Am. J. Hum. Genet. 75:1094–1105, 2004

Genetic Inheritance of Gene Expression in Human Cell Lines

S. A. Monks,[1,2,4] A. Leonardson,[4] H. Zhu,[4] P. Cundiff,[3] P. Pietrusiak,[5] S. Edwards,[4] J. W. Phillips,[6] A. Sachs,[4] and E. E. Schadt[4]

[1]Department of Statistics, Oklahoma State University, Stillwater, OK; Departments of [2]Biostatistics and [3]Pharmacology, University of Washington, and [4]Rosetta Inpharmatics LLC, Seattle; [5]Department of Epidemiology, Johns Hopkins University, Baltimore; and [6]Merck Research Laboratories, Merck & Co., Rahway, NJ

Combining genetic inheritance information, for both molecular profiles and complex traits, is a promising strategy not only for detecting quantitative trait loci (QTLs) for complex traits but for understanding which genes, pathways, and biological processes are also under the influence of a given QTL. As a primary step in determining the feasibility of such an approach in humans, we present the largest survey to date, to our knowledge, of the heritability of gene-expression traits in segregating human populations. In particular, we measured expression for 23,499 genes in lymphoblastoid cell lines for members of 15 Centre d'Etude du Polymorphisme Humain (CEPH) families. Of the total set of genes, 2,340 were found to be expressed, of which 31% had significant heritability when a false-discovery rate of 0.05 was used. QTLs were detected for 33 genes on the basis of at least one *P* value <.000005. Of these, 13 genes possessed a QTL within 5 Mb of their physical location. Hierarchical clustering was performed on the basis of both Pearson correlation of gene expression and genetic correlation. Both reflected biologically relevant activity taking place in the lymphoblastoid cell lines, with greater coherency represented in Kyoto Encyclopedia of Genes and Genomes database (KEGG) pathways than in Gene Ontology database pathways. However, more pathway coherence was observed in KEGG pathways when clustering was based on genetic correlation than when clustering was based on Pearson correlation. As more expression data in segregating populations are generated, viewing clusters or networks based on genetic correlation measures and shared QTLs will offer potentially novel insights into the relationship among genes that may underlie complex traits.

Introduction

In 1980, Botstein et al. proposed that sequence differences be treated as markers, in order to map genes involved in inherited traits. Since that time, the number of genes mapped to positions in the human genome has grown exponentially. Mapping these genes for inherited traits has been extremely successful for simple Mendelian diseases; however, finding such genes for diseases—and their associated risk traits—that are of large public health interest has proven difficult. Reasons for this difficulty include disease heterogeneity (disease subtypes with some or no overlapping genetic causes), misclassification (from using discrete classifications of disease from thresholds and combinations of thresholds), and unaccounted-for environmental influences. With the advent of technology to measure changes in molecular profiles—for example, changes in mRNA transcript abundance, protein levels, and metabolite levels—it should be possible to unravel some of the complexity of these complex diseases. In particular, gene expression can be viewed as a more refined phenotype, since it is a measure of phenotypic variation at the molecular level. In addition, each gene-expression phenotype provides annotation, pathway, and genome location data. Combining these data with genetic-inheritance information, for both molecular profiles and complex traits, is a promising strategy not only for detecting QTLs for complex traits but for understanding which genes, pathways, and biological processes are also under the influence of a given QTL.

Jansen and Nap (2001) were among the first to suggest the use of expression profiles in segregating populations. They discussed the power of using well-developed methods and designs available for dissecting quantitative traits along with the rapidly expanding collection of methods for large-scale sets of phenotypes. They provided an illustration that combined linkage data from a set of genes with known genomic locations, to construct a putative pathway. Jin et al. (2001) studied the contributions of sex, genotype, and age on transcription in *Drosophila melanogaster* through a study of two inbred lines of *Drosophila*. They observed a large sex effect on expression and less of an effect due to

Received July 9, 2004; accepted for publication October 1, 2004; electronically published October 21, 2004.

Address for correspondence and reprints: Dr. Stephanie A. Monks, Department of Statistics, 301G Mathematics, Statistics, and Computer Science Building, Oklahoma State University, Stillwater, OK 74078-1056. E-mail: stephanie.monks@okstate.edu

0002-9297/2004/7506-0014$15.00

genotype and age, although there was evidence for sex-by-genotype interactions. Brem et al. (2002) and Yvert et al. (2003) provided an in-depth exploration into the genetics of gene expression in yeast. These studies indicated significant control of gene expression by genetic variation with both *cis*- and *trans*-acting mechanisms. In addition, support was provided for linkage "hotspots" that controlled large sets of functionally related genes by a single QTL. Cowles et al. (2002) explored the role of *cis*-acting QTLs in mice and found evidence for regulatory variation, some of which was tissue specific. Schadt et al. (2003) provided a survey of the genetics of gene expression in maize, mice, and humans. This study further supported significant genetic control of gene expression in both *cis*- and *trans*-acting regulatory mechanisms. In addition, gene expression was utilized to subphenotype mice such that the underlying genetics for each subtype could be dissected. Data were also provided to support heritable influences on gene expression in human lymphoblastoid cell lines.

Yan et al. (2002) took one of the initial steps for extending these studies to include humans. Their technique compared two alleles of the same gene within the same cellular sample, to identify differences in expression between the two alleles. Thirteen genes were studied in 96 individuals. Of the 13 genes, 6 showed differences in expression due to allelic variation. In addition, they presented three families with expression levels segregating according to allelic variation. Another study in lymphoblastoid cell lines further established familial aggregation of gene expression and related functional classification to expression (Cheung et al. 2003).

The goal of the present article is to move beyond family aggregation, or heritability, and to more fully explore the genetic component of gene expression in humans through the use of lymphoblastoid cell lines in a sample of CEPH families (Dausset et al. 1990). Our study includes (1) determination of expressed genes, (2) estimation of the heritability of gene expression, (3) linkage analysis to establish oligogenic effects, and (4) characterization of *cis* and *trans* effects of detected QTLs. In addition, gene annotation will be studied in the context of each of the steps above, and we provide an example establishing the extra information, with regard to biological pathways, that is obtained by considering shared genetic influences. This study presents the largest survey to date, to our knowledge, of the heritability of gene-expression traits in segregating human populations.

Material and Methods

Families

Fifteen families from the CEPH/Utah family collection were selected for profiling. The family identifiers were 1334, 1340, 1345, 1346, 1349, 1350, 1358, 1362, 1375, 1377, 1408, 1418, 1421, 1424, and 1477. These families were selected because of the availability of genotypes and lymphoblastoid cell lines for all three generations and because of their large numbers of children. In total, the families represent 210 individuals. Of these, 167 individuals provided adequate quantity and quality of RNA for expression profiling.

Tissue Growth, Processing, and Profiling

Lymphoblastoid cell lines were obtained from Coriell Repositories and propagated. All cell lines were grown in media and supplements purchased from the Invitrogen Corporation. The culture media consisted of RPMI supplemented with 15% fetal bovine serum, 1% penicillin/streptomycin, and 0.5% sodium pyruvate. To minimize variability between experiments, all fetal bovine serum used was from lot number 10082147 1129480. The cell lines were grown at 37°C in humidified incubators, in an atmosphere of 5% CO_2.

Experiment series were set up by seeding 25-ml cultures in T25 flasks at a density of 2.5×10^5 cells/ml. Each culture was grown for 48 h or until the cell density was at least 780,000 cells/ml. To harvest the cells, the cultures were centrifuged, the media was decanted, and 500 μl of guanidine isothiocynate cell lysis buffer (Buffer RLT, Qiagen) was added. Cell lysates were then transferred to 96-well block format and stored at -80°C.

Total RNA was isolated using RNeasy 96 kits (Qiagen) with the following protocol modifications. Harvesting of cells was performed in 500 μl, instead of in the 150 μl specified by the protocol. To eliminate DNA contamination, the appended DNase protocol was used in concert with the isolation protocol. DNase was added to the membrane after the first 350-μl RW1 wash (guanidinium thiocyanate and ethanol) and was allowed to sit on an RNeasy membrane for 30 min. An additional 350-μl RW1 buffer wash and an additional 500-μl RPE buffer wash were performed.

To quantitate and perform quality control on the experiments, the A260/A280 ratio was taken through use of a Spectramax spectrophotometer (Molecular Devices). Samples whose A260/A280 ratio deviated ± 0.2 from the accepted ratio value of 2.0 were excluded. Formaldehyde gels (1.2%) were run on each sample to ensure that ribosomal RNA bands were intact and that significant degradation had not occurred. Samples that met the minimal mass requirement of 13 μg (for two replicates) and whose ribosomal bands were visible in the QC gel were transferred from the 96-well block and aliquoted into microcentrifuge tubes by use of a Multiprobe II EX (Packard BioScience Company). For samples of individuals that were to be used in the pool, 46 μg of RNA was allocated by use of the same procedure. In total, 167 individuals

in 15 pedigrees provided adequate quantity and quality of RNA for expression profiling.

The microcentrifuge tubes were vacuum dried and stored at −80°C before processing. Dried total RNA samples were reconstituted, and 3 μg of total RNA was used from each sample for subsequent RT-PCR–in vitro transcription amplification using the T7 promoter, which produced allyl-UTP–labeled single-stranded complementary RNA (sscRNA) (Hughes et al. 2001). Amplified cRNA was purified using the RNeasy purification kit (Qiagen) and was coupled with either cy3 or cy5 (Hughes et al. 2001). Purified cy3/cy5-labeled cRNA was fragmented using a ZnOAc/EDTA addition and was hybridized to at least two DNA microarray slides with fluor reversal for 24 h in a hybridization chamber, washed, and scanned using a laser confocal scanner (Hughes et al. 2001). Arrays were quantified on the basis of the intensity of each spot relative to background, by use of the Qhyb program (Rosetta Inpharmatics) (Marton et al. 1998).

Expression profiling of lymphoblastoid cell lines was performed using a 25K human gene oligonucleotide microarray. All individuals were compared with a common pool created from equal portions of RNA from all samples that passed quality control and were from founders within the 15 pedigrees (Gene Expression Omnibus Web site). Sequences for the microarray were selected from the RefSeq database (NCBI Reference Sequence Web site; see the Electronic-Database Information section for genes and accession numbers) and EST contigs (van't Veer et al. 2002).

Genotype Data and Genetic Maps

Genotype data for 346 autosomal genetic markers for 210 of the pedigree members were obtained from the CEPH genotype database, version 9.0 (CEPH Genotype Database Web site). Genetic markers were selected from the 14,404 markers represented in the full database, so that at least 75% of the pedigrees had genotypes available for at least 75% of the families. The median intermarker distance was 11 cM, on the basis of the deCODE genetic map (Kong et al. 2002). Marker-allele frequencies available from the CEPH genotype database were used for estimating identity-by-descent probabilities.

Statistical Methods

For each profile, genes were tested to assess differential expression relative to the pool, by use of procedures described elsewhere (Hughes et al. 2000). The color displays given in figure 4 show $\log_{10}$(expression ratio) as (1) purple, when an individual's expression is up-regulated relative to the pool; (2) blue, when an individual's expression is down-regulated relative to the pool; (3) black, when the $\log_{10}$(expression ratio) is close to zero; and (4) gray, when data from one or both of the channels for a given probe is unreliable. For each gene, the $\log_{10}$(expression ratio) is measured as the gene expression for an individual compared with that of the pool.

Variance-components methodology was used to estimate the overall and QTL-specific heritabilities of gene expression and to test for linkage across the genome at 4-cM steps, as described below (Almasy and Blangero 1998). For consistency, we follow the notation of Almasy and Blangero (1998). Consider a phenotype denoted by y. A linear model is used to relate variation in y to covariates, QTLs, polygenic background, and random error:

$$y = \mu + \mathbf{X}'\boldsymbol{\beta} + \sum_{i=1}^{n} \gamma_i + g + e \ ,$$

where μ is the grand mean; $\mathbf{X}'$ is a vector of covariates, with $\boldsymbol{\beta}$ being the associated vector of regression coefficients; γ_i is the effect of the ith of n QTLs under the additive model; g represents the polygenic background; and e corresponds to individual-specific random error. It is assumed that γ_i, g, and e are uncorrelated random effects with expectation 0. Hence, the total variance for y is $\sum_{i=1}^{n} \sigma^2_{\gamma_i} + \sigma^2_g + \sigma^2_e$, where $\sigma^2_{\gamma_i}$ is the additive genetic variance for QTL i, σ^2_g is the variance due to polygenic effects, and σ^2_e is the residual variance. Let $\mathbf{y} = (y_1, y_2, \ldots, y_t)'$ be the phenotype vector for a pedigree with t members. Under the assumption of multivariate normality, $\mathbf{y}$ can be modeled as a multivariate normal random variable with mean $\mu + \mathbf{X}\boldsymbol{\beta}$, where $\mathbf{X}$ is now a matrix of covariates with row i containing the covariates for individual i. The covariance matrix is equal to

$$\boldsymbol{\Omega} = \sum_{i=1}^{n} \hat{\boldsymbol{\Pi}}_i \sigma^2_{\gamma_i} + 2\boldsymbol{\Phi}\sigma^2_g + \mathbf{I}\sigma^2_e \ ,$$

where $\hat{\boldsymbol{\Pi}}_i$ contains elements $(\hat{\pi}_{ijk})$ such that the entry for the jth row and kth column is an estimate of the proportion of genes that individual j and k share at QTL i, $\boldsymbol{\Phi}$ is the kinship matrix, and $\mathbf{I}$ is a t-dimensional identity matrix. The matrix $\hat{\boldsymbol{\Pi}}_i$ is computed using a regression-based approach that is a function of the estimated IBD sharing for a set of markers and the distances from those markers to the location being modeled as a QTL (Fulker et al. 1994; Almasy and Blangero 1998). For heritability analyses, no QTLs are modeled, and the maximum-likelihood estimate of σ^2_g provides an estimate of heritability given the constraint $\mathrm{Var}(y) = 1$. Tests of heritability correspond to the following null and alternative hypotheses: $H_0: \sigma^2_g = 0$ versus $H_1: \sigma^2_g > 0$. For linkage analyses, a single QTL is modeled with all other variation due to genetic effects accounted for by σ^2_g. QTL-specific heritabilities are estimated using maximum-likelihood techniques. Tests of linkage correspond to the following null

and alternative hypotheses: $H_0:\sigma^2_{\gamma_1} = 0$ versus $H_1:\sigma^2_{\gamma_1} > 0$, where γ_1 is the single QTL being modeled.

In addition, bivariate segregation analyses were conducted using variance-components models (Almasy et al. 1997; Williams et al. 1999). Consider a vector,

$$\mathbf{z} = \begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix},$$

such that **x** and **y** are the pedigree trait vectors for two traits. The mean vector for **z** consists of the piecewise mean vectors for the two traits,

$$\boldsymbol{\mu}_z = \begin{bmatrix} \boldsymbol{\mu}_x \\ \boldsymbol{\mu}_y \end{bmatrix}.$$

The covariance matrix can be written as

$$\boldsymbol{\Omega}_z = \begin{bmatrix} \boldsymbol{\Omega}_x & \boldsymbol{\Omega}_{xy} \\ \boldsymbol{\Omega}_{xy} & \boldsymbol{\Omega}_y \end{bmatrix},$$

where $\boldsymbol{\Omega}_x$ and $\boldsymbol{\Omega}_y$ are the univariate covariance matrices for traits x and y. The matrix $\boldsymbol{\Omega}_{xy}$ consists of the cross-covariances for the two traits and can be written as

$$\boldsymbol{\Omega}_{xy} = 2\boldsymbol{\Phi}\sigma^2_{gxy} + \mathbf{I}\sigma^2_{exy},$$

where σ^2_{gxy} is the genetic covariance between traits x and y. Likewise, σ^2_{exy} is the residual covariance between traits x and y. Since we are interested in the genetic correlation (GC), ρ_{gxy}, and the residual correlation, ρ_{exy}, the covariances are reparameterized as

$$\sigma^2_{gxy} = \sigma_{gx}\sigma_{gy}\rho_{gxy}$$

and

$$\sigma^2_{exy} = \sigma_{ex}\sigma_{ey}\rho_{exy},$$

where σ_{gx} (σ_{gy}) corresponds to the square root of the total genetic variance for trait x (trait y). In addition, σ_{ex} (σ_{ey}) corresponds to the square root of the residual variance for trait x (trait y).

All analyses, with the exception of the bivariate segregation analyses and Pearson correlation (PC), were adjusted for age, sex, and age-by-sex interaction, and likelihood-ratio tests were utilized for tests of heritability, linkage, and GC. Variance-components models were analyzed using the software package SOLAR, the sequential oligogenic linkage analysis routines (Almasy and Blangero 1998).

Multiple testing for significance of total heritability was taken into account through the use of false-discovery rate procedures (Benjamini and Hochberg 1995). Here, P values are computed for the set of 2,430 genes that were selected on the basis of expression data alone. These P values are ordered such that $P_1 \leq P_2 \leq \ldots \leq P_{2,430}$. The first k tests are significant where k is the largest i such that $P_i \leq (i/2{,}340)\alpha$. This rule controls the false-discovery rate at level α.

For genes with available annotation in the Proteome BioKnowledge Library (Incyte), key phrases were compared between differentially expressed genes and the full set of 23,499 genes by use of a Fisher's exact test. A total of 4,783 categories are represented in the full set of genes. Hence, tests were conducted for each of the 4,783 possible categories represented in the full set of genes, and significance was assessed using a Bonferroni correction for a familywise type I error rate of 0.05.

Results

Genes Expressed in Lymphoblast Cell Lines

Of the 23,499 genes represented on the microarray, 2,430 were differentially expressed (type I error rate 0.05) in at least half of the children. Nine key phrases were enriched within the differentially expressed set (testwise type I error rate 0.05/4,783); these include "immune response," "response to viruses," and "inflammatory response." Table 1 contains a list of the nine key phrases, along with P values and corresponding occurrence counts.

Heritability of Gene Expression

Heritability analysis was conducted for the set of differentially expressed genes. When a false-discovery rate of 0.05 was used, 762 genes (31%) were detected as heritable. The median heritability for the 762 genes is 0.34. Figure 1 provides a summary of the distribution for the 762 heritability estimates. Results were consistent with our power calculations. The sample size provides ~28% power to detect $h^2 = 0.1$, 63% power to detect $h^2 = 0.2$, 85% power to detect $h^2 = 0.3$, 94% power to detect $h^2 = 0.4$, and 100% power to detect $h^2 \geq 0.5$. It is noted that heritability estimates were >0.44 for only 25% of the genes. Given that adjustments have been made for age and sex, this implies a large environmental or nongenetic influence on expression for the majority of genes. Of the 762 heritable genes, 705 had a median fold change between 1 and 2, 46 had a median fold change between 2 and 3, 6 had a median fold change between 3 and 4, and 5 had a median fold change >4. Annotation, obtained from the Proteome BioKnowledge Library (Incyte), was compared between genes with significant heritability and those that were differentially expressed. Each of the 4,783 categories that were represented in the full set of genes was considered. No significant differences were detected (testwise type I error rate 0.05/4,783). Comparison of annotation for genes

Table 1

Results of Comparison between Annotation for Differentially Expressed Genes and the Full Set of Genes

Phrase	No. of Occurrences among Differentially Expressed Genes (N = 2,430)	No. of Occurrences among Full Set of Genes (N = 23,499)	Fisher's Exact Test P Value
Immune response	104	460	4.14×10^{-12}
Response to viruses	36	132	4.98×10^{-7}
Integral to plasma membrane	194	1,243	2.40×10^{-6}
Immune cell chemotaxis	20	59	4.99×10^{-6}
Cytokine and chemokine mediated signaling pathway	28	100	5.23×10^{-6}
Cytokine activity	21	64	5.25×10^{-6}
Inflammatory response	75	392	5.44×10^{-6}
Plasma membrane	164	1,037	7.72×10^{-6}
Immediate hypersensitivity response	8	12	9.95×10^{-6}

with significant heritability and the full set of genes yielded results comparable to those shown in table 1; however, a couple of differences do exist. In particular, only two categories are statistically significant: immune response (P = .000000256) and defense/immunity protein activity (P = .00000386). Several of the categories found to be enriched in the differentially expressed genes are no longer among the most enriched categories for the subset of heritable genes.

We previously conducted a heritability pilot study on four CEPH families (Schadt et al. 2003). Although the reference pool utilized in the pilot study was substantially different from the reference pool used in the present study, we expected to see some consistency between the two. For the 440 genes found to be differentially expressed and heritable in the pilot sample (false-discovery rate 0.05), 65% were confirmed in the present study.

Expression QTLs

Multipoint-based identity-by-descent sharing was computed and utilized in a linkage analysis at 4-cM steps across all autosomal chromosomes. Figure 2 summarizes the linkage results for three levels of pointwise significance: .0005, .00005, and .000005. There were 33 genes with significant linkage defined by at least one P value $\leqslant$.000005, 50 defined by at least one P value $\leqslant$.00005, and 132 defined by at least one P value $\leqslant$.0005. Not surprisingly, genes with significant linkages correspond well to those genes that were found to have significant heritability and that are associated with immune-related functions.

For the 33 genes with significant linkages at the .000005 level, there was minimal correlation among expression levels. The maximum absolute correlation was 0.61; however, the third quartile of all pairwise correlations was 0.29. Twenty-two of these genes have significant expression QTLs even after a Bonferroni correction is applied to genomewide significance levels to obtain a familywise error rate of 0.05. That is, when significance is assessed for each of the 2,430 differentially expressed genes on the basis of a genomewide significance level of 0.05/2,430, 20 genes have significant expression QTLs (LOD score threshold 6.53). Interestingly, 8 of these genes have QTLs that overlap with their physical location within 5 Mb. This is in contrast to 13 of 33, 18 of 50, and 25 of 132 genes with QTLs significant at the pointwise level of .000005, .00005, and .0005, respectively. For genes with significant linkages at the .000005 level, most (25 of 33) had only a single QTL detected from the linkage scan. Six genes had 2 QTLs, one gene had 3 QTLs, and one had 15 QTLs. Figure 3 provides a summary of the QTL-specific heritabilities for the 55 QTLs. All detectable QTLs accounted for at least 50% of the trait variance, with 75% of the QTLs having heritabilities >0.76. This is consistent with the power to detect QTLs at a type I error rate of .000005 for a randomly selected sample of 15 pedigrees (results not shown).

Lack of Evidence for Linkage Hotspots

Previous studies have detected linkage "hotspots" in studies of the genetics of gene expression (Brem et al. 2002; Schadt et al. 2003; Yvert et al. 2003). Our linkage analyses were conducted at 4-cM steps, for a total of 816 positions along the autosomal genome. At the pointwise significance level of .000005, there were 586 locations with no linkage hits, 159 with one linkage hit, 59 with two linkage hits, 6 with three linkage hits, 3 with five linkage hits, and 3 with six linkage hits. Simulations were used to study the distribution of linkage counts per location under the assumption that linkages are distributed randomly through the genome. On the basis of 60,000 simulations, the probabilities of seeing three, four, five, or six linkage hits at some location in the genome were estimated to be 0.4488, 0.04505, 0.00315, and 0.0001666667, respectively. Hence, the locations with five and six linkages are consistent with nonrandom clusters of QTLs. In addition, the QTLs are

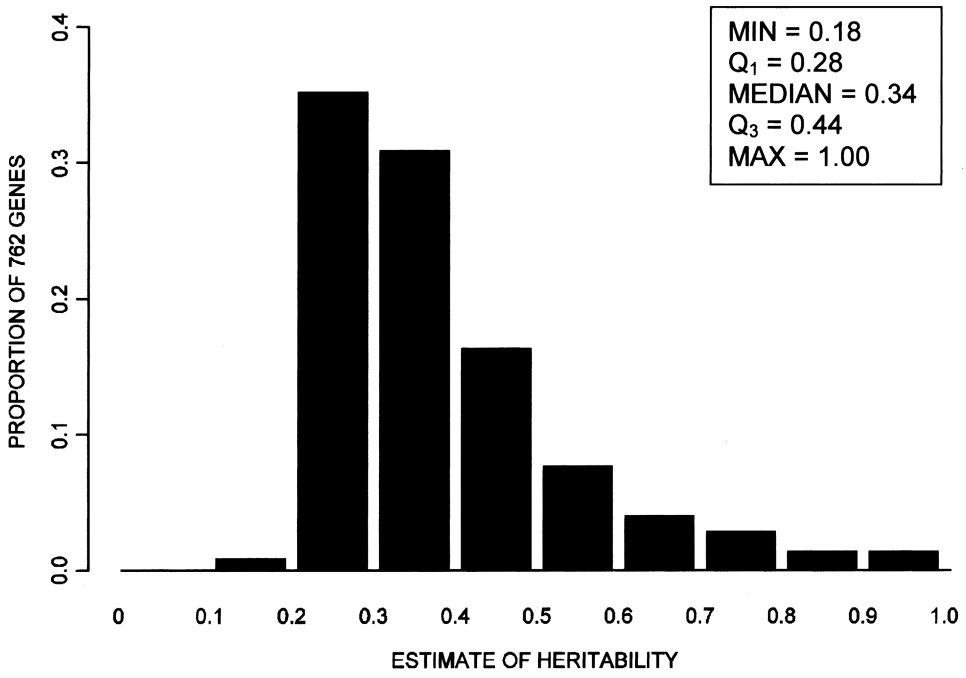


Figure 1 Effect sizes for heritable genes. The figure shows a histogram for estimates of heritability for those genes that are both differentially expressed and significantly heritable, when a false-discovery rate of 0.05 is used.

all located in a single area on chromosome 6 and correspond to linkages for six genes. Four of the genes (*HLA-DPB1, HLA-DRB3, HLA-DRB5,* and *HLA-G*) correspond to the major histocompatibility complex. One of the transcripts corresponds to an EST that is highly similar to a *Homo sapiens* major histocompatibility complex, class II, DR51 haplotype, and the last of the six genes is the cubilin gene. Two of the HLA genes, *HLA-DPB1* and *HLA-DRB3,* are located close to the shared linkage segment on chromosome 6; however, two factors cast doubt on whether the linkages are due to pleiotropic effects. First, HLA genes are highly polymorphic; therefore, probe selection without regard to such variation is likely to yield a probe that is subject to more-extensive SNP variation than would be realized in other genomic regions. For example, the 60-mer probe for *HLA-DRB5* has nine SNPs, that for *HLA-DPB1* has seven SNPs, and that for *HLA-DRB3* has five SNPs, on the basis of the dbSNP database (dbSNP Home Page). Hence, it is likely that the presence of genetic variation in the probe location could mimic expression patterns similar to those for genetic inheritance of a QTL. Hughes et al. (2001) demonstrated that variation in probe intensities realized on the microarray platform used in the present study were not significant if the probe contained fewer than five mismatches to the corresponding RNA sequence, but significant variation was observed for probes containing five or more mismatches. Second, HLA genes are highly similar to one another, making gene-specific probe selection difficult and resulting expression measures subject to significant cross-hybridization.

Clustering of Genes on the Basis of GC

Estimating GC between any two traits provides a measure of the extent of variation between two traits explained by common genetic components. Toward this end, we identified a set of genes by taking into account gene-expression activity criteria, heritability measures, and linkage analysis. Genes found to be transcriptionally active in at least 10% of the CEPH samples and that had a statistically significant heritability component (type I error rate 0.01) or at least one QTL with an associated LOD score $\geqslant 3$ were identified for further analysis. This resulted in a set of 574 genes that was carried forward into a bivariate analysis performed on each pair of traits in the set, to estimate GC. In addition, PCs for all gene pairs were calculated. For each correlation measure, agglomerative hierarchical clustering was applied to the gene-expression and experiment (individuals from the CEPH families) dimensions (Hastie et al. 2001). Color matrix displays, in addition to the experiment and gene-expression cluster trees generated from this procedure, are shown in figure 4. The two clusters share many common features, but there are also important differences. First, the GC-based cluster is seen

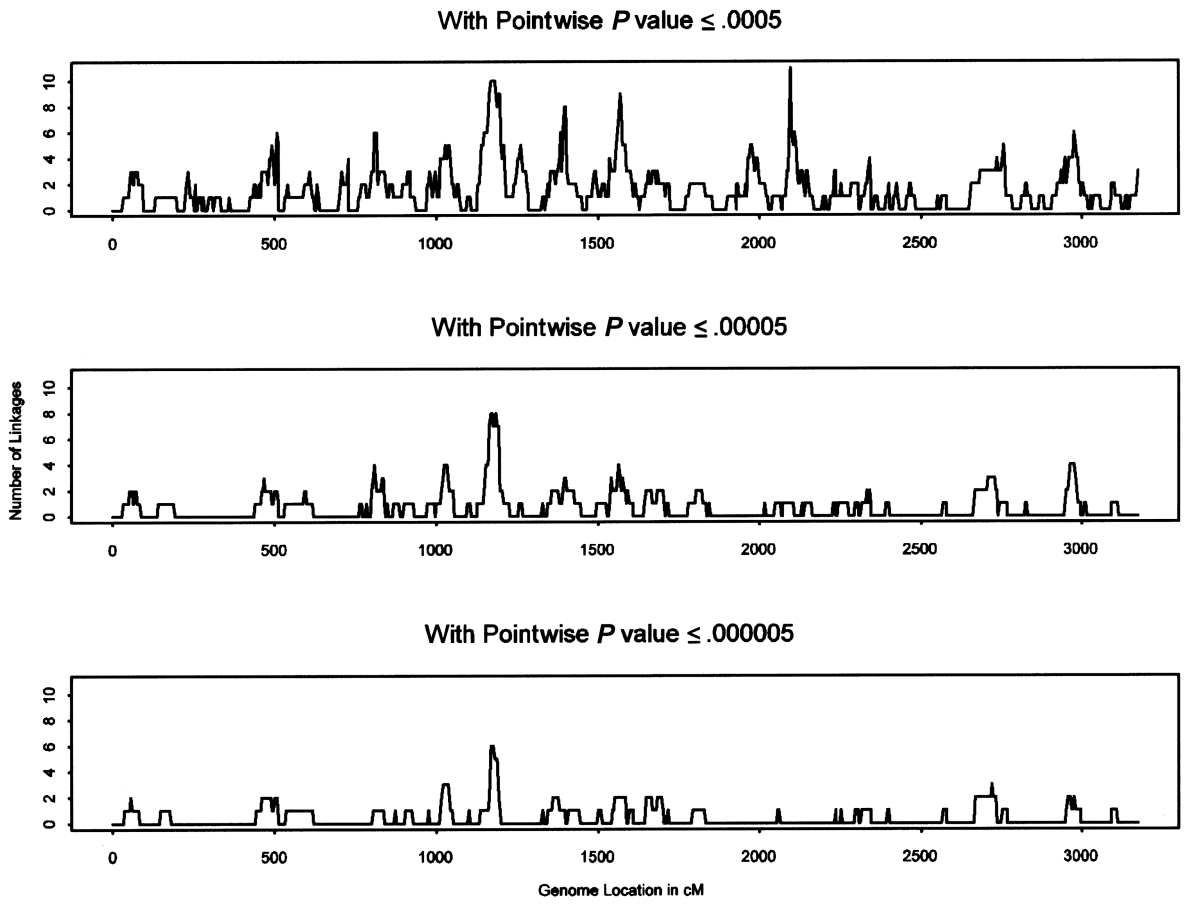


Figure 2 Linkage results for 2,430 differentially expressed genes. Multipoint linkage analysis was conducted every 4 cM over the autosomal genome. This figure summarizes the results by counting the number of genes with *P* values ≤.0005, ≤.00005, and ≤.000005, for each of the 4-cM locations.

to have smaller distances between two gene-expression traits on average, indicating a higher mean correlation measure than was observed for the PC-based cluster. Although the significances of the GCs were generally less than those observed for the PCs, the higher GC measures are consistent with the way in which this set of genes was selected. Another difference is the extent of pathway coherence represented by each cluster.

One measure of whether the GC- or PC-based cluster is providing more meaningful information is to examine the extent to which genes in known pathways are seen as clustering more closely together. Using the Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway databases (Gene Ontology Consortium Web site; KEGG Genes Database Web site), we determined that 164 and 32 genes, respectively, mapped to pathways represented in the GO and/or KEGG databases by ≥2 genes in the 574-gene set. To assess the extent of pathway coherence represented in the set of genes that mapped to these databases, we computed the average distance between genes in the cluster tree that mapped to the same pathway, for each pathway having ≥2 genes represented from the 574 gene set. For the PC clustering, the median distance for the GO pathways was 21 (minimum and maximum distances were 2 and 50, respectively), and the median distance for the KEGG pathways was 20 (minimum and maximum distances were 8 and 45, respectively). After 1,000 rounds of Monte Carlo simulation, the *P* values associated with the GO and KEGG pathway median distances were estimated to be .03 and .007, respectively.

The above results indicate that the pathway information in these databases does reflect biologically relevant activity taking place in the lymphoblastoid cell lines, with greater coherency represented in the KEGG pathways than in the GO pathways. Given this, we wanted to assess whether the GC clustering provided for increased coherency over all pathways represented in the 574-gene set, compared with PC clustering based on the observed expression values. The median distances between genes in the GC clusters were 23 for the GO pathways and 15 for the KEGG pathways. Although the

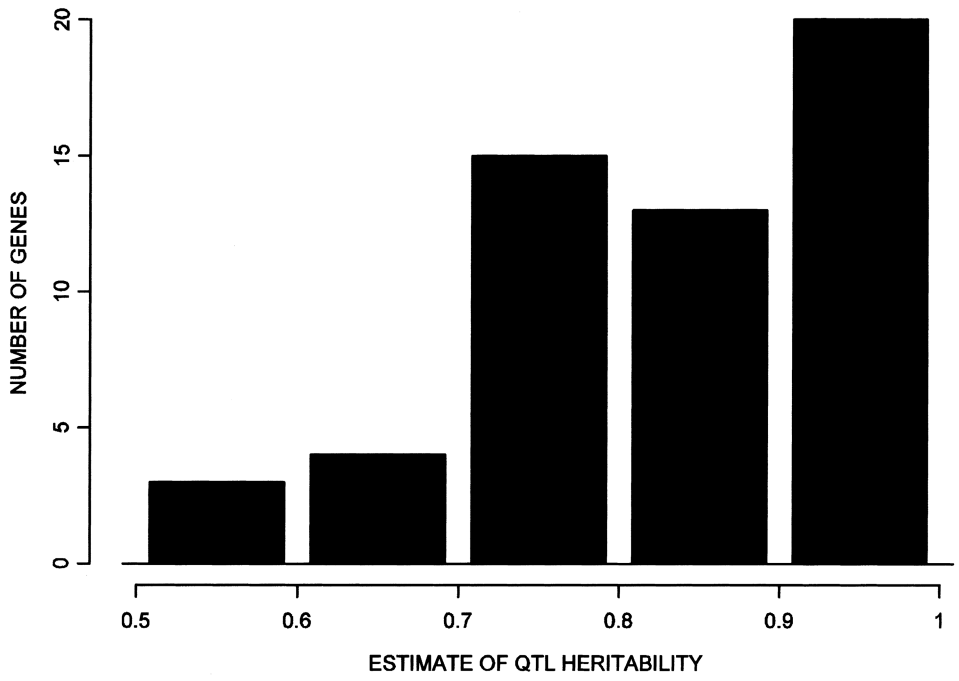


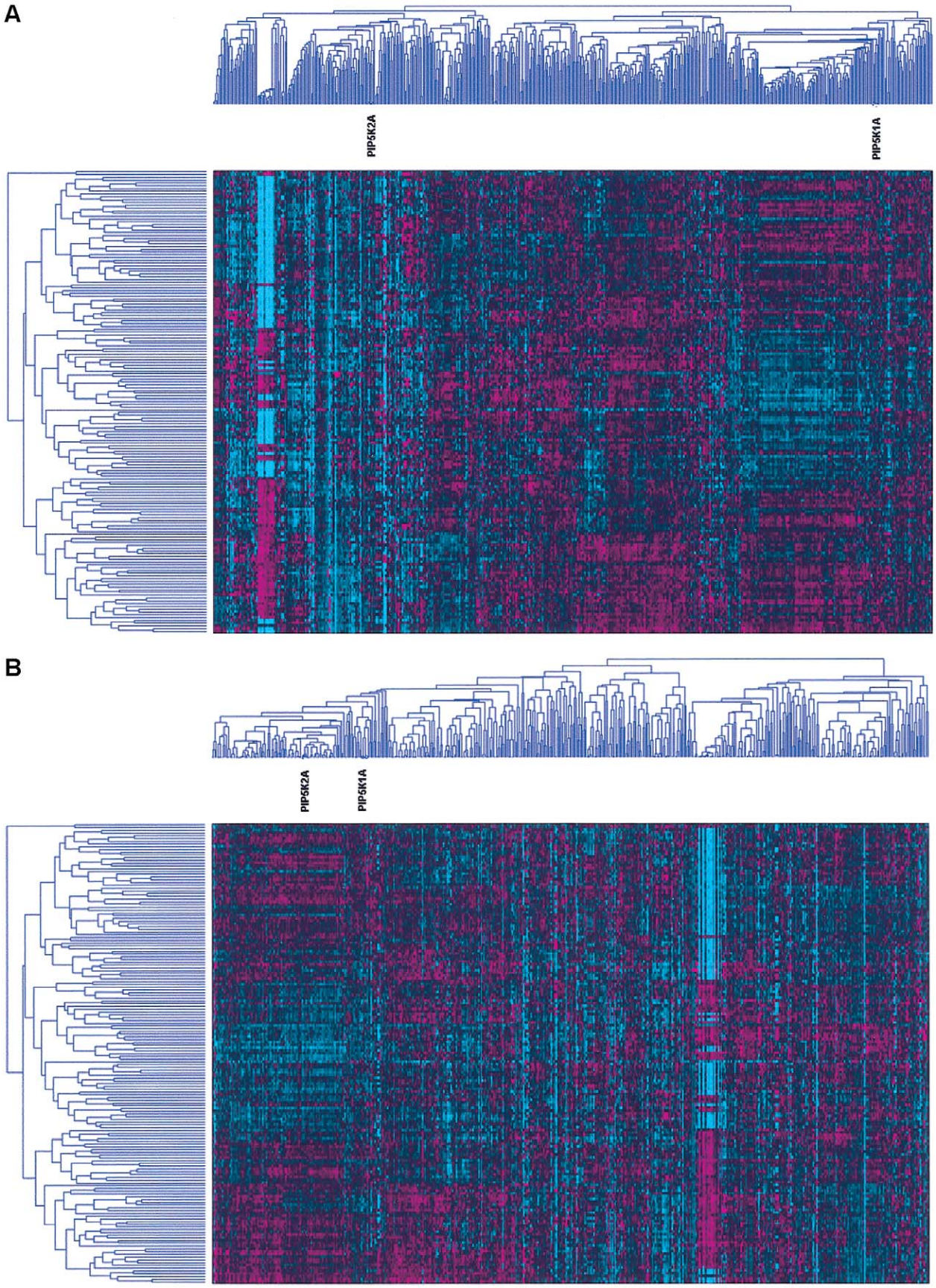
Figure 3 Effect sizes for QTLs detected at a pointwise significance level of .000005

median distance was higher for the GO pathways, it was interesting to note that, of the 67 GO pathways represented by ⩾2 genes in the 574-gene set, 29 had smaller distance measures in the GC cluster, compared with 30 having smaller distance measures in the PC cluster (8 had identical measures). In light of this, the degree of pathway coherence with respect to the GO pathways does not appear to be significantly different between the GC and PC clusters. This may reflect the more general nature of the GO pathways represented. However, the increased pathway coherence observed in the KEGG pathways is well reflected by the GC cluster; not only did the GC cluster show increased coherency for the KEGG pathways, but 29 of the 39 KEGG pathways represented in the 574-gene set had smaller distance measures than in the PC cluster, compared with 8 having smaller distance measures in the PC cluster.

An example of two genes represented in a KEGG pathway that grouped more tightly together on the basis of genetic correlation is highlighted in figure 4 for the PC and GC clusters. These two genes, *Pip5K1a* and *Pip5K2a,* are involved in phosphatidylinositol signaling and are members of the same KEGG pathway. Inositol signaling is known to be active in lymphoblastoid cell lines (Belmaker et al. 2002), and our data suggest that genes in this pathway are transcriptionally active in our cell lines as well. However, as can be seen in figure 4*A*, *Pip5K1a* and *Pip5K2a* are clustered at completely opposite ends of the cluster tree (shortest path connecting the two genes: 27), indicating that the PC cluster did not establish a relationship between these two genes. On the other hand, the GC cluster established a stronger connection between these two genes, by bringing them relatively close together in the cluster tree (shortest path connecting the two genes: 20).

Discussion

We have demonstrated that there is a genetic component to the control of gene expression in human lymphoblastoid cell lines. In fact, of the differentially expressed genes, 31% were heritable, on the basis of a false-discovery rate of 0.05. These genes were enriched for immunity-related functions, including immune response, defense/immunity protein activity, response to virus, and inflammatory response. Estimates of heritability were on the same order as that observed for complex traits. This is perhaps not surprising, given that the cell lines utilized were created for use in genotyping. Hence, a large amount of experimental variation could be diminishing any existing genetic effects. In particular, each family may not have been collected at a comparable time or with comparable sampling methods. Further, it is possible that the cell lines have not undergone the necessary procedures for immortalization at the same time or in the same experimental environment. Despite these potential problems, linkage analysis yielded several QTLs controlling for gene expression, even when a conservative Bonferroni correction was used to maintain


A
PIP5K2A
PIP5K1A
B
PIP5K2A
PIP5K1A

the false-positive rate, across all linkage analyses for all genes, at a level of .05. For the given sample size, the study is powered to detect major QTLs only. In fact, QTLs detected at a pointwise significance level of .000005 account for ⩾50% of the trait variance, with 75% of the QTLs having heritabilities >0.76.

It is of note that other studies of the genetics of gene expression have detected "hotspots" in which a single QTL controls the expression of a large set of genes. Our results did not detect such a phenomenon. It may be that our study was not powered to pick up such effects. However, other studies have focused on genetic crosses of inbred lines that had been selected to be divergent for a particular phenotype. One could argue that this type of ascertainment would enrich for a small set of QTLs that are underlying the observed phenotypic differences among the resulting offspring. Given that the resulting phenotypic differences are due to many changes at the molecular level, one would expect to see these types of hotspots for expression QTLs. Our sample was not ascertained on the basis of a particular phenotype. Given a random sample of families, one might not expect to see hotspots for control of gene expression.

One advantage of generating gene-expression data in a segregating population is the ability to decompose the relationship between any two traits into genetic and environmental components. Estimating genetic correlation between any two traits provides a direct measure of the extent of variation between two traits explained by common genetic components. It is this level of information that allows for more-direct causal inferences among the expression traits and clinical phenotypes of interest (e.g., disease-related phenotypes). The GC information among the gene-expression traits could be incorporated into standard Bayesian network reconstruction models, as a way of formalizing the reconstruction of genetic networks underlying complex phenotypes of interest. Successes achieved using more heuristic methods of incorporating genetic information into network reconstruction processes (Zhu et al. 2004) suggest that this is a worthy area of research to investigate. Here, we conducted analyses that showed clusters of genes based on GC corresponded well to biological pathways, and we provided an example of two genes involved in the same pathway that have been shown to be active in lymphoblastoid cell lines. These genes were more tightly clustered when GC, as opposed to PC, was utilized. Although there are similar examples in which PC offers a tighter association among genes of a given pathway, clustering based on GC offers a different view of the data that may enhance the information that can be derived from the clusters of gene-expression data and that has not previously been exploited. As more expression data in segregating populations are generated, viewing clusters or networks based on GC measures will offer potentially novel insights into the relationship among genes that may underlie complex traits.

Our results establish the existence of genetic control of gene expression and include a description of what this control looks like in a random sample of families. In addition, clustering based on GC provided groupings of genes that are consistent with biological pathways. For a sample ascertained for the study of a complex trait, such information could provide in-depth functional information that could be overlaid with inheritance data for the complex trait. Studies in maize, mice, and yeast are starting to provide such examples (Brem et al. 2002; Schadt et al. 2003; Yvert et al. 2003). More studies are needed to determine the utility of such an approach in humans. For instance, what tissues are amenable to sampling? What types of traits could be studied with such tissues? Peripheral blood is easy to obtain, but for what diseases or risk factors will this be relevant? Studies have already shown influences of age, sex, time of sample draw, blood cell count, and health status in peripheral blood (Whitney et al. 2003). Another question of interest is, to what extent does expression in cell lines relate to expression in the original tissue? Also, should the focus be on expression in the tissue or, perhaps, changes in expression due to a challenge?

The present study suffers from many of the same limitations of gene-expression studies, in that a large number of variables are tested. We tried to minimize multiple testing, by focusing on genes that we deemed to be expressed in the lymphoblastoid cell lines. Of course, there are other ways to determine a set of differentially expressed genes; however, we expect results based on alternative gene-selection methods to be comparable to

Figure 4 *A,* Two-dimension agglomerative hierarchical cluster constructed in the experiment (*Y*-axis) and gene-expression (*X*-axis) dimensions, using PC as the similarity measure, on the 574 genes described in the text. Several clusters in the gene-expression dimension are apparent from this color matrix display. However, two genes, *Pip5k1a* and *Pip5k2a,* known to function in the phosphatidylinositol signaling pathway, are seen here to cluster into two completely separate clusters in the gene-expression tree. *B,* Two-dimension agglomerative hierarchical cluster constructed in the experiment (*Y*-axis) and gene expression (*X*-axis) dimensions, using GC as the similarity measure, on the 574 genes described in the text. Although there are clearly patterns of expression that are highly similar to those shown in panel A, there are differences that serve to highlight the information that can be derived from the genetics dimension. In this instance, the two genes indicated in panel A as operating in the same pathway but clustering far away from each other cluster relatively closely together.

those presented here. Although results were presented for 167 profiled samples, only 15 families were utilized. Given such a large number of tests on this sample size, it is likely that the asymptotic distributions, on which *P* values are based, are not appropriate for all genes. Permutation-based test statistics could be used to estimate such *P* values; however, the computation time required makes this approach infeasible in a reasonable amount of time. Regardless of these limitations, the results presented here establish the existence of genetic control of gene expression and provide a glimpse into the possibilities of using such an approach to better understand complex traits.

Acknowledgment

Rosetta Inpharmatics is a wholly owned subsidiary of Merck & Co., Inc.

Electronic-Database Information

Accession numbers and URLs for data presented herein are as follows:

CEPH Genotype Database, http://www.cephb.fr/cephdb/

dbSNP Home Page, http://www.ncbi.nlm.nih.gov/SNP/index.html

Gene Expression Omnibus, http://www.ncbi.nlm.nih.gov/geo/ (for expression data for the 167 individuals utilized in the present study [GEO accession number GSE1726])

Gene Ontology Consortium, http://www.geneontology.org/ (for the GO database)

Kyoto Encyclopedia of Genes and Genomes (KEGG) Genes Database, http://www.genome.ad.jp/kegg/genes.html

NCBI Reference Sequence, http://www.ncbi.nlm.nih.gov/RefSeq/ (for *Pip5K2a* [accession number NM_005028], *Pip5K1a* [accession number NM_003557], cubilin [accession number NM_001081], *HLA-DRB5* [accession number NM_002125], *HLA-G* [accession number NM_002127], *HLA-DPB1* [accession number NM_002121], and *HLA-DRB3* [accession number V00522])

References

Almasy L, Blangero J (1998) Multipoint quantitative-trait linkage analysis in general pedigrees. Am J Hum Genet 62:1198–1211

Almasy L, Dyer TD, Blangero J (1997) Bivariate quantitative trait linkage analysis: pleiotropy versus co-incident linkages. Genet Epidemiol 14:953–958

Belmaker RH, Shapiro J, Vainer E, Nemanov L, Ebstein RP, Agam G (2002) Reduced inositol content in lymphocyte-derived cell lines from bipolar patients. Bipolar Disord 4: 67–69

Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc B 57:289–300

Botstein D, White RL, Skolnick M, Davis RW (1980) Construction of a genetic linkage map in man using restriction fragment length polymorphisms. Am J Hum Genet 32:314–331

Brem RB, Yvert G, Clinton R, Kruglyak L (2002) Genetic dissection of transcriptional regulation in budding yeast. Science 296:752–755

Cheung VG, Conlin LK, Weber TM, Arcaro M, Jen KY, Morley M, Spielman RS (2003) Natural variation in human gene expression assessed in lymphoblastoid cells. Nat Genet 33: 422–425

Cowles CR, Hirschhorn JN, Altshuler D, Lander ES (2002) Detection of regulatory variation in mouse genes. Nat Genet 32:432–437

Dausset J, Cann H, Cohen D, Lathrop M, Lalouel JM, White R (1990) Centre d'etude du polymorphisme humain (CEPH): collaborative genetic mapping of the human genome. Genomics 6:575–577

Fulker DW, Chernew SS, Cardon LR (1994) Multipoint interval mapping of quantitative trait loci, using sib pairs. Am J Hum Genet 56:1224–1233

Hastie T, Tibshirani R, Friedman JH (2001) The elements of statistical learning. Springer-Verlag, New York

Hughes TR, Mao M, Jones AR, Burchard J, Marton MJ, Shannon KW, Lefkowitz SM, Ziman M, Schelter JM, Meyer MR, Kobayashi S, Davis C, Dai H, He YD, Stephaniants SB, Cavet G, Walker WL, West A, Coffey E, Shoemaker DD, Stoughton R, Blanchard AP, Friend SH, Linsley PS (2001) Expression profiling using microarrays fabricated by an inkjet oligonucleotide synthesizer. Nat Biotechnol 19:342–347

Hughes TR, Roberts CJ, Dai H, Jones AR, Meyer MR, Slade D, Burchard J, Dow S, Ward TR, Kidd MJ, Friend SH, Marton MJ (2000) Widespread aneuploidy revealed by DNA microarray expression profiling. Nat Genet 25:333–337

Jansen RC, Nap JP (2001) Genetical genomics: the added value from segregation. Trends Genet 17:388–391

Jin W, Riley RM, Wolfinger RD, White KP, Passador-Gurgel G, Gibson G (2001) The contributions of sex, genotype and age to transcriptional variance in *Drosophila melanogaster*. Nat Genet 29:389–395

Kong A, Gudbjartsson DF, Sainz J, Jonsdottir GM, Gudjonsson SA, Richardsson B, Sigurdardottir S, Barnard J, Hallbeck B, Masson G, Shlien A, Palsson ST, Frigge ML, Thorgeirsson TE, Gulcher JR, Stefansson K (2002) A high-resolution recombination map of the human genome. Nat Genet 31:241–247

Marton MJ, DeRisi JL, Bennett HA, Iyer VR, Meyer MR, Roberts CJ, Stoughton R, Burchard J, Slade D, Dai H, Bassett DE, Jr., Hartwell LH, Brown PO, Friend SH (1998) Drug target validation and identification of secondary drug target effects using DNA microarrays. Nat Med 4:1293–1301

Schadt EE, Monks SA, Drake TA, Lusis AJ, Che N, Colinayo V, Ruff TG, Milligan SB, Lamb JR, Cavet G, Linsley PS, Mao M, Stoughton RB, Friend SH (2003) Genetics of gene expression surveyed in maize, mouse and man. Nature 422: 297–302

van't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AA, Mao M, Peterse HL, van der Kooy K, Marton MJ, Witteveen AT, Schreiber GJ, Kerkhoven RM, Roberts C, Linsley PS, Bernards R, Friend SH (2002) Gene expression profiling

predicts clinical outcome of breast cancer. Nature 415:530–536

Whitney AR, Diehn M, Popper SJ, Alizadeh AA, Boldrick JC, Relman DA, Brown PO (2003) Individuality and variation in gene expression patterns in human blood. Proc Natl Acad Sci USA 100:1896–1901

Williams JT, Van Eerdewegh P, Almasy L, Blangero J (1999) Joint multipoint linkage analysis of multivariate qualitative and quantitative traits. I. Likelihood formulation and simulation results. Am J Hum Genet 65:1134–1147

Yan H, Yuan W, Velculescu VE, Vogelstein B, Kinzler KW (2002) Allelic variation in human gene expression. Science 297:1143

Yvert G, Brem RB, Whittle J, Akey JM, Foss E, Smith EN, Mackelprang R, Kruglyak L (2003) Trans-acting regulatory variation in Saccharomyces cerevisiae and the role of transcription factors. Nat Genet 35:57–64

Zhu J, Lum PY, Lamb J, GuhaThakurta D, Edwards SW, Thieringer R, Berger JP, Wu MS, Thompson J, Sachs AB, Schadt EE (2004) An integrative genomics approach to the reconstruction of gene networks in segregating populations. Cytogenet Genome Res 105:363–374