Methods

Genome-wide search for genetic modulators in gene regulatory pathways: Weighted window-based peak identification algorithm

Eunjee Lee a,1,2, Kyunga Kim b,1, Taesung Park a,c,*

a Interdisciplinary Program in Bioinformatics, Seoul National University, Seoul, Republic of Korea
b Department of Statistics, Sookmyung Women’s University, Seoul, Republic of Korea
c Department of Statistics, Seoul National University, Seoul, Republic of Korea

A R T I C L E   I N F O

Article history:
Received 9 November 2010
Accepted 11 March 2011
Available online 17 March 2011

Keywords:
Gene regulatory pathway
Expression quantitative trait loci (eQTL)
Peak identification algorithm

A B S T R A C T

Genome-wide gene expression and genotype data have been integratively analyzed in expression quantitative trait loci (eQTL) studies to elucidate the genetics of gene transcription. Most eQTL analyses have focused on identifying polymorphic genetic variants that influence the expression levels of individual genes, and such analyses may have limitations in explaining gene regulatory pathways that are likely to involve multiple genes and their genetic and/or non-genetic modulators. We have developed a novel two-step method for identifying potential genetic modulators of transcription processes for multiple genes in a biological pathway. We proposed a new weighted window-based peak identification algorithm to improve the detection of genetic modulators for individual genes and employed a Poisson-based test to search for master genetic modulators of multiple genes. Here, we have illustrated this two-step approach by analyzing the gene expression data in the Centre d’Etude du Polymorphisme Humain (CEPH) lymphoblast cells and single nucleotide polymorphism chip data.

© 2011 Elsevier Inc. All rights reserved.

1. Introduction

Understanding the structure, function, and evolution of genomes is a central goal of genetics [1]. One way to pursue this goal is to unravel the complex mechanism that controls the mRNA transcription rate of multiple genes in a certain pathway. Recently, high-throughput technologies, such as gene expression microarray and single nucleotide polymorphism (SNP) chip, have made it possible to simultaneously detect variations in gene expression level and DNA sequence in a large number of genes. The gene expression data have been combined with the genotype variation data to elucidate the genetics of gene expression in experimental populations of model organisms such as yeast, mice, and humans [2–6]. These studies are known as “expression quantitative trait locus (eQTL) mapping” or “genetical genomics” [7] and have revealed the global role of sequence variation in controlling gene expression. Understanding the genetic mechanisms controlling gene expression provides a way to discover complex networks of biological processes, which underlie complex traits such as common human multi-factorial diseases [8–10].

Recent eQTL studies have treated gene expression levels as a phenotype to identify polymorphic genetic variants that influence differences in expression level by targeting individual genes. According to the locations of QTLs, an eQTL can be distinguished into 2 classes of genetic variations: a cis-acting variation or a trans-acting variation. A polymorphism in cis-regulatory regions (cis-acting variation) can account for part of the genetic variation in expression by altering a functional motif in the promoter region, or by the activity of the gene product. In contrast, a polymorphism in one gene (trans-acting variation) can affect the expression of other genes by triggering feedback loops or by changing the coding sequence of a trans-acting factor [11]. However, conducting a large number of tests to determine associations between gene expression levels and genotype variations leads to statistical issues, such as multiple testing and effective sample size problems [12]. In particular, when subtle to modest alterations in gene expression occur, depending on the genotype variation, these statistical issues present obstacles to differentiating noise from actual differences in the level of gene expression across genotypes. Although gene expression levels can be measured on a genomic scale, eQTLs have been detected for a single gene at a time, and thus, biological interpretation based on this single-gene approach may have limitations in explaining complex traits. Most complex traits, such as clinical and physiological phenotypes, are multi-factorial, and hence, their manifestation is usually regulated by multiple genes and/or environmental factors. Phenotypic variation is unlikely to result from a simple variation in DNA sequence and/or mRNA expression level of a single...
gene, because genes operate in networks via complicated mechanisms, such as redundancy, feedback, and compensation [13].

Several approaches have been suggested to increase the statistical power to detect differentially expressed genes [14,15]. One promising approach was based on an idea that alterations in gene expression might manifest at the level of biological pathways or co-ordinated gene sets, rather than individual genes. Subtle but coordinated changes in expression might be detected more readily by combining measurements across a set of multiple genes [14]. By employing the concept of a “gene set” in genetical genomics research, one can identify genetic modulators for mRNA transcription levels of multiple genes in a biological pathway to characterize genetic variations that influence an entire process. Such identified genetic modulators can indicate that a perturbation in a particular gene induces a cascade of physiologic events that affect all or many of the other genes in that particular pathway.

In order to identify genetic modulators controlling functional pathways in lymphoblastoid cells, Lee et al. [16] employed a peak identification algorithm (PIA) [17–19]. First, they examined the association between the genetic variation at a genetic marker and the transcriptional variation of every single gene. Then, a PIA was used to detect genetic markers (hereafter referred to as peaks) that have significant effects on the expression levels of each gene in a particular gene set. The PIA is based on smoothing methods, which use the moving average of “neighboring peaks” to enhance peaks and remove spurious peaks. When constructing sliding windows for the moving average and variance, they used the same fixed numbers of adjacent markers for every marker. Fisher’s exact test [20] was then applied to identify common peaks that affect the expression levels of multiple genes in the gene set, under the assumption that the peaks observed for a specific marker have the same distribution as those for all other markers.

While this method provides a way to detect genetic regulators of a whole pathway, it has 2 limitations. First, this PIA is based on windows containing the same fixed numbers of adjacent markers for every marker, and hence does not take into account the uneven marker distribution that is common in whole genome research. Thus, direct application of the PIA may not be appropriate for obtaining genotype variation data measured at unevenly spaced markers across chromosomes or an entire genome. Since uneven marker distribution represents unequal distance between markers, the windows for different markers can have different sizes (i.e., widths) and can even contain “neighboring” markers that are located far from the marker. Although those distant located markers may not play any biological role as “neighbors,” they can have significant effects on the calculation of moving average and variance. In addition, PIA tends to be highly dependent on tuning parameters such as the variance or the number of neighboring markers for obtaining the moving average. Second, this method cannot consider the possible linkage between a true common regulatory locus of a gene set and its adjacent markers because the single marker-based Fisher’s exact test is used. Because of this linkage, it is likely that false peaks for some individual genes in a gene set can be detected in the neighborhood of the true common regulatory locus at the first step. Focusing only on individual markers rather than marker regions, each of which contains adjacent multiple markers, Lee et al. [16] used the single marker-based Fisher’s exact test to identify common peaks for a gene set, and hence failed to consider the linkage between a real regulatory locus and adjacent markers.

In this study, we propose a new two-step method to overcome the 2 limitations of the previous approach suggested by Lee et al. [16]. The genomic regions detected using our procedure can act as potential genetic modulators whose sequence variation can influence the transcriptional regulation of multiple genes involved in a biological function. In the first step, we developed a weighted peak identification algorithm (WPIA), which has 2 features that differentiate it from the PIA. First, the moving average and variance for a specific marker are computed from its neighboring markers located within a fixed width of sliding windows. Because the WPIA takes into account distances between markers by using the fixed size of windows, it would be appropriate for analysis of genotype variation data, particularly when markers are not evenly distributed. Second, the use of the distance-based weights reduces the effect of distant “neighboring” markers on the calculation of moving average and variance, and thus, the WPIA tends to select the most robust window size. In the second step, an existing method based on a Poisson distribution [5] was employed to detect common regulatory regions containing genetic modulators of transcription processes for multiple genes in a gene set. The simultaneous analysis of adjacent genetic markers allows us to properly address the problem that peaks for individual genes can be detected for different markers because of their linkage with a common regulatory region.

In the following sections, we present the proposed two-step method using WPIA and apply this method to an integrative dataset of gene expression [4] and SNP genotypes [21]. Furthermore, biological interpretation of the real data application demonstrates that the proposed method successfully detects potential genetic modulators affecting transcription processes of multiple genes in biological pathways.

2. Results and discussion

The proposed two-step procedure based on WPIA was applied to CEPH gene expression [4] and SNP data [21] for identifying common genetic modulators for expression rates of multiple genes in gene sets. The expression data of 23,880 genes were measured from lymphoblastoid cell lines of 167 individuals in 15 families from the CEPH/Utah family collection. For each individual, genotyping was conducted at 2,756 autosomal SNP markers. Using external pathway information, we previously searched for 454 gene sets that are related to biological pathways. In order to apply the two-step procedure, the significance of the association between gene expression and marker genotype was calculated on the basis of a multipoint genome-wide linkage analysis for each gene whose expression level is treated as a trait.

2.1. All 454 pathways

Using pathway information on Homo sapiens available from the Kyoto Encyclopedia of Genes and Genomes (KEGG) [22], Gene Map Annotator and Pathway Profiler (GenMAPP) [23], Pharmacogenetics and Pharmacogenomics Knowledge Base (PharmGKB) [24], and BioCarta (http://www.biocarta.com), we found 454 biological pathways, each of which contains 2–167 genes (14 genes on average) profiled in the CEPH gene expression dataset. The proposed two-step method was applied to each of these 454 pathways as a gene set to identify common genetic modulators for multiple genes in the pathway.

Using the WPIA, potential peaks were identified for individual genes in each of the 454 pathways. In order to identify regions with evidence of clustering of genetic marker-expression associations for each pathway, we grouped all SNPs into 1488 disjoint 2-Mbp-long intervals on the whole genome according to their positions. The potential peaks were then counted in each genomic interval across genes in a pathway, and the statistical significance of the genetic modulator was determined using a Poisson distribution. Using a family-wise error rate (FWER) of 0.05 with Bonferroni correction [25] (p-value = 7.40 × 10−8) and false discovery rate (FDR) of 1.17 × 10−5 [26], we found 60 pathways, each of which has 1–38 potential regulatory regions containing important genetic modulators. For example, “chaperones modulate interferon signaling pathway” appear to be regulated by a potential regulatory region (20.7–22.7 Mbp) in chromosome 21 (Fig. 1A). Among the 16 profiled genes within this pathway, 8 genes were found to have a potential genetic modulator (i.e., peak) at this regulatory region.
These potential regulatory regions were further investigated for regulatory-region classification, hotspots, and pathway complexity. First, we classified the potential regulatory regions into those acting "in cis" and those acting "in trans." A "cis-acting" region is defined as a region that contains a genotypic polymorphism affecting the expression level of a gene (say A), and lies within gene A itself or its regulatory region. A "trans-acting" region is defined as a region that contains a genotypic polymorphism affecting the expression level of gene A, and is not usually located near gene A. For example, this region could potentially lie in a different gene (say gene B) encoding a transcription factor of gene A. For each pathway, we classified a potential regulatory region as either a cis-acting region, if it lies within...
Each pathway contains at least one gene whose expression level is closely linked to a significant regulatory region within 10 Mb of its own location. The genes which are placed in upper cascade and known to regulate activities of other genes within the pathways.

Table 1
16 pathways having cis-acting regulatory regions.

<table>
<thead>
<tr>
<th>Pathway</th>
<th>Location</th>
<th>Gene</th>
</tr>
</thead>
<tbody>
<tr>
<td>Akt signaling pathway</td>
<td>17.7 (1)</td>
<td>Casp9 (caspase 9, apoptosis-related cysteine peptidase)</td>
</tr>
<tr>
<td>ALK in cardiac myocytes</td>
<td>30.1 (14)</td>
<td>Bmp3 (bone morphogenetic protein 4)</td>
</tr>
<tr>
<td>Antiarrhythmic drug pathways</td>
<td>108.5 (12)</td>
<td>Atp2a2 (ATPase, Ca++ transporting, cardiac muscle, slow twitch 2)</td>
</tr>
<tr>
<td>Cell cycle</td>
<td>17.7 (1)</td>
<td>Ef2 (EF transcription factor 2)</td>
</tr>
<tr>
<td>Chaperones modulate interferon signaling pathway</td>
<td>20.7 (21)</td>
<td>Ifngr2 (interferon gamma receptor 2)</td>
</tr>
<tr>
<td>Cytokines and inflammatory response</td>
<td>193.7 (1)</td>
<td>Tgfb2 (transforming growth factor, beta 2)</td>
</tr>
<tr>
<td>Electron transport chain</td>
<td>118.5 (12)</td>
<td>Cox6a1 (cytochrome c oxidase subunit Va polypeptide 1)</td>
</tr>
<tr>
<td>Erythrocyte differentiation pathway</td>
<td>193.7 (1)</td>
<td>Tgfb2 (transforming growth factor, beta 2)</td>
</tr>
<tr>
<td>Glycocon metabolism</td>
<td>30.1 (14)</td>
<td>Pfcl (phosphorylase, glycogen, liver)</td>
</tr>
<tr>
<td>GPCRs class A Rhodopsin-like</td>
<td>17.7 (1)</td>
<td>Cnr2 (cannabinoid receptor 2), Htrid1 (5-hydroxytryptamine receptor 1D), Htr6 (5-hydroxytryptamine receptor 6), Htr7 (5-hydroxytryptamine receptor 7)</td>
</tr>
<tr>
<td></td>
<td>33.7 (1)</td>
<td>Hctr1 (hypocretin receptor 1)</td>
</tr>
<tr>
<td></td>
<td>6.3 (2)</td>
<td>Ntr5s2 (neurotensin receptor 2), Opn1lw (opsin 1 long-wave-sensitive), Opn1mw (opsin 1, medium-wave-sensitive), Opn1sw (opsin 1, short-wave-sensitive)</td>
</tr>
<tr>
<td></td>
<td>12.3 (2)</td>
<td>Ntr5s2 (neurotensin receptor 2), Opn1lw (opsin 1, long-wave-sensitive)</td>
</tr>
<tr>
<td></td>
<td>4.2 (4)</td>
<td>Adra2a (adrenergic, alpha-2B-receptor), Drd5 (dopamine receptor D5)</td>
</tr>
<tr>
<td></td>
<td>88.0 (6)</td>
<td>Cnri1 (cannabinoid receptor 1), Htrie (5-hydroxytryptamine receptor 1E), Htrif (5-hydroxytryptamine receptor 1F)</td>
</tr>
<tr>
<td></td>
<td>91.0 (11)</td>
<td>Mthr1b (melatonin receptor 1B)</td>
</tr>
<tr>
<td></td>
<td>10.5 (12)</td>
<td>C5ar1 (complement component 3a receptor 1), C5ar1 (complement component 3a receptor 1), Gpr19 (G protein-coupled receptor 19), Lpar5 (lysophosphatidic acid receptor 5)</td>
</tr>
<tr>
<td></td>
<td>108.5 (12)</td>
<td>Cmklr1 (chemokine-like receptor 1)</td>
</tr>
<tr>
<td></td>
<td>30.1 (14)</td>
<td>Ptger2 (prostaglandin D2 receptor), Ptger2 (prostaglandin E receptor 2)</td>
</tr>
<tr>
<td></td>
<td>74.9 (16)</td>
<td>Mchr1 (melanocortin 1 receptor)</td>
</tr>
<tr>
<td></td>
<td>16.1 (20)</td>
<td>Sstr4 (somatostatin receptor 4)</td>
</tr>
<tr>
<td></td>
<td>46.1 (20)</td>
<td>Mchr3 (melanocortin 3 receptor)</td>
</tr>
<tr>
<td></td>
<td>54.1 (20)</td>
<td>Hrhr3 (histamine receptor H3), Mchr3 (melanocortin 3 receptor)</td>
</tr>
<tr>
<td></td>
<td>30.1 (14)</td>
<td>Map2k5c (mitogen-activated protein kinase kinase kinase kinase 5)</td>
</tr>
<tr>
<td></td>
<td>176.0 (3)</td>
<td>Pkcs3a (phosphoinositide-3-kinase, catalytic, alpha polypeptide)</td>
</tr>
<tr>
<td></td>
<td>46.0 (20)</td>
<td>Mchr3 (melanocortin 3 receptor)</td>
</tr>
<tr>
<td></td>
<td>20.7 (21)</td>
<td>Carg (phosphoribosylglycinamide formyltransferase), Gda (guanine deaminase)</td>
</tr>
<tr>
<td></td>
<td>15.3 (17)</td>
<td>Nts5 (5'3'-nucleotidase, mitochondrial)</td>
</tr>
<tr>
<td></td>
<td>30.1 (14)</td>
<td>Bmp4 (bone morphogenetic protein 4)</td>
</tr>
</tbody>
</table>

Each pathway contains at least one gene whose expression level is closely linked to a significant regulatory region within 10 Mb of its own location. The genes which are placed in upper cascade and known to regulate activities of other genes within the pathways.
element, and play a key role in the assembly of the preinitiation complex during transcription [35,36].

Third, the complexity of a pathway was investigated by looking at the number of potential regulatory regions that regulate the pathway as well as the number of genes contained in the pathway. Out of the 60 pathways, 35 had a single potential regulatory region, and 16 had more than 2 regions. The most complex pathway was that of “GPCRs class A rhodopsin-like,” which contains 167 genes and is regulated by 38 potential regulatory regions. “Antiarrhythmic drug,” “cell cycle,” “electron transport chain,” and “ribosomal protein” pathways also show high complexity in that at least 10 significant regulatory regions were identified for each pathway. These observations support the fact that gene regulatory pathways are multi-factorial.

2.2. Sixty pathways related to the immune system

The CEPH gene expression data were measured using mRNA samples extracted from lymphoblastoid cells. Because lymphoblastoid cells are components of the immune system, we further investigated the results of the two-step analysis for each of the 60 pathways related to the immune system. In 21 of these 60 pathways, we identified at least 1 potential regulatory region with significant genetic modulators (Supplementary Table 2), given an FWER of 0.05 with Bonferroni correction ($p$-value = $1.11 \times 10^{-7}$, FDR = $5.02 \times 10^{-4}$). Among these pathways, the “B cell survival pathway” seems to be regulated by a potential regulatory region containing a T cell receptor beta locus (TRB), which lies at 140.5–142.5 Mbp in chromosome 7. T cell receptor beta loci are rearranged and expressed in response to several foreign antigens. Furthermore, cytokines secreted by T cells help B cells multiply and mature into antibody producers [37]. There are 13 other pathways whose potential transcription factors are located within their potential trans-acting regions.

2.3. Discussion

By analyzing the CEPH gene expression and SNP data for randomly selected genes, we compared the proposed WPIA with the PIA [16]. First, the PIA was applied with several different settings for 2 parameters (i.e., the numbers of adjacent markers used for obtaining the moving average and the variance). For a particular gene, namely CHUK, the number of potential peaks highly depends on both parameters (Fig. 3A). Similar patterns have been observed in other genes (not shown here). Since it is critical for the PIA to choose the right parameters, a further investigation of the parameter choice is necessary.

Fig. 2. Hotspot search on 71 potential regulatory regions across the genome. The number of pathways for which a significant genetic modulator was identified at each potential regulatory region was displayed across the genome. The genome is divided into 1488 disjoint 2-Mbp-long intervals, shown in chromosomal order from the start of chromosome 1 to the end of chromosome X. Of the 71 potential regulatory regions, 29 are related to more than one pathway, and 9 out of the 29 regions are good candidates for hotspots in that they were found to have more than 5 locus–pathway associations.

Fig. 3. Number of potential peaks identified via PIA and WPIA for the CHUK gene. (A) PIA was applied with 2 parameters ($N_{\text{avg}}$ = the number of selected markers for obtaining the moving average and $N_{\text{var}}$ = the number of selected markers for obtaining the variance). Each line represents the number of peaks computed with $N_{\text{avg}}$ as 3, 7, 11, 15, and 19 using various $N_{\text{var}}$ values. (B) WPIA was applied with various window sizes. The line represents the number of peaks that were computed with various window sizes.
needed. Second, we applied the window-based WPIA to the same dataset. The number of potential peaks seemed unaffected by the window width for most genes, including CHUK (Fig. 3B). Therefore, the WPIA tends to be the robust parameter choice.

However, there are some limitations to the proposed method. First, this method did not take into account the different number of markers within the fixed windows when detecting common peaks for a gene set in the second step. Because the different number of markers within the fixed windows gives rise to different probabilities of getting peaks in the first step, further considerations should be taken into account for the different number of markers within the fixed windows in the second step. Second, the proposed method is based on the assumption that the number of peaks within a given marker interval has a Poisson distribution. Although the Poisson assumption is being commonly used in “hotspot” analyses, it still needs to be verified. We plan to further investigate this problem in the future.

3. Conclusions

In this study, we have proposed a two-step method to identify potential regulatory regions of a gene set that is defined as a set of genes included in a biological pathway. Because multiple genes in a common pathway do not usually function in isolation, their expression levels are likely to be correlated with each other. This correlation among the expression levels of multiple genes in a pathway could result from the fact that common genetic modulators often exist and have an impact on the expression levels of multiple genes in various ways. For example, a common modulator can be a transcription factor (TF) that binds to TF-binding sites of multiple genes, and thus it has a direct effect on regulating those genes. Another example would be a genetic modulator of a gene whose transcriptional change induces a cascade perturbation of other genes in the same biological pathway. In other words, a genetic modulator can regulate genes within a pathway in direct and/or indirect manners. Thus, the impact of the genetic regulators would vary across the genes, and it can be subtle on some genes that are indirectly regulated or less variable. When a classical linkage analysis with a strict threshold is conducted for determining the expression levels of individual genes, it is difficult to detect the subtle yet coordinated perturbation in the expression level of multiple genes induced by common genetic modulators [38]. In order to identify common genetic regulators that influence a group of genes working in the same pathway, we used a gene set approach in which we evaluated the impact of sequence variations on the expression levels of multiple genes.

For this purpose, we extended a peak identification algorithm (PIA), originally suggested by Lee et al. [16], for detecting pathway regulators. Using the PIA, one can first detect genetic markers with relatively high and similar significance patterns in a set. Common peaks for a gene set are then identified via Fisher’s exact test. While Lee’s method made it possible to detect markers that influence the expression of the whole pathway, it has some limitations when employed to genotype variation data. In order to overcome these limitations, we developed a more appropriate PIA called the weighted window-based PIA (WPIA) for genotype variation data; WPIA takes into account distances for the moving average and variance using windows of a pre-defined size with distance-based weights. In the second step, we proposed to detect regulatory regions using adjacent genetic markers simultaneously on the basis of Poisson distribution, rather than a specific marker using Fisher’s exact test, so that we could detect regulatory regions highly linked with the actual regulatory region of the expression level of genes.

Unlike standard eQTL analyses, our method could identify the regulatory regions that affect the expression level of multiple genes in the same pathway, even though the effect is moderate for some genes because of indirect regulation or less variable gene characteristics. Furthermore, our method is a knowledge-based approach in that gene sets were identified using prior knowledge such as biological pathway information in the literature and public databases. This approach enabled us to obtain modulators of the whole pathway rather than individual genes and to obtain a more functional inference. Such identified genetic modulators can indicate that a perturbation in a particular gene induces a cascade of physiologic events that affect all or many of the other genes in that particular pathway. The expression of a complex trait is usually multi-factorial and affected by multiple genes, some of which may be regulated by a common modulator. Therefore, our approach is more appropriate for explaining complex traits.

Our analysis of real data demonstrates that this new two-step method can successfully identify regulatory regions that control multiple genes in pathways. By dissecting the identified regulatory regions using gene information, we have identified transcription factors that are adjacent to correlated SNPs and possibly linked to the regulation of target pathway genes. Furthermore, our procedure suggests the possible presence of additional cis-acting modulators that may have an indirect influence on the expression of other genes in the same pathway, even when expression data for some of those other genes are unavailable.

4. Materials and methods

4.1. Two-step procedure to identify genetic modulators of multiple gene expression

In order to identify genetic modulators controlling the expression levels of multiple genes, we propose a stepwise procedure using association significance. In the first step, we identified potential genetic modulators whose genetic variations have a significant association with the expression levels of individual genes. Those potential genetic modulators are identified as “peaks” via a new weighted peak identification algorithm (WPIA) based on a fixed window size. Because the association significances at 2 close genetic markers are likely to be dependent on one another, spurious peaks can be falsely identified near a true genetic modulator. Smoothing methods with sliding windows can be used to remove these spurious peaks [17,18]. To determine the expression level of an individual gene, we examined the existence of a peak at a marker by considering association significances at the marker and its neighboring markers within the fixed window size. The effects of distant neighboring markers are penalized via distance-based weights. In the second step, we identified genetic modulators of expression levels for multiple genes. The number of peaks identified for multiple genes within a specific marker interval is assumed to have a Poisson distribution [5]. We tested the significance of sharing peaks within a specific marker interval on the basis of the Poisson assumption.

Let G be a “gene set” of genes whose expression levels are assumed to be regulated together in a particular pathway, and let M be a set of genetic markers that have been genotyped using a SNP chip. We denote $S_{(g,m)}$ as an association or linkage significance between the expression level of a gene $g \in G$ and the genotype of a genetic marker $m \in M$. For example, $S_{(g,m)}$ can be the negative logarithm of the $p$-value at marker $m \in M$ from a linkage analysis in which the gene expression level of gene $g \in G$ is treated as a trait. $S_{(g,m)}$, $\forall m \in M$ are ordered according to the marker position on the whole genome. Higher values of $S_{(g,m)}$ indicate a stronger effect of the genotype variation of marker $m$ on the expression level of gene $g$.

4.1.1. STEP 1: identification of genetic modulators for individual genes by using the weighted peak identification algorithm (WPIA)

In order to account for both measurement errors and dependency among adjacent association significances, we proposed a test statistic based on weighted moving average of the significances of a marker
and its adjacent genetic markers within a predetermined window size. For a given \(g \in G\) and \(m \in M\), the weighted moving average test statistic is as follows:

\[
T_{g,m} = \frac{S_{g,m}}{\sqrt{\text{Var}(S_{g,m})}}
\]

(1)

where the moving average and variance are calculated as

\[
S_{g,m} = \sum_{i \in I_m} w_i S_{g,i}
\]

and

\[
\text{Var}(S_{g,m}) = \frac{\sum w_i^2 (S_{g,i} - S_{g,m})^2}{\left(\sum w_i^2\right)^2}.
\]

respectively.

\(I_m\) is the index set of markers within a fixed size of the sliding window for marker \(m\), and \(w_i\) is the weight for marker \(i\). Assuming that the distance \(d_i\) between markers \(i\) and \(m\) has an exponential distribution, we develop the weight proportional to the probability density of the distance:

\[
w_i = \lambda \exp(-\lambda \times d_i),
\]

(2)

where \(\lambda\) represents the number of markers within 1 Mb.

On the basis of a certain threshold \(c\), we assigned an initial peak at marker \(m\) for a given gene \(g\) as below:

\[
P(g, m) = \begin{cases} 
1 & \text{if } T_{g,m} > c \text{ and } S_{g,m} > S_{g,m'} \text{ for all } m' \in I_m \\
0 & \text{otherwise}
\end{cases}
\]

(3)

For example, a threshold could be determined via permutations depending on the significance level. The initial peaks are further screened via Ripley’s K function [39] to exclude falsely detected peaks [40]. In other words, an initial peak at a specific marker is identified as a potential peak if its association significance is higher than the “average” association significance of its neighboring markers.

4.1.2. STEP 2: identification of common genetic modulators for a gene set

For each gene set, we detected common peaks in a genomic region containing multiple markers, on the basis of the potential peaks for individual genes in the gene set. Genomic regions are then identified using the evidence of genetic modulators affecting the transcription processes of multiple genes in a pathway. We consider these identified genomic regions as gene regulatory regions because their sequence variation could possibly play a regulatory role in the transcription process.

First, we divided the whole genome into genomic intervals of equal length and computed the number of peaks identified at any markers within each genomic interval \((R)\) for any gene in the gene set \((G)\):

\[
N_{R,G} = \sum_{m \in R} \sum_{g \in G} P(g, m)
\]

(4)

Second, we applied the approach suggested by Morley et al. [5] to test the significance of the sharing peaks within intervals for a gene set \(G\) by assuming that \(N(R, G)\) is distributed as a Poisson distribution. This procedure allows for simultaneous analysis of adjacent peaks that are detected for individual genes in a gene set at STEP 1, and that may be linked with a real regulatory locus. For multiplicity correction, we employed false discovery rate (FDR) [26] along with Bonferroni’s correction [25] of family-wise error rate (FWER).

4.2. Data and linkage analysis

4.2.1. CEPH gene expression and SNP data

The expression data were obtained from the study performed by Monks et al. [4] and were assessed for 15 families from the CEPH/Utah family collection (CEPH 1334, 1340, 1345, 1346, 1349, 1350, 1358, 1362, 1375, 1377, 1408, 1418, 1421, 1424, and 1477). The mRNA expression levels of 23,880 genes were measured in lymphoblastoid cell lines. While the 15 families contain 3 generations and 210 individuals, data for only 167 individuals were provided because of the quantity and quality of RNA samples used for expression profiling. For each of the individuals whose lymphoblastoid cells were phenotyped, 2,756 autosomal SNP markers were genotyped [21], and these SNP genotype data are available at the SNP Consortium database of the SNP Consortium Linkage Map Project (http://snp.cshl.org/downloads/index.html). Most SNPs in the SNP Consortium database are clustered in very closely linked sets (2 or 3 SNPs within 100 kilo base pairs) with an average intercluster distance of approximately 3 mega base pairs (Mb).

4.2.2. Linkage analysis

In order to determine the association significance between gene expression and marker genotype, we used the negative logarithm of the p-value at a marker position for each gene whose expression level was treated as a trait in a multipoint genome-wide linkage analysis. In particular, the analysis was performed using SIBPAL, a subroutine of the software package SAGE [41], which is supported by a U.S. Public Health Service Resource Grant (RR03655) from the National Center for Research Resources. Pairwise phenotype differences between siblings were weighted using the option (W4) of SIBPAL.

Supplementary materials related to this article can be found online at doi:10.1016/j.ygeno.2011.03.006.

Acknowledgments

The authors thank to J.H. Woo for his technical assistance in finding pathways related with the profiled genes based on the pathway information on Homo sapiens. The work was supported by the National Research Laboratory Program of Korea Science and Engineering Foundation (M10500000126), the National Research Foundation (KRF-2008-313-C00086), and the Brain Korea 21 Project of the Ministry of Education.

References


