DACE: Differential Allelic Co-Expression test for estimating regulatory associations of SNP and biological pathway

Jung Hoon Woo

Macrogen Inc., Seoul, Korea
E-mail: hyde83@snu.ac.kr

Tian Zheng

Department of Statistics,
Columbia University,
New York, NY, USA
E-mail: tzheng@stat.columbia.edu

Ju Han Kim*

Seoul National University Biomedical Informatics (SNUBI),
Seoul National University College of Medicine,
Seoul 110-799, Korea
E-mail: juhan@snu.ac.kr
*Corresponding author

Abstract: To identify genomic regions associated with individual gene's expressions, genetical genomic approaches have been developed. The approach treats each gene expression value as a trait to determine the genetic factor that explains the variance of the mRNA expression. However, genes often demonstrate coordinated activities and the patterns and levels of coordination are also regulated. In this research, we present a method, the Differential Allelic Co-Expression (DACE) test that identifies the regulatory association derived by alteration of co-expression patterns within a molecular pathway according to allelic difference of a certain SNP.

Keywords: genetical genomics; gene sets; co-expression; regulator.

Reference to this paper should be made as follows: Woo, J.H., Zheng, T. and Kim, J.H. (2008) 'DACE: Differential Allelic Co-Expression test for estimating regulatory associations of SNP and biological pathway', *Int. J. Functional Informatics and Personalised Medicine*, Vol. 1, No. 4, pp.407–418.

Biographical notes: Jung Hoon Woo received the BS Degree in Biochemical Engineering from Seoul National University in 2005 and the MS Degree in Bioinformatics in 2007. He is currently a research scientist in the Bioinformatics Research Center at Macrogen, Inc.

Tian Zheng received her PhD in Statistics from the Department of Statistics at Columbia University in 2002. She was an assistant professor of Statistics at Columbia from 2002–2007 and has been an associate professor of Statistics at Columbia since July 2007.

Ju Han Kim received his MD and PhD Degrees from Seoul National University (SNU) and completed his residency training in neuro-psychiatry at SNU Hospital in 1996. He obtained an MS Degree in Biomedical Informatics at MIT. He was an Assistant Professor of Medicine in Biomedical Informatics, Children's Hospital Informatics Program at Harvard Medical School. He is currently an Associate Professor of Seoul National University College of Medicine.

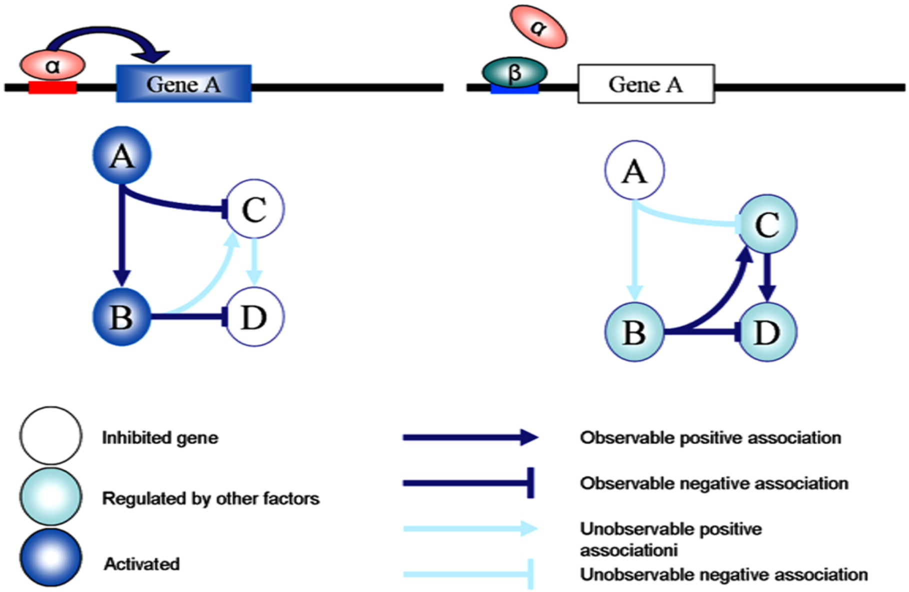
1 Introduction

Investigating the complex genetic interactions underlying molecular phenotypic variations has been a focal point in the field of biomedical research (Cheung and Spielman, 2002). Recently, the 'genetical genomics approach', which integrates large scale genotype data and high-throughput genomics data, like micro-array molecular profiles, has been established and offers a new perspective. This method treats gene expression profiles as quantitative traits and searches for genomic variation which can explain the variance of the molecular traits (Jansen and Nap, 2001; Brem et al., 2002; Schadt et al., 2003; Morley et al., 2004; Bystrykh et al., 2005). The scheme of this approach is very similar to most common micro-array analyses. Assuming that genes are individually expressed, this approach has shown insufficiency when investigating the complicated regulatory mechanism of transcriptome (Mootha et al., 2003; Lan et al., 2006).

To take regulatory conditions for functionally related genes into consideration, several methods have been proposed in DNA micro-array data analysis. For instance, Mootha et al. (2003) introduced the Gene Set Enrichment Analysis (GSEA) test, which focuses on gene sets, groups of genes that share common biological function, to identify genes sets differentially expressed for corresponding stimuli. Similar approaches have been applied in the area of genetical genomics, such as Lan et al. (2006) and Ghazalpour et al. (2005). With genome-wide SNP genotype data, they performed Quantitative Trait Locus (QTL) mapping for each gene expression trait, first. It was followed by applying GSEA (Ghazalpour et al., 2005) or Gene Ontology (GO) (Ashburner et al., 2000) enrichment analysis (Lan et al., 2006) when excessive numbers of expression traits had linkage or association with a single locus. Lan et al. (2003) also utilised Principal Component Analysis (PCA) to summarise expressions of multiple genes and identified SNP regulator which explained variance of the principal component. The studies addressed the issue of the single gene approach and intended to identify the regulator of related multiple genes. Both of the studies focused on identifying the co-regulators of a set of genes or the regulators of common variations of multiple genes. However, regulators of a molecular pathway may affect not only the expression levels of these genes but also the extent to which they are inter-correlated.

In this analysis, we focused on the correlation structure derived from the expression profiles of multiple genes. The level of correlation between the expression values of two transcripts is now commonly used to define co-expressions of functionally related genes (Eisen et al. 1998). Transcripts that are regulated by a common regulator will show high correlation under the functioning variant of the regulator but not so under the non-functioning variant. We hypothesise that if a genetic regulator affects a gene set, the correlation tendencies among the set would vary as the genotypic differences of the regulator (Figure 1). In this study, we designed a Differential Allelic Co-Expression (DACE) test to identify the regulatory association between a SNP marker and a-priori gene set chosen based on previous knowledge. We chose Pearson's *r* as the measure of co-expression and modelled it by an allelic linear model depending on the genotypes of a given marker. Through the DACE test, we showed putative genetic regulators which affect co-expression in biological pathways.

Figure 1 A hypothetical pathway with an upper cascade regulator. Gene A is in the upper cascade of a four-gene pathway. When one has the variant *red* in the promoter region of gene A on the genome, the Transcription Factor (TF)-α binds and leads to the expression of gene A, which, in turn, regulates genes B, C and D and their interactions. When the variant *blue* is present in the promoter of A, TF-ß binds and gene A is suppressed. Genes B, C and D will show different interacting patterns due to other regulation they receive (see online version for colours)


α
Gene A
β
α
Gene A
A
C
B
D
A
C
B
D
Inhibited gene
Regulated by other factors
Activated
Observable positive association
Observable negative association
Unobservable positive associationi
Unobservable negative association

2 Methods

2.1 Differential Allelic Co-Expression (DACE) test

The DACE test is performed between a given gene set and a given SNP marker. It tests correlation structure differentiation of mRNA transcripts due to a SNP's genotype. Assume that expression phenotypes and SNP genotypes are measured on *n* subjects.

Given a SNP, subjects are divided into its *G* genotypes. To study a set of *p* transcripts (expression phenotypes), first compute, within each genotype group, the *Pearson correlation coefficients* (Pearson, 1900) of the expression levels between all pairs of transcripts in a set. Denote r_{ijg} as the correlation between transcripts *i* and *j* within genotype group *g*. Correlation coefficients are not normally distributed (Fisher, 1921). For our test, we perform the 'Fisher's z' transformation' (Fisher, 1921) on the original correlation values as follows:

$$\begin{aligned} &Z_{ijg} = \frac{1}{2}\ln\left(\frac{1+r_{ijg}}{1-r_{ijg}}\right) \\ &i = 1, \ldots, p-1,\ j = i+1, \ldots, p, \\ &g = 1, \ldots, G. \end{aligned} \tag{1}$$

To test whether the SNP under study has a significant effect on the levels of correlation among these genes, we adopt the general framework of a linear model:

$$Z_{ijg} = B_0 + B_1 X_g + \varepsilon_{ijg} \tag{2}$$

where z is defined as above and Xg represents genotype variation. For an SNP with alleles A and B, we code Xg 0 for genotype AA, 1 for genotype AB and 2 for genotype BB. Therefore, the t test for H_0: ß$_1 = 0$ vs. H_1: ß$_1 = 0$ searches for significant *allelic association* between an SNP and the trait of interest (correlations).

2.2 Compilation of gene set

Publicly available pathway information was used to group genes. These data were obtained from publicly available pathway resources (BioCarta, http://www.biocarta.com/genes/allPathways.asp; KEGG, Kanehisa et al., 2004) for mapping genes to pathways. A total of 276 pathways were present for the 12,488 probes on the Affymetrix U74Av2 array.

2.3 Integrated data set of BXD RI strains

We targeted BXD recombinant inbred strains derived from 2 parental inbred strains, C57BL/6 (B6) and DBA/2. The usefulness of the strains in genetic mapping has led to genetic studies of a wide variety of phenotypes (Chesler et al., 2005). We downloaded DNA micro-array data sets which measure transcriptome expression in Hematopoietic Stem Cell (HSC) across 22 BXD RI strains (Bystrykh et al., 2005), from the National Center for Biotechnology Center (NCBI) Gene Expression Omnibus (GEO). Affymetrix U74Av2 arrays were used for HSC transcriptome profiling. For each chip, we computed 12,488 gene expression values for the Affymetrix data using the Robust Multichip Average (RMA) algorithm (Irizarry et al., 2003) which uses background adjustment, quantile normalisation and summarisation. The genotype data of the 22 BXD RI strains of mice were downloaded from webQTL (Chesler et al., 2004) on the GeneNetwork website (http://www.genenetwork.org). The original BXD genotype data file included a set of 3795 markers, both SNPs and micro-satellites. We chose 3033 SNP genotypes for our study.

3 Result

We assembled both gene expression data and genotype data, assuming that the genotype of each RI strain was fixed, and compiled a new BXD RI data set as a meta data for applying DACE test (see Methods). We hypothesised the regulatory association between a genomic locus and gene set when the extent of inter-correlations significantly varied between genotypes. The DACE test described in the Methods section was used to identify the regulatory association in the BXD RI data set.

A total of 276 gene sets was configured by pathway information (see Methods). Due to computational concerns, we restricted ourselves to pathways with fewer than 30 gene members. As a result, the test was performed using 233 pathways and 3033 SNP markers. The False Discovery Rate (FDR) (Benjamini and Hochberg, 1995) was controlled at < 0.001. Fifteen pathways showed significant association with at least one SNP locus. Five of these, some of which play critical roles in HSC function, are listed in Table 1. A complete list of the association found between the 15 pathways and 132 loci is provided in Supplement Table 1.

Table 1 Gene sets with significantly associated region from the DACE test

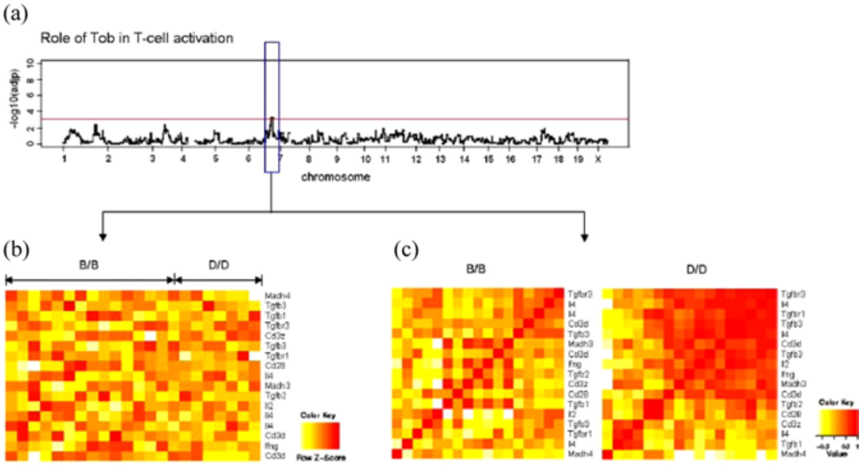
Gene set (= *pathway*)	*Chr*	*cM*	*p-value*	*FDR*
IL-10 Anti-inflammatory signalling	13	1.22 – 4.69	6.59e–07	1.53e–04
B Lymphocyte cell surface molecules	5	0.28 – 16.89	2.097e–07	1.82e–04
Role of Tob in T-cell activation	6	84.03 – 88.47	8.99e–07	2.47e–4
IL22 Soluble receptor signalling	7	39.60 – 41.93	1.19e–07	4.00e–05
Msp/Ron receptor signalling	2	138.89 – 139.78	3.84e–06	2.77e–04
	4	2.44 – 6.29	6.24e–06	3.57e–04
	6	82.28 – 102.25	7.89e–12	4.79e–09
	9	5.18 – 5.19	1.40e–11	5.31e–09
	17	37.99 – 48.75	2.76e–07	2.14e–05

For comparison, we performed the widely used single gene tests on the same data set, in which the association between gene and locus detected by the DACE test cannot be detected. Therefore, the set-wise co-expression regulators, identified using our approach, are different from (and possibly complementary to) the gene set (or GO) enriched regulatory regions studied by Lan et al. (2006) and Ghazalpour et al. (2005). For example, using the DACE test, the *role of Tob in the T-cell activation pathway* showed significant association with SNP markers within the region 84.03-88.47 cM on chromosome 6. On the other hand, none of the 17 genes comprising this pathway showed association with the same SNPs in single gene tests. These results indicate that none of the 17 transcripts have differentiated expression levels across the genotype groups of this locus (Figure 2(b)). Rather, the extent and patterns of correlation among the transcripts are significantly different between the two genotype groups (Figure 2(c)).

In this paper, we detected SNPs that were associated with the difference in the correlation structure among genes in a pathway. To elucidate the biological meaning for the significant association between the genomic loci identified and the correlation relation among the genes in a specific pathway, we examined their positional relation on the genome and in the regulatory network.

We hypothesise, as one possible biological interpretation, that if a certain genomic locus regulates the expression of genes in the upper cascade of a specific pathway, this locus might show significant association with the pathway in the DACE test. Therefore, we examined, in the regulatory network of known biological interactions, the positional relation between the significant genomic regulatory loci for a specific pathway identified by the DACE test and the genes in this pathway. More specifically, for a pathway with a significant genomic regulatory locus returned by the DACE test, we located the members of this pathway on the genome and identified those that were adjacent to the identified regulatory locus. Here, we empirically defined *adjacent genes* as those located within 10 Mbp from the physical location of the identified regulatory locus. Among the 15 significant pathways with significant loci identified, seven pathways have at least one gene member that is adjacent to the regulatory genomic loci returned by the DACE test. The results are provided in Supplement Table 2.

Figure 2 Genome-wide DACE test results for the role of Tob in the T-cell activation pathway. (a) Genome-wide distribution of negative log10 of p-values from the DACE test. The horizontal red line is our threshold (FDR < 0.001); (b) Heatmap for the gene expression matrix. There is no significant change of expression level between two groups when each gene is considered individually and (c) Heatmaps for the correlation matrices. Hierarchical clustering was done on the genes to show clearer patterns (see online version for colours)



We found that most of the adjacent genes were positioned as candidate regulators on the upper layer of corresponding pathway. In the case of the *Cyclin E Destruction pathway*, the “chr2. 149122813 – 159760664” locus was detected as a putative genomic regulator of the pathway by the DACE test (p value < 2.59E–06). The E2F1 gene, a member of the *Cyclin E Destruction pathway*, has been physically located within the locus of “chr2. 149122813 – 159760664”, next to the identified regulatory locus (Figure 3). By the single gene association test, the E2F1 gene and the regulatory locus did not show significant association after correcting for multiple comparisons (minimum p-value = 0.0639).

In this study, we designed the DACE test based on ordinary simple linear regression and the regression coefficient was used as a parameter to identify differentiation of correlation structure. We, therefore, used the permutation procedure to obtain empirical p-values for regression coefficients since the distribution of statistics may deviate from normality. For each permutation, we randomly reassigned the genotypes of a given SNP to the 22 samples and repeated the DACE tests, measuring the association between the correlation structure of the mRNA transcripts and the shuffled genotypes of the SNP. The procedure was repeated 1000 times to get empirical p-values of the original test results. We performed the permutation test for the 15 previously identified significant pathways and FDR was controlled at <0.05. Not all associations originally identified by the DACE test were significant in the permutation procedure, but strong signals were still captured by the permutation test (Table 2).

Figure 3 Physical map of the genomic regulator identified for the *Cyclin E Destruction pathway* (see online version for colours)

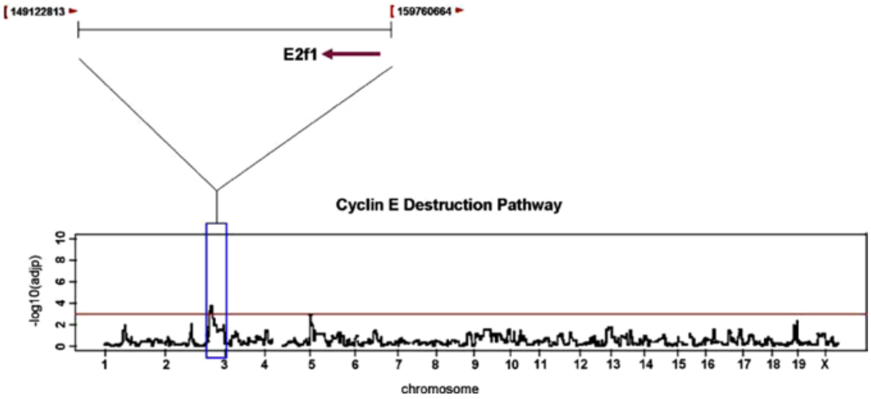


Table 2 Significant association identified by permutation procedure

Gene set (= pathway)	*Chr*	*cM*	*Perm p*	*FDR*
Cytokines and inflammatory response	6	85.48–86.38	0.0	0.0
The Co-stimulatory signal during T-cell activation	3	96.78–97.06	0.0	0.0
Role of tob in T-cell activation	6	84.61	0.0	0.0
The role of FYVE-finger proteins in vesicle transport	6	91.14–96.24	0.0	0.0
	12	57.72	0.0	0.0
Regulation of hematopoiesis by cytokines	7	80.62	0.0	0.0
IL22 soluble receptor signalling pathway	7	40.51	0.0	0.0
Cyclin E destruction pathway	2	93.84–105.82	0.0	0.0
	5	0.28	0.0	0.0
E2F1 destruction pathway	18	102.66–102.67	0.0	0.0
	2	93.85–104.67	0.0	0.0
	5	0.28–2.79	0.0	0.0
Visceral fat deposits and the metabolic syndrome	5	66.21–66.49	0.0	0.0
	7	50.97–55.97	0.0	0.0
Regulation of p27 Phosphorylation during cell cycle progression	2	86.43–104.65	0.0	0.0
	5	0.29–10.84	0.0	0.0
	9	15.96	0.0	0.0
Angiotensin-converting enzyme 2 regulates heart function	6	64.75–72.38	0.0	0.0
CBL mediated ligand-induced down regulation of EGF receptors	6	91.14–92.72	0.0	0.0
	10	87.31–89.36	0.0	0.0
	15	18.39–23.87	0.0	0.0

Chesler et al. (2005) and Morley et al. (2004) suggested that functionally related genes are genetically correlated. In the previous studies, a certain locus was identified as regulator of multiple genes' expression, and the common locus was referred to as *regulatory hotspot*. In this study, we identified a common locus for multiple gene sets, and thought that the identified locus represents genetic similarities among mapped gene sets. We performed hierarchical clustering of the 178 pathways used in previous DACE tests based on their association signal at individual SNP loci. The clustering was done using results of the DACE test which were 178 negative log10 *p*-value vectors length 3033. In other words, these 178 pathways were clustered by their similarity in significant genetic regulators of their within set co-expressions. Observed vertical patterns in the Figure 4(a) indicate the genetic similarity of pathways at certain genomic loci. For example, 5 pathways made the vertical bands at chromosome 6, and it might suggest that the pathways have genetic similarity corresponding to a locus at chromosome 6 (Figure 4(b)). These pathways *includes The role of FYVE-finger proteins in vesicle transport, CBL mediated ligand-induced down regulation of EGF receptors, Gamma-aminobutyric Acid Receptor Life Cycle, Role of Tob in T-cell activation*, and *Angiotensin-converting enzyme 2 regulates heart function*. We permuted the DACE result matrix to determine whether the vertical patterns occurred randomly or not. We shuffled columns within each row randomly, and reclustered the permuted matrix. Comparing clustering of the permuted matrix (Figure 5(b)) with clustering of the raw result matrix (Figure 5(a)), we could not identify distinct patterns in the permuted case. Furthermore, variation of tree depth of hierarchical clustering was disappeared after permutation.

Figure 4 Genetic similarity of pathways. (a) Hierarchical clustering for identifying genetic similarity among pathways. Row-wise clustering was done using the result matrix of the DACE test. In the matrix, column means SNPs, row means gene sets and each cell means association signal (–log10 *p*) between corresponding row and column. White arrow points at one of vertical band at chromosome 6 and (b) The vertical bands observed in heatmaps represents common associative locus among five pathways. That is, the five pathways have joint association with a locus at chromosome 6. The pathways are The role of FYVE-finger proteins in vesicle transport, CBL mediated ligand-induced down regulation of EGF receptors, Gamma-aminobutyric Acid Receptor Life Cycle, Role of Tob in T-cell activation, and Angiotensin-converting enzyme 2 regulates heart function (see online version for colours)

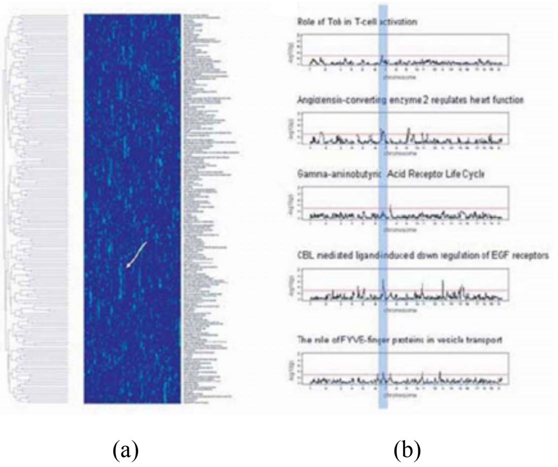
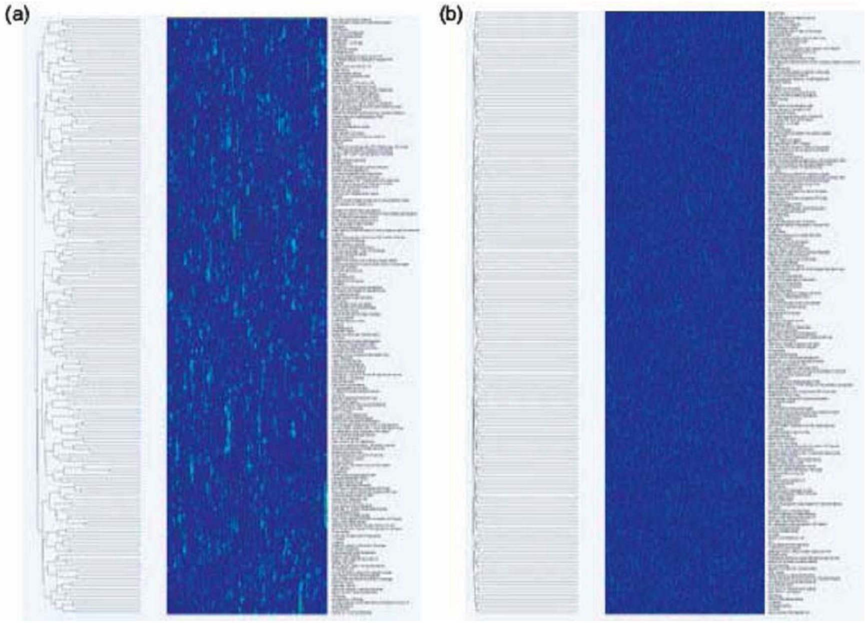


Figure 5 Heatmaps using clustering of the raw data matrix and clustering of the permuted data matrix. We permuted the values of the 178 pathways (shuffle column within each row) and recluster the data matrix to determine whether the original patterns occurred randomly or not. (a) Heatmap and dendrogram generated using clustering of raw result matrix and (b) Heatmap and dendrogram generated using clustering of permuted matrix (see online version for colours)



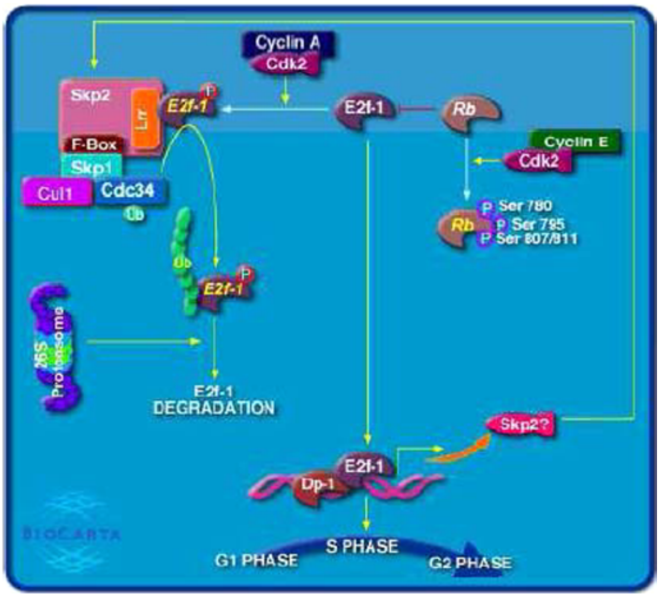
4 Discussion and conclusion

Previous single gene analyses assumed that all genes were individually expressed and looked for regulatory factors for each individual gene, an insufficient approach. In contrast, we developed a set-wise approach (the DACE test) to reveal the genomic loci involved in the regulation of co-expression, or of the inter-correlation of mRNA expressions, in gene sets, such as those involved in pathways. The DACE test treats a gene set as an analytical unit and examines factors associated with correlated activities within a given pathway. This relationship, between a genomic locus and a gene set, cannot be detected by single gene tests. Having configured our gene set using biological knowledge, rather than through an optimising approach, to generate a random set, our results are more biologically relevant. Using biological knowledge of pathways also narrows the computational scope of such studies, leading to higher power and better efficiency.

It was shown that the SNP regulators identified by the DACE test were not associated with the actual expression level of genes. Genotype differentiation of the SNPs did not explain the variance of gene expression profiles, as the SNP-gene relationships were not identified as significant in single gene tests as a single gene test. Instead, they regulated the level of correlations between genes and were only detectable when the set of genes was studied as a unit (See Figure 2 for an example using the role of Tob in T-cell activation pathway). This clearly demonstrates the utility of our approach for biological discovery, i.e., studying the regulation of interactive activities within pathways, which cannot be achieved with traditional genetical genomics approaches. It should also be noted that our approach is different from studying gene-gene regulation within a pathway, which focuses on the interactive activities of individual gene pairs genes within a pathway.

It is well known that complex functions of living cells are performed through the concerted efforts of many genes (Segal et al., 2003). A biological pathway is defined as a series of molecular interactions and reactions. If there are subtle changes in the expression level of a few genes located in the upper cascade of a pathway, they may alter the overall co-expression patterns of the pathway. This hypothesis is plausible when explaining our DACE results. We found, for 7 of the 15 significant pathways, at least one upper cascade gene that was adjacent to the identified regulatory loci of the pathways. In the case of the Cyclin E destruction pathway, the E2F1 gene is high in the upper cascade of the pathway (Figure 6) and the gene was located right next to the significant regulatory locus of the pathway (Figure 3).

Figure 6 The Cyclin E destruction pathway (see online version for colours)



In this study, we first demonstrated genetic similarity among gene sets and its patterns across genomes. Unlike the previous study, about common regulators for multiple genes, we focused on gene sets and they might have common regulators as well as individual genes. In the clustering of the DACE result matrix, there were several vertical patterns, representing common regulatory loci, across genomes (Figure 4(a)). Genes located in flanking regions of the common loci might be hubs of the gene regulatory network. Among 15 significant pathways, some regulatory loci were common in association patterns of the pathways, but some loci were not. It might suggest that pathways are affected not only by a common regulator but also by a unique regulator. The true situation is more complicated; however, our results may provide insights for elucidating regulatory networks in HSCs.

While there have been several studies concerned with co-expression or co-regulation of genes under external stimuli, these studies have yet to take into consideration natural variations, such as inherited genetic variation. Our procedure, designed to measure the influence of genetic variation on the co-regulation of gene expression at a genome-wide scale, offers important evidence of the heritability of mRNA co-expression between individuals.

Acknowledgements

This study was supported by grants from the Korea Health 21 R&D Project, Ministry of Health & Welfare, Republic of Korea (A040163 and A050558); and in part by NIH grant (R01 GM070789). Jung Hoon Woo and Tian Zheng contributed equally to this paper.

References

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M. and Sherlock, G. (2000) 'Gene ontology: tool for the unification of biology. The gene ontology consortium', *Nat. Genet.*, Vol. 25, pp.25–29.

Benjamini, Y. and Hochberg, Y. (1995) 'Controlling the false discovery rate – a practical and powerful approach to multiple testing', *Journal of the Royal Statistical Society Series B-Methodological*, Vol. 57, pp.289–300.

Brem, R.B., Yvert, G., Clinton, R. and Kruglyak, L. (2002) 'Genetic dissection of transcriptional regulation in budding yeast', *Science*, Vol. 296, pp.752–755.

Bystrykh, L., Weersing, E., Dontje, B., Sutton, S., Pletcher, M.T., Wiltshire, T., Su, A.I., Vellenga, E., Wang, J.T., Manly, K.F., Lu, L., Chesler, E.J., Alberts, R., Jansen, R.C., Williams, R.W., Cooke, M.P. and de Haan, G. (2005) 'Uncovering regulatory pathways that affect hematopoietic stem cell function using 'genetical genomics'', *Nat. Genet.*, Vol. 37, pp.225–232.

Chesler, E.J., Lu, L., Shou, S.M., Qu, Y.H., Gu, J., Wang, J.T., Hsu, H.C., Mountz, J.D., Baldwin, N.E., Langston, M.A., Threadgill, D.W., Manly, K.F. and Williams, R.W. (2005) 'Complex trait analysis of gene expression uncovers polygenic, pleiotropic networks that modulate nervous system function', *Nat. Genet.*, Vol. 37, pp.233–242.

Chesler, E.J., Lu, L., Wang, J.T., Williams, R.W. and Manly, K.F. (2004) 'WebQTL: rapid exploratory analysis of gene expression and genetic networks for brain and behavior', *Nat. Neurosci.*, Vol. 7, pp.485–486.

Cheung, V.G. and Spielman, R.S. (2002) 'The genetics of variation in gene expression', *Nat. Genet.*, Vol. 32, pp.522–525.

Eisen, M.B., Spellman, P.T., Brown, P.O. and Botstein, D. (1998) 'Cluster analysis, display of genome-wide expression patterns', *Proc. Natl. Acad. Sci. USA*, Vol. 95, pp.14863–14868.

Fisher, R.A. (1921) 'On the 'probable error' of a coefficient of correlation deduced from a small sample', *Metron*, Vol. 1, pp.1–32.

Ghazalpour, A., Doss, S., Sheth, S.S., Ingram-Drake, L.A., Schadt, E.E., Lusis, A.J. and Drake, T.A. (2005) 'Genomic analysis of metabolic pathway gene expression in mice', *Genome Biol.*, Vol. 6, pp.R59.

Irizarry, R.A., Bolstad, B.M., Collin, F., Cope, L.M., Hobbs, B. and Speed, T.P. (2003) 'Summaries of affymetrix genechip probe level data', *Nucleic Acids Res.*, Vol. 31, pp.e15.

Jansen, R.C. and Nap, J.P. (2001) 'Genetical genomics: the added value from segregation', *Trends Genet.*, Vol. 17, pp.388–391.

Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y. and Hattori, M. (2004) 'The KEGG resource for deciphering the genome', *Nucleic Acids Res.*, Vol. 32, Database issue, pp.D277–80.

Lan, H., Stoehr, J.P., Nadler, S.T., Schueler, K.L., Yandell, B.S. and Attie, A.D. (2003) 'Dimension reduction for mapping mRNA abundance as quantitative traits', *Genetics*, Vol. 164, pp.1607–1614.

Lan, H., Chen, M., Flowers, J.B., Yandell, B.S., Stapleton, D.S., Mata, C.M., Mui, E.T., Flowers, M.T., Schueler, K.L., Manly, K.F., Williams, R.W., Kendziorski, C. and Attie, A.D. (2006) 'Combined expression trait correlations, expression quantitative trait locus mapping', *PLoS Genet*., Vol. 2, No. 1, pp.e6.

Mootha, V.K., Lindgren, C.M., Eriksson, K.F., Subramanian, A., Sihag, S., Lehar, J., Puigserver, P., Carlsson, E., Ridderstrale, M., Laurila, E., Houstis, N., Daly, M.J., Patterson, N., Mesirov, J.P., Golub, T.R., Tamayo, P., Spiegelman, B., Lander, E.S., Hirschhorn, J.N., Altshuler, D. and Groop, L.C. (2003) 'PGC-1 alpha-responsive genes involved in oxidative phosphorylation are coordinately down regulated in human diabetes', *Nat. Genet.*, Vol. 34, pp.267–273.

Morley, M., Molony, C.M., Weber, T.M., Devlin, J.L., Ewens, K.G., Spielman, R.S. and Cheung, V.G. (2004) 'Genetic analysis of genome-wide variation in human gene expression', *Nature*, Vol. 430, pp.743–747.

Pearson, K. (1900) 'Mathematical contributions to the theory of evolution VIII on the correlation of characters not quantitatively measurable', *Proc. R. Soc. London*, Vol. 66, pp.241–244.

Schadt, E.E., Monks, S.A., Drake, T.A., Lusis, A.J., Che, N., Colinayo, V., Ruff, T.G., Milligan, S.B., Lamb, J.R., Cavet, G., Linsley, P.S., Mao, M., Stoughton, R.B. and Friend, S.H. (2003) 'Genetics of gene expression surveyed in maize, mouse and man', *Nature*, Vol. 422, pp.297–302.

Segal, E., Shapira, M., Regev, A., Pe'er, D., Botstein, D., Koller, D. and Friedman, N. (2003) 'Module networks: identifying regulatory modules, their condition-specific regulators from gene expression data', *Nat. Genet.*, Vol. 34, pp.166–176.