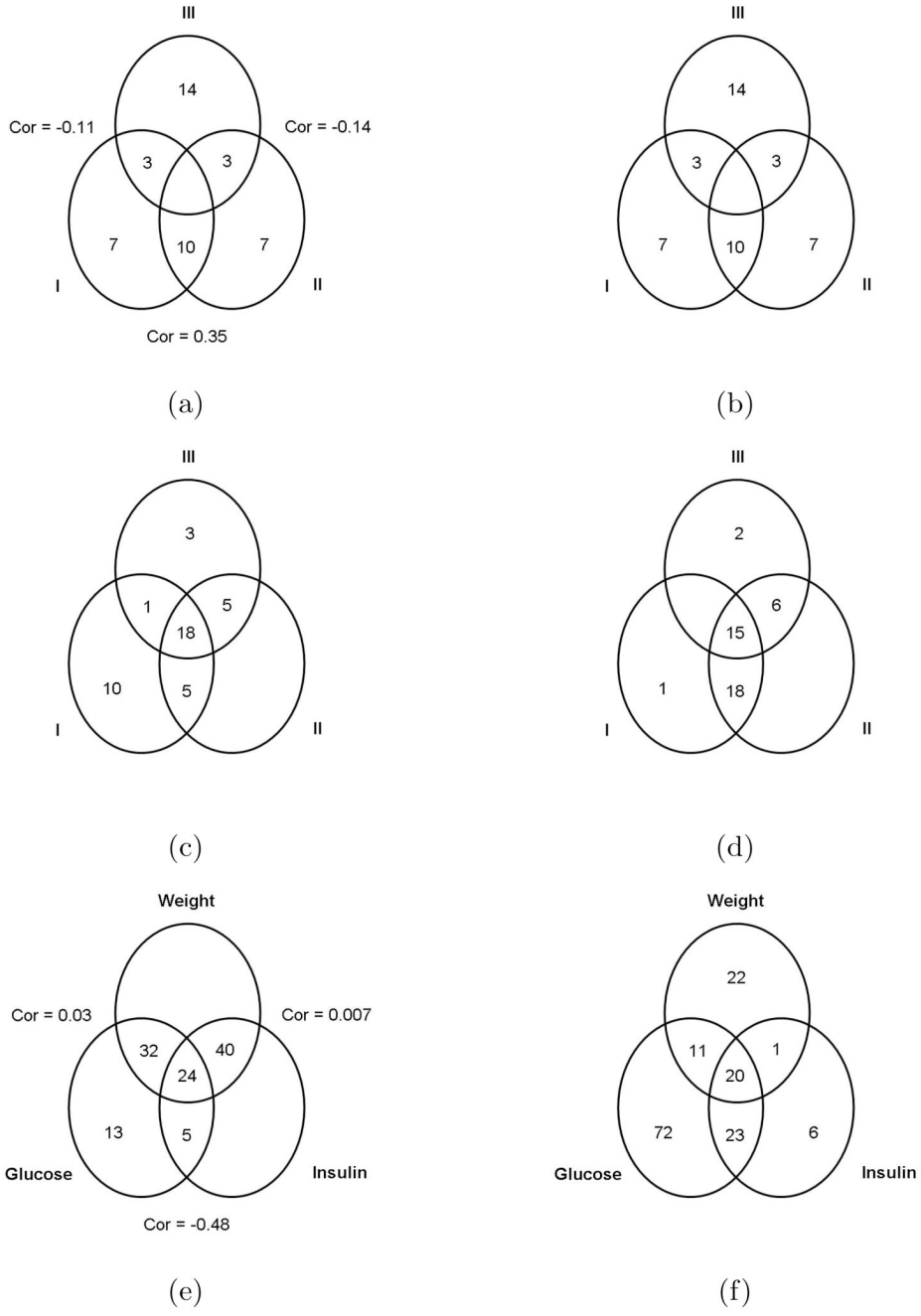

Fig. 1. Venn diagrams

(a) The numbers of genes related to the 3 simulated phenotypes and the pairwise correlations of the 3 simulated phenotypes. (b) The numbers of genes declared to be relevant to the 3 simulated phenotypes by method 1. (c) The numbers of genes declared to be relevant to the 3 simulated phenotypes by method 2. (d) The numbers of genes declared to be relevant to the 3 simulated phenotypes by method 3. (e) The numbers of genes declared to be relevant to the 3 obesity phenotypes by method 1 and the pairwise correlations of the 3 obesity phenotypes. (f) The numbers of genes declared to be relevant to the 3 obesity phenotypes by method 3.

Table 1

Type I and type II error rates for three analytical methods in simulated studies

	Method		
	1	**2**	**3**
α	0.000	0.300	0.100
β	0.000	0.016	0.014

Table 2

Type I and type II error rates for the new method when different residual variances were used for simulated studies

	Residual variance σ^2								
	0.01	**0.04**	**0.09**	**0.16**	**0.25**	**0.36**	**0.49**	**0.64**	**0.81**
α	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
β	0.00	0.00	0.00	0.00	0.00	0.11	0.26	0.43	0.58

Table 3

The numbers of genes identified by method 1, method 3 and both

	Method		
	1	**3**	**1&3**
Glu	74	126	60
Ins	69	50	19
Wt	96	54	33
All	114	155	79