Dimension Reduction for Mapping mRNA Abundance as Quantitative Traits

Hong Lan,[*] Jonathan P. Stoehr,[†] Samuel T. Nadler,[†] Kathryn L. Schueler,[*] Brian S. Yandell[‡] and Alan D. Attie[*,1]

[*]Department of Biochemistry, [†]Medical Scientist Training Program and [‡]Departments of Statistics and Horticulture, University of Wisconsin, Madison, Wisconsin 53706

Manuscript received August 23, 2002
Accepted for publication April 5, 2003

ABSTRACT

The advent of sophisticated genomic techniques for gene mapping and microarray analysis has provided opportunities to map mRNA abundance to quantitative trait loci (QTL) throughout the genome. Unfortunately, simple mapping of each individual mRNA trait on the scale of a typical microarray experiment is computationally intensive, subject to high sample variance, and therefore underpowered. However, this problem can be addressed by capitalizing on correlation among the large number of mRNA traits. We present a method to reduce the dimensionality for mapping gene expression data as quantitative traits. We used a blind method, principal components, and a sighted method, hierarchical clustering seeded by disease relevant traits, to define new traits composed of a small collection of promising mRNAs. We validated the principle of our approach by mapping the expression levels of metabolism genes in a population of F_2-*ob/ob* mice derived from the BTBR and C57BL/6J strains. We found that lipogenic and gluconeogenic mRNAs, which are known targets of insulin action, were closely associated with the insulin trait. Multiple interval mapping and Bayesian interval mapping of this new trait revealed significant linkages to chromosome regions that were contained in loci associated with type 2 diabetes in this same mouse sample. As a further statistical refinement, we show that principal component analysis also effectively reduced dimensions for mapping phenotypes composed of mRNA abundances.

INDIVIDUAL susceptibility to complex diseases, such as type 2 diabetes, has a strong inherited component. Genetic mapping and positional cloning of genes underlying quantitative trait loci (QTL) offer promise for understanding the molecular mechanism of the etiology and provide new therapeutic targets. However, such efforts are usually hindered by the fact that many complex diseases, including type 2 diabetes, are etiologically heterogeneous (McCarthy and Froguel 2002). The primary traits used in the mapping studies, such as plasma glucose concentrations, insulin levels, triglycerides, or body mass index, are usually the outcomes of many different genes interacting with each other and with environmental factors. Only a few physiological phenotypes can usually be scored for a disease. In a given cross, there might be many other unanticipated covert phenotypes that are also segregating. This wealth of information would be missed without additional specific phenotype assays.

Messenger RNA (mRNA) abundance can be used as a surrogate phenotype in mapping studies. Microarray technology made it possible to score simultaneously the mRNA levels of thousands of genes (Lander 1999), and it is now feasible to use a genome-wide genetic linkage approach to map the determinants of variation in gene expression (Brem *et al.* 2002; Cheung and Spielman 2002). This has facilitated exploration of the association of gene expression with multifactorial phenotypes of interest. However, integrating the multidimensional information from microarrays into genome-wide mapping poses a great challenge. We may choose to map each mRNA as a distinct trait, providing that proper control of the false discovery rate is addressed (Storey 2002). However, we anticipate substantial benefits by first reducing the dimensionality to a few "supergenes" that capture the majority of variation in expression data so that they would be quickly mapped. Ideally, some supergenes will reflect regulatory networks with controlling loci that influence many covarying mRNAs. These regulatory networks can be further explored by mapping the individual mRNAs composing the supergenes and by investigating the functions of these individual genes.

A geneticist may first wish to study the physiological connection between individual genes and traits and then use the expression levels of these genes as surrogate phenotypes. However, our knowledge about gene-disease connections is usually incomplete, due to the simultaneous effects of other genes, environmental factors, and complex interactions. In this scenario, we seek new traits in the form of combinations of correlated genes that segregate in an experimental mouse cross and appear to control aspects of the biochemical processes of diabetes and obesity.

How does an investigator select these new traits in an

[1] *Corresponding author:* Department of Biochemistry, University of Wisconsin, 433 Babcock Dr., Madison, WI 53706.
E-mail: attie@biochem.wisc.edu

objective way? Such a method should be straightforward and capture a low-dimensional data snapshot. Several methods have been used to select or combine mRNAs on the basis of their patterns of expression, including clustering (Eisen *et al.* 1998) and principal component analysis (West *et al.* 2001). Although dimension reduction methods provide objective criteria for organizing individual mRNAs, caution is warranted when a complex data structure is reduced to just a few dimensions. In practice, a comprehensive analysis necessitates a wide range of methods (Mahler *et al.* 2002).

Previously, we investigated inheritance of type 2 diabetes susceptibility loci segregating in a population of F$_2$-*ob/ob* mice derived from the BTBR and C57BL/6J (B6) mouse strains (Stoehr *et al.* 2000). Whereas the linkage to fasting plasma glucose and insulin levels meet statistical criteria, we desired new traits that may control expression of biochemical pathways manifested during the pathogenesis of the disease. We examined liver expression of a set of genes encoding metabolic enzymes in a panel of 108 F$_2$-*ob/ob* mice. Here, we demonstrate a method for identifying a few genes from the original data set and prospectively mapping them to putative controlling loci in the genome with enhanced linkage significance.

MATERIALS AND METHODS

Animals: The 108 F$_2$-*ob/ob* mice were a subset of the F$_2$ cross between B6 and BTBR strains that were previously used to study QTL associated with obesity and diabetes (Stoehr *et al.* 2000). Genotypes of 192 microsatellite markers across the mouse genome, as well as physiological phenotypes including fasting glucose, fasting insulin, and body weight, were described in the previous study (Stoehr *et al.* 2000).

Quantitation of mRNA: The mRNA abundance in the liver was estimated using the real-time quantitative reverse transcriptase-PCR (RT-PCR) assay. Representative genes encoding transcriptional factors or enzymes in major metabolic pathways were studied, including sterol regulatory element binding protein 1 (*SREBP1*), peroxisome proliferator activated receptor gamma (*PPAR*γ), fatty acid synthase (*FAS*), stearoyl CoA desaturase 1 (*SCD1*), glycerol-3-phosphate acyl transferase (*GPAT*), phosphoenolpyruvate carboxykinase (*PEPCK*), and acyl CoA oxidase (*ACO*). The housekeeping gene β-*actin* was used as a normalization control. Oligonucleotide primers were designed on the basis of their mRNA sequences in GenBank. The primer sequences are as follows: β-*actin* (M12481), forward 5′-CCATCCTGCGTCTGGACTTG, reverse 5′-TTCCCTCTCAGCTGTGGTGG; *SREBP1* (AF374266), forward 5′-AACCACCGTCACTTCCAGCTAG, reverse 5′-TGGTCCTGATTGCTTGTCAGG; *PPAR*γ (NM011146), forward 5′-TGAACGTGAAGCCCATCGAG, reverse 5′-CTTGGCGAACAGCTGAGAGG; *FAS* (AF127033), forward 5′-TCCTGGGAGGAATGTAAACAGC, reverse 5′-CACAAATTCATTCACTGCAGCC; *SCD1* (NM_009127), forward 5′-CTTCTTCTCTCACGTGGGTTGG, reverse 5′-TCGGCTTTCAGGTCAGACATGT; *GPAT* (NM_008149), forward 5′-TCTTGTTTCTGCCGGTGCAC, reverse 5′-ATTGCCCGAGGCGATGTAC; *PEPCK* (NM_011044), forward 5′-CCCCTTGTCTATGAAGCCCTCA, reverse 5′-GCCCTTGTGTTCTGCAGCAG; *ACO* (AF006688), forward 5′-TCTTCTTGAGACAGGGCCCAG, reverse 5′-GTTCCGACTAGCCAGGCATG.

Total RNA was isolated from frozen liver tissues using RNA-Zol (Tel-Test) and was purified using RNeasy columns (QIAGEN, Valencia, CA). First-strand cDNA was synthesized from 1 μg of total RNA using Super Script II reverse transcriptase (Invitrogen, San Diego) primed with a mixture of oligo(dT) and random hexamers. Reactions lacking the reverse transcriptase served as a control for amplification of genomic DNA. The reaction was carried out in a 25-μl volume in 1× SYBR Green PCR core reagents (Applied Biosystems, Foster City, CA) containing cDNA template from 10 ng of total RNA and 6-pmol primers. Quantitative PCR was performed on an ABI GeneAmp 5700 sequence detection system in 96-well plates. For each sample, duplicate amplifications were performed and the average measurements were used for data analysis. A regression analysis showed that the measurements for *SCD1* expression across the F$_2$ samples were highly reproducible ($Y = 1.0041$, $R^2 = 0.929$). The linearity of the real-time PCR procedure was also checked by carrying out four serial dilutions of four liver RNA samples derived from each parental strain. The signals of *SCD1* measured by real-time PCR followed a linear function that precisely matched the extent of dilution throughout each series (data not shown). We determined the cycle number at which the abundance of the accumulated PCR product crosses a specific threshold, the threshold cycle (C_{T}) for each reaction. The difference in average C_{T} values between β-*actin* and a specific mRNA was calculated for each individual and termed ΔC_{T}. The ΔC_{T} value, which is comparable to the log-transformed, normalized mRNA abundance, was used as the phenotype for follow-up analysis.

Hierarchical clustering: The ΔC_{T} values for each mRNA and each individual were gender adjusted and standardized using PROC STDIZE in SAS (1999). Cluster analysis included phenotypic measurements on each mouse, namely 8- and 10-week values of fasting plasma glucose, insulin, and body mass. The goal of "seeding" clusters is to identify subsets of expressed genes that are highly correlated with physiological traits of primary interest for subsequent mapping. Hierarchical clustering with oblique principal components was performed using PROC VARCLUS. Other hierarchical clustering approaches using PROC CLUSTER (*e.g.*, Ward's method) were examined to verify patterns of clustering.

Mapping gene clusters in F$_2$-*ob/ob* mice: A total of 192 microsatellite markers spanning the 19 mouse autosomes were genotyped and assembled into a framework map using MAPMAKER/EXP (Lander *et al.* 1987). We used multiple-trait interval mapping, JZMAPQTL developed by Jiang and Zeng (1995) and implemented in QTL Cartographer (Basten *et al.* 1994), to detect initial linkages between locus and mRNA trait and to investigate which mRNAs contributed to the composite linkage signal. Interval mapping of single mRNA traits in MAPMAKER/QTL was used to verify the statistical significance of the linkages.

The maximum-likelihood interval mapping analysis was verified and extended by methods that allow for multiple QTL across the genome, namely multiple interval mapping [Kao *et al.* 1999; in QTL Cartographer (Basten *et al.* 1994)] and Bayesian interval mapping (Satagopan *et al.* 1996; Gaffney 2001; see http://www.stat.wisc.edu/~yandell/qtl/software/Bmapqtl). Bayesian interval mapping uses Bayes' factors, which are conceptually similar to the Bayesian information criterion used for model selection in classical interval mapping. Bayes' factors used in interval mapping can be fairly insensitive to choice of priors by employing empirical Bayes methods (Gaffney 2001). We estimate the marginal posterior probability that a QTL is found at a locus allowing for an

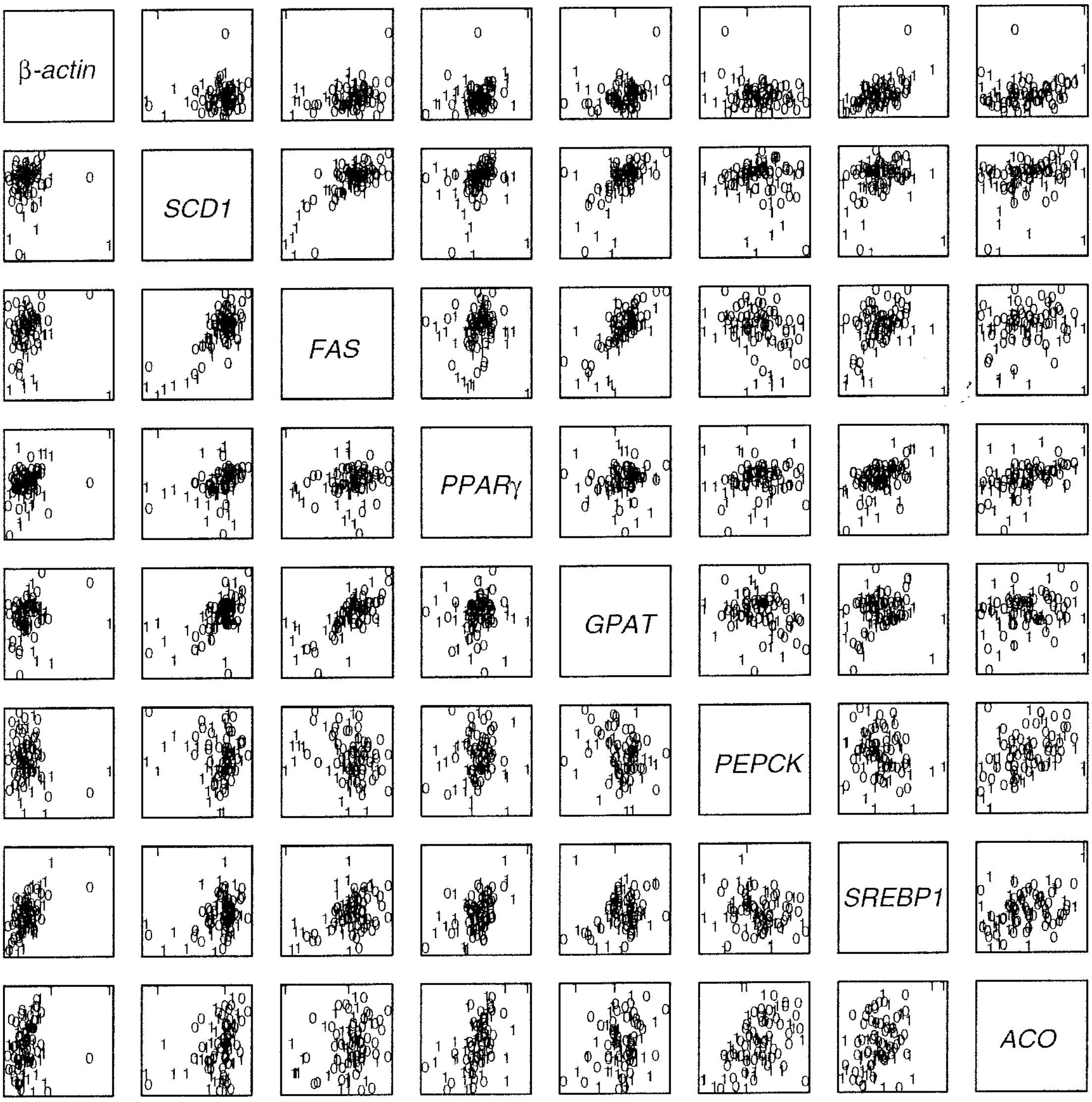

Figure 1.—Pairwise bivariate scatter plots of the raw C_T values for β-*actin* and ΔC_T values for the seven other mRNAs in the livers of (BTBR × B6) F_2-*ob/ob* mice at 14 weeks of age. The figure is symmetric, meaning that the bottom left and the top right halves are identical. For each plot, the *x*-axis plots the measurements for the gene in the row and the *y*-axis plots the values for the gene in the corresponding column. Individuals are plotted by gender (0, females; 1, males).

arbitrary number of other multiple QTL across the genome. This posterior is inversely proportional to the positive FDR (Storey 2002) for QTL discovery. That is, false positives are more probable in regions of low posterior probability. See appendix for details.

Mapping principal components: Some authors have used singular value decomposition or principal component analysis to reduce dimensionality for microarray data analysis (West *et al.* 2001); others have used principal components generated from morphological phenotypes as quantitative traits for multiple interval mapping in genetic crosses (Liu *et al.* 1996; Zeng *et al.* 2000) or for association analysis in pedigrees (Chase *et al.* 2002). Each principal component is a linear combination, or weighted average, of the original data values. The first few principal components capture most of the information in the original mRNA values as new, uncorrelated supergenes. We constructed principal components using the standardized ΔC_T values of the mRNA in PROC PRINCOMP in SAS (SAS 1999). We then mapped the principal components using both multiple interval mapping and Bayesian interval mapping.

RESULTS

We analyzed liver abundance of seven metabolic mRNAs from a population of 108 F_2-*ob/ob* mice using RT-PCR. In clustering analysis, we intentionally included some data that we would not expect to be connected to the disease process (raw C_T values of β-*actin*) and some genes that may or may not have a connection with the pathogenesis of type 2 diabetes at the level of mRNA expression. Pairwise bivariate scatter plots of the ΔC_T values for each mRNA yield clues to the correlation structure of the data set: there are a few pairs of highly

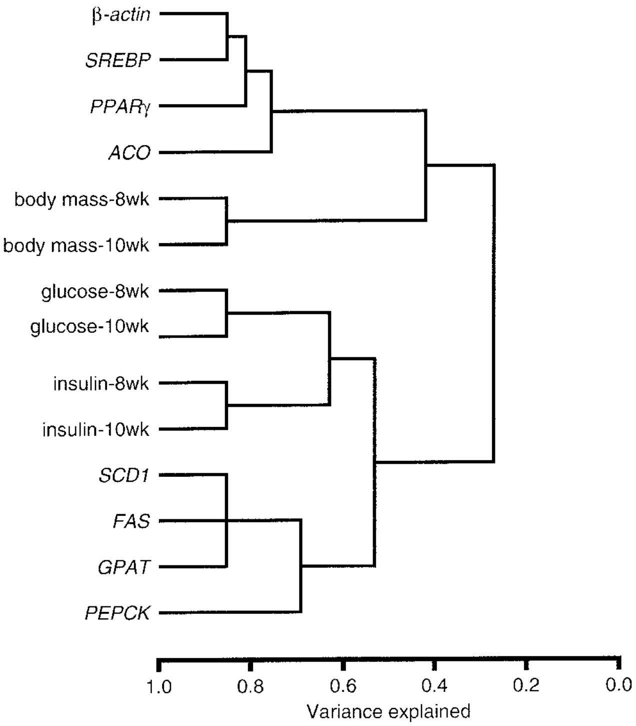

Figure 2.—Hierarchical clustering of eight mRNA traits, seeded with the 8- and 10-week levels of fasting plasma glucose, insulin, and body mass using oblique principal components via PROC VARCLUS. The raw ΔC_T values (C_T values for β-*actin*) were first standardized to mean 0 and variance 1 for each gene and gender.

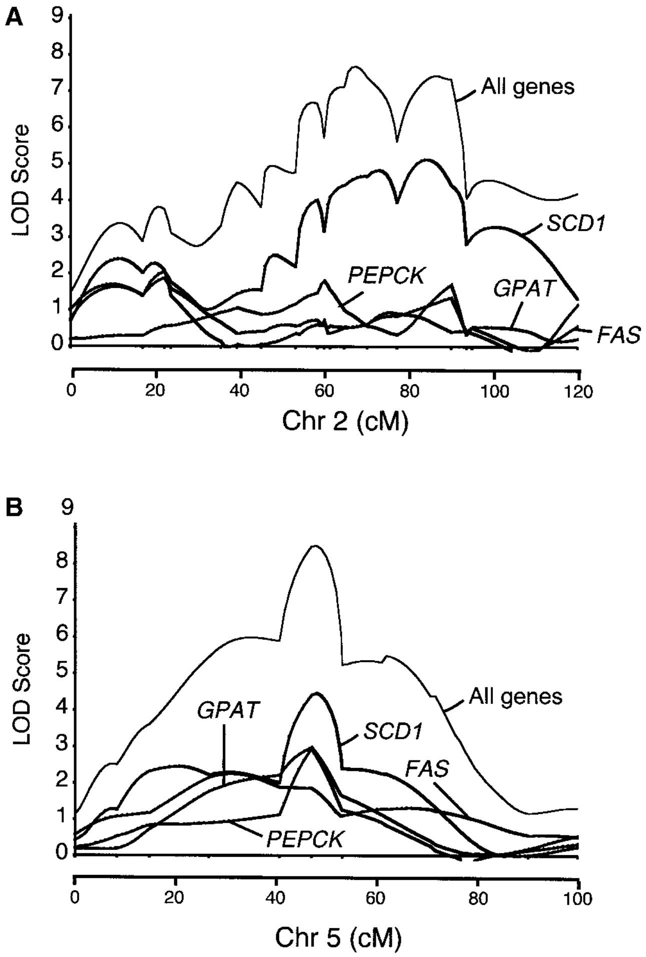

Figure 3.—Multiple-trait interval mapping for the four-gene cluster containing *SCD1, FAS, GPAT,* and *PEPCK* onto chromosome 2 (A) and chromosome 5 (B). LOD scores for each individual gene are depicted by the traces. The composite linkage signal for all of the genes together is depicted by the topmost trace. The tick marks on the *x*-axes represent genetic markers. The rulers below the *x*-axes represent the genetic distances from the centromeres calculated by MAPMAKER/EXP.

correlated mRNAs in the data (Figure 1). For example, *SCD1* and *FAS* are closely correlated.

We performed hierarchical clustering on the data set of mRNA seeded with glucose, insulin, and body mass traits. The results demonstrate that the eight mRNAs segregated into two distinct groups by their statistical correlation with the seeded traits (Figure 2). One group, composed of β-*actin*, *PPARγ*, *SREBP*, and *ACO*, showed poor correlation to the diabetes phenotypes. By linear regression, the first principal component of these four genes explains only 1.2% of the variance in the first principal component of 8- and 10-week fasting glucose and insulin; furthermore, they showed no significant QTL when we attempted multiple-trait interval mapping (data not shown).

The other group, *SCD1*, *FAS*, *GPAT*, and *PEPCK*, showed strong association with the insulin trait. While we primarily used oblique principal components in PROC VARCLUS, various other hierarchical clustering methods (*e.g.*, using PROC CLUSTER) consistently found *SCD1, FAS,* and *GPAT* clustered with the insulin traits. Multiple-trait interval mapping (MTM) revealed two loci with high LOD scores for the composite of the four mRNAs, which we named *Diabetes mRNA Cluster 1* and *2* (*DMC1* and *2*). *DMC1* is on chromosome 2: the MTM peak linkage LOD of 7.7 is reached at the marker *D2Mit106* (Figure 3A). The region overlaps with *t2dm3*, a locus previously shown to associate with fasting insulin levels in the same population of mice (Stoehr *et al.* 2000). The majority of the linkage signal at *DMC1* is accounted for by *SCD1* mRNA alone, because *FAS*, *GPAT*, and *PEPCK* do not yield LOD scores above 2.0 in this region. The result suggests that *FAS*, *GPAT*, and *PEPCK* are not regulated by *DMC1*. At the *DMC1* LOD peak, each BTBR allele increases the *SCD1* ΔC_T value by 0.8 cycles, corresponding to a 3.4-fold difference in mRNA abundance between *B6* and *BTBR* homozygotes. The heterozygotes have ΔC_T values 0.6 cycles greater than the mean of the homozygotes, indicating that the BTBR allele is dominant over B6 to elevate the *SCD1* expression level. These estimates agree with prior studies of *t2dm3*, which show that the *BTBR* allele acts in a dominant fashion to raise fasting insulin levels.

DMC2 is located on chromosome 5 (Figure 3B), near

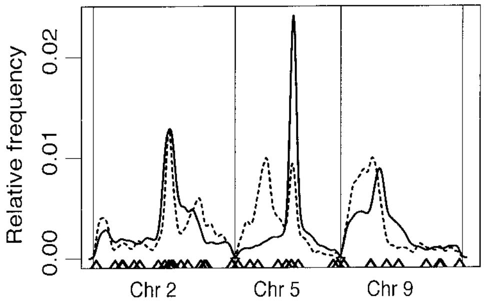

Figure 4.—Bayesian interval mapping of *SCD1* mRNA abundance (solid line) and the first principal component from all the mRNA traits (dashed line), using a model allowing for three or more QTL. Curves are density function or relative frequency histograms across the genome. Triangles on the *x*-axis represent genetic markers. The method detects two regions corresponding to *DMC1-2* and a third region on chromosome 9, named *DMC3*.

an unnamed suggestive linkage to fasting glucose levels previously observed in (BTBR × B6) F_2-*ob/ob* mice (Stoehr *et al.* 2000). The MTM peak LOD of 8.5, near the marker *D5Mit240*, is also mostly attributable to the *SCD1* expression level (LOD = 4.5), although those of *FAS* and *GPAT* also show weaker linkage. *DMC2* exhibits a mode of inheritance similar to that of *DMC1*: the *BTBR* allele acts dominantly to increase *SCD1* mRNA abundance.

Since *SCD1* mRNA abundance emerged as a strong trait contributing to *DMC1* and *DMC2*, we applied multiple interval mapping (MIM) and Bayesian interval mapping (BIM) to refine the genome-wide linkage of *SCD1* mRNA abundance. MIM found QTL at 66 cM on chromosome 2 and at 60 cM on chromosome 5 (69% heritability), a suggestion of another QTL at 10 cM, and epistasis between chromosome 2 and chromosome 5 loci (78% heritability, 5% due to epistasis). BIM, which does not at present estimate epistasis, found similar results, with QTL at 63 cM on chromosome 2, 48 cM on chromosome 5, and 31 cM on chromosome 9, and two suggested QTL on chromosome 2 at 10 and 75 cM. Pairscan using R/qtl (Broman *et al.* 2003) found evidence for QTL on chromosomes 2, 5, and 9 and epistasis between chromosomes 2 and 5 and between chromosomes 5 and 9. However, both MIM and pairscan seem to have difficulty localizing QTL on chromosomes 2 and 5 (data not shown). Thus, regardless of the analytical approach, there is evidence for complicated genetic architecture, with multiple QTL and possibly epistatic interactions. The results from Bayesian interval mapping are shown in Figure 4. In addition to *DMC1* and *DMC2*, the methods detected strong evidence of a third *SCD1* mRNA abundance QTL on chromosome 9, which is designated *DMC3* (Figure 4). The *B6* allele of *DMC3* acts in complete dominance to raise *SCD1* mRNA levels.

As an alternative to clustering, singular value decomposition is another way to reduce dimensionality of genome-wide expression data (West *et al.* 2001). We computed principal components from the standardized ΔC_T values for the seven mRNAs (Figure 5) and mapped the first two principal components. Both multiple interval mapping and Bayesian interval mapping detected linkages between the first principal component (PC1) and all three *DMC* loci (Figure 4). Bayes' factors supported at least three QTL for *SCD1* and PC1, all located on chromosomes 2, 5, and 9. Thresholds of 0.0047 for *SCD1* and 0.0070 for PC1 correspond to 50% high posterior density and genome-wide positive false discovery rates (Storey 2002) for QTL of 13.2 and 4.2%, respectively (see appendix). We checked if a small number of genes are accounting for this principal component by computing the correlation to each of the individual mRNAs. PC1 was largely accounted for by *SCD1* and *FAS* mRNA abundances (Figure 5). Thus, both clustering and principal component analysis lead us to the same conclusion: *SCD1* mRNA abundance links to loci that are associated with type 2 diabetes.

DISCUSSION

Messenger RNA abundance offers new insight in genetic mapping studies. Genetic variation in a *cis*-acting sequence might lead to changes in gene expression. This could result in a link to the location of the gene itself in a mapping study. Alternatively, variation in gene expression could result from genetic variation in *trans*-acting factors that segregate in a cross. By using mRNA levels as quantitative traits, it may be possible to map such *trans*-acting factors in genetic crosses. If multiple mRNAs map to a single locus, novel pathways of coordinate regulation might be inferred. As an example of such a study, Dumas *et al.* (2000) mapped the expression levels of several genes encoding heat-shock proteins (*Hsp*) in a panel of recombinant inbred rat strains. A region on rat chromosome 7 showed significant linkages to the mRNA abundance of these *Hsp* genes, and the region harbors an *Hsp* transcription factor (Dumas *et al.* 2000). Recently, R. W. Williams, S. Shou, L. Lu, Y. Qu, J. Wang, K. Manly, E. Chesler, H. C. Hsu, J. Mountz and D. Threadgill (unpublished data; see http://www.complextrait.org/ctc2002/start.html) exploited recombinant inbred mouse strains in combination with microarrays to map large sets of *cis*- and *trans*-acting modulators of transcriptional activity in brain. This was also accomplished in yeast (Brem *et al.* 2002).

Simply mapping individual mRNA abundance measurements, although feasible in a small-scale experiment, is not an optimal method. Detection of true linkage could be buffered by mRNAs that either play no role in the disease process or are not variable within the population. As noted by Jiang and Zeng (1995), statistical power tends to be lower and sampling variances of parameter estimates tend to be higher when multiple

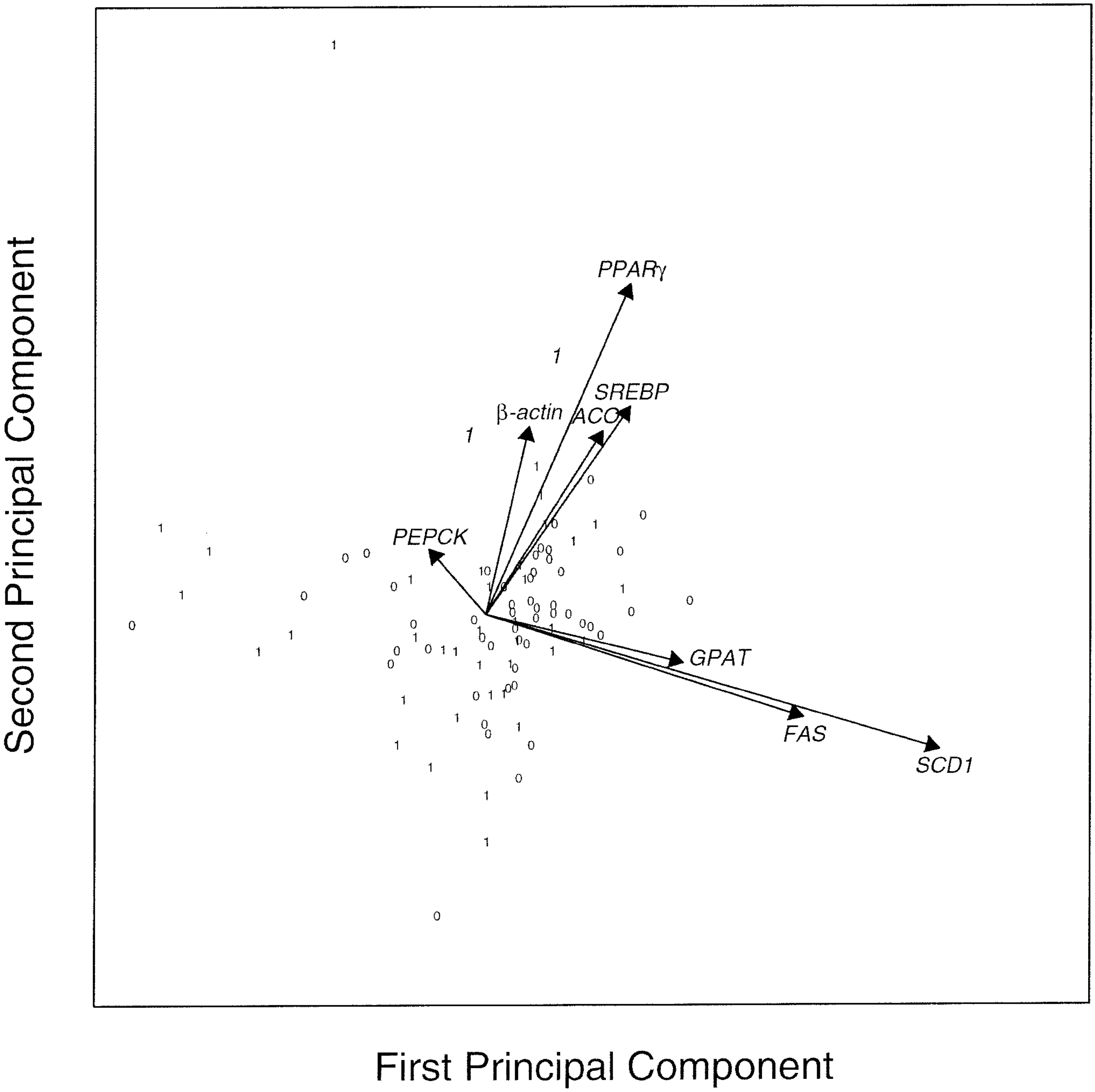

Figure 5.—Principal components analysis of the eight mRNA traits. Again, ΔC_T values (C_T values for β-*actin*) were standardized to mean 0 and variance 1. Individuals are plotted by gender (0, females; 1, males) on the plane of the first and second principal components of the data set. Arrows depict the loadings of each mRNA trait into each principal component.

correlated traits are mapped individually. While there are no established guidelines for genome-wide multiple-trait analysis, the multiple-trait LOD threshold should increase with the number of included traits. Thus multiple-trait analysis is an effective approach with a modest number of correlated traits, but can be counterproductive when used with many uncorrelated traits.

Model selection for a large number of individual traits necessitates proper multiple testing control, such as the positive FDR (Storey 2002). However, it is important to allow for multiple QTL for each trait, and this is extremely computationally intensive with today's methodology. Further, we want to capitalize on the correlation among traits, particularly with gene expression data, when selecting a model of controlling loci. Thus, individual trait mapping can be most effective as a confirmation of discoveries gleaned by multivariate methods.

We proposed in this proof-of-principle study an approach to reduce the dimensionality of the gene expression data before applying the data to mapping analysis. By including the disease traits in the analysis, clustering may help to exclude genes that contribute little or no information about the disease process being studied. Principal component analysis reduces the expression of thousands of individual genes in a microarray study to only a handful of "superphenotypes," each of which captures a composite picture of vast tracts of the microarrays (West *et al.* 2001). We computed principal components from the four mRNAs in the *SCD1-FAS-GPAT-PEPCK* cluster and performed genome-wide interval mapping. The first principal component, which was primarily accounted for by *SCD1*, showed linkages with *DMC1* and *DMC2* (data not shown). However, the LOD scores were not as high as those yielded by *SCD1* alone. As we anticipated, the linkage signal was diluted by other genes in the cluster. It is worth pointing out that caution should be taken when using data reduction techniques, because information is lost during the process of reduction. Sometimes it may be difficult to distinguish whether or not a superphenotype is associated with a disease of interest. In this case, it may in fact be more fruitful to map individual gene expression values. Thus the techniques should be applied with care and with awareness of potential shortcomings. A comprehensive study should consider data reduction combined with individual trait analysis (Mahler *et al.* 2002).

It is not a surprise that *SCD1*, *FAS*, *GPAT*, and *PEPCK* showed strong association with insulin in the clustering analysis. The first three genes encode lipogenic enzymes; the fourth gene, *PEPCK*, encodes a rate-limiting enzyme in the gluconeogenesis pathway. All the genes are known targets of insulin regulation (Shimomura *et al.* 2000). It is known that *SREBP1* is a major mediator of insulin function on hepatic lipogenesis (Shimomura *et al.* 1999), but *SREBP1* mRNA failed to cluster with the insulin trait and lipogenic genes. A possible reason is that the difference of *SREBP1* expression levels in the parental strains was not large enough, so that the variance in *SREBP1* expression is affected by nongenetic factors. By including only the genes that are cosegregating with the seed trait in the F_2 population, we obtained an enhanced linkage signal as we were mapping the concerted action of these genes. With only 108 individuals, we were able to produce LOD scores that had required more than triple the number of mice in the previous study (Stoehr *et al.* 2000).

DMC1 on chromosome 2 was primarily accounted for by *SCD1*. This region may harbor a new regulator controlling the expression of the *SCD1* gene. Recently, we (Ntambi *et al.* 2002) and others (Cohen *et al.* 2002) have shown that loss of *SCD1* function (*SCD1*$-/-$) protects mice against adiposity, in part by raising the total metabolic rate. We have also shown that *SCD1*$-/-$ mice demonstrate improved oral glucose tolerance compared to wild-type controls. Here, we show that liver expression of *SCD1* mRNA significantly links to a region of chromosome 2 that is associated with susceptibility to type 2 diabetes in a genetic system known to segregate insulin resistance alleles (Ranheim *et al.* 1997; Stoehr *et al.* 2000). Our previous study (Stoehr *et al.* 2000) found that the BTBR allele of *t2dm3* increases the level of fasting insulin. The present study shows that the BTBR allele of *DMC1* (*SCD1*) also increases *SCD1* mRNA.

DMC2 on chromosome 5 regulates the concerted action of all four genes in the insulin cluster. The gene in *DMC2* may be a general regulator of lipogenesis and gluconeogenesis. The Bayesian interval mapping of the first principal component of the eight mRNAs revealed two LOD peaks on chromosome 5, one accounted for by *SCD1* and the other by *FAS* (Figure 4). It is unclear whether either of the genes is the same as the one that produced a suggestive linkage of fasting glucose in approximately the same region (Stoehr *et al.* 2000).

In summary, we have shown how to use clustering and principal components analysis to combine mRNA abundance traits to form new traits that can be genetically mapped. This proof-of-principle experiment can be scaled to microarray experiments to map disease phenotypes composed of gene expression levels.

We thank W. F. Dove for his strong encouragement and his helpful comments on this manuscript. The work was funded by National Institute of Diabetes and Digestive and Kidney Diseases grant DK58037 and by Xenon Genetics, Inc.

LITERATURE CITED

Basten, C. J., B. S. Weir and Z-B. Zeng, 1994 Zmap: a QTL cartographer, pp. 65–66 in *Proceedings of the 5th World Congress on Genetics Applied to Livestock Production: Computing Strategies and Software*, edited by E. B. Burnside. Organizing Committee, 5th World Congress on Genetics Applied to Livestock Production, Guelph, Ontario, Canada.

Benjamini, Y., and Y. Hochberg, 1995 Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. Ser. B **85:** 289–300.

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

Broman, K. W., H. Wu, S. Sen and G. A. Churchill, 2003 R/qtl: QTL mapping in experimental crosses. Bioinformatics **19:** 889–890.

Chase, K., D. R. Carrier, F. R. Adler, T. Jarvik, E. A. Ostrander *et al.*, 2002 Genetic basis for systems of skeletal quantitative traits: principal component analysis of the canid skeleton. Proc. Natl. Acad. Sci. USA **99:** 9930–9935.

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

Cohen, P., M. Miyazaki, N. D. Socci, A. Hagge-Greenberg, W. Liedtke *et al.*, 2002 Role for stearoyl-CoA desaturase-1 in leptin-mediated weight loss. Science **297:** 240–243.

Dumas, P., Y. Sun, G. Corbeil, S. Tremblay, Z. Pausova *et al.*, 2000 Mapping of quantitative trait loci (QTL) of differential stress gene expression in rat recombinant inbred strains. J. Hypertens. **18:** 545–551.

Eisen, M. B., P. T. Spellman, P. O. Brown and D. Botstein, 1998 Cluster analysis and display of genome-wide expression patterns. Proc. Natl. Acad. Sci. USA **95:** 14863–14868.

Gaffney, P., 2001 An efficient reversible jump Markov chain Monte Carlo approach to detect multiple loci and their effects in inbred crosses. Ph.D. Thesis, University of Wisconsin, Madison, WI.

Jiang, C., and Z-B. Zeng, 1995 Multiple trait analysis of genetic mapping for quantitative trait loci. Genetics **140:** 1111–1127.

Kao, C. H., Z-B. Zeng and R. D. Teasdale, 1999 Multiple interval mapping for quantitative trait loci. Genetics **152:** 1203–1216.

Lander, E. S., 1999 Array of hope. Nat. Genet. **21:** 3–4.

Lander, E. S., P. Green, J. Abrahamson, A. Barlow, M. J. Daly *et al.*, 1987 MAPMAKER: an interactive computer package for constructing primary genetic linkage maps of experimental and natural populations. Genomics **1:** 174–181.

Liu, J., J. M. Mercer, L. F. Stam, G. C. Gibson, Z-B. Zeng *et al.*, 1996 Genetic analysis of a morphological shape difference in the male genitalia of *Drosophila simulans* and *D. mauritiana.* Genetics **142:** 1129–1145.

Mahler, M., C. Most, S. Schmidtke, J. P. Sundberg, R. Li *et al.*, 2002 Genetics of colitis susceptibility in IL-10-deficient mice: backcross versus F2 results contrasted by principal component analysis. Genomics **80:** 274–282.

McCarthy, M. I., and P. Froguel, 2002 Genetic approaches to the molecular understanding of type 2 diabetes. Am. J. Physiol. Endocrinol. Metab. **283:** E217–E225.

Ntambi, J. M., M. Miyazaki, J. P. Stoehr, H. Lan, C. M. Kendziorski *et al.*, 2002 Loss of stearoyl-CoA desaturase-1 function protects mice against adiposity. Proc. Natl. Acad. Sci. USA **99:** 11482–11486.

Ranheim, T., C. Dumke, K. L. Schueler, G. D. Cartee and A. D. Attie, 1997 Interaction between BTBR and C57BL/6J genomes produces an insulin resistance syndrome in (BTBR × C57BL/6J) F1 mice. Arterioscler. Thromb. Vasc. Biol. **17:** 3286–3293.

SAS, 1999 *SAS Version 8.00.* SAS Institute, Cary, NC.

Satagopan, J. M., B. S. Yandell, M. A. Newton and T. C. Osborn, 1996 A bayesian approach to detect quantitative trait loci using Markov chain Monte Carlo. Genetics **144:** 805–816.

Shimomura, I., Y. Bashmakov, S. Ikemoto, J. D. Horton, M. S. Brown *et al.*, 1999 Insulin selectively increases SREBP-1c mRNA in the livers of rats with streptozotocin-induced diabetes. Proc. Natl. Acad. Sci. USA **96:** 13656–13661.

Shimomura, I., M. Matsuda, R. E. Hammer, Y. Bashmakov, M. S. Brown *et al.*, 2000 Decreased IRS-2 and increased SREBP-1c lead to mixed insulin resistance and sensitivity in livers of lipodystrophic and ob/ob mice. Mol. Cell **6:** 77–86.

Stoehr, J. P., S. T. Nadler, K. L. Schueler, M. E. Rabaglia, B. S. Yandell *et al.*, 2000 Genetic obesity unmasks nonlinear interactions between murine type 2 diabetes susceptibility loci. Diabetes **49:** 1946–1954.

Storey, J. D., 2002 A direct approach to false discovery rates. J. R. Stat. Soc. Ser. B **64:** 479–498.

West, M., C. Blanchette, H. Dressman, E. Huang, S. Ishida *et al.*, 2001 Predicting the clinical status of human breast cancer by using gene expression profiles. Proc. Natl. Acad. Sci. USA **98:** 11462–11467.

Zeng, Z-B., J. Liu, L. F. Stam, C. H. Kao, J. M. Mercer *et al.*, 2000 Genetic architecture of a morphological shape difference between two Drosophila species. Genetics **154:** 299–310.

Communicating editor: G. Churchill

APPENDIX

The positive false discovery rate provides an estimate of the percentage of false positives. It has been used recently for analysis of gene expression in microarrays (see Storey 2002 and http://www.stat.berkeley.edu/~storey). While we support such a use, our application is somewhat distinct, as we provide percentage of false positive for the gene mapping of putative QTL for those mRNAs. Suppose that only one QTL is controlling an mRNA, with either *cis*- or *trans*-action. Given a marginal posterior density for a QTL, pr(λ | data), and a highest probability density (HPD) region (say 50%) that thresholds down from the peak, the positive FDR is

$$\mathrm{pr}(H = 0|\lambda \text{ in HPD, data}) = \frac{\mathrm{pr}(H = 0|\text{data})\mathrm{pr}(\lambda \text{ in HPD}|H = 0, \text{data})}{\mathrm{pr}(\lambda \text{ in HPD}|\text{data})},$$

with $H = 0$ being the event that no QTL is at locus λ. The conditional probability that λ is in the HPD when $H = 0$ is simply the relative length of the HPD region. The unconditional probability that λ is in the HPD is essentially the marginal posterior density for a QTL, since in most experiments the posterior probability that there is no QTL is negligible.

That is, the concern is not whether there are any QTL, but whether the threshold approach for Bayesian HPD regions has a high chance of making mistakes, *i.e.*, false positive detection. When there are multiple QTL, the above idea can be extended by using the joint posterior for multiple QTL. We choose instead to consider the marginal posterior that a QTL is found at a locus allowing for an arbitrary number of other multiple QTL, as presented in Figure 4. This approximately captures the margins of joint posterior for multiple QTL when they are not too closely linked and can be a useful diagnostic.

The conservative choice of $\mathrm{pr}(H = 0|\text{data}) = 1$ due to Benjamini and Hochberg (1995) can be improved (Storey 2002) by using the lowest probability density (LPD) region. That is, threshold in the opposite direction, from 0 upward, and estimate $\mathrm{pr}(H = 0|\text{data})$ as the proportion the LPD is to its relative genome length as the threshold tends to 0. In our experiment, this proportion estimated the probability that any locus on chromosomes 2, 5, or 9 is not a QTL as 36.3% for *SCD1* and 13.1% for the PC supergene.