OMICS A Journal of Integrative Biology
Volume 11, Number 1, 2007

DOI: 10.1089/omi.2006.0007

Evolutionary-conserved Gene Expression Response Profiles Across Mammalian Tissues

JI CHEN,[1] THOMAS W. BLACKWELL,[1] DAMIAN FERMIN,[1] RAJASREE MENON,[1] YILI CHEN,[1] JING GAO,[1] ANGEL W. LEE,[2] and DAVID J. STATES[1,3]

ABSTRACT

Gene expression responses are complex and frequently involve the actions of many genes to effect coordinated patterns. We hypothesized these coordinated responses are evolutionarily conserved and used a comparison of human and mouse gene expression profiles to identify the most prominent conserved features across a set of normal mammalian tissues. Based on data from multiple studies across multiple tissues in human and mouse, 13 gene expression modes across multiple tissues were identified in each of these species using principal component analysis. Strikingly, 1-to-1 pairing of human and mouse modes was observed in 12 out of 13 modes obtained from the two species independently. These paired modes define evolutionarily conserved gene expression response modes (CGEMs). Notably, in this study we were able to extract biological responses that are not overwhelmed by laboratory-to-laboratory or species-to-species variation. Of the variation in our gene expression dataset, 84% can be explained using these CGEMs. Functional annotation was performed using Gene Ontology, pathway, and transcription factor binding site over representation. Our conclusion is that we found an unbiased way of obtaining conserved gene response modes that accounts for a considerable portion of gene expression variation in a given dataset, as well as validates the conservation of major gene expression response modes across the mammals.

INTRODUCTION

MOST OF OUR PRESENT KNOWLEDGE of gene response patterns comes from pathway databases, such as KEGG, BioCarta, and STKE. There are also gene expression analysis studies based on this knowledge, such as Gene Set Enrichment Analysis (GSEA) (Subramanian et al., 2005). These gene response patterns are basically derived from the literature, however, which introduces ascertainment bias (i.e., some genes that are part of a pathway might not be annotated as such simply because they are not well studied). Evolutionary conservation offers a powerful, unbiased alternative approach to the definition of gene response patterns.

Biological systems exist in dynamic environments requiring reaction to concurrent and complex stimuli through coordinated multigene expression responses. Functional constraints on variation lead to evolution-

[1]Bioinformatics Program, [2]Department of Pharmacology, [3]Department of Human Genetics, University of Michigan, Ann Arbor, Michigan.

ary conservation of gene expression, as well as gene sequences (Jordan et al., 2004, 2005; O'Brien and Fraser, 2005). Previously, many studies of molecular evolution, such as clusters of orthologous groups (COG) of proteins, focused on primary sequence (Doolittle, 2005; Li, 1997; Tatusov et al., 2003; von Mering et al., 2003). With the availability of gene expression profile data from widely applied microarray technology, we have seen more and more studies on evolutionary analysis of gene expression patterns (Jimenez et al., 2003). However, the details of these expression patterns vary from study to study, and many studies focus on a specific tissue (Adjaye et al., 2004), a specific process (Andersson et al., 2005; McCarroll et al., 2004), or, in particular, a specific type of cancer (Andersson et al., 2005; Fang et al., 2005).

Several previous studies on global gene expression evolution focus on the evolution model of gene expression instead of using evolution as a tool to define gene expression patterns (Enard et al., 2002; Jordan et al., 2005; Liao and Zhang, 2005; Oleksiak et al., 2002; Yanai et al., 2004). Stuart et al. (2003), however, have done analysis defining pairs of genes that are coexpressed from humans, files, worms, and yeast and thus have discovered some conserved genetic modules. In our work, instead of using pairwise gene expression comparisons as Stuart and co-workers have, we demonstrate a new, systematic way of defining conserved gene expression modes. We use a simple numeric method—principal component analysis (PCA)—to find conserved gene expression response modes (CGEMs) prominent in a given dataset between two species; these CGEMs can be characterized as a tool to better understand the functional role and mechanisms underlying gene expression responses.

Principal component analysis is a widely used exploratory data analysis tool that is able to identify structure in complex multidimensional data. The technique has been used in microarray analysis in different ways, such as summarizing experimental conditions or selecting gene markers in cancer search (Alter et al., 2000; Bicciato et al., 2003; Raychaudhuri et al., 2000; Wang and Gehan, 2005; Yeung and Ruzzo, 2001). Of particular relevance to biological interpretation, several studies have examined the relationship of function to PCA component loading in gene expression responses. For example, Crescenzi and Giuliani (2001) assigned biological themes to the major components, and Misra et al. (2002) used component loading in identification of tissue-specific gene expression patterns. These studies demonstrate that the modes of gene expression responses defined by PCA can be interpreted in biological terms. Our study might be the first to utilize PCA in an evolutionary context and find conservation between human and mouse principal components. We examine those components (modes) that are conserved between species to strengthen the connection to the functional biological processes.

There are several advantages to using PCA analysis to identify conserved gene expression modes. First, the mathematics of PCA ensures that we choose the modes that represent directions of largest amount of variation in gene expression dataset, thus giving us the most prominent patterns from the current dataset. Each PCA mode represents a specific combination of tissue specificity; for example, a certain mode might represent the scenario where genes have very high expression in lung, medium-low expression in liver, very low expression in brain, and medium expression in other tissues. This combination captures maximal portion of gene expression variation not explained by the preceding modes. Traditional tissue specificity studies have focused on gene expression variation in individual tissues, but do not examine coordinated, quantified tissue specificity present as major patterns in datasets across different types of tissues. Our work, in this aspect, stands out distinctly from other work (Liao and Zhang, 2005). Second, studies on finding global conserved gene expression patterns using data from heterogeneous sources are often confounded by technical variation, resulting in strong within-laboratory and within-species correlations (Hampson and Hughes, 2001; Irizarry et al., 2003b). For example, such technical variation sometimes masks the conservation of gene expression patterns between human and mouse while conservation is expected because the two species share common phylogeny and ontogeny (Yanai et al., 2004). PCA, by extracting major biological patterns, can make this interspecies conservation discernable, as it did in our study. Third, PCA has an advantage over clustering and classification, which are used in many studies, in that genes are not artificially constrained to belong in a single mode. Since the selection of representative genes for each mode is based on loadings on each principal component, it is possible that a certain gene can be selected in several response modes instead of one, which can be the reality in biological systems.

In this work, we developed a method that identifies evolutionary conserved modes of expression that are prominent in a given dataset based on matching of principal components of gene expression between hu-

man and mouse. We demonstrate our method on a dataset integrated from multiple studies across multiple tissues from the two species and find 12 CGEMs that are associated with functions fundamental to the two species and conserved across evolution. Also, the promoter regions of the genes are studied to reveal transcription factors that may be related to these biological processes.

MATERIALS AND METHODS

Gene expression data assembly

First, gene expression data were gathered for both human and mouse. We restricted our analysis to Affymetrix GeneChip data to reduce technical variations between samples and between species. For human we used four datasets: (1) Novartis Research Foundation's dataset of normal tissues and cell lines (GEO ID: GSE96) (Barrett et al., 2005; Edgar et al., 2002; Su et al., 2002), which has 80 samples; (2) Genenote (gene normal tissue expression) dataset (GEO ID: GSE803) (Shmueli et al., 2003), also of normal tissues, which has 24 samples; (3) a blood dataset (Feezor et al., 2003) with control and heat-killed SAC treated cells, which has 9 samples; and (4) a dendritic cell dataset (Messmer et al., 2003) with control and cells treated with LPS, CD40L, or CyC (TNF-a, IL-1b, IL-6+PGE-2), which has 32 samples. For mouse we used two datasets: (1) Novartis Research Foundation's dataset of normal tissues and cell lines (GEO ID: GSE97) (Su et al., 2002), which has 90 samples; and (2) a 32D cell dataset with control and cells treated with Csf1 or IL3, which has 12 samples.

For the procedure to obtain the 32D cell mouse dataset, interleukin-3 (IL-3)-dependent murine 32D myeloid clones stably expressing the wildtype colony-stimulating factor-1 receptor (Lee and States, 2000) were deprived of IL-3 for 6 h, followed by the addition (or not) of IL-3 for 18 h. Cells were washed and RNA was harvested using the RNAeasy kit (Qiagen). Two independently selected clones were used and either three or four sets of data were collected for starved or IL-3-treated cells on 3 separate days. Processing and hybridization to the Affymatrix U74A chip were performed by the Washington University Medical School (St. Louis) Gene Chip Core Facility. Altogether there were 145 human samples and 102 mouse samples.

Processing and integration of datasets

After we obtained the original gene expression datasets in .CEL files, we processed them using the RMA method (Bolstad et al., 2003; Irizarry et al., 2003a, 2003b) as implemented in the R Bioconductor package (Gentleman et al., 2004). Multiple species and multiple datasets were used; for human we have data from both version 1 and 2 of U95A chip, and for mouse we have data from version 1 and 2 of U74A chip. We used Bolstad's mixed CDF environment, which takes only probe sets that appear on both chip types (<http://stat-www.berkeley.edu/users/bolstad/mixtureCDF/MixtureCDF.html>), and were able to pool the data into two groups: human data from U95A chip series, and mouse data from U74A chip series. After that RMA preprocessing was carried out in R package.

Pairing genes and tissues between species

To compare the two species, orthologous genes were assigned using the TIGR database (<http://pga.tigr.org/AnalysisTools.shtml>) (Lee et al., 2002; Tsai et al., 2001). We only chose gene pairs that have expression information in the microarray data we use. Similar results are obtained using the NCBI Homologene and ENSMEBL orthologs assignments.

Tissue correspondence also has to be assigned before comparing the two species. Anatomic origin and similarity in orthologous gene expression profiles were both used to pair samples from human and mouse. Samples were paired if three criteria were met: (1) if the Manhattan distance between orthologous gene expression profiles was in the upper 10% of all sample pairs; (2) the two samples fell within the reciprocal 10 top hits respectively; and (3) pairing could be validated by anatomic naming. Manhattan distance is defined as the distance between two points measured along axes at right angles (i.e., in a plane with p_1 at (x_1, y_1) and p_2 at (x_2, y_2), Manhattan distance is $|x_1 - x_2| + |y_1 - y_2|$). Using this process we assigned the 13

pairs of tissue sets in Table 1. Gene expression values were averaged in each tissue set within each species. By now we have aligned human and mouse dataset with regard to genes and tissues.

Principal component analysis and clustering

The gene expression model we are assuming is

$$E_{gt} = \sum_{p} \alpha_{gp}\rho_{pt},$$

where subscript g indicates gene, t indicates tissue, and p indicates profile. E is expression intensity, α_{gp} is projection of gene g on profile p, and ρ_{pt} is weight of tissue t on profile p. Written in matrix form $E = AP$, where A is the projection matrix and P is the matrix connecting profiles and tissues. PCA is one kind of transformation we can perform to get A from E. We performed PCA on human and mouse data separately (the two datasets had been aligned regarding genes and tissues, as mentioned above). PCA was performed using "prcomp" in the R "stat" package with singular value decomposition of the data matrix rather than "eigen" on the covariance matrix. This method is recommended for numerical accuracy. After that, principal components obtained using PCA from the two species were clustered. For clustering we chose an unsupervised method—hierarchical clustering, also implemented in R. We found that 1 to 1 pairing of human and mouse components is observed for 12 out of 13 principal components obtained from the two species independently, and thus we defined 12 CGEMs.

Reconstructing gene expression using average loadings

Next we assessed how much of the gene expression variation could be accounted for by the conserved part of principal components. From our expression model

$$E_{gt} = \sum_{p} \alpha_{gp}\rho_{pt},$$

the mathematics of PCA guarantees we can reconstruct E_{gt} from α_{gp} and ρ_{pt}. In this step, however, instead of using α_{gp} from the concerned species, we used $(\alpha^h_{gp} + \alpha^m_{gp})/2$ where superscript h and m indicated human and mouse, respectively. Thus for human we applied the formula

$$E^h_{gt}\,(predicted) = \sum_{p} (\alpha^h_{gp} + \alpha^m_{gp})/2{*}\rho^h_{pt}$$

to calculate predicted values of gene expression, and for mouse the formula we used was

$$E^m_{gt}\,(predicted) = \sum_{p} (\alpha^h_{gp} + \alpha^m_{gp})/2{*}\rho^m_{pt}.$$

Note that considering most of the variation could be explained by the first principal component, which represented average gene expression across all tissues; this was subtracted from the data during reconstruction.

Annotating gene expression response profiles

We then chose representative genes for each CGEM based on loadings of genes on corresponding human and mouse principal components. We calculated products of the positive loadings of genes on human and mouse components and selected the top 5% genes with highest products of loadings.

We annotated CGEMs based on functions of these representative genes. For functional annotation, we examined enriched GO terms, pathways, and "words" in representative genes for each CGEM. GO terms for genes were obtained from hgu74av2 and mgu74av2 packages of bioconductor (Gentleman et al., 2004). We used information from these packages to build an association file in running a software termfinder (Boyle et al., 2004), which calculates a p value using the hypergeometric distribution and outputs over-represented GO terms. When identifying over-represented Kegg pathways, we drew on information from NCBI Entrez Gene records and used Fisher's exact test with a cut-off of 0.05. When identifying over-represented

words, we again obtained information from Entrez Gene records and used Fisher's exact test with a more stringent cut-off of 0.001.

We also examined transcription factor enrichment for representative genes of CGEMs. Two approaches were used: one was using TRANSFAC, the other was using GSEA database. In the first approach, tran-

Table 1. Paired Tissue Groups between Human and Mouse

	Human		*Mouse*	
Tissue group	*Sample*	*Dataset*	*Sample*	*Dataset*
Neural	CEREBELLUM	N	Frontal cortex	N
	CEREBELLUM	N	Lower spinal cord	N
	CORTEX	N	Lower spinal cord1	N
	AMYGDALA1	N	Upper spinal cord	N
	CAUDATE NUCLEUS	N	Upper spinal cord1	N
	CAUDATE NUCLEUS1	N	Cerebral cortex	N
	FETAL BRAIN1	N	Cerebral cortex1	N
	BRAIN	G	Frontal cortex	N
	BRAIN1	G	Striatum	N
Prostate	PROSTATE	G	Prostate	N
	PROSTATE1	G	Prostate1	N
Bone marrow	BONE MARROW	G	Bone marrow	N
	BONE MARROW1	G	Bone marrow1	N
Lymphocyte	THYMUS	N	Lymphnode	N
	THYMUS1	N	Lymphnode1	N
	SPLEEN	N	Spleen	N
	SPLEEN1	N	Spleen1	N
	SPLEEN2	G	Thymus	N
	SPLEEN3	G	Thymus1	N
	THYMUS1	G		
	RAJI	N		
Myeloid	PROMYELOCYTIC LEUKEMIA	N	32Dcell	D
	PROMYELOCYTIC LEUKEMIA1	N	32Dcell1	D
	PROMYELOCYTIC LEUKEMIA2	N	32Dcell2	D
			32Dcell3	D
Liver	LIVER	G	Liver	N
	LIVER1	G	Liver1	N
	LIVER2	N		
	LIVER3	N		
Lung	LUNG	G	Lung	N
	LUNG1	G	Lung1	N
Placenta	PLACENTA	N	Placenta	N
	PLACENTA1	N	Placenta1	N
Heart	HEART	N	Heart	N
	HEART2	G	Heart1	N
	HEART3	G		
	HEART4	N		
Kidney	KIDNEY	G	Kidney	N
	KIDNEY1	G	Kidney1	N
Salivary gland	SALIVARY GLAND	N	Salivary gland	N
	SALIVARY GLAND1	N	Salivary gland1	N
Uterus	UTERUS	N	Uterus1	N
Testis	TESTIS	N	Testis1	N
	TESTIS1	N		

G, Genonote dataset; N, novartis dataset; D, 32D cell dataset.

scription factor binding site analysis was performed by obtaining sequences from ENSEMBL database (Birney, 2003), and scanning the gene region with a flanking sequence of 1 kb using TRANSFAC MATCH (Kel et al., 2003; Matys et al., 2003; Wingender et al., 2000). In the second approach, we used the motif-based gene set of GSEA (Subramanian et al., 2005) and only chose those annotated motifs. Fisher's exact test was used and a cut-off of 0.05 was applied to select over-represented transcription factors or motifs.

RESULTS

Pairing human and mouse tissues and cell lines

To assess the conserved gene expression patterns between human and mouse, tissue correspondence between the two species needs to be established. From the datasets we analyzed, we identified 13 groups of corresponding tissues (Table 1). Many of the groups contained the same tissue as indicated by anatomic name for human and mouse, but some groups contained samples from functionally related tissues where the naming used in one species did not match that used in the other. For example, the first group included neural tissues from both species but was labeled with a number of different names. This pairing involved both matching across species and clustering within a species, resulting in a many-to-many mapping of sets from each species.

Principal component analysis of gene expression responses

To assess the major expression patterns in the dataset, we employed principal components analysis. Based on the averaged expression profile within each of the 13 groups listed in Table 1, we performed principal component analysis on human and mouse expression data, respectively, with orthologous gene pairs aligned. Shown in Figure 1 is a plot of the eigen values of these components. For human data, the first principal component accounted for about 87% of the total variability while for mouse data it accounted for about

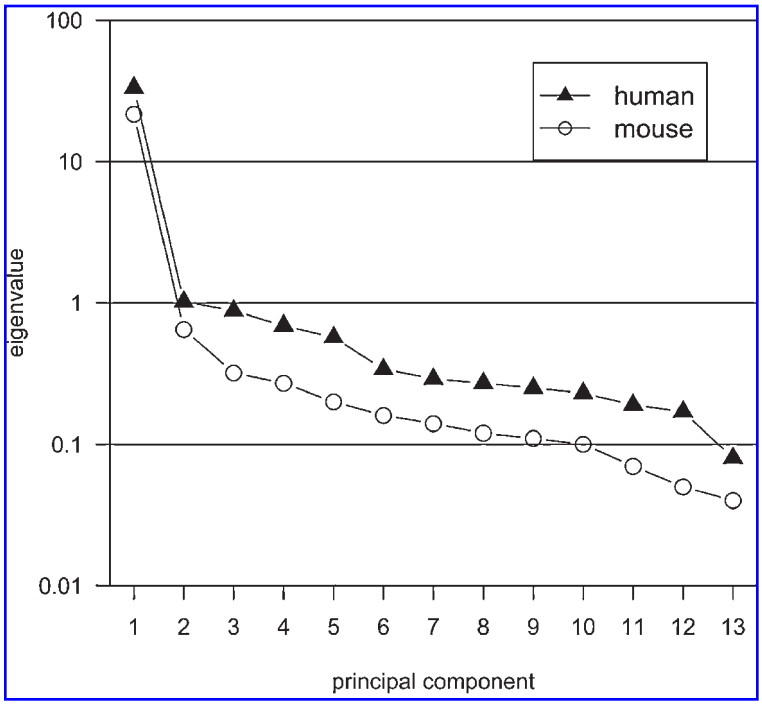

FIG. 1. Plot of Eigen values of the principal components for human and mouse species. Most of the variance in the data is contained in the first principal components.

91% of the total variability. The first component represents the average expression level across tissues. For detailed information of eigen values and the variance they explain, see Table 2.

Comparison of species-specific principal components

After we performed PCA separately on human and mouse data, the loadings on each of the 2×13 principal components were clustered to determine if there was a correspondence between principal components between the two species. As Figure 2 shows, there is highly significant pairing between human and mouse principal components based on gene loadings, indicating that there is a high degree of evolutionary conservation with the gene expression modes revealed by PCA. We found that there were 12 pairs of principal components in which the terminal binary nodes of the tree contain exactly one human and one mouse terminal leaf. Apart from the second component of human data and the second component of mouse data, all other components had their pairing components in the other species. Each pair of components represents a conserved gene expression response mode (CGEM) between human and mouse. The likelihood of a pairing this good occurring at random is calculated by randomizing the orthologous relationships. After 10,000 iterations, we found that such pairing occurs randomly about six times. We conclude that the correspondence of human and mouse gene expression modes is highly significant.

Conservation between species is reflected in correspondence between loadings; high loadings on both corresponding human and mouse components are of interest. By calculating products of the positive loadings, we could select genes with high loadings on both species for a specific principal component pair. Figure 3 shows an example for a component pair H4M3 (CGEM H4M3), which consists of the fourth principal component of human and the third principal component of mouse. From the distribution of cross-species loadings shown in Figure (3A) we find a few genes with large product values and a large number of genes making little contribution to the cross-species component loading dot product. The genes that have conserved expression profiles with high loadings on this component pair are selected for further analysis. We used a cut-off of 5% to find those highly conserved high loading gene set for each CGEM; the actual gene sets are provided in Appendix 1. Figure 3B gives us a more direct view of the products of loadings, and we can actually see the sparse clusters of genes with conserved higher loadings. The larger the product, the stronger the conserved component of the expression response for the gene.

Table 2. Eigen Values with Percentage of Variance and Cumulative Variance

	Human			*Mouse*		
Component no.	*Eigen value*	*Variance (%)*	*Cumulative variance (%)*	*Eigen value*	*Variance (%)*	*Cumulative variance (%)*
1	33.29	86.99	87	21.61	90.62	90.6
2	1.02	2.68	89.7	0.65	2.71	93.3
3	0.88	2.29	92	0.32	1.34	94.7
4	0.69	1.79	93.7	0.27	1.14	95.8
5	0.57	1.49	95.2	0.2	0.85	96.7
6	0.34	0.9	96.1	0.16	0.68	97.3
7	0.29	0.76	96.9	0.14	0.58	97.9
8	0.27	0.71	97.6	0.12	0.51	98.4
9	0.25	0.67	98.3	0.11	0.46	98.9
10	0.23	0.59	98.9	0.1	0.43	99.3
11	0.19	0.51	99.4	0.07	0.29	99.6
12	0.17	0.43	99.8	0.05	0.21	99.8
13	0.08	0.2	100	0.04	0.18	100

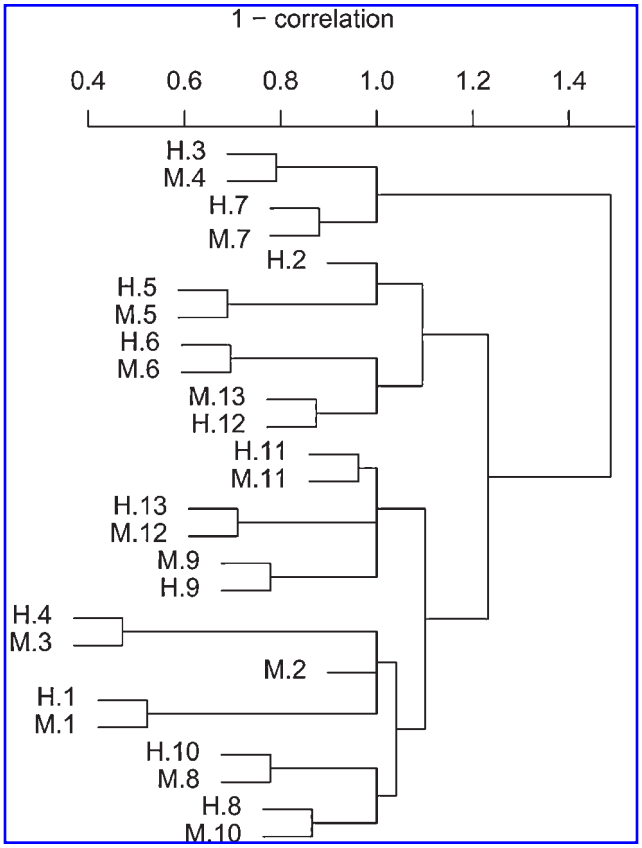

FIG. 2. Hierarchical clustering of 2×13 principal components from human and mouse. A hierarchical clustering was performed based on the loadings. Similar principal components were clustered together. Principal components from human, H.; components from mouse, M.

Reconstructing gene expression using conserved part of component loadings

To assess how much of the variation in gene expression profiles can be explained by evolutionarily conserved components, we attempted to reconstruct gene expression using average loadings (from human and mouse) on components. The mathematics of PCA guarantees that we can reconstruct the gene expression profile exactly using the species-specific rotation matrix and loadings on the components. Instead of using the loadings on components from each species independently, we used the average loading values from human and mouse as an indication of the evolutionarily conserved components of gene expression. During the process, the first principal component, which represents the average gene expression level across tissues, is subtracted from the data.

Using these evolutionarily conserved components, we were able to reconstruct an average of 84% of the variation in gene expression levels over all of the different tissue groups (the variation explained by the first principal component aside). Shown in Figure 4 is the correlation between predicted (calculated) and observed gene expression in liver, a representative tissue example. Figure 4A and B are for human and mouse liver tissue, respectively. Observed values, which is gene expression intensity, and predicted values, which is calculated from average loadings, have a good linear relationship. The correlation coefficient is 0.89 and 0.82, respectively, which implies that most of gene expression can be explained by these CGEMs. Table 3 lists the correlation of observed and predicted values for each tissue in both human and mouse. From Table 3 it can be seen that in most cases conserved gene expression contributes significantly to overall gene expression. Interestingly, the predicted values for human tend to have a better correlation with the

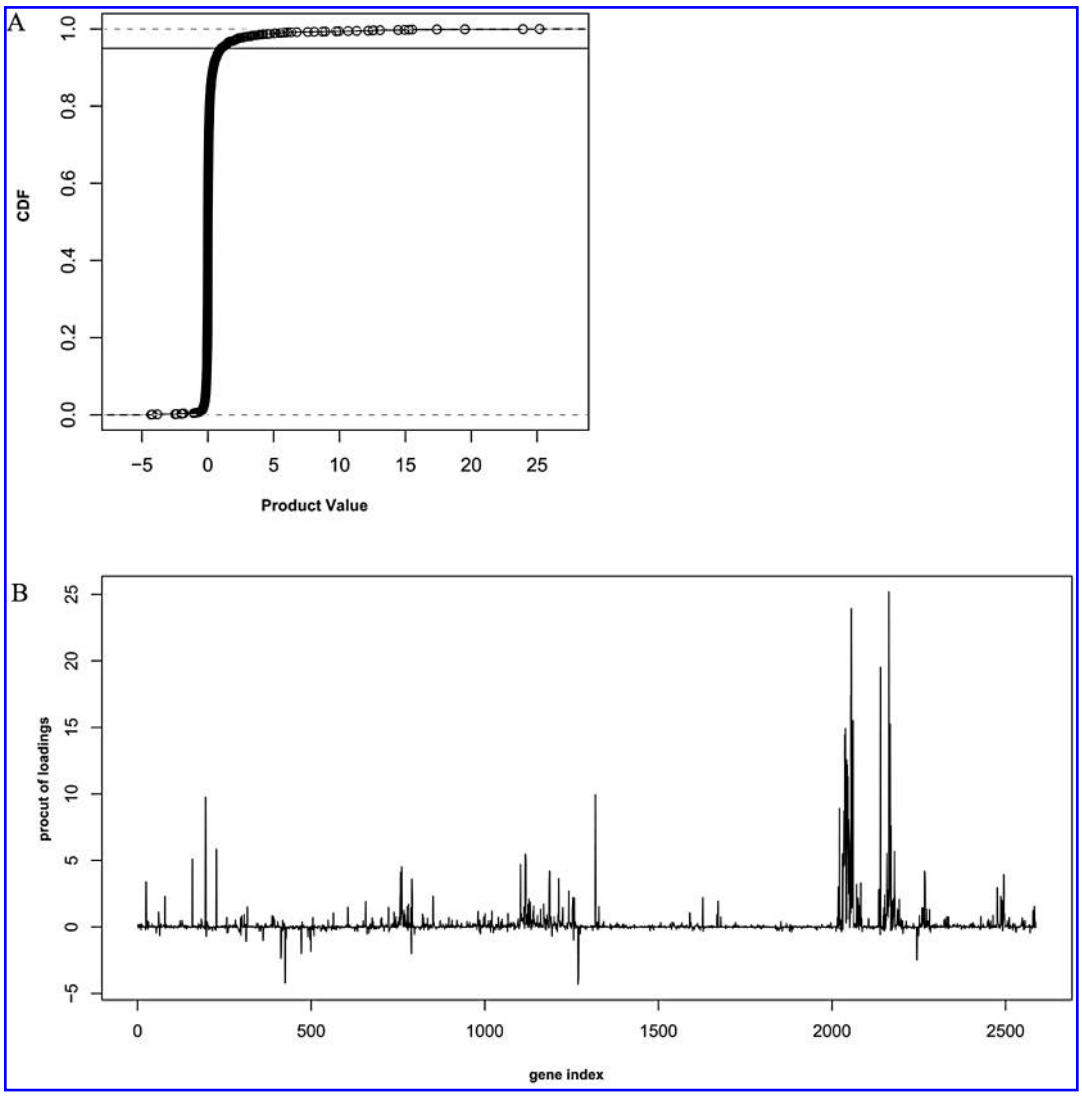

FIG. 3. Distribution of the products of loadings between human and mouse on CGEM H4M3. (**A**) Empirical cumulative density function of the products. The horizontal solid line indicates cut-off for selecting gene set. (**B**) Intensity of the products for each gene.

observed value than do those for mouse ($p < 0.001$). The reason for this difference is not obvious, but technical variation in the GeneChips used for the different species cannot be excluded.

Detecting over-represented functions using gene sets

For each representative gene set of orthologous genes with high loadings on a specific principal component pair, we were able to find a GO term, as well as pathway over-representation that we used to annotate the dominant functions of the CGEM. Because the annotation of orthologous genes in the two species frequently differed, we focused on conserved functions between the two species. Figure 5 uses CGEM H4M3 as an example. The over-represented human and mouse GO terms are listed according to their statistical significance of over-representation; lines connect the same term occurring in the two lists. This analysis shows that many dominant functions are conserved, and in most cases the rankings of degree of dominance do not change dramatically. GO has a hierarchical structure and many of the functions are related.

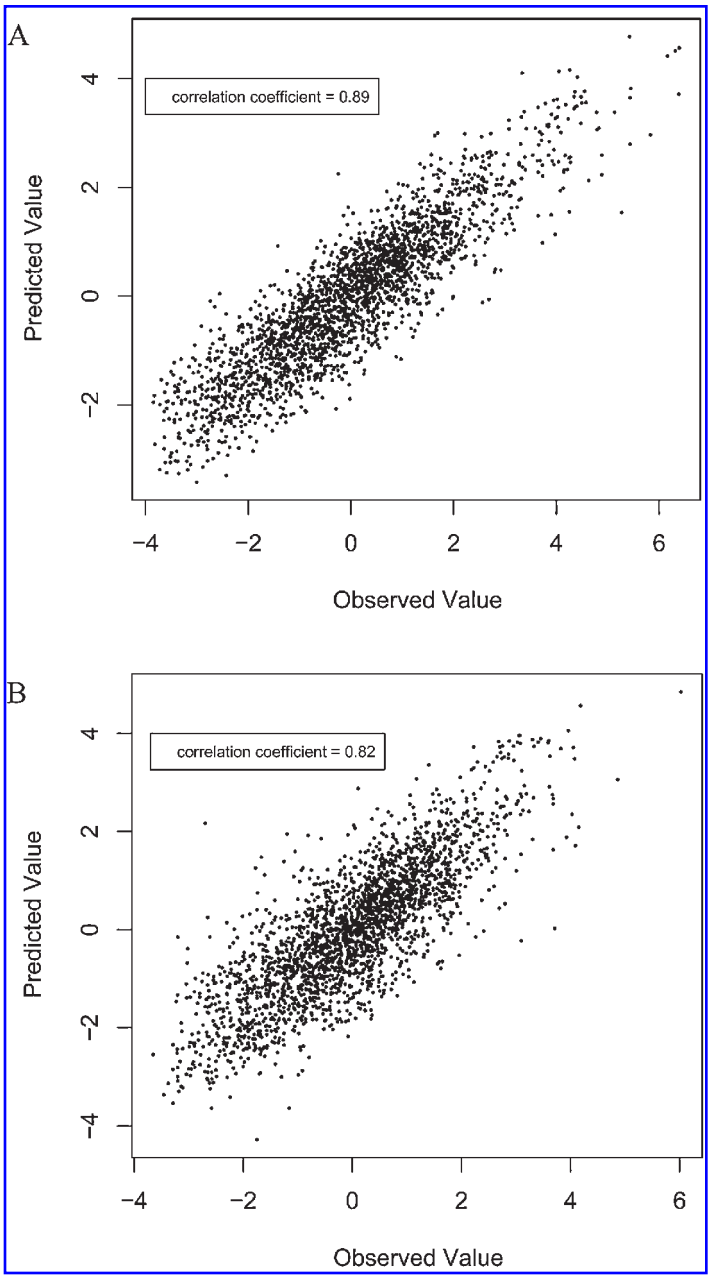

FIG. 4. Comparison between observed and predicted gene expression intensity. The x-axis shows observed gene expression intensity and the y-axis shows predicted gene expression intensity calculated from average component loadings of human and mouse. (**A**) human liver; (**B**) mouse liver.

We have not attempted to collapse parent/child relationships when both are over-represented. In Appendix 1, we summarize the functions conveyed by the GO terms.

We also searched for conserved pathway over-representation (results also listed in App. 1) and found that for most cases there is a good consistency between GO terms and pathway. For example, in CGEM H4M3, over-represented GO terms include blood coagulation, acute-phase response, and related processes. The pathway search identifies complement and coagulation cascades, fatty acid metabolism, and caprolactam degradation, all of which are functionally related to the above GO terms. To further seek validation,

Table 3. Correlation Coefficient of Observed and Predicted Gene Expression Intensities

	Correlation coefficient	
Tissue group	*Human*	*Mouse*
Neural	0.89	0.79
Prostate	0.9	0.81
Bone marrow	0.88	0.8
Lymphocyte	0.89	0.81
Myeloid	0.88	0.85
Liver	0.89	0.82
Lung	0.89	0.78
Placenta	0.88	0.68
Heart	0.89	0.8
Kidney	0.9	0.82
Salivary gland	0.89	0.8
Uterus	0.89	0.81
Testis	0.88	0.69

we explored words that are enriched in the NCBI description. For our example H4M3, words such as apolipoprotein, apom, apoe, and fatty supported our GO term and pathway findings.

Association of transcription factors recognition sites with gene expression response profiles

It is interesting to know whether specific transcription factors are associated with each gene set and how this might relate to functional themes. We explored the transcription factor binding sites around the transcription starting sites of the high-loading genes, and for most gene sets we found that there were transcription factors that were over-represented in both species. A literature survey revealed evidence for relations between those transcription factors and the major function theme of the gene set. Continuing to use CGEM H4M3 as an example, Figure 6 shows the list of transcription factors over-represented in both species for the gene set. Though most transcription factors over-represented in each species have a connection with the dominant functions of the gene set, there are just a handful that are consistently over-represented in both species, which implies subtle differences between human and mouse gene regulation, even for similar pathways and similar functions. We also searched the relationship between our gene sets and GSEA-annotated motif-based gene sets and found enriched representation of GSEA gene sets for each of our CGEM. Both results for transcription factors and GSEA gene sets are listed in Appendix 1.

DISCUSSION

Annotating CGEMs

CGEM H4M3 is associated with blood coagulation, stress response, and other factors, and there is a good agreement between over-represented GO terms and pathways. Transcription factors over-represented in the promoter regions are hepatocyte nuclear factors (HNF4, HNF1) and estrogen receptor (ER). The connections between blood coagulation and ER (Farsetti et al., 1998; Moverare et al., 2004) / HNF-4 (Farsetti et al., 1998) have been reported. HNF-1 was reported to be involved in some stress response (Leu et al., 2001). H4M3 is statistically significantly correlated with GSEA motif sets HNF-1 and estrogen-receptor related receptors (ERR).

CGEM H1M1 is associated with the GO term *homophilic cell adhesion*, but there is no significant over-represented pathway, which is expected because this mode is strongly correlated with average gene expression across all tissues. Also we were not able to identify connections between homophilic cell adhe-

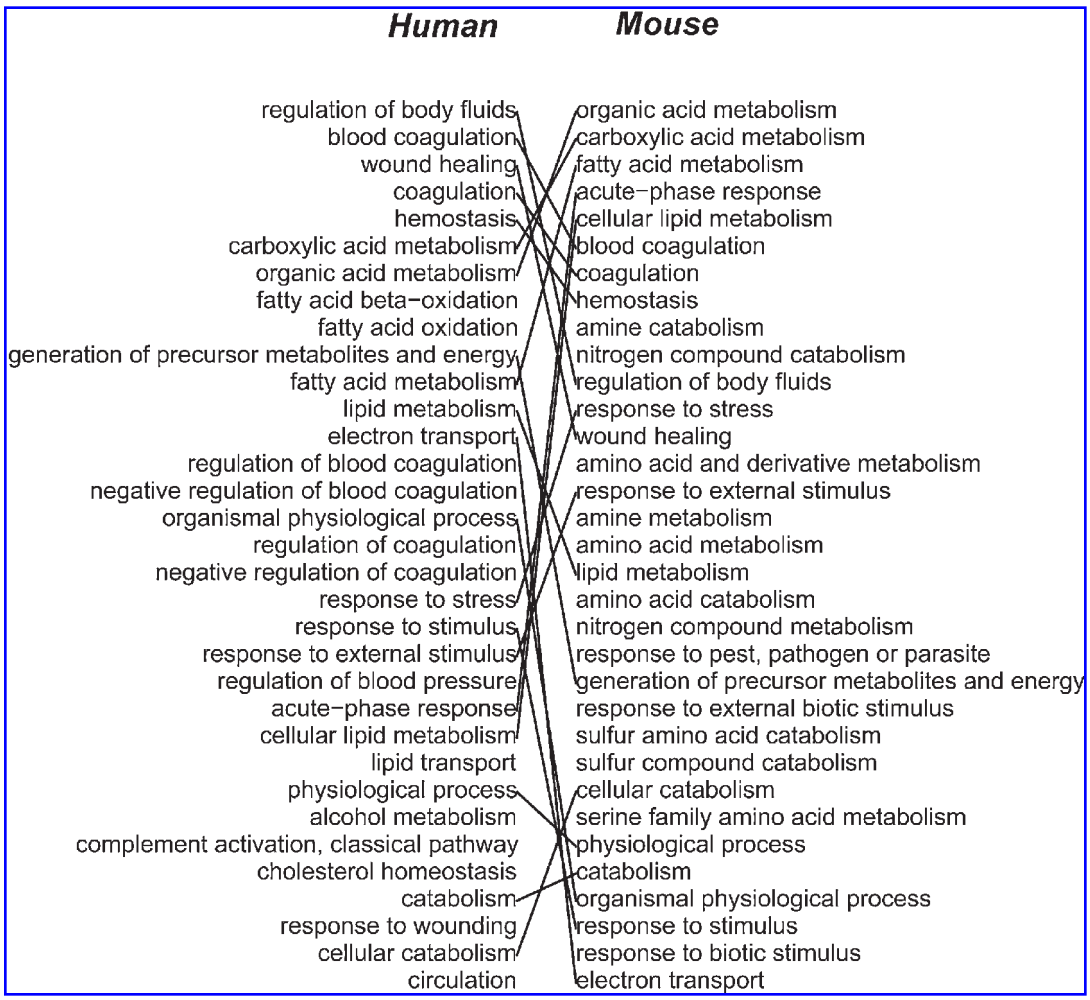

FIG. 5. Over-represented GO terms in CGEM H4M3. The straight lines connect overlapping GO terms, which are sorted in order of dominance. Left column, human; right column, mouse.

sion and the over-expressed transcriptional factor octamer binding factor 1 (OCT1_B) or any correlated GSEA motif sets (AP4, GFI1, NRF1, or SF1).

CGEM H5M5 is associated with energy generation coupled consistently with the citric acid cycle in both the GO terms and pathway over-representation. The over-represented transcription factors include splicing factor 1 (SF1) and CREB. CREB is involved in metabolism (Koo et al., 2005). This mode is statistically significantly correlated with GSEA motif sets such as YY1 transcription factor, nuclear respiratory factor 1 (NRF1), and estrogen-related receptor alpha (ERR1), whose binding sites are all found in some genes involved in regulating mitochondrial energy metabolism (Chinenov et al., 2000; Sladek et al., 1997).

For CGEM H6M6, we did not find common over-represented GO terms between human and mouse; however, extracellular matrix (ECM)-receptor interaction comes up in the pathway analysis. There is literature showing E2F1 and SOX9, which are enriched GSEA motifs for this mode, are involved in ECM regulation (Davies et al., 2002; Tsuboi et al., 2000).

CGEM H13M12 is associated with immune responses. The over-represented transcription factor is MYC, which is known to play a role in immune processes (Hayday et al., 1984). This mode is statistically significantly correlated with GSEA motif set acute myeloid leukemia 1 (AML1), which is involved in development of leukemic dendritic cells (Houtenbos et al., 2005); *v-ets* avian erythoblastosis virus E26 oncogene homolog2 (ETS2), which is involved in acute myelogenous leukemia (Le Beau et al., 1986; Sacchi et al.,

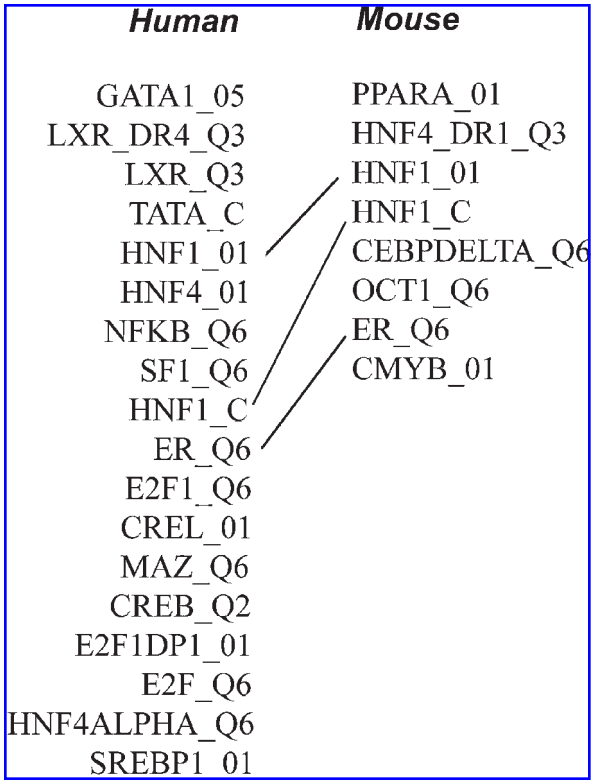

FIG. 6. Over-represented transcription factor binding sites (TFBS). The straight lines connect overlapping TFBSs. TFBSs are sorted in the order of significance of over-representation.

1988); and GA-binding protein, which works in concert with other transcription factors, including PU.1, regulates immune-related genes (Rosmarin et al., 1995, 1998; Shimokawa and Ra, 2005).

CGEM H9M9 is associated with acid and proton transport. The over-represented transcription factor is GATA4, which is related with Na+/Ca2+ (NCX) exchanger (Hudecova et al., 2004). This mode is statistically significantly correlated with paired-like homeodomain transcription factor 2 (PITX2).

CGEM H10M8 is associated with energy generation coupled with pyruvate metabolism and glycolysis. GO term over-representation search also suggests cell adhesion. Although TRANSFAC does not yield clues through enriched transcription factors, GSEA does suggest that motif MEF2 is enriched in this gene set and evidence shows MEF2 involved in energy charge (Holmes and Dohm, 2004).

CGEM H3M4 is associated with the cell cycle. Transcription factor nuclear factor Y (NFY) is over-represented in the promoter regions, which is consistent with the fact that most genes regulated by NFY play a regulatory role in the cell cycle (Zhou et al., 2005). Gene sets regulated by ETS domain transcription factor (ELK1), nuclear respiratory factor (NRF), and specificity protein 1 (Sp1) besides NFY are also shown to be correlated with this mode. There is evidence that Sp1 protein is involved in regulating cell cycle genes (Safe and Abdelrahim, 2005), and NRF is found to be a coregulator of a large number of target genes of E2F that play an important role in cell cycle (Cam et al., 2004).

CGEM H8M10 is associated with cation homeostasis and macromolecule metabolism, such as glycolysis/gluconeogenesis. GSEA analysis reveals that GATA, which is involved in hematologic disease, is over-represented (Cantor, 2005; Crispino, 2005).

CGEM H12M13 is associated with toll-like receptor signaling pathway and the gamma-hexachlorocyclohexane degradation pathway. The over-represented transcription factors from TRANSFAC (CREL, NFKB, HNF1) are also immune-related (Li and Verma, 2002).

From the GO term search, CGEM H7M7 is associated with microtubule polymerization; however, from pathway search, only complement and coagulation cascades are over-represented. Enriched transcription

factors (HNF1, CDX2, CEBPDELTA) revealed that this mode is more stress response related (Alam et al., 1992; German et al., 1994; Gilpin et al., 1996; Leu et al., 2001).

CGEM H11M11 is associated with protein biosynthesis. The over-represented transcription factor is HEB (a helix-loop-helix protein related to E2A and ITF2), and this mode is statistically significantly correlated with SRF. Both are transcription factors that are important in muscle gene regulation (Hu et al., 1992; Wang et al., 2004).

Our functional annotations for the different components sometimes overlap with each other. For example, H5M5 and H10M8 are both associated with oxidative phosphorylation. However, these are coupled to different functions. Oxidative phosphorylation is coupled to citric acid cycle in H5M5 but coupled with pyruvate metabolism and other functions in H10M8. Since biological systems are very complex and functions are sometimes interwoven, this phenomenon is expected. It helps to look at functions from different aspects, e.g., the GO term, as well as pathway, to understand function in a more comprehensive way.

A challenge in annotating the CGEMs is that in some cases they can be multi-factorial responses integrating multiple functions. As a result, in these cases there is not a simple 1-to-1 mapping of CGEM to historically defined biological functions or pathways. This phenomenon is illustrated in Figure 7, which shows the number of genes shared between each CGEM representative set and the COG functional families (Tatusov et al., 2003). A number of associations between gene membership in a CGEM and gene membership in an annotated function are observed. In the lower left corner of Figure 7, we see a nearly 1-to-1 association between CGEM H11M11 (which is associated with protein biosynthesis) and COG functional class J (translation, ribosomal structure and biogenesis) and between CGEM H5M5 (which is associated energy generation) and C (energy production and conversion). In other cases, such as CGEM H6M6, several COGs appear to be associated with the mode but none is dominant. This can be caused by discrepancy between sequence space, which COGs are based on, and expression space, which CGEMs are based on, or it can be caused by the different gene space of CGEM and COG (right now COG only has 860 proteins conserved across all species); it can also be that CGEM H6M6 is a mode with a complex response that involves a combination of the 20 or so simple COG functional categories. Nevertheless, these associations provide a useful aspect in annotating CGEMs (App. 1).

Another note-worthy point to make is that CGEMs delineate multi-tissue coordination in biological responses and thus there is not a simple 1-to-1 mapping of CGEM to tissue. Figure 8 shows the association between CGEMs and tissues. In the upper right-hand corner, a 1-to-1 association between CGEM H13M12, which is associated with immune response, and lymphocyte is observed. However, in most cases a CGEM seems to be positively associated with several tissues. For example, H6M6, which is mainly associated with extracellular matrix–receptor interaction, is positively associated with several tissues but none is dominant. So the majority of the cGEMs involve responses across several tissues, and the representative genes of those responses are not single tissue–specific genes, but rather genes that are highly expressed in a number of tissues and low expressed in other tissues.

As defined by us, CGEMs are connected to fundamental functions in human and mouse species. We base our analysis on a set of normal tissues such as lung, liver, and brain, so much of the variation of gene expression across these samples can be attributed to these tissues coordinating with one another and performing different important biological functions. Our approach for defining CGEMs is a general one that can also be applied to other datasets for different purposes, depending on the specific dataset. For example, it would be of interest to apply our method to a series of treated samples from two species and examine conserved stimulus response patterns. Note that it is possible to apply our approach to compare two distant organisms since we mainly use an objective pairing method based on expression profiles of orthologous genes to pair up tissues; thus, if only orthologous genes between two organisms can be identified, tissue correspondence can be assigned using our pairing method. Principal components can then be calculated and clustered for these two organisms.

Evolutionary implications

The evolution of gene expression is an interesting and controversial topic. One view is that gene expression patterns evolve through a neutral model (Khaitovich et al., 2005, 2004) in which most changes in

gene expression are not related to fitness and changes in gene expression accumulate randomly over time. Other studies suggest that a neutral model accompanied by selective constraint is likely (Jordan et al., 2005). Based on expression similarity among 32 human and mouse tissues (Su et al., 2002), Yanai et al. (2004), found incongruent expression profiles between human and mouse othologous genes and they supported neutral evolution of transcription control. They also showed that expression of human and mouse tissues was clustered into two species-specific clades. By using a measurement called relative mRNA abundance among tissues, however, Liao and Zhang (2005) observed that orthologous tissues between species are more similar than nonorthologous tissues in terms of expression profile, thus supporting conservation of gene expression. In our study, 12 of the 13 modes we found from each species form a binary pair with a sample from the other species. Thus, by using PCA, we are able to find "orthologous modes" between species, which again validates the conservation of gene expression theory. In terms of the relationship between conservation of gene expression and biological function, a previous report (Yanai et al., 2004) mentioned "examples of orthologous profiles where the tissues of conserved expression relate to gene function, whereas divergent expression does not." This statement is consistent with our finding that the conserved expression patterns are linked with specific functions.

For one half of the CGEMs defined here, the magnitudes of the component loadings of the paired components appear in the same order in human and mouse (PCA components are ordered by variance). In one third of the CGEMs, the orders of the mode loadings differ by one (e.g., cell cycle–related genes dominate the third component of human but the fourth component of mouse). In two CGEMs, the orders differ by two. This implies that expression pattern variation within genes with similar functions differs slightly between human and mouse in relative magnitude. This reshuffling of order might be interpreted as slightly different rates of evolution between genes of different functions.

CONCLUSIONS

By integrating data from multiple tissues and two mammalian species, by executing PCA analysis separately on two species, and by matching principal components across species, it is possible to find evolutionary conserved gene expression modes, and these modes can be associated with specific biological functions and pathways. Since no prior knowledge is used, these modes represent an unbiased way of finding gene expression response patterns. In our study, we mostly used a set of normal tissues and found 12 CGEMs related to functions vital to human and mouse, including energy metabolism, immune response, protein biosynthesis, and cell cycle regulation, among others. Representative genes for each mode are also identified. Compared to traditional tissue specificity studies, our analysis is relatively less sensitive to the intrinsic high noise of microarray data and ensures that these CGEMs explain most of the variation in our dataset. Also our result further validates the conservation of many major gene response patterns between human and mouse.

FIG. 7. Relationship of CGEM with COG functional classes. Rows show CGEM modes; columns, COG functional classes. Elements in the matrix are colored according to the number of genes in the CGEM representative set that are assigned to each COG functional class. The pseudocolor scale runs from black to grey to white corresponding to zero to maximal overlap. A, RNA processing and modification; B, chromatin structure and dynamics; C, energy production and conversion; D, cell cycle control, cell division, chromosome partitioning; E, amino acid transport and metabolism; F, nucleotide transport and metabolism; G, carbohydrate transport and metabolism; H, coenzyme transport and metabolism; I, lipid transport and metabolism; J, translation, ribosomal structure and biogenesis; K, transcription; M, cell wall/membrane/envelope biogenesis; O, post-translational modification, protein turnover, chaperones; P, inorganic ion transport and metabolism; Q, secondary metabolites biosynthesis, transport, and catabolism; R, general function prediction only; S, function unknown; T, signal transduction mechanisms; U, intracellular trafficking, secretion, and vesicular transport; V, defense mechanisms; W, extracellular structures; Z, cytoskeleton.

FIG. 8. Relationship of CGEM to tissues. Rows show tissues; columns, CGEM modes. Elements in the matrix are colored according to the correspondence between CGEMs and tissues based on rotation matrix calculated using PCA. The pseudocolor scale runs from black to green to yellow corresponding to low to high correspondence.

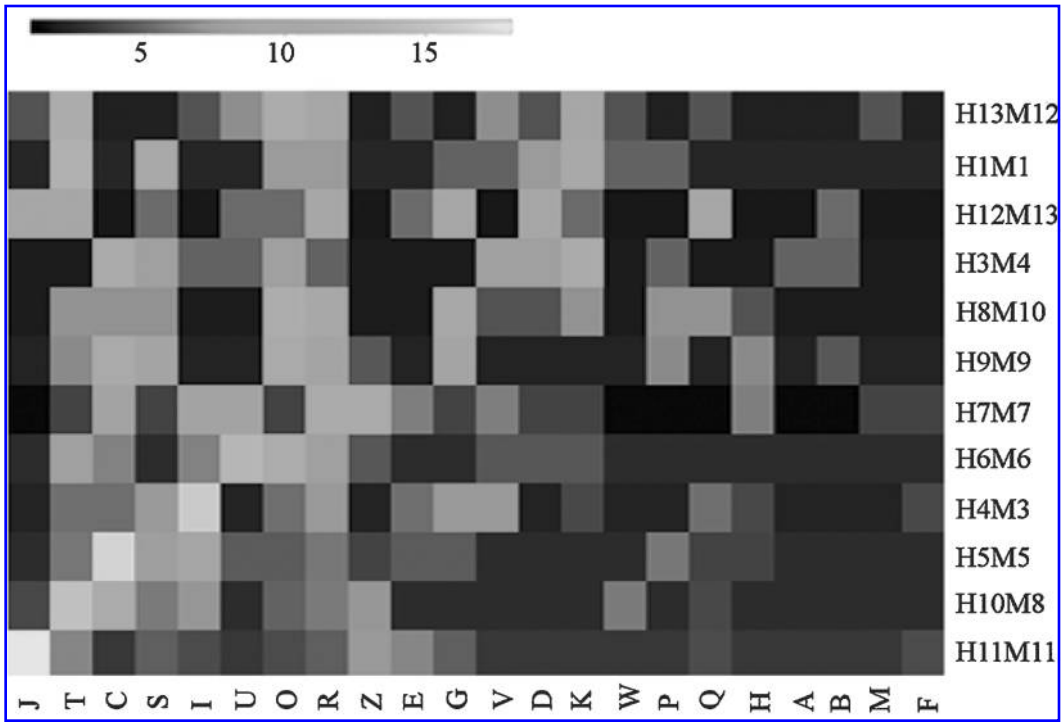
5
10
15
H13M12
H1M1
H12M13
H3M4
H8M10
H9M9
H7M7
H6M6
H4M3
H5M5
H10M8
H11M11
J T C S I U O R Z E G V D K W P Q H A B M F

FIG. 7.

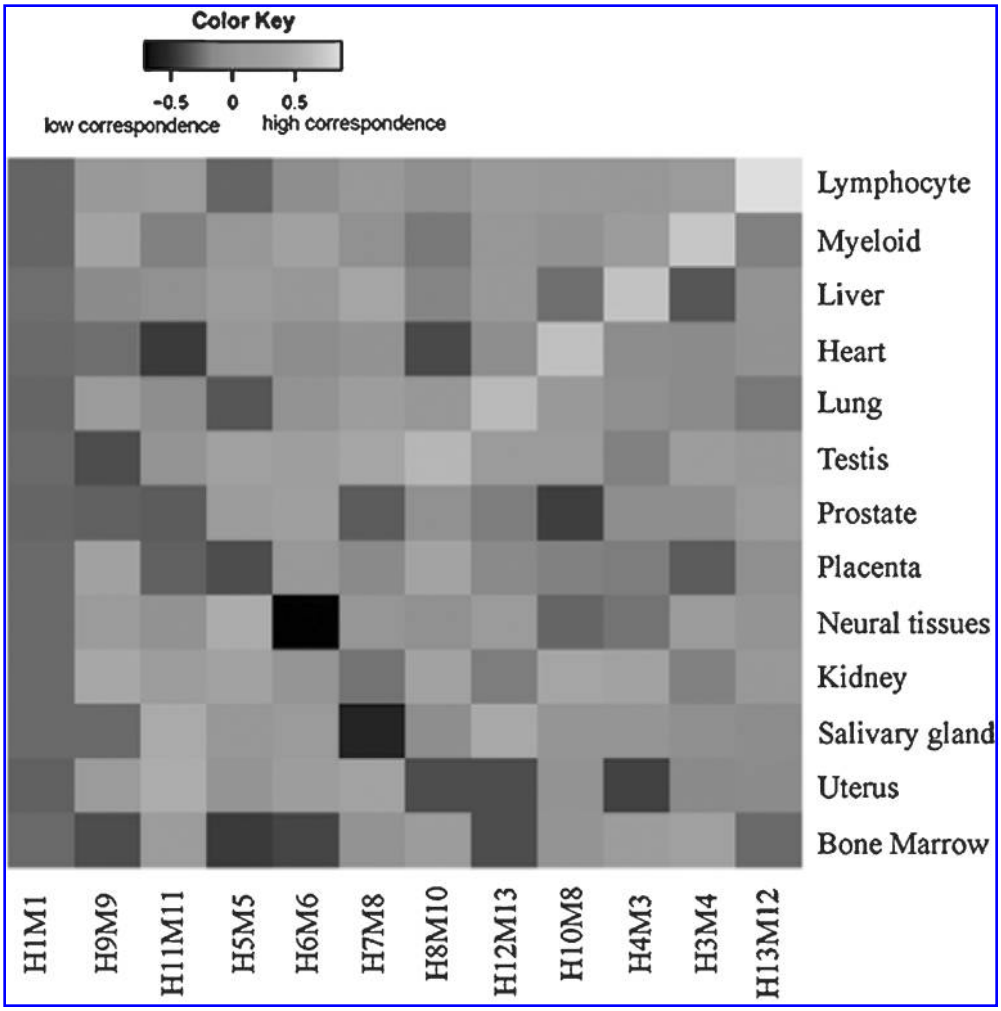
Color Key
-0.5
0
0.5
low correspondence
high correspondence
Lymphocyte
Myeloid
Liver
Heart
Lung
Testis
Prostate
Placenta
Neural tissues
Kidney
Salivary gland
Uterus
Bone Marrow
H1M1
H9M9
H11M11
H5M5
H6M6
H7M8
H8M10
H12M13
H10M8
H4M3
H3M4
H13M12

FIG. 8.

ACKNOWLEDGMENTS

This work was supported in part by grants R01 LM008106, R01 CA85368, U54 DA021519, and P41 RR018627 from the National Institutes of Health. We thank Drs. Lyle Moldawer and Davorka Messmer for providing datasets, and we thank Dr. Jianzhi Zhang (Liao and Zhang, 2005) for discussing the manuscript with us.

REFERENCES

ADJAYE, J., HERWIG, R., HERRMANN, D., WRUCK, W., BENKAHLA, A., BRINK, T. C., et al. (2004). Cross-species hybridisation of human and bovine orthologous genes on high density cDNA microarrays. BMC Genomics **5,** 83.

ALAM, T., AN, M.R., and PAPACONSTANTINOU, J. (1992). Differential expression of three C/EBP isoforms in multiple tissues during the acute phase response. J. Biol. Chem. **267,** 5021–5024.

ALTER, O., BROWN, P.O., and BOTSTEIN, D. (2000). Singular value decomposition for genome-wide expression data processing and modeling. Proc. Natl. Acad. Sci. USA **97,** 10101–10106.

ANDERSSON, M., KARLSSON, L., SVENSSON, P.A., ULFHAMMER, E., EKMAN, M., JERNAS, M., et al. (2005). Differential global gene expression response patterns of human endothelium exposed to shear stress and intraluminal pressure. J. Vasc. Res. **42,** 441–452.

BARRETT, T., SUZEK, T.O., TROUP, D.B., WILHITE, S.E., NGAU, W.C., LEDOUX, P., et al. (2005). NCBI GEO: mining millions of expression profiles—database and tools. Nucleic Acids Res. **33,** D562–566.

BICCIATO, S., LUCHINI, A., and DI BELLO, C. (2003). PCA disjoint models for multiclass cancer analysis using gene expression data. Bioinformatics **19,** 571–578.

BIRNEY, E. (2003). Ensembl: a genome infrastructure. Cold Spring Harbor Symp. Quant. Biol. **68,** 213–215.

BOLSTAD, B.M., IRIZARRY, R.A., ASTRAND, M., and SPEED, T.P. (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics **19,** 185–193.

BOYLE, E.I., WENG, S., GOLLUB, J., JIN, H., BOTSTEIN, D., CHERRY, J.M., et al. (2004). GO::TermFinder—open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. Bioinformatics **20,** 3710–3715.

CAM, H., BALCIUNAITE, E., BLAIS, A., SPEKTOR, A., SCARPULLA, R.C., YOUNG, R., et al. (2004). A common set of gene regulatory networks links metabolism and growth inhibition. Mol. Cell **16,** 399–411.

CANTOR, A.B. (2005). GATA transcription factors in hematologic disease. Int. J. Hematol. **81,** 378–384.

CHINENOV, Y., COOMBS, C., and MARTIN, M.E. (2000). Isolation of a bi-directional promoter directing expression of the mouse GABPalpha and ATP synthase coupling factor 6 genes. Gene **261,** 311–320.

CRESCENZI, M., and GIULIANI, A. (2001). The main biological determinants of tumor line taxonomy elucidated by a principal component analysis of microarray data. FEBS Lett. **507,** 114–118.

CRISPINO, J.D. (2005). GATA1 in normal and malignant hematopoiesis. Semin. Cell Dev. Biol. **16,** 137–147.

DAVIES, S.R., SAKANO, S., ZHU, Y., and SANDELL, L.J. (2002). Distribution of the transcription factors Sox9, AP-2, and [delta]EF1 in adult murine articular and meniscal cartilage and growth plate. J. Histochem. Cytochem. **50,** 1059–1065.

DOOLITTLE, R.F. (2005). Evolutionary aspects of whole-genome biology. Curr. Opin. Struct. Biol. **15,** 248–253.

EDGAR, R., DOMRACHEV, M., and LASH, A.E. (2002). Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res. **30,** 207–210.

ENARD, W., KHAITOVICH, P., KLOSE, J., ZOLLNER, S., HEISSIG, F., GIAVALISCO, P., et al. (2002). Intra- and interspecific variation in primate gene expression patterns. Science **296,** 340–343.

FANG, H., TONG, W., PERKINS, R., SHI, L., HONG, H., CAO, X., et al. (2005). Bioinformatics approaches for cross-species liver cancer analysis based on microarray gene expression profiling. BMC Bioinformatics **6,** suppl 2, S6.

FARSETTI, A., MORETTI, F., NARDUCCI, M., MISITI, S., NANNI, S., ANDREOLI, M., et al. (1998). Orphan receptor hepatocyte nuclear factor-4 antagonizes estrogen receptor alpha-mediated induction of human coagulation factor XII gene. Endocrinology **139,** 4581–4589.

FEEZOR, R.J., OBERHOLZER, C., BAKER, H.V., NOVICK, D., RUBINSTEIN, M., MOLDAWER, L.L., et al. (2003). Molecular characterization of the acute inflammatory response to infections with gram-negative versus gram-positive bacteria. Infect. Immunol. **71,** 5803–5813.

GENTLEMAN, R.C., CAREY, V.J., BATES, D.M., BOLSTAD, B., DETTLING, M., DUDOIT, S., et al. (2004).

Bioconductor: open software development for computational biology and bioinformatics. Genome Biol. **5,** R80.

GERMAN, M.S., WANG, J., FERNALD, A.A., ESPINOSA, R., III, LE BEAU, M.M., and BELL, G.I. (1994). Localization of the genes encoding two transcription factors, LMX1 and CDX3, regulating insulin gene expression to human chromosomes 1 and 13. Genomics **24,** 403–404.

GILPIN, D.A., HSIEH, C.C., KUNINGER, D.T., HERNDON, D.N., and PAPACONSTANTINOU, J. (1996). Effect of thermal injury on the expression of transcription factors that regulate acute phase response genes: the response of C/EBP alpha, C/EBP beta, and C/EBP delta to thermal injury. Surgery **119,** 674–683.

HAMPSON, R., and HUGHES, S.M. (2001). Muscular expressions: profiling genes in complex tissues. Genome Biol. **2,** 1033.

HAYDAY, A.C., GILLIES, S.D., SAITO, H., WOOD, C., WIMAN, K., HAYWARD, W.S., et al. (1984). Activation of a translocated human c-myc gene by an enhancer in the immunoglobulin heavy-chain locus. Nature **307,** 334–340.

HOLMES, B., and DOHM, G.L. (2004). Regulation of GLUT4 gene expression during exercise. Med. Sci. Sports Exer. **36,** 1202–1206.

HOUTENBOS, I., WESTERS, T.M., OSSENKOPPELE, G.J., and VAN DE LOOSDRECHT, A.A. (2005). Employing the immunological synapse in AML: development of leukemic dendritic cells for active specific immunization. Immunobiology **210,** 249–257.

HU, J.S., OLSON, E.N., and KINGSTON, R.E. (1992). HEB, a helix-loop-helix protein related to E2A and ITF2 that can modulate the DNA-binding ability of myogenic regulatory factors. Mol. Cell Biol. **12,** 1031–1042.

HUDECOVA, S., STEFANIK, P., MACEJOVA, D., BRTKO, J., and KRIZANOVA, O. (2004). Retinoic acid increased expression of the Na+/Ca2+ exchanger in the heart and brain. Gen. Physiol. Biophys. **23,** 417–422.

IRIZARRY, R.A., BOLSTAD, B.M., COLLIN, F., COPE, L.M., HOBBS, B., and SPEED, T.P. (2003a). Summaries of Affymetrix GeneChip probe level data. Nucleic Acids Res. **31,** e15.

IRIZARRY, R.A., HOBBS, B., COLLIN, F., BEAZER-BARCLAY, Y.D., ANTONELLIS, K.J., SCHERF, U., et al. (2003b). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics **4,** 249–264.

JIMENEZ, J.L., MITCHELL, M.P., and SGOUROS, J.G. (2003). Microarray analysis of orthologous genes: conservation of the translational machinery across species at the sequence and expression level. Genome Biol. **4,** R4.

JORDAN, I.K., MARINO-RAMIREZ, L., and KOONIN, E.V. (2005). Evolutionary significance of gene expression divergence. Gene **345,** 119–126.

JORDAN, I.K., MARINO-RAMIREZ, L., WOLF, Y.I., and KOONIN, E.V. (2004). Conservation and coevolution in the scale-free human gene coexpression network. Mol. Biol. Evol. **21,** 2058–2070.

KEL, A.E., GOSSLING, E., REUTER, I., CHEREMUSHKIN, E., KEL-MARGOULIS, O.V., and WINGENDER, E. (2003). MATCH: a tool for searching transcription factor binding sites in DNA sequences. Nucleic Acids Res. **31,** 3576–3579.

KHAITOVICH, P., PAABO, S., and WEISS, G. (2005). Toward a neutral evolutionary model of gene expression. Genetics **170,** 929–939.

KHAITOVICH, P., WEISS, G., LACHMANN, M., HELLMANN, I., ENARD, W., MUETZEL, B., et al. (2004). A neutral model of transcriptome evolution. PLoS Biol. **2,** e132.

KOO, S.H., FLECHNER, L., QI, L., ZHANG, X., SCREATON, R.A., JEFFRIES, S., et al. (2005). The CREB coactivator TORC2 is a key regulator of fasting glucose metabolism. Nature **437,** 1109–1111.

LE BEAU, M.M., ROWLEY, J.D., SACCHI, N., WATSON, D.K., PAPAS, T.S., and DIAZ, M.O. (1986). Hu-ets-2 is translocated to chromosome 8 in the t(8;21) in acute myelogenous leukemia. Cancer Genet. Cytogenet. **23,** 269–274.

LEE, A.W., and STATES, D.J. (2000). Both src-dependent and -independent mechanisms mediate phosphatidylinositol 3-kinase regulation of colony-stimulating factor 1-activated mitogen-activated protein kinases in myeloid progenitors. Mol. Cell Biol. **20,** 6779–6798.

LEE, Y., SULTANA, R., PERTEA, G., CHO, J., KARAMYCHEVA, S., TSAI, J., et al. (2002). Cross-referencing eukaryotic genomes: TIGR Orthologous Gene Alignments (TOGA). Genome Res. **12,** 493–502.

LEU, J.I., CRISSEY, M.A., LEU, J.P., CILIBERTO, G., and TAUB, R. (2001). Interleukin-6-induced STAT3 and AP-1 amplify hepatocyte nuclear factor 1-mediated transactivation of hepatic genes, an adaptive response to liver injury. Mol. Cell Biol. **21,** 414–424.

LI, Q., and VERMA, I.M. (2002). NF-kappaB regulation in the immune system. Nat. Rev. Immunol. **2,** 725–734.

LI, W.-H. (1997). *Molecular Evolution*. (Sinauer Associates, Sunderland, MA).

LIAO, B.Y., and ZHANG, J. (2005). Evolutionary conservation of expression profiles between human and mouse orthologous genes. Mol. Biol. Evol.

MATYS, V., FRICKE, E., GEFFERS, R., GOSSLING, E., HAUBROCK, M., HEHL, R., et al. (2003). TRANSFAC: transcriptional regulation, from patterns to profiles. Nucleic Acids Res. **31,** 374–378.

MCCARROLL, S.A., MURPHY, C.T., ZOU, S., PLETCHER, S.D., CHIN, C.S., JAN, Y.N., et al. (2004). Comparing genomic expression patterns across species identifies shared transcriptional profile in aging. Nat. Genet. **36,** 197–204.

MESSMER, D., MESSMER, B., and CHIORAZZI, N. (2003). The global transcriptional maturation program and stimuli-specific gene expression profiles of human myeloid dendritic cells. Int. Immunol. **15,** 491–503.

MISRA, J., SCHMITT, W., HWANG, D., HSIAO, L.L., GULLANS, S., STEPHANOPOULOS, G., et al. (2002). Interactive exploration of microarray gene expression patterns in a reduced dimensional space. Genome Res. **12,** 1112–1120.

MOVERARE, S., SKRTIC, S., LINDBERG, M.K., DAHLMAN-WRIGHT, K., and OHLSSON, C. (2004). Estrogen increases coagulation factor V mRNA levels via both estrogen receptor-alpha and -beta in murine bone marrow/bone. Eur. J. Endocrinol. **151,** 259–263.

O'BRIEN S.J., and FRASER, C.M. (2005). Genomes and evolution: the power of comparative genomics. Curr. Opin. Genet. Dev. **15,** 569–571.

OLEKSIAK, M.F., CHURCHILL, G.A., and CRAWFORD, D.L. (2002). Variation in gene expression within and among natural populations. Nat. Genet. **32,** 261–266.

RAYCHAUDHURI, S., STUART, J.M., and ALTMAN, R.B. (2000). Principal components analysis to summarize microarray experiments: application to sporulation time series. Pac. Symp. Biocomput. 455–466.

ROSMARIN, A.G., CAPRIO, D.G., KIRSCH, D.G., HANDA, H., and SIMKEVICH, C.P. (1995). GABP and PU.1 compete for binding, yet cooperate to increase CD18 (beta 2 leukocyte integrin) transcription. J. Biol. Chem. **270,** 23627–23633.

ROSMARIN, A.G., LUO, M., CAPRIO, D.G., SHANG, J., and SIMKEVICH, C.P. (1998). Sp1 cooperates with the ets transcription factor, GABP, to activate the CD18 (beta2 leukocyte integrin) promoter. J. Biol. Chem. **273,** 13097–13103.

SACCHI, N., CHENG, S.V., TANZI, R.E., GUSELLA, J.F., DRABKIN, H.A., PATTERSON, D., et al. (1988). The ETS genes on chromosome 21 are distal to the breakpoint of the acute myelogenous leukemia translocation (8;21). Genomics **3,** 110–116.

SAFE, S., and ABDELRAHIM, M. (2005). Sp transcription factor family and its role in cancer. Eur. J. Cancer **41,** 2438–2448.

SHIMOKAWA, T., and RA, C. (2005). C/EBPalpha functionally and physically interacts with GABP to activate the human myeloid IgA Fc receptor (Fc alphaR, CD89) gene promoter. Blood **106,** 2534–2542.

SHMUELI, O., HORN-SABAN, S., CHALIFA-CASPI, V., SHMOISH, M., OPHIR, R., BENJAMIN-RODRIG, H., et al. (2003). GeneNote: whole genome expression profiles in normal human tissues. C. R. Biol. **326,** 1067–1072.

SLADEK, R., BADER, J.A., and GIGUERE, V. (1997). The orphan nuclear receptor estrogen-related receptor alpha is a transcriptional regulator of the human medium-chain acyl coenzyme A dehydrogenase gene. Mol. Cell Biol. **17,** 5400–5409.

STUART, J.M., SEGAL, E., KOLLER, D., and KIM, S.K. (2003). A gene-coexpression network for global discovery of conserved genetic modules. Science **302,** 249–255.

SU, A.I., COOKE, M.P., CHING, K.A., HAKAK, Y., WALKER, J.R., WILTSHIRE, T., et al. (2002). Large-scale analysis of the human and mouse transcriptomes. Proc. Natl. Acad. Sci. USA **99,** 4465–4470.

SUBRAMANIAN, A., TAMAYO, P., MOOTHA, V.K., MUKHERJEE, S., EBERT, B.L., GILLETTE, M.A., et al. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. USA **102,** 15545–15550.

TATUSOV, R.L., FEDOROVA, N.D., JACKSON, J.D., JACOBS, A.R., KIRYUTIN, B., KOONIN, E.V., et al. (2003). The COG database: an updated version includes eukaryotes. BMC Bioinformatics **4,** 41.

TSAI, J., SULTANA, R., LEE, Y., PERTEA, G., KARAMYCHEVA, S., ANTONESCU, V., et al. (2001). RESOURCERER: a database for annotating and linking microarray resources within and across species. Genome Biol. **2,** SOFTWARE0002.

TSUBOI, N., YOSHIDA, H., KAWAMURA, T., FURUKAWA, Y., HOSOYA, T., and YAMADA, H. (2000). Three-dimensional matrix suppresses E2F-controlled gene expression in glomerular mesangial cells. Kidney Int. **57,** 1581–1589.

VON MERING, C., ZDOBNOV, E.M., TSOKA, S., CICCARELLI, F.D., PEREIRA-LEAL, J.B., OUZOUNIS, C.A., et al. (2003). Genome evolution reveals biochemical networks and functional modules. Proc. Natl. Acad. Sci. USA **100,** 15428–15433.

WANG, A., and GEHAN, E.A. (2005). Gene selection for microarray data analysis using principal component analysis. Stat. Med. **24,** 2069–2087.

WANG, Z., WANG, D.Z., HOCKEMEYER, D., MCANALLY, J., NORDHEIM, A., and OLSON, E.N. (2004). Myocardin and ternary complex factors compete for SRF to control smooth muscle gene expression. Nature **428,** 185–189.

WINGENDER, E., CHEN, X., HEHL, R., KARAS, H., LIEBICH, I., MATYS, V., et al. (2000). TRANSFAC: an integrated system for gene expression regulation. Nucleic Acids Res. **28,** 316–319.

YANAI, I., GRAUR, D., and OPHIR, R. (2004). Incongruent expression profiles between human and mouse orthologous genes suggest widespread neutral evolution of transcription control. OMICS **8,** 15–24.

YEUNG, K.Y., and RUZZO, W.L. (2001). Principal component analysis for clustering gene expression data. Bioinformatics **17,** 763–774.

ZHOU, D., MASRI, S., YE, J.J., and CHEN, S. (2005). Transcriptional regulation of the mouse PNRC2 promoter by the nuclear factor Y (NFY) and E2F1. Gene **361,** 89–100.

Address reprint requests to:
David States
Bioinformatics Program
The University of Michigan
Palmer Commons 2035
100 Washtenaw Rd.
Ann Arbor, MI 48109

E-mail: dstates@umich.edu

Appendix 1. Detailed Information of CGEMs

A. Representative Genes

Mode	*Human genes*				*Mouse genes*			
	ABCC2	C8G	GC	PCBD1	1700013L23Rik	Azgp1	Fbp2	Nnt
	ACADM	C9	GK	PDCD8	1810014L12Rik	C4	Fga	Orm1
	ACOX1	CDH16	GNMT	PDZK1	2610205H19Rik	C4bp	Fgb	Pcbd1
	ACSL1	CDO1	GRHPR	PGM1	Abcc2	C9	Fgg	Pdcd8
	ADH1C	CHUK	HADH2	PHYH	Acadm	Cdh16	Gc	Pdzk1
	ADH5	COMT	HGD	PLG	Acox1	Cdo1	Gnmt	Pgm2
	AGT	CRYZ	HGFAC	RBP4	Acsl1	Cfi	Grhpr	Phyh
	AHCY	CYP1A2	HRG	RNASE4	Adh1	Chuk	Gyk	Rpb4
	AK2	CYP2E1	HSPA9B	SCP2	Adh5	Cml1	Hadh2	Rnase4
	ALB	DBI	HSPD1	SDHB	Agt	Comt	Hgd	Scp2
	ALDH9A1	DKFZP564B167	HSPE1	SERPINA1	Ahcy	Cryz	Hgfac	Sdhhb
H4M3	ALDOB	DLD	IF	SERPINA6	Ak2	Cyp1a2	Hrg	Serpina1a
	AMBP	ECH1	ITIH4	SERPINC1	Alb1	Cyp2e1	Hspa9a	Serpina6
	APOC4	ECHS1	KLKB1	SERPINF2	Aldh9a1	Dbi	Hspd1	Serpinc1
	APOE	ENPEP	KNG1	SHMT1	Aldob	Dld	Hspe1	Serpinf2
	APOH	F12	LCAT	SLC22A1	Amp	Ech1	Itih3	Shmt1
	APOM	F5	LYPLA1	SLC2A2	Apoc4	Echs1	Itih4	Slc22a1
	ARG1	FABP1	MAC30	SLC37A4	Apoe	Enpep	Klkb1	Slc2a2
	ASGR1	FASN	MST1	TXN	Apoh	F12	Kng1	Slc37a4
	ASL	FBP1	NAT8	UMOD	Apom	F5	Lcat	Txn1
	AZGP1	FGA	NDUFV2	VTN	Arg1	Fabp1	Lypla1	Umod
	C4A	FGB	NNT		Asgr1	Fasn	Mst1	Vtn
	C4BPA	FGG	ORM1		Asl1	Fbp1	Ndufv2	
H1M1	ATRX	DSPG3	MATN3	RB1	2600011C06Rik	Dspg3	Matn3	Prkca
	BDNF	EDNRB	MEOX2	RBM25	4833408C14Rik	Ednrb	Meox2	Ptpro
	BNC1	ELAVL4	MGAT2	REL	Bdnf	Elavl4	Mgat2	Rb1
	BUB1	FGF7	MKRN3	SIM1	Bnc1	Fgf7	Mkrn3	Rel
	CALB1	FOLH1	MMP13	SLC10A2	Bub1	Folh1	Mlf1	Sim1
	CCNE2	FUT9	MTP	SLC16A1	Calb1	Fut9	Mmp13	Slc10a2
	CCR5	GAD2	MYBL1	T	Ccr5	Gad2	Mttp	Slc16a1
	CDH11	HAPLN1	NOX1	TANK	Cdh11	Hapln1	Myb	T
	CDH8	HGF	NPY2R	TPH1	Cdh8	Hgf	Mybl1	Tank
	COPS2	IL2	NR5A2	TYR	Cstf2	Il2	Npy2r	Tph1
	CLCL10	IL7	OR1	ZFX	Cxcl10	Il7	Nr5a2	Tyr
	CYP24A1	KCNJ3	PCDH7	ZW10	Cyp24a1	Kcnj3	Pawr	Zfa
	CYP7A1	KITLG	PHEX		Cyp7a1	Kitl	Pcdh7	Zfx
	DKK1	KLRG1	PLA2G4A		Dkk1	Klrg1	Phex	Zw10
	DSC2	LEPR	PTPRO		Dsc2	Lepr	Pla2g4a	
	ACADM	CYCS	MAPT	PHYH	1110030L0Rik	Cox7a2	Ina	Pgam2
	ACO2	DBI	MGC8721	PIPPIN	1810045K07Rik	Cryz	Kifap3	phyh
	AHCYL1	DKFZP564B167	NCDN	PMM1	2610205H19Rik	Cycs	Klk6	Ptgds
	ANK3	DLD	NCKAP1	PTGDS	2900054D09Rik	D10Bwg0791e	Mapt	Pthr1
	APP	ECH1	NDUFA5	PTHR1	Acadm	Dbi	Mccc1	Scp2
	ATP1A1	ECHS1	NDUFAB1	SCP2	Aco2	Dld	Ncdn	Sdhb
	ATP1A2	EGF	NDUFB8	SDHB	Ahcyl1	Dp1	Nckap1	Sdhc
	ATP5A1	ESD	NDUFC1	SDHC	Ank3	Ech1	Ndufa5	Slc4a4
	ATP5C1	FEZ1	NDUFS1	SLC4A4	App	Echs1	Ndufab1	Spnb2
	ATP5L	GAD1	NDUFS5	SPTBN2	Atp1a1	Egf	Ndufb8	Spnb3
H5M5	ATP6V1A	GJA1	NDUFV2	SUCLA2	Atp1a2	Esd	Ndufc1	Sucla2
	ATP6V1D	GK	NNT	TDE2	Atp5a1	Fez1	Ndufs5	Tde2
	C5orf18	GPM6B	PAFAH1B1	THY1	Atp5c1	Gad1	Ndufv2	Thy1
	CALB1	HSPA8	PARK7	TSPYL4	Atp5l	Gja1	Nnt	Uchl1
	CDH16	HSPA9B	PCCA	UBPH	Atp6v1a1	Gpm6b	Pafah1b1	Umod

Appendix 1. Detailed Information of CGEMs (Cont'd)

A. Representative Genes

Mode	*Human genes*				*Mouse genes*			
	CDH2	HSPD1	PDCD8	UCHL1	Atp6v1d	Gyk	Park7	Vsnl1
	CGI-51	HSPE1	PDHB	UMOD	Calb1	Hspa8	Pdcd8	
	COX5B	IGFBP5	PDZK1	VSNL1	Cdh16	Hspa9a	Pdhb	
	COX7A1	INA	PEG3		Cdh2	Hspd1	Pdzk1	
	COX7A2	KIFAP3	PFKM		Cox5b	Hspe1	Peg3	
	CRYZ	KLK1	PGAM2		Cox7a1	Igfbp5	Pfkm	
	AKAP4	DNAJB9	LUM	SDC4	1110014C03Rik	Cyr61	Lgals3	Rpn2
	ANXA4	EPHB4	LYPLA1	SEC11L1	1810014L12Rik	Dbi	Lgals3bp	S100a10
	ATP6V0E	FN1	MAC30	SEC61B	Akap4	Dlk1	Lum	Sdc4
	C1orf8	FUCA1	MFAP5	SERPINE1	Anxa4	Dnajb9	Lypla1	Sec11I1
	CALU	GADD45G	NAGLU	SLC39A6	Atp11a	Fn1	Mfap5	Sec11I3
	CCT5	GM2A	ODF2	TACSTD1	Atp6v0e	Fuca1	Naglu	Sec61b
	CD151	GNG5	PABPC1	TAF9	Calu	Gadd45g	ORF18	Serpine 1
	CLDN3	GRP58	PRM1	TAGLN	Cct3	Gm2a	Odf2	Slc39a6
H6M6	COL3A1	HMGN1	PROCR	TFPI2	Cct5	Gng5	P4hb	Tacstd1
	COL4A1	HSPA5	PRSS8	TMP21	Cd151	Grp58	Pabpc1	Taf9
	COL4A2	KDELR3	RHOC	TMPRSS2	Cldn3	Hmgn1	Prm2	Tagln
	CSRP1	KRT18	RNASE4	TRAM1	Col3a1	Hspa5	Prm3	Tfpi2
	CTSL2	LAPTM4A	RNP24	VIL2	Col4a1	Kdelr3	Procr	Tmprss2
	CYR61	LDHC	RPN2	WEE1	Col4a2	Krt1-18	Prss8	Tram1
	DBI	LGALS3	S100A10	WWTR1	Csrp1	Laptm4a	Rhoc	Vil2
	DLK1	LGALS3BP	SCC-112		Ctsl	Ldh3	Rnase4	Wee1
H13M12	ADRBK2	CD74	HLA-DOB	PSMB8	6330406L22Rik	Cel	Jak1	Psmb9
	APOE	CD8A	JAK1	PSMB9	Apoe	Cr2	Klk6	Rac2
	ARHGEF1	CEL	KLK1	RAC2	Arhgef1	Csk	Lcp1	Rag2
	BIRC2	CR2	LCP1	RAG2	Birc2	Cst3	Lcp2	Rasgrp1
	C1QB	CSK	LCP2	RASGRP1	C1qb	Ctsl	Lgmn	Rpl26
	C4A	CTSL2	LGMN	RPS16	C4	Cxcl9	Ltb	Rps16
	C7orf32	CXCL9	LOC92482	SELL	Ccl5	Cxcr4	Ly86	Sell
	CCL5	CXCR4	LTB	STAT1	Ccr7	Cyp1b1	Marcks	Sh2d1a
	CCR7	CYP1B1	LY86	TAP1	Cd19	Ddit4	Mlp	Stat1
	CD19	DDIT4	MARCKS	TCF12	Cd2	Dntt	Ms4a1	Tap1
	CD2	DNTT	MARCKSL1	TNFRSF1B	Cd28	Dusp2	Mx2	Tcf12
	CD28	DUSP2	MS4A1	TRAM1	Cd3d	Fcer2a	Mxd4	Tnfrsf1b
	CD3D	FCER2	MX1	UBXD2	Cd3e	Fkbp5	Pdcd4	Tram1
	CD3E	FKBP5	MXD4	UGCG	Cd3z	Gmfg	Pfc	Ubxd2
	CD3Z	GMFG	PFC	VCAM1	Cd48	Gnb2-rs1	Pla2g7	Ugcg
	CD48	GNB2L1	PLA2G7		Cd53	H2-Ob	Pscdbp	Vcam1
	CD53	HA-1	PSCDBP		Cd8a	Ii	Psmb8	
	ABCC2	CSH1	ICAM1	SCGB1A1	2900054D09Rik	Cdh16	Gpx3	Ptprd
	AHCY	CSHL1	IGFBP5	SDC4	A030007L17Rik	Cln2	Gyk	Ptprf
	AHCYL1	DAO	KCNAB2	SEMA3B	Abcc2	Cml1	H2afz	Ptprs
	ALDOB	DDX17	KIBRA	SLC12A3	Ahcy	Cryz	Icam1	Scgb1a1
	ANK3	EGF	KNG1	SLC1A1	Ahcyl1	Dao1	Igfbp5	Sdc4
	AQP4	ENPEP	LGMN	SLC22A6	Alcam	Ddx5	Ii	Sema3b
	ATP1B3	FBP1	LIPA	SLC3A2	Aldob	Egf	Kcnab2	Slc12a3
	ATP6V1A	FTH1	MRLC2	SLC4A4	Ank3	Enpep	Kng1	Slc12a1
H9M9	ATP6V1B1	FTL	MS4A1	SLC7A7	Agp4	Fbp1	Lgmn	Slc22a6
	ATP6V1B2	GAD1	MSN	SMPDL3A	Atp11a	Fbp2	Lip1	Slc3a2
	C7orf2	GH1	NAT8	TKT	Atp1b3	Fth1	Ms4a1	Slc4a4
	CALB1	GK	PDZK1	TPP1	Atp6v1a1	Flt1	Msn	Slc7a7
	CD48	GM2A	PPIA	TXN	Atp6v1b2	Gad1	Mylc2b	Smpdl3a

(*continued*)

Appendix 1. Detailed Information of CGEMs (Cont'd)

A. Representative Genes

Mode	*Human genes*				*Mouse genes*			
	CD53	GPC4	PRKCZ	UCHL1	BC037006	Gh	Pdzk1	Tkt
	CD74	GPM6B	PTHR1	UMOD	Calb1	Gm2a	Ppia	Txn1
	CDH16	GPX3	PTPRD	VIL2	Cd48	Gpc4	Prkcz	Uchl1
	CRYZ	H2AFZ	RNF24	YWHAZ	Cd53	Gpm6b	Pthr1	Umod
H10M8	ACADM	CYR61	ICAM2	PKIG	1110030L07Rik	Cycs	Ier3	Pdlim1
	ALDOB	DKFZP564B167	IER3	PKP2	1500010M16Rik	Cyr61	Ii	Pkig
	ATP6V03	DLD	ITGB1	PPGB	2610205H19Rik	Dld	Itgb1	Pkp2
	C1QR1	DLK1	LDHC	PPIC	Acadm	Dlk1	Ldh3	Pltp
	CA4	ECH1	LGALS3	PTHR1	Aldob	Ech1	Lgals3	Ppic
	CD74	ENG	LOC112714	RBMS1	Atp6v0e	Eng	Lum	Pthr1
	CDH13	ENPEP	LUM	SDHB	C1qr1	Enpep	Lyzs	Rbms1
	CDH16	F2R	LYZ	SLC12A3	Car4	F2r	Mapkapk2	Sdhb
	CDH2	FABP1	MAPKAPK2	SLC7A7	Cdh13	Fabp1	Mef2	Slc12a3
	CGI-51	FABP3	MEF2A	SOD2	Cdh16	Fabp3	Myh2	Slc7a7
	COL4A1	FTH1	MYH6	TM4SF7	Cdh2	Fth1	Myh6	Sod2
	COL4A2	GC20	NDUFA5	TNCC1	Col4a1	Gja1	Myh7	Tm4sf7
	COX6A2	GJA1	NDUFAB1	TUBA2	Col4a2	Gja4	Ndufa5	Tnnc1
	COX7A1	GJA4	NDUFC1	UBPH	Cox6a2	Gng11	Ndufa9	Tuba1
	COX7A2	GK	NDUFS1	UMOD	Cox7a1	Gng5	Ndufab1	Tuba3
	CRIP1	GNG11	NDUFV2	VCAM1	Cox7a2	Gpx3	Ndufc1	Umod
	CSRP3	GNG5	NNT	WWTR1	Crip1	Gyk	Ndufv2	Vcam1
	CYC1	GPX3	PDHB		CSRP3	HBA-A1	NNT	WWTR1
	CYCS	HBA1	PDLIM1		Cyc1	Icam2	Pdhb	
	ACYP1	CDKN3	LDHC	RAN	1810014L12Rik	Cct	Hspca	Prm3
	AF1Q	CETN3	LYPLA1	SCC-112	2610205H19Rik	Cdc2a	Hspcb	Psip1
	AKAP4	CHC1	MAC30	SLC7A5	573049M16Rik	Cdkn3	Impdh2	Psma6
	ASNS	COPS5	MARCKSL1	SMC4L1	AI839562	Cetn3	Ldh3	Psme4
	ATP1B3	COX7A2	NDUFA5	SNRPD2	Acyp1	Chc1	Lypla1	Ran
	BUB1	CSDA	NEK2	SRP14	Akap4	Cops5	Mlp	Slc7a5
H3M4	BUB3	DKFZP564B167	NRD1	SRPK1	Asns	Cox7a2	Ndufa5	Smc4l1
	C18orf10	ERH	ODF2	TAF9	Atp1b3	Csda	Nek2	Snrpd2
	CCNA1	EZH2	PABPC1	TBPL1	Bub1	Dnajc2	Nrd1	Srp14
	CCNBA2	FDFT1	PRM1	TRIM28	Bub3	Erh	Odf2	Srpk1
	CCT4	H2AFZ	PSIP1	UCHL1	Ccna1	Ezh2	Osbpl9	Srpk2
	CCT5	HSPCA	PSMA6	ZRF1	Ccnb2	Fdft1	Pabpc1	Stk23
	CDC2	IMPDH2	PSME4		Cct4	H2afz	Prm2	Taf9
H8M10	ALDOB	EPB42	LMO2	S100A9	1110014C03Rik	Epb4.2	Ltf	Serpine1
	ARFGEF1	EZH2	LTF	SCC-112	Aldob	Ezh2	Mcm5	Slc12a3
	ATP6V0C	FBP1	MMP9	SERPINE1	Arfgef1	Fbp1	Mmp9	Slc22a6
	BPGM	FLT1	MPP1	SLC12A3	Atp6v0c	Fbp2	Mpp1	Slc3a2
	CALB1	GALC	NAT8	SLC22A6	Bpgm	Flt1	Pabpc1	Slc4a4
	CCNB2	GK	PABPC1	SLC3A2	Calb1	Gns	Pdzk1	Slc7a5
	CCR2	GNS	PDZK1	SLC4A4	Ccnb2	Gpr56	Peg3	Slc7a7
	CCR5	GPR56	PEG3	SLC7A5	Cd59a	Gpx3	Ppgb	Tfpi2
	CD59	GPX3	PPGB	SLC7A7	Cdh16	Gyk	Prss11	Tfrc
	CDH16	HBA1	PRSS11	TFDP1	Cdr2	Hba-a1	Prtn3	Umod
	CDR2	HMS	PRTN3	TFPI2	Cldn4	Hmbs	Psen1	Vil2
	CLDN4	HMOX1	PSEN1	TFRC	Cml1	Hsd17b2	Pthr1	Xpo7
	CTSL2	HSD17B2	PTHR1	TMP21	Ctsl	Igfbp1	Rab11fip5	
	CYP17A1	IGFBP1	RAB11FIP5	UMOD	Cyp17a1	Kng1	Rcor1	
	DAO	KNG1	RCOR1	VIL2	Dao1	Krt1-18	Rhced	
	DLK1	KRT18	RHCE	XPO7	Dlk1	Lgmn	S100a8	
	DNAJB1	LGMN	S100A8		Dnajb1	Lmo2	S100a9	

Appendix 1. Detailed Information of CGEMs (Cont'd)

A. Representative Genes

Mode	*Human genes*				*Mouse genes*			
	AGER	CYP2A13	ITGA6	RARS	1110014C03Rik	Cpd	Galnt3	Notch4
	ALDH3A1	CYP4B1	KLK1	RDX	Ager	Csf3	Gp38	Nr1d2
	AMY2A	DOCK9	LAMB3	SAA1	Alcam	Cyp1a1	Gpsm3	Nr4a2
	ANSN	EIF2AK2	LMO7	SCC-112	Aldh3a1	Cyp2a4	Gzmb	ORF31
	ATP11A	ELF3	LOC92482	SCGB1A1	Amy2	Cyp4b1	Icam1	Ocln
	ATP2A1	ELL2	LYZ	XED61B	Asns	D3Ucla1	Icam2	Pabpc1
	CAMK1G	EMP2	MAPK11	SERP1	Atp11a	Dock9	Il6	Pdcd4
H12M13	CASP4	F2R	NAP1L1	SOX10	Atp2a1	Egfl8	Itga6	Pon3
	CD14	F3	NFKBIA	T1A-2	Atp2a2	Elf3	Klk6	Ppt2
	CD9	FGG	NR1D2	TMP21	Atp2a3	Ell2	Lamb3	Prkr
	CDO1	GALNT3	NR4A3	TRAM1	Casp11	Emp2	Lmo7	Ptprm
	CLDN3	GZMB	OCLN	VIPR1	Cd14	F2r	Lyzs	Rars
	CPD	ICAM1	PABPC1		Cd9	F3	Mapk12	Rdx
	CSF3	ICAM2	PON3		Cdo1	Fgg	Nap1I1	Rnf5
	CYP1A1	IL6	PTPRM		Cldn3	Fkbpl	Nfkbia	Saa3
H7M7	ADD1	FGA	NDN	TCEB1	1700093E07Rik	Dpp6	Macf1	Tde2
	ALB	FGB	PDGFRA	TCF8	4930542G03Rik	Fabp1	Ndn	Tuba1
	AMBP	FGG	PEA15	TDE2	5430432P15Rik	Fez1	Pdgfra	Tuba3
	APOE	FLJ13052	PEG3	TNKS	Alb1	Fga	Pea15	Tubb2
	APOH	FN1	PENK	TSPYL4	Ambp	Fgb	Peg3	Tubb5
	APP	FTH1	PLG	TUBA2	Apoe	Fgg	Penk1	Tyro3
	AQP4	FYN	PLS3	TUBA3	Apoh	Fn1	Pls3	Ugcg
	ARG1	GC	PTGDS	TUBB	App	Fth1	Ptgds	Vsnl1
	ATP6V1B2	GLUL	RAB31	TUBB2	Agp4	Fyn	Rab6ip1	Vtn
	C4BPA	HRG	RAB6IP1	TYRO3	Arg1	Gc	Ranbp1	Wfs1
	C9	HSPCA	RANBP1	UGCG	Atp6v1b2	Glul	Rbp4	Zfhx1a
	CDC2L1	KIFAP3	RBP4	VSNL1	BC004012	Hrg	Rohn	
	CRMP1	LCAT	RND2	VTN	C4bp	Hspca	Serpina1a	
	DIA1	LOC112714	SERPINA1	WFS1	C9	Hspcb	Serpinc1	
	DPP6	LUM	SERPINC1		Crmp1	Kifap3	Serpine1	
	FABP1	MACF1	SERPINE1		D10Bwg0791e	Lcat	Serpinf1	
	FEZ1	MMP23B	SERPINF1		Dia1	Lum	Tceb1	
	ACTG1	CYP1B1	RBP1	SFRP1	Actb	Esd	Rbp1	Slc22a6
	ACTR3	EPHB4	RNASE4	SLC12A3	Actr3	Fgl2	Rnase4	Slc7a7
	ADRBK2	ESD	RPL10	ASLC22A6	Aldh1a2	Hnrph1	Rpl10a	Svil
	ALDH1A2	IER3	RPL13	SLC7A7	Amy2	Ier3	Rpl13	Tacstd2
	AMPD3	IF	RPL19	SVIL	Arg1	Igfbp5	Rpl19	Tagln
	AMY2A	IGFBP5	RPL23A	TACSTD2	Atp5a1	Igi	Rpl23a	Tgfb1i1
	ARG1	IGJ	RPL30	TAGLN	Calb1	Impdh2	Rpl30	Thy1
	ASS	IMPDH2	RPL5	TEAD3	Ccl11	LOC436061	Rpl5	Tnc
	CALB1	LTF	RPL6	TGFB1I1	Ccl5	Lamr1	Rpl6	Trip6
H11M11	CCL11	LUM	RPL7	THY1	Cct3	Ltf	Rpl7	Umod
	CCL5	MAF	RPS10	TNC	Cd2	Lum	Rps10	Wdr1
	CCR2	MFAP5	RPS11	TRIP6	Cd3d	Mfap5	Rsp12	Wnt5a
	CCR5	MYH11	RPS12	UMOD	Cdh16	Myh11	Rps19	Wnt5b
	CD2	NAT8	RPS19	WDR1	Cfi	Myh9	Rps3	Wt1
	CD3D	PADI2	RPS3	WNT5A	Cml1	Padi2	Rps3a	Ywhaq
	CDH16	PDZK1	RPS3A	WT1	Cspg2	Pdzk1	Rps7	
	CSPG2	PLAT	RPS7	YWHAQ	Csrp1	Plat	Serpina1a	
	CSRP1	PTHR1	RPSA		Ctsl	Pthr1	Sfrp1	
	CTSL2	PTMA	SERPINA1		Cyp1b1	Ptma	Slc12a3	

Appendix 1. Detailed Information of CGEMs (Cont'd)

B. Functional Annotation and Transcription Factor Binding Information

Mode	*Common over-represented GO terms*	*Common over-represented pathways*	*Common over-represented words*	*Common over-represented TFs*	*Correlated GSEA motifs*	*Over-represented COG functional classes*
H4M3	Blood coagulation, acute-phase response	Complement and coagulation cascades, fatty acid metabolism, glycolysis/ gluconeogenesis, caprolactam degradation	Alcohol, nitrogenous, ikk-alpha, apolipoprotein, apoe, lcat, none, apoe, dehydrogenase, fatty, metabolism, lipoprotein	HNF4, HNF1, ER	ERR1:ERR1 motif, HNF1:HNF1 motif	Lipid transport and metabolism
H1M1	Homophilic cell adhesion		bdnf, phex, cells, c-rel, tyrosinase, leptin, bile, blast, none, unigene	OCT1_B	SF1:Splicing factor 1	Transcription
H5M5	Electron transport, ATP biosynthesis, energy metabolism	Oxidative phosphorylation, citrate cycle (TCA cycle), ATP synthesis, propanoate metabolism	Dpidermal, n-cadherin, beta-amyloid, precursor, amyloid, abeta, connexin, communication	SF1, CREB	CHX10;Ceh-10 homeo domain containing homolog (*C. elegans*), ERR1:ERR1 motif, NRF1:Nuclear respiratory factor 1, PU1:PU1 motif, SF1:Splicing factor 1, TATA:TATA motif YY1:YY1 transcription factor	Energy production and conversion
H6M6		ECM-receptor interaction	Ezrin, plasminogen, activator	PAX6, USF, E2F1	SOX9:SRY (sex determining region Y)-box 9 (campomelic dysplasia, autosomal sex-reversal)	Intracellular trafficking, secretion, vesicular transport
H13M12	Immune response	Toll-like receptor signaling pathway, cytokine-cytokine receptor interaction	Cells, l-selectin, chemokine, apoliprotein, antigen, costimulatory, blast rantes, adhesion, none,	MYC ETS2:V-ets	AML:RUNX1—Runt-related transcription factor 1 (acute myeloid leukemia 1; aml1 oncogene),	

			apoe, antibody, complement, unigene		erythroblastosis virus, E26 oncogene homolog 2 (avian), GABP:GA-binding protein, GFI1:Growth factor independent 1, PU1:PU1 motif, MYC:V-myc myelocytomastosis viral oncogene homolog	
H9M9	Organic acid transport, proton transport	Carbon fixation, pentose phosphate pathway	Epidermal, prkcz, ezrin, hormone, growth	GATA4	PITX2:Paired-like homeodomain transcription factor 2	
H10M8	Cell adhesion, generation of energy	Oxidative phosphorylation, pyruvate metabolism, glycolysis/gluconeogenesis	n-cadherin, blast, adhesion, superoxide, intercellular, integrin, manganese, connexin,		ETS2:V-ets erythroblastosis virus, E26 oncogene homolog 2 (avian), MEF2: myocyte enhancer factor 2	Signal transduction mechanisms, extracellular structures
H3M4	Mitotic cell cycle	Cell cycle	Cyclin, cycle	NFY	ELK1:ELK1, member of ETS oncogene family, NFY:nuclear factor Y, NRF1:Nuclear respiratory factor 1, SP1:Sp1 transcription factor	Transcription
H8M10	Cation homeostasis, cellular macromolecule catabolism,	Glycolysis/gluconeogenesis	Gamma-secretase, matrix, blast, none, ezrin, presenilin, plasminogen, unigene, activator	ZTA	GATA:GATA motif	
H12M13		Gamma-Hexachlorocyclohexane degradation, toll-like receptor signaling pathway	ikappabalpha, thrombin, occludin, inflammatory, rage, g-csf, blast, none, unigene	SREBP1, CREL, NFKB, HNF1	AP4:NA E4F1:E4F transcription factor 1, NFKB:NFKB motif	

(*continued*)

Appendix 1. Detailed Information of CGEMs (Cont'd)

B. Functional Annotation and Transcription Factor Binding Information

Mode	*Common over-represented GO terms*	*Common over-represented pathways*	*Common over-represented words*	*Common over-represented TFs*	*Correlated GSEA motifs*	*Over-represented COG functional classes*
H7M7	Microtubule cytoskeleton organization and biogenesis	Complement and coagulation cascades	Precursor, apolipoprotein, amyloid, lipid, beta-amyloid, lcat, none, apoe, amyloid-beta, vitronectin, plasminogen,abeta, unigene, activator	HNF1, CDX2, CEBPDELTA	HNF3:HNF3 motif, NFY:nuclear facator Y, PAX4:Paired box gene 4	Cytoskeleton
H11M11	Protein biosynthesis	Ribosome	Ribosomal, rantes, eotaxin, plasminogen	HEB	SRF:Serum response factor (c-fos serum response element-binding transcription factor)	Cytoskeleton, energy production and conversion, amino acid transport and metabolism, translation, ribosomal structure and biogenesis

This article has been cited by:

1. Jeremy C Collette, Xiao-Ning Chen, Debra L Mills, Albert M Galaburda, Allan L Reiss, Ursula Bellugi, Julie R Korenberg. 2009. William's syndrome: gene expression is related to parental origin and regional coordinate control. *Journal of Human Genetics* **54**:4, 193-198. [CrossRef]