Artificial Intelligence in Medicine xxx (2010) xxx–xxx

Contents lists available at ScienceDirect

Artificial Intelligence in Medicine

journal homepage: www.elsevier.com/locate/aiim

Identifying regulatory relationships among genomic loci, biological pathways, and disease

Jung Hoon Woo [a,b,1], Sung Bum Cho [a,1], Eunjee Lee [c], Ju Han Kim [a,d,*]

[a] Seoul National University Biomedical Informatics (SNUBI), Seoul National University College of Medicine, 28 Yongon-dong Chongno-gu, Seoul 110799, Republic of Korea
[b] Macrogen Inc., 60-24, Gasan-dong, Seoul 153-781, Republic of Korea
[c] Department of Biological Science, Columbia University, New York, NY 10027, United States
[d] Division of Biomedical Informatics, Seoul National University College of Medicine, 28 Yongon-dong Chongno-gu, Seoul 110799, Republic of Korea

ARTICLE INFO

Article history:
Received 15 March 2009
Received in revised form 18 September 2009
Accepted 8 March 2010

Keywords:
Genetical genomics
Differential allelic co-expression test
Principal component analysis
Integrative approach
Relationship among sequence level variation
Functional pathway and disease outcome

ABSTRACT

Objective: Elucidating genetic factors of complex diseases is one of the most important challenges in biomedical research. Recently, a genetical genomics approach of mapping genotype to transcripts has been used in complex disease analysis. This approach treats messenger ribonucleic acid (mRNA) expression as a quantitative trait and identifies putative regulatory loci for the expression of an individual gene. However, the single-gene approach could not detect single nucleotide polymorphisms (SNP's) associated with the concerted activity of multiple genes. Since complex diseases result from the interactions of multiple genes, it is important to consider associations between genotype and multiple gene expression.

Methods and materials: We developed the differential allelic co-expression (DACE) that identifies regulatory loci that affect the inter-correlation structure of multiple genes or a gene set. We applied DACE to two benchmark datasets: the normal human lymphoblastoid cell dataset and the glioblastoma dataset. These datasets consist of both SNPs and mRNA expression profiles for each sample. When analyzing the lymphoblastoid cell dataset, principal component analysis (PCA) was compared with the DACE test.

Results: While PCA identified associations found by single-gene analysis, the DACE test detected associations not identified by the single-gene approach. Using the DACE test, seven putative regulatory loci of immune-related pathways were identified in lymphoblastoid cells after controlling for family-wise error rate. In the glioblastoma dataset, DACE identified 4582 SNPs associated with six pathways. In 231 of the 4582 SNPs, patient survival length was correlated significantly with the SNP genotype. This finding suggests that our integrative approach may provide a biological explanation for the putative relationship between sequence level variation and disease outcome, via expression of a functional pathway.

Conclusion: The DACE test shows promise for finding regulatory relationships between a genomic locus and sets of genes which may be related to disease outcome.

1. Background

While many genomic data obtained by genome-wide association studies (GWAS) and gene expression microarray studies have been separately analyzed, researchers are now in need of developing integrative genomic analysis approaches. Genetical genomics, the study of the genetic basis of gene expression, is a newly introduced approach that treats messenger ribonucleic acid (mRNA) expression of a gene as a quantitative trait. While traditional genetics has focused on identifying loci linked to (or associated with) conventional traits such as Mendelian disease or common disease, genetical genomics has focused on identifying loci that regulate the mRNA expression of certain genes [1–3].

The general scheme in genetical genomics is quite similar to that of common microarray single-gene analysis. For each gene, a linkage or association test is performed on its expression level and several genetic markers to detect loci that regulate the gene's mRNA expression. The main assumption of the single-gene approach is that genes are expressed independently [4]. The complex functions of a living cell, however, are often carried out through the concerted activity of related genes [5]. For example,

* Corresponding author at: Seoul National University Biomedical Informatics, Seoul National University College of Medicine, Yongon-dong, Seoul 110-799, Republic of Korea. Tel.: +82 2 740 8320; fax: +82 2 742 5947.
E-mail addresses: hyde83@snu.ac.kr (J.H. Woo), csb1749@snu.ac.kr (S.B. Cho), el2380@columbia.edu (E. Lee), juhan@snu.ac.kr (J.H. Kim).
[1] These authors contributed equally to this paper.

0933-3657/$ – see front matter © 2010 Published by Elsevier B.V.
doi:10.1016/j.artmed.2010.03.001

genes that function in cell-signaling pathways act simultaneously rather than independently. Furthermore, certain pathway structures, such as cascade reactions, make identifying the regulatory elements of some of the pathway's genes difficult. This is because functional pathways consist of many sequential steps, and genes located further down the cascade are likely to be less affected by regulators that stimulate expression of pathway components at the top of the cascade. The single-gene approach is therefore not adequate to test genes when alterations in gene expression are modest.

With the data produced by Morley et al. [2], we focused on identifying regulatory loci for multiple genes, especially with regard to functional pathways, rather than for a single-gene. To this end, we applied two different approaches to the lymphoblastoid data sets. First, we applied principal component analysis (PCA), which has previously been used by Lan et al. [6] and Ghazalpour et al. [7]. They used PCA to reduce the dimension of the expression of multiple genes in a specific pathway and regarded a principal component (PC) as representative. Genomic loci associated with the PC were deemed to be regulatory loci of the pathway. Second, in our previous work [8], we hypothesized that the regulator of a gene set may affect not only the expression levels of the member genes but also the degree of their inter-correlation. We developed a new approach, the differential allelic co-expression (DACE) test, to identify genetic regulators of co-expression of a gene set. We applied both approaches to the data and compared the results. We also applied the DACE test to the human glioblastoma data sets. In the present study, we utilized a co-localizing strategy to identify single nucleotide polymorphisms (SNP's) associated with survival times of glioblastoma patients to identify functional pathways that connect sequence level variation with survival phenotype. Our approach is based on a two-step process: first, associations between SNPs and biological pathways are identified. Second, SNPs having a significant effect on survival times are selected from the SNPs that were significant in the first step. With our integrative approach, we identified not only the putative SNPs associated with disease survival but also the pathways providing biological explanation of the underlying process of SNP–survival association.

2. Methods

2.1. Centre d'Etude du Polymorphisme Humain (CEPH) datasets containing gene profiles

We selected gene expression and genotype data from the 56 independent individuals in the Genetic Analysis Workshop (GAW) 15 data set provided by Morley et al. [2]. We used genotype data of 2882 genome-wide SNPs and expression data of 8793 mRNA's across 56 samples for this analysis. mRNA expression was measured with Affymetrix Human Focus Arrays. We normalized the raw microarray data with the robust multichip average (RMA), comprising background adjustment, quantile normalization, and probe summarization.

2.2. Glioblastoma datasets

SNP and mRNA expression datasets of glioblastoma were downloaded from the Broad Institute website (www.broad.mit.edu). We selected 34 samples that were used in both the SNP chip and mRNA microarray experiments. The platforms of the SNP chip and mRNA microarray were the Affymetrix 100K SNP chip and U133A, respectively. In this analysis, any probes containing missing values (absent calls) in SNP dataset were excluded. In total, 52,885 SNP probes were used.

2.3. Mapping genes to pathways

We used pathway information to compile the gene set. These data were obtained from publicly available pathway resources, including KEGG [9], GenMAPP [10], and BioCarta (http://www.biocarta.com), for mapping genes to pathways. In the U133A platform, probes were mapped to 468 pathways. Among the pathways, we used 437 pathways having five or more probes.

2.4. SNP–pathway association analysis using principal component analysis

Principal component analysis was used to summarize the expression of multiple genes for the pathways. Since principal component 1 (PC1) explains most of the variance in expression levels within a pathway, PC1 for each pathway was extracted and used as a quantitative trait in the association test. The procedure of the association test was followed using simple linear regression.

2.5. SNP/pathway association analysis using the DACE test

We used the DACE test [8] to identify associations between a given SNP and a given gene set. It tests for differences in the structure of the correlation between multiple mRNA transcript levels that are associated with a SNP's genotype. Given a SNP, samples are grouped according to their genotype. First, for samples with the same genotype, we computed the Pearson correlation coefficients between the expression levels of all pairs of transcripts in a gene set. As the correlation coefficients are not normally distributed, the procedure includes "Fisher's z transformation". To test whether the SNP under study has significant effects on the levels of correlation among those genes, it adopts the general framework of a linear model.

2.6. SNP/single-gene association analysis

For comparison with set-wise approaches, we also tested the association between a gene's expression level and a SNP for all SNP/gene pairs. Simple linear regression was used for the association test. Expression values across 56 independent samples became the response variable for each expression trait, and individual SNP genotypes are the independent variables in the regression model. We applied this association test to only two groups of genes that were of specific interest. One group consisted of the members of six significant pathways identified by PCA. The other group consisted of members of seven significant pathways identified by DACE.

3. Results

3.1. SNP/pathway association test

Since principal component analysis (PCA) was conventionally used for reducing the dimensions of multiple gene expression, we applied both PCA and the DACE test (proposed in our previous paper) [8], on the lymphoblastoid dataset to assess the relevance of the two methodologies for identifying associations between SNPs and pathways.

We first extracted the first principal component (PC1) from the expression profiles of genes comprising a certain pathway. We regarded PC1 as a quantitative trait and performed an association test with each of the SNP markers. Six PC1s corresponding to six pathways showed significant association with at least one SNP after controlling for family-wise error rate (Table 1). For the purpose of comparing the PCA approach to conventional single-gene association analysis, we independently tested the association between genome-wide SNPs and the genes comprising the six

Please cite this article in press as: Woo JH, et al. Identifying regulatory relationships among genomic loci, biological pathways, and disease. Artif Intell Med (2010), doi:10.1016/j.artmed.2010.03.001

Table 1
Shared associations among pathways and major contributors.

Pathway[a]	Major contributor[b]	SNP	Source DB
Role of Parkin in the ubiquitin-proteasomal	SNCA	rs638113, rs1476049	BioCarta
Alkaloid biosynthesis II	ABP1	rs638113, rs703612, rs1476049	KEGG
Histidine metabolism	ABP1	rs638113, rs1476049	KEGG
Alpha-synuclein and Parkin-mediated proteolysis in Parkinson's disease ribosomal proteins	SNCA	rs638113, rs1478292, rs1472672, rs1476049, rs746101	BioCarta
Ribosomal proteins	RPS4Y1	rs530629	GenMAPP
Translation factors	EIF1AY	rs530629	GenMAPP

[a] Each pathway expression matrix reduces to principal component 1, which was used for the association test.
[b] Major contributor: a gene that gives the largest contribution to principal component 1.

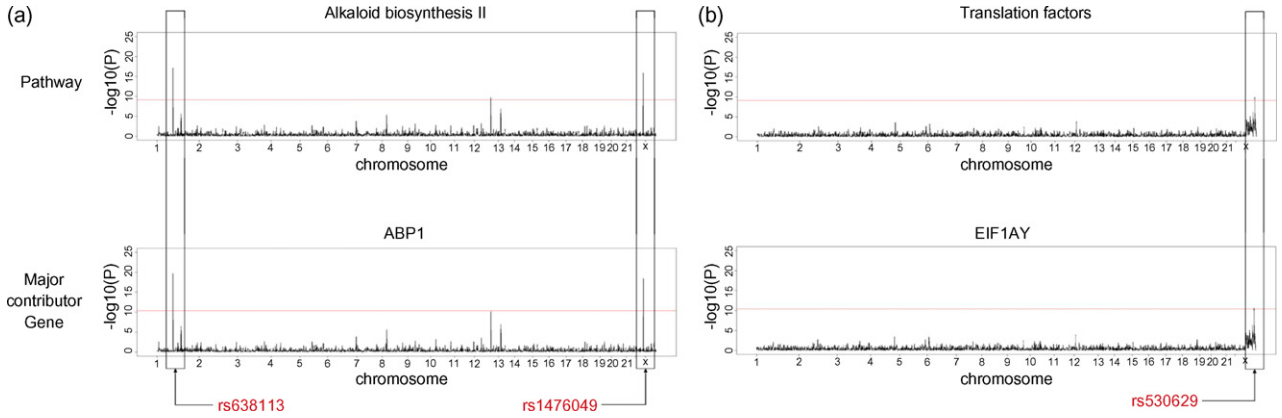

Fig. 1. Concordance of genome-wide association results between the first principal component of a pathway and its major contributor. (a) Principal components (PC1) extracted from both the *Alkaloid biosynthesis II* pathway and the *ABP1* gene revealed evidence of association with the same SNP's, *rs638113* and *rs1476049*. (b) The PC1s of the *Translation factors* pathway and the *EIF1AY* gene showed common association with the SNP *rs530629*.

significant pathways. Interestingly, some of the genes showed exactly the same significant association with the SNPs that PC1s showed. We found that these genes gave the largest contribution to the each of the six PC1s (we refer to these genes as major contributors). For instance, the PC1s of the *Alkaloid biosynthesis II* and *Translation factors* pathways showed almost the same *P*-value distributions as those of two major contributors, *ABP1* and *EIF1AY*, respectively (Fig. 1).

Secondly, we applied the DACE test to the same data set. Since RNA samples were extracted from lymphoblastoid cells for gene expression data, we chose seven immune system-related pathways for further analysis (Table 2). We independently tested the association between genome-wide SNPs and all genes comprising the seven significant pathways for comparison. Unlike PCA, the DACE test detected associations that were not identified by single-gene association analysis. For example, we found that the inflammatory response pathway showed significant evidence for association with a SNP, *rs1294028*, by the DACE test (Fig. 2(b)). By contrast, none of the 23 genes comprising this pathway showed evidence of association with *rs1294028* by single-gene association analysis. A heatmap of the gene expression matrix of the inflammatory response pathway supported this result. There is no significant change in expression level among three genotype groups when considering each gene independently (Fig. 2(c)). There is, however, an obvious alteration in correlation structure among genes in the pathway, and this observation demonstrated that *rs1294028* has significant association with the correlations among the 23 genes (Fig. 2(d)). The *rs1294028* polymorphism is located exactly in the intron region of the *splA/ryanodine receptor domain and SOCS box containing 1* (*SPSB1*) gene (Fig. 2(a)).

3.2. Identifying SNP/pathway/disease associations

Among 23,110,745 SNP–pathway pairs, 4582 SNP–pathway pairs were significant with Bonferroni's adjusted *P*-value (=2.16e−09). The 4582 pairs contained six different pathways, including *Aspirin Blocks Signaling pathway involved in platelet activation*, *Eicosanoid metabolism*, *G-protein coupled receptors*

Table 2
Pathways with significantly associated SNPs as indicated by the DACE test.

Pathway	SNP	*P*-value	Corrected *P*-value*	Source DB
IFN alpha signaling	rs1884910	7.77e−15	8.80e−09	BioCarta
Human Cytomegalovirus and MAP kinase	rs2135047	4.48e−14	5.07e−08	BioCarta
IL22 soluble receptor signaling	rs1889279	8.14e−13	9.22e−07	BioCarta
IL12 and Stat4 dependent signaling	rs213006	6.01e−13	6.81e−07	BioCarta
	rs1327532	7.18e−12	8.13e−06	
T cytotoxic cell surface molecules	rs1414944	1.09e−12	1.24e−06	BioCarta
Inflammatory response	rs1294028	1.72e−12	1.95e−06	GenMAPP
TACI and BCMA stimulation of B cell immune responses	rs2222976	1.70e−12	1.92e−06	BioCarta

* *P*-value adjusted by Bonferroni correction for multiple hypothesis testing.

Please cite this article in press as: Woo JH, et al. Identifying regulatory relationships among genomic loci, biological pathways, and disease. Artif Intell Med (2010), doi:10.1016/j.artmed.2010.03.001

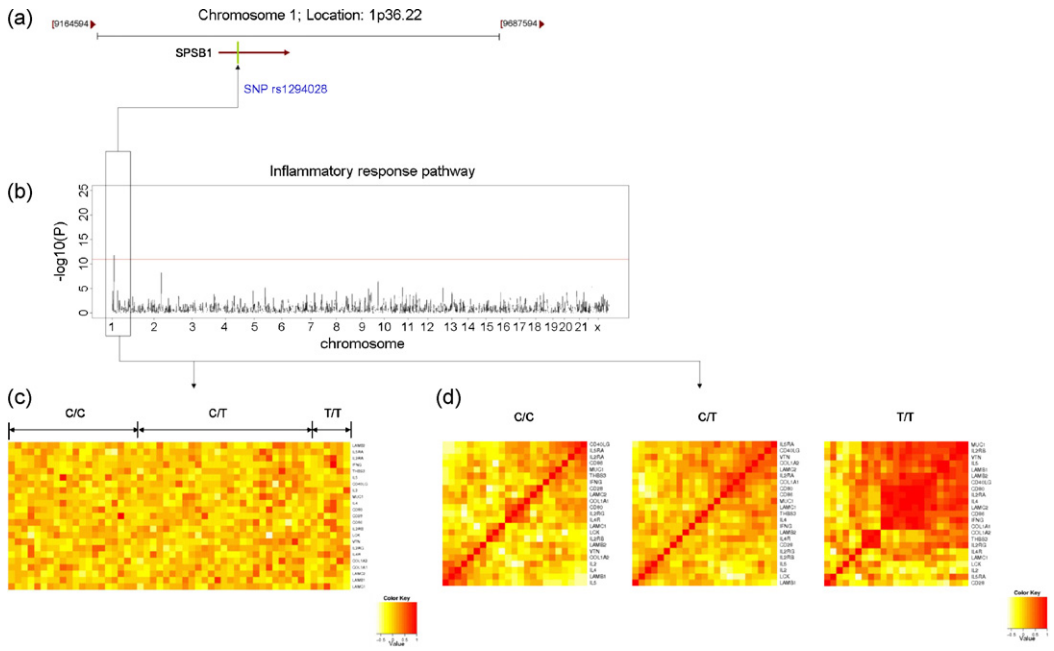

Fig. 2. Genome-wide DACE test results for the *Inflammatory response* pathway. (a) Physical map of rs129408 showing association with the *Inflammatory response* pathway (from NCBI Gene View). The polymorphism is located exactly in intron 1 of a gene named *SPRY domain-containing SOCS box protein* (*SSB-1*). (b) Genome-wide negative log 10 of *P*-value distribution for the *Inflammatory response* pathway. The horizontal red line is our threshold ($P < 7.430e{-}12$; Bonferroni corrected $P < 0.00001$) for evidence of significant association. (c) Heatmap for the gene expression matrix. There is no significant change in expression level among the three groups when each gene is considered independently. (d) Heatmaps for correlation matrices. The three heatmaps show significant changes in correlation tendencies of the *Inflammatory response* pathway for genotype differences in a given SNP.

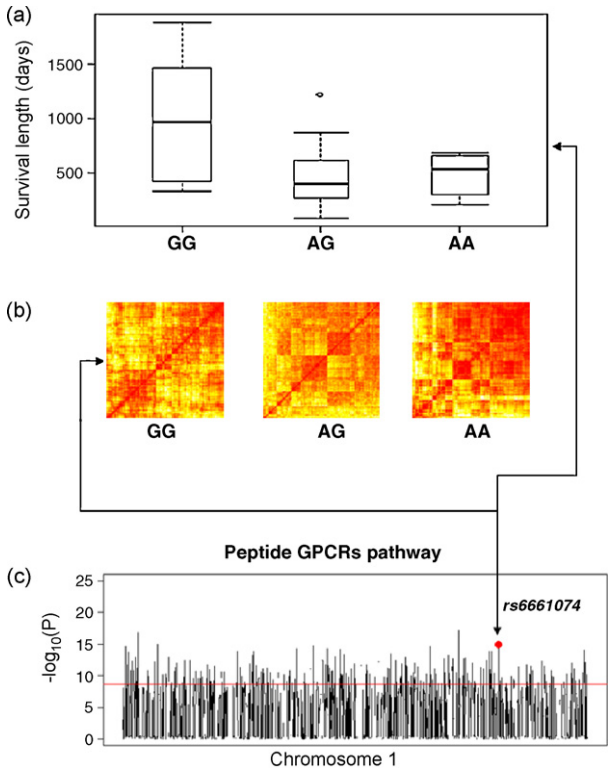

Fig. 3. Co-localization of results from the SNP–pathway association and the SNP–survival association. (a) Survival differences in patients with glioblastoma according to the genotype of *rs6661074*. A statistical difference in overall survival times across three genotypes was observed ($P = 6.99e{-}3$ by ANOVA). (b) Heatmaps of correlation matrices. There were also significant changes in correlation patterns of the *Peptide GPCRs* pathway for genotype differences in SNP *rs6661074*. (c) DACE test results on chromosome 1 for the *Peptide GPCRs* pathway. The horizontal red line is our threshold ($P = 2.16e{-}09$) and the red dot shows DACE results of *rs6661074* ($P = 1.34e{-}15$). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

(*GPCRs*) *class A rhodopsin-like*, *GPCRs other*, *Peptide GPCRs* and *Tryptophan metabolism* pathways. It is notable that the co-expression pattern of the *Peptide GPCRs* pathway significantly varied with 4566 SNPs. We then tested whether survival length varied according to the significant SNPs' genotypes. Since survival length was the only supported information, instead of the conventional log-rank test, we used ANOVA to compare survival lengths between different genotypes for each SNP. The ANOVA test of difference in survival length according to genotype identified 231 (of the 4582 previously identified SNPs) significant SNPs with survival time ($P < 0.05$). When the *P* values from the DACE and ANOVA tests were sorted, *rs666107* was the first SNP ranked within the top 100 in both tests (DACE *P*-value = 1.34e−15, ANOVA *P*-value = 6.99*e−3, Fig. 3). As the SNP/pathway association and SNP/disease survival association can be integrated by an anchoring common SNP, we describe the result as the 'co-localizing approach'. As shown in Fig. 3, both the intra-correlational structure of the *Peptide GPCR* pathway and survival lengths among glioblastoma patients varied significantly across the *rs666107* genotypes.

4. Discussion and conclusion

In the analysis of the lymphoblastoid dataset, we compared two genetical genetics approaches for identifying associations between SNPs and pathways. The first approach was principal component analysis. This approach was found to be insufficient because the pattern of variation of each first principal component was biased toward that of the specific gene with the largest contribution to that component. To make up for the shortcomings not only of PCA but also of conventional single-gene association analysis when attempting to identify the regulatory loci of functionally related genes, we also applied the DACE test. This test identified novel associations between transcripts and SNPs. For example, none of the 23 genes comprising the inflammatory response pathway associated with *rs1294028* by single-gene association analysis, while a significant difference in the correlation structure of this

Please cite this article in press as: Woo JH, et al. Identifying regulatory relationships among genomic loci, biological pathways, and disease. Artif Intell Med (2010), doi:10.1016/j.artmed.2010.03.001

pathway was seen in the results of the DACE test. Furthermore, we identified a link between an underlying biological mechanism of glioblastoma and its prognosis. For example, as shown in Fig. 3, among glioblastoma patients, those with the GG genotype at *rs6661074* showed significantly lower intra-correlation among genes involved in the *Peptide GPCRs* pathway, and also had better survival. Therefore, we conclude that a single nucleotide change at SNP *rs666107* might create differential interactions of the pathway and eventually result in changes in the survival time of glioblastoma patients. Even though an association between *rs666107* and glioblastoma survival has not yet been reported, the GPCRs pathway is already known to mediate the metastasis of several malignant tumors and to play a role in supporting glioblastoma cell survival and to promote their production [11]. Our results are the first to suggest a regulatory association between *rs666107* and the *Peptide GPCRs* pathway in glioblastoma. The conventional approach of genetic association studies is to discover associations between DNA-level variation (i.e., SNP) and phenotype. However, results stemming from this approach often suffered from a lack of biological explanation. For example, it is hard to explain the effect of SNPs located in introns or SNPs located in intergenic regions (i.e., *rs666107*). In this regard, our integrative analysis, using set-wise genetical genomics to identify SNP/pathway associations, might provide more relevant explanations than current genetic or genomic complex disease studies.

Acknowledgements

This study was supported by a grant of the Korea Health 21 R&D Project, Ministry of Health & Welfare (A040002) and of KOSEF, Ministry of Sciences and Technology (M10729070001-07N2907-00110 and MM10641000104-06N4100-1040), Republic of Korea. J.H.W. was supported in part by a GAW grant, R01 GM031575.

References

[1] Jansen RC, Nap JP. Genetical genomics: the added value from segregation. Trends in Genetics 2001;17(7):388–91.
[2] Morley M, Molony CM, Weber TM, Devlin JL, Ewens KG, Spielman RS, et al. Genetic analysis of genome-wide variation in human gene expression. Nature 2004;430(7001):733–7.
[3] Cheung VG, Spielman RS, Ewens K, Weber TM, Morley M, Burdick JT. Mapping determinants of human gene expression by regional and genome-wide association. Nature 2005;437:1365–9.
[4] Mootha VK, Lindgren CM, Eriksson KF, Subramanian A, Sihag S, Lehar J, et al. PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. Nature Genetics 2003;34:267–73.
[5] Segal E, Shapira M, Regev A, Pe'er D, Botstein D, Koller D, et al. Module networks: identifying regulatory modules and their condition-specific regulators from expression data. Nature Genetics 2003;34(2):166–76.
[6] Lan H, Stoehr JP, Nadler ST, Schueler KL, Yandell BS, Attie AD. Dimension reduction for mapping mRNA abundance as quantitative traits. Genetics 2003;164(4):1607–14.
[7] Ghazalpour A, Doss S, Sheth SS, Ingram-Drake LA, Schadt EE, Lusis AJ, et al. Genomic analysis of metabolic pathway gene expression in mice. Genome Biology 2005;6(7):R59.
[8] Woo JH, Zhengh T, KimF J.H.. DACE: differential allelic co-expression test for estimating regulatory associations of snp and biological pathway. International Journal of Functional Informatics and Personalized Medicine 2008;1(4):407–18.
[9] Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M. The KEGG resource for deciphering the genome. Nucleic Acids Research 2004;32:D277–80.
[10] Dahlquist KD, Salomonis N, Vranizan K, Lawlor SC, Conklin BR. GenMAPP, a new tool for viewing and analyzing microarray data on biological pathways. Nature Genetics 2002;31(1):19–20.
[11] Huang J, Chen K, Gong W, Zhou Y, Le Y, Bian X, et al. Receptor "hijacking" by malignant glioma cells: a tactic for tumor progression. Cancer Letter 2008;267(2):254–61.

Please cite this article in press as: Woo JH, et al. Identifying regulatory relationships among genomic loci, biological pathways, and disease. Artif Intell Med (2010), doi:10.1016/j.artmed.2010.03.001