Estimating regulatory associations of genomic loci and biological pathways in lymphoblastoid cells

Jung Hoon Woo[1,2], Eunjee Lee[3], Sung Bum Cho[1], Ju Han Kim[1,4§]
[1]*Seoul National University Biomedical Informatics (SNUBI), Seoul National University College of Medicine, Seoul 110-799, Korea*
[2]*Macrogen Inc., Seoul, Korea*
[3]*Department of Biological Science, Columbia University, New York, New York, United States*
[4]*Human Genome Research Institute, Seoul National University College of Medicine, Seoul National University, Seoul, Korea*
[§]*Corresponding author*

JHW: hyde83@snu.ac.kr
EL: el2380@columbia.edu
SBC: csb1749@snu.ac.kr
JHK: juhan@snu.ac.kr

Abstract

Genetical genomics has been established to study genetic variation of gene expression. It treats transcript expression as a quantitative trait and identifies putative regulatory loci for the expression of each gene. It is, however, well known that biological functions are often carried out through concerted activity of multiple genes. Therefore, we studied genetic regulators of biological pathways in lymphoblastoid cells. To identify the association of genomic loci and pathways, we applied two genetical genomics approaches, principal component analysis (PCA) and the differential allelic co – expression (DACE) test. We found that PCA is inadequate to identify regulatory loci for functional pathways. Association between a principal component, which summarizes the expression of a certain pathway, and a given marker was observed only when a direct association existed between the marker and the gene contributing most to that principal component. Another approach was the DACE test, a method that identifies regulatory loci that affect the inter-correlation structure of a gene set. Utilizing this test, we identified seven putative regulatory loci of immune-related pathways in lymphoblastoid cells after controlling for family-wise error rate.

1. Introduction

Recently, genetical genomics, the study of the genetic basis of gene expression, has been established. It treats mRNA expression of a gene as a quantitative trait [1-3]. The general scheme followed by established works in genetical genomics is quite similar to that of common microarray single-gene analysis. For a certain gene, a linkage or association test is performed on its expression and several genetic markers to detect loci that regulate that gene's mRNA expression, and the same process is performed in parallel on other genes. The 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, genes that contribute to the function of the signalling pathways of cells act simultaneously rather than independently. Therefore, using publicly available the 15th Genetic Analysis Workshop (GAW) data, we focused on identifying the regulatory loci for multiple genes, especially with regard to functional pathways, rather than single regulators of a gene. To this end, we applied two different approaches to the lymphoblastoid data sets.

Firstly, we applied principal component analysis (PCA), which has previously been used in Lan et al. [6] and Ghazalpour et al. [7]. They used PCA to reduce

978-0-7695-3452-7/08 $25.00 © 2008 IEEE
DOI 10.1109/BIBM.2008.13

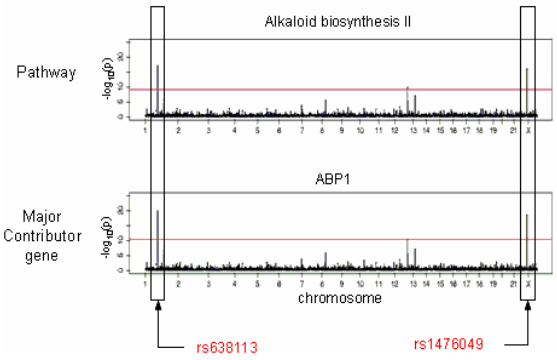

Figure 1. Concordance of genome-wide association results between the first principal component of a pathway and its major contributor. Both of the first principal components (PC1) extracted from the Alkaloid biosynthesis II pathway and the ABP1 gene revealed evidence of association with the same SNP, rs638113 and rs1476049

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 determined as regulatory loci of the pathway. Secondly, in our previous work, Woo et al. [8], we hypothesized that the regulator of a gene set may affect not only the expression of levels of those genes but also the extent of inter-correlation. Therefore, a new approach, the differential allelic co – expression DACE test, was applied to identify genetic regulators of co-expressions in a gene set. To elucidate the regulatory association between genomic loci and transcriptome expression in lymphoblastoid cells, we applied the above two approaches to the data and compared the results.

2. Materials and Methods

2.1. Gene expression data and SNP Genotype data

We selected gene expression data and genotype data of the 56 independent individuals in the 15th Genetic Analysis Workshop (GAW) data set provided by Morley et al. [2]. We were concerned only with the independent samples among the 194 CEPH individuals. We used genotype data of 2,882 genome-wide SNP and 8,793 mRNA expressions across 56 samples for this analysis. mRNA expression was measured by Affymetrix Human Focus Arrays. We additionally computed expression using the robust multichip average (RMA), which uses background adjustment, quantile normalization, and summarization.

2.2. Compilation of gene set

We used pathway information to compile the gene set. These data were obtained from publicly available major pathway resources, including KEGG [9], GenMAPP [10], and BioCarta (http://www.biocarta.com), for mapping genes to pathways. A total of 467 pathways were found when considering the 8,793 probes in the Affymetrix array. The 467 pathways were used in following principal component analysis and DACE test.

2.3. Differential allelic co–expression (DACE) test

We used the DACE test [8] to identify association 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 caused by an SNP's genotype. Given an 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 the "Fisher's z transformation". To test whether the SNP under study has significant effects on the levels of correlation among these genes, it adopts the general framework of a linear model.

3. Results

3.1. Identifying regulatory loci for pathways using PCA

We applied PCA to find the regulatory loci of functionally related genes. We extracted the first principal component (PC1) in all 467 pathways. We regarded PC1 as a quantitative trait and performed an association test with each of 2,882 SNP markers. Six PC1s corresponding to six pathways showed significant association with at least one SNP after controlling for family-wise error rate (data is not shown). Then, we independently tested the association between genome-wide SNPs and the genes comprising the six significant pathways. We used simple linear regression for the single-gene association analysis.

Interestingly, some of the genes showed the exactly same pattern of association when comparing pattern of six PC1s (Figure1). 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, PC1 of the Alkaloid biosynthesis II pathways

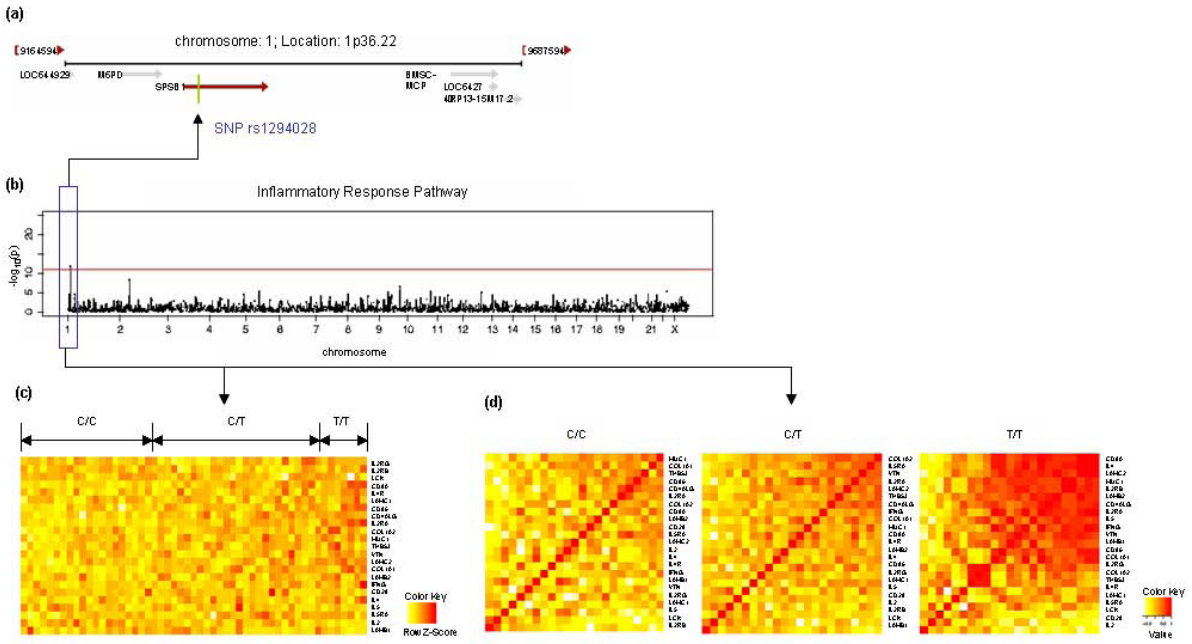

Figure 2. Genome-wide DACE test results for the Inflammatory Response Pathway (a) Physical map of rs129408 showing association with the Inflammatory Response Pathway from NCBI Gene View. The polymorphism is exactly located in intron 1 of a gene named SPRY domain-containing SOCS box protein SSB-1. (b) Genome-wide negative log10 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 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

showed almost the same P-value distributions as its major contributors, *ABP1* gene (Figure 1).

To see whether the major contributors were the only reason why the six PC1s showed association with the SNP, we repeated the same procedure without the major contributor genes for six pathways. None of the six PC1s without major contributor genes showed association with any of SNPs.

3.2. Identifying regulatory loci for pathways using the DACE test

We applied the DACE test to identify regulatory loci affecting coordinated alteration of multiple genes. We performed the test between 2,882 genome-wide SNPs and 467 pathways. A total of 103 pathways showed evidence of association with at least one marker using our criterion (Bonferroni corrected P < 0.00001). As RNA samples were extracted from lymphoblastoid cells for gene expression data, we listed seven immune system-related pathways in Supplement table 1 [14].

We independently tested the association between genome-wide SNPs and all genes comprising the seven significant pathways for comparison. Unlike in the previous section, the associations determined by the DACE test were not detectable by single-gene association analysis. For example, we identified that the inflammatory response pathway showed significant evidence for association with an SNP, *rs1294028*, by the DACE test (Figure 2(b)). Otherwise, 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 the result. There is no significant change in expression level among three genotype groups when considering each gene independently (Figure 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 23 genes (Figure 2(d)). The rs1294028 polymorphism is exactly located in the intron region of the splA/ryanodine receptor domain and SOCS box containing 1 (SPSB1) gene (Figure 2(a)).

4. Discussions and Conclusions

To find regulatory associations between SNPs and pathways in lymphoblastoid cells, we applied two established approaches to the GAW 15 data. The first approach was PCA, which was found to be insufficient since the pattern of variation of each first principal component was biased toward that of the specific gene with the largest contribution to that component.

We applied the DACE test to make up for the shortcomings of the principal component approach when attempting to identify the regulatory loci of functionally related genes. We identified novel associations between transcripts and SNPs by the DACE test. For example, none of the 23 genes comprising the inflammatory response pathway revealed association with *rs1294028* by single-gene association analysis, but a significant difference in correlation structure was seen in the result of DACE test. This observation, which shows an association between the polymorphism located within the *SPSB1* gene and the inflammatory response pathway, is consistent with previous studies [11-13].

5. Acknowledgements

This study was supported by a grant of the Korea Health 21 R&D Project, Ministry of Health & Welfare (A050558) and of KOSEF, Ministry of Sciences and Technology (MM10641000104 - 06N4100 - 1040), Republic of Korea J.H.W. was supported in part by GAW grant, R01 GM031575.

6. References

[1] R. C. Jansen and J. P. Nap, "Genetical genomics: the added value from segregation." *Trends in Genetics*, 2001, vol. 17, pp. 388–391.

[2] M. Morley, C. M. Molony, T. M. Weber, J. L. Devlin, K. G. Ewens, R. S. Spielman, and V. G. Cheung, "Genetic analysis of genome-wide variation in human gene expression." *Nature*, 2004, vol. 430, pp. 743–747.

[3] V. G. Cheung, R. S. Spielman, K. Ewens, T. M. Weber, M. Morley, J. T. Burdick, "Mapping determinants of human gene expression by regional and genome-wide association." *Nature*, 2005, vol. 437, pp. 1365-1369.

[4] V. K. Mootha, C. M. Lindgren, K. F. Eriksson, A. Subramanian, S. Sihag, J. Lehar, P. Puigserver, E. Carlsson, M. Ridderstrale, E. Laurila, N. Houstis, M. J. Daly, N. Patterson, J. P. Mesirov, T. R. Golub, P. Tamayo, B. Spiegelman, E. S. Lander, J. N. Hirschhorn, D. Altshuler, and L. C. Groop, "Pgc-1 alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes." *Nature Genetics*, 2003, vol. 34, pp. 267–273.

[5] E. Segal, M. Shapira, A. Regev, D. Pe'er, D. Botstein, D. Koller, and N. Friedman, "Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data." *Nature Genetics*, 2003, vol. 34, pp. 166–176.

[6] H. Lan, J. P. Stoehr, S. T. Nadler, K. L. Schueler, B. S. Yandell, and A. D. Attie, "Dimension reduction for mapping mrna abundance as quantitative traits." *Genetics*, 2003, vol. 164, pp. 1607–1614.

[7] A. Ghazalpour, S. Doss, S. S. Sheth, L. A. Ingram-Drake, E. E. Schadt, A. J. Lusis, and T. A. Drake, "Genomic analysis of metabolic pathway gene expression in mice." *Genome Biology*, 2005, vol. 6, p. R59.

[8] J. H. Woo, T. Zheng, and J. H. Kim, "Identifying genomic regulators of set-wise co-expression." *Proceedings of the IEEE BIBE2007,* 2008, pp. 433-439.

[9] M. Kanehisa, S. Goto, S. Kawashima, Y. and Okuno, M. Hattori, "The KEGG resource for deciphering the genome." *Nucleic Acids Research*, 2004*,* vol. 32, pp. D277–D280.

[10] K. D. Dahlquist, N. Salomonis, K. Vranizan, S. C. Lawlor, and B. R. Conklin, "GenMAPP, a new tool for viewing and analyzing microarray data on biological pathways." *Nature Genetics*, 2002*,* vol. 31(1), pp. 19–20.

[11] W. S. Alexander and D. J. Hilton, "The role of suppressors of cytokine signaling (SOCS) proteins in regulation of the immune response." *Annual Review of Immunology*, 2004*,* vol. 22, pp. 503–529.

[12] D. Wang, Z. Li, S. R. Schoen, E. M. Messing, and G. Wu, "A novel MET-interacting protein shares high sequence similarity with RanBPM, but fails to stimulate MET-induced Ras/Erk signaling." *Biochemical and Biophysical Research Communications*, 2004*,* vol. 313(2), pp. 320–326.

[13] L. Trusolino, and P. M. Comoglio, "Scatter-factor and semaphorin receptors: cell signaling for invasive growth." *Nature Review Cancer*, 2002*,* vol. 2(4), pp. 289–300.

[14] J. H. Woo, E. Lee, S. B. Cho, and J. H. Kim, "Supplement table 1: Pathways with significantly associated SNPs as indicated by the DACE test." http://www.snubi.org/publication/BIBM08/Supplement_Table_1.xls.