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. STATES1,3

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-
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, \( \alpha_{gp} \) is projection of gene \( g \) on profile \( p \), and \( \rho_{pt} \) 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 \( \alpha_{gp} \) and \( \rho_{pt} \). In this step, however, instead of using \( \alpha_{gp} \) from the concerned species, we used \((\alpha_{gp}^h + \alpha_{gp}^m)/2 \) where superscript \( h \) and \( m \) indicated human and mouse, respectively. Thus for human we applied the formula

\[ E_{gt}^h (predicted) = \sum_p (\alpha_{gp}^h + \alpha_{gp}^m)/2 \rho_{pt}^h \]

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

\[ E_{gt}^m (predicted) = \sum_p (\alpha_{gp}^h + \alpha_{gp}^m)/2 \rho_{pt}^m. \]

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

<table>
<thead>
<tr>
<th>Tissue group</th>
<th>Sample</th>
<th>Dataset</th>
<th>Sample</th>
<th>Dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human</td>
<td></td>
<td></td>
<td>Mouse</td>
<td></td>
</tr>
<tr>
<td>Neural</td>
<td>CEREBELLUM</td>
<td>N</td>
<td>Frontal cortex</td>
<td>N</td>
</tr>
<tr>
<td></td>
<td>CEREBELLUM</td>
<td>N</td>
<td>Lower spinal cord</td>
<td>N</td>
</tr>
<tr>
<td></td>
<td>CORTEX</td>
<td>N</td>
<td>Lower spinal cord1</td>
<td>N</td>
</tr>
<tr>
<td></td>
<td>AMYGDAL1</td>
<td>N</td>
<td>Upper spinal cord</td>
<td>N</td>
</tr>
<tr>
<td></td>
<td>CAUDATE NUCLEUS</td>
<td>N</td>
<td>Upper spinal cord1</td>
<td>N</td>
</tr>
<tr>
<td></td>
<td>CAUDATE NUCLEUS1</td>
<td>N</td>
<td>Cerebral cortex</td>
<td>N</td>
</tr>
<tr>
<td></td>
<td>FETAL BRAIN1</td>
<td>N</td>
<td>Cerebral cortex1</td>
<td>N</td>
</tr>
<tr>
<td></td>
<td>BRAIN</td>
<td>G</td>
<td>Frontal cortex</td>
<td>N</td>
</tr>
<tr>
<td></td>
<td>BRAIN1</td>
<td>G</td>
<td>Striatium</td>
<td>N</td>
</tr>
<tr>
<td>Prostate</td>
<td>PROSTATE</td>
<td>G</td>
<td>Prostate</td>
<td>N</td>
</tr>
<tr>
<td></td>
<td>PROSTATE1</td>
<td>G</td>
<td>Prostate1</td>
<td>N</td>
</tr>
<tr>
<td>Bone marrow</td>
<td>BONE MARROW</td>
<td>G</td>
<td>Bone marrow</td>
<td>N</td>
</tr>
<tr>
<td></td>
<td>BONE MARROW1</td>
<td>G</td>
<td>Bone marrow1</td>
<td>N</td>
</tr>
<tr>
<td>Lymphocyte</td>
<td>THYMIUS</td>
<td>N</td>
<td>Lymphnode</td>
<td>N</td>
</tr>
<tr>
<td></td>
<td>THYMIUS1</td>
<td>N</td>
<td>Lymphnode1</td>
<td>N</td>
</tr>
<tr>
<td></td>
<td>SPLEEN</td>
<td>N</td>
<td>Spleen</td>
<td>N</td>
</tr>
<tr>
<td></td>
<td>SPLEEN1</td>
<td>N</td>
<td>Spleen1</td>
<td>N</td>
</tr>
<tr>
<td></td>
<td>SPLEEN2</td>
<td>G</td>
<td>Thymus</td>
<td>N</td>
</tr>
<tr>
<td></td>
<td>SPLEEN3</td>
<td>G</td>
<td>Thymus1</td>
<td>N</td>
</tr>
<tr>
<td></td>
<td>THYMIUS1</td>
<td>G</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>RAJI</td>
<td>N</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Myeloid</td>
<td>PROMYELOCYTIC LEUKEMIA</td>
<td>N</td>
<td>32Dcell</td>
<td>D</td>
</tr>
<tr>
<td></td>
<td>PROMYELOCYTIC LEUKEMIA1</td>
<td>N</td>
<td>32Dcell1</td>
<td>D</td>
</tr>
<tr>
<td></td>
<td>PROMYELOCYTIC LEUKEMIA2</td>
<td>N</td>
<td>32Dcell2</td>
<td>D</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Liver</td>
<td>LIVER</td>
<td>G</td>
<td>Liver</td>
<td>N</td>
</tr>
<tr>
<td></td>
<td>LIVER1</td>
<td>G</td>
<td>Liver1</td>
<td>N</td>
</tr>
<tr>
<td></td>
<td>LIVER2</td>
<td>N</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>LIVER3</td>
<td>N</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lung</td>
<td>LUNG</td>
<td>G</td>
<td>Lung</td>
<td>N</td>
</tr>
<tr>
<td></td>
<td>LUNG1</td>
<td>G</td>
<td>Lung1</td>
<td>N</td>
</tr>
<tr>
<td>Placenta</td>
<td>PLACENTA</td>
<td>N</td>
<td>Placenta</td>
<td>N</td>
</tr>
<tr>
<td></td>
<td>PLACENTA1</td>
<td>N</td>
<td>Placenta1</td>
<td>N</td>
</tr>
<tr>
<td>Heart</td>
<td>HEART</td>
<td>N</td>
<td>Heart</td>
<td>N</td>
</tr>
<tr>
<td></td>
<td>HEART2</td>
<td>G</td>
<td>Heart1</td>
<td>N</td>
</tr>
<tr>
<td></td>
<td>HEART3</td>
<td>G</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>HEART4</td>
<td>N</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kidney</td>
<td>KIDNEY</td>
<td>G</td>
<td>Kidney</td>
<td>N</td>
</tr>
<tr>
<td></td>
<td>KIDNEY1</td>
<td>G</td>
<td>Kidney1</td>
<td>N</td>
</tr>
<tr>
<td>Salivary gland</td>
<td>SALIVARY GLAND</td>
<td>N</td>
<td>Salivary gland</td>
<td>N</td>
</tr>
<tr>
<td></td>
<td>SALIVARY GLAND1</td>
<td>N</td>
<td>Salivary gland1</td>
<td>N</td>
</tr>
<tr>
<td>Uterus</td>
<td>UTERUS</td>
<td>N</td>
<td>Uterus1</td>
<td>N</td>
</tr>
<tr>
<td>Testis</td>
<td>TESTIS</td>
<td>N</td>
<td>Testis1</td>
<td>N</td>
</tr>
<tr>
<td></td>
<td>TESTIS1</td>
<td>N</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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 $2 \times 13$ 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>
<thead>
<tr>
<th>Component no.</th>
<th>Eigen value</th>
<th>Variance (%)</th>
<th>Cumulative variance (%)</th>
<th>Eigen value</th>
<th>Variance (%)</th>
<th>Cumulative variance (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human</td>
<td></td>
<td></td>
<td></td>
<td>Mouse</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>33.29</td>
<td>86.99</td>
<td>87</td>
<td>21.61</td>
<td>90.62</td>
<td>90.6</td>
</tr>
<tr>
<td>2</td>
<td>1.02</td>
<td>2.68</td>
<td>90.7</td>
<td>0.65</td>
<td>2.71</td>
<td>93.3</td>
</tr>
<tr>
<td>3</td>
<td>0.88</td>
<td>2.29</td>
<td>92</td>
<td>0.32</td>
<td>1.34</td>
<td>94.7</td>
</tr>
<tr>
<td>4</td>
<td>0.69</td>
<td>1.79</td>
<td>93.7</td>
<td>0.27</td>
<td>1.14</td>
<td>95.8</td>
</tr>
<tr>
<td>5</td>
<td>0.57</td>
<td>1.49</td>
<td>95.2</td>
<td>0.2</td>
<td>0.85</td>
<td>96.7</td>
</tr>
<tr>
<td>6</td>
<td>0.34</td>
<td>0.9</td>
<td>96.1</td>
<td>0.16</td>
<td>0.68</td>
<td>97.3</td>
</tr>
<tr>
<td>7</td>
<td>0.29</td>
<td>0.76</td>
<td>96.9</td>
<td>0.14</td>
<td>0.58</td>
<td>97.9</td>
</tr>
<tr>
<td>8</td>
<td>0.27</td>
<td>0.71</td>
<td>97.6</td>
<td>0.12</td>
<td>0.51</td>
<td>98.4</td>
</tr>
<tr>
<td>9</td>
<td>0.25</td>
<td>0.67</td>
<td>98.3</td>
<td>0.11</td>
<td>0.46</td>
<td>98.9</td>
</tr>
<tr>
<td>10</td>
<td>0.23</td>
<td>0.59</td>
<td>98.9</td>
<td>0.1</td>
<td>0.43</td>
<td>99.3</td>
</tr>
<tr>
<td>11</td>
<td>0.19</td>
<td>0.51</td>
<td>99.4</td>
<td>0.07</td>
<td>0.29</td>
<td>99.6</td>
</tr>
<tr>
<td>12</td>
<td>0.17</td>
<td>0.43</td>
<td>99.8</td>
<td>0.05</td>
<td>0.21</td>
<td>99.8</td>
</tr>
<tr>
<td>13</td>
<td>0.08</td>
<td>0.2</td>
<td>100</td>
<td>0.04</td>
<td>0.18</td>
<td>100</td>
</tr>
</tbody>
</table>
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
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.
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,

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 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-1 (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-

---

**TABLE 3.** Correlation Coefficient of Observed and Predicted Gene Expression Intensities

<table>
<thead>
<tr>
<th>Tissue group</th>
<th>Human</th>
<th>Mouse</th>
</tr>
</thead>
<tbody>
<tr>
<td>Neural</td>
<td>0.89</td>
<td>0.79</td>
</tr>
<tr>
<td>Prostate</td>
<td>0.9</td>
<td>0.81</td>
</tr>
<tr>
<td>Bone marrow</td>
<td>0.88</td>
<td>0.8</td>
</tr>
<tr>
<td>Lymphocyte</td>
<td>0.89</td>
<td>0.81</td>
</tr>
<tr>
<td>Myeloid</td>
<td>0.88</td>
<td>0.85</td>
</tr>
<tr>
<td>Liver</td>
<td>0.89</td>
<td>0.82</td>
</tr>
<tr>
<td>Lung</td>
<td>0.89</td>
<td>0.78</td>
</tr>
<tr>
<td>Placenta</td>
<td>0.88</td>
<td>0.68</td>
</tr>
<tr>
<td>Heart</td>
<td>0.89</td>
<td>0.8</td>
</tr>
<tr>
<td>Kidney</td>
<td>0.9</td>
<td>0.82</td>
</tr>
<tr>
<td>Salivary gland</td>
<td>0.89</td>
<td>0.8</td>
</tr>
<tr>
<td>Uterus</td>
<td>0.89</td>
<td>0.81</td>
</tr>
<tr>
<td>Testis</td>
<td>0.88</td>
<td>0.69</td>
</tr>
</tbody>
</table>
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 erythroblastosis virus E26 oncogene homolog2 (ETS2), which is involved in acute myelogenous leukemia (Le Beau et al., 1986; Sacchi et al.,

---

**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.
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⁺/H⁺/Ca²⁺/H⁺ exchanger (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

![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.](image-url)
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 exam-
ple, 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 func-
tions 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 inte-
grating multiple functions. As a result, in these cases there is not a simple 1-to-1 mapping of CGEM to his-
torically 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 be-
tween CGEM H11M11 (which is associated with protein biosynthesis) and COG functional class J (trans-
lation, ribosomal structure and biogenesis) and between CGEM H5M5 (which is associated energy gener-
ation) 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
caued 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 re-
sponses 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 tis-
sues 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 ex-
pression across these samples can be attributed to these tissues coordinating with one another and per-
forming 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 exam-
ple, it would be of interest to apply our method to a series of treated samples from two species and exam-
ine conserved stimulus response patterns. Note that it is possible to apply our approach to compare two dis-
tant 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 ex-
pression 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.
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


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

<table>
<thead>
<tr>
<th>Mode</th>
<th>Human genes</th>
<th>Mouse genes</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ABCC2</td>
<td>C8G</td>
</tr>
<tr>
<td></td>
<td>ACADM</td>
<td>C9</td>
</tr>
<tr>
<td></td>
<td>ACOX1</td>
<td>CDH16</td>
</tr>
<tr>
<td></td>
<td>ACSL1</td>
<td>CD01</td>
</tr>
<tr>
<td></td>
<td>ADH1C</td>
<td>CHUK</td>
</tr>
<tr>
<td></td>
<td>ADH5</td>
<td>COMT</td>
</tr>
<tr>
<td></td>
<td>AGT</td>
<td>CRYZ</td>
</tr>
<tr>
<td></td>
<td>AHCY</td>
<td>CYP1A2</td>
</tr>
<tr>
<td></td>
<td>AK2</td>
<td>CYP2E1</td>
</tr>
<tr>
<td></td>
<td>ALB</td>
<td>DBI</td>
</tr>
<tr>
<td></td>
<td>ALDH9A1</td>
<td>DGFZP564B167</td>
</tr>
<tr>
<td>H4M3</td>
<td>ALDOB</td>
<td>DLD</td>
</tr>
<tr>
<td></td>
<td>AMBP</td>
<td>ECH1</td>
</tr>
<tr>
<td></td>
<td>APOC4</td>
<td>ECHS1</td>
</tr>
<tr>
<td></td>
<td>APOE</td>
<td>ENPEP</td>
</tr>
<tr>
<td></td>
<td>APOH</td>
<td>F12</td>
</tr>
<tr>
<td></td>
<td>APOM</td>
<td>F5</td>
</tr>
<tr>
<td></td>
<td>ARG1</td>
<td>FABP1</td>
</tr>
<tr>
<td></td>
<td>ASGR1</td>
<td>FASN</td>
</tr>
<tr>
<td></td>
<td>ASL</td>
<td>FBP1</td>
</tr>
<tr>
<td></td>
<td>AZGP1</td>
<td>FGA</td>
</tr>
<tr>
<td></td>
<td>C4A</td>
<td>FGB</td>
</tr>
<tr>
<td></td>
<td>C4BPA</td>
<td>FGG</td>
</tr>
<tr>
<td>H1M1</td>
<td>ATRX</td>
<td>DSPG3</td>
</tr>
<tr>
<td></td>
<td>BDNF</td>
<td>EDNRB</td>
</tr>
<tr>
<td></td>
<td>BNC1</td>
<td>ELAVL4</td>
</tr>
<tr>
<td></td>
<td>BUB1</td>
<td>FGF7</td>
</tr>
<tr>
<td></td>
<td>CALB1</td>
<td>FOLH1</td>
</tr>
<tr>
<td></td>
<td>CCNE2</td>
<td>FUT9</td>
</tr>
<tr>
<td></td>
<td>CCR5</td>
<td>GAD2</td>
</tr>
<tr>
<td></td>
<td>CDH11</td>
<td>HAPLN1</td>
</tr>
<tr>
<td></td>
<td>CDH8</td>
<td>HGF</td>
</tr>
<tr>
<td></td>
<td>COP2S</td>
<td>IL2</td>
</tr>
<tr>
<td></td>
<td>CLCL10</td>
<td>IL7</td>
</tr>
<tr>
<td></td>
<td>CYP23A1</td>
<td>KCN3</td>
</tr>
<tr>
<td></td>
<td>CYP7A1</td>
<td>KLTLG</td>
</tr>
<tr>
<td></td>
<td>DKK1</td>
<td>KLRG1</td>
</tr>
<tr>
<td></td>
<td>DSC2</td>
<td>LEPR</td>
</tr>
<tr>
<td></td>
<td>ACADM</td>
<td>CYCS</td>
</tr>
<tr>
<td></td>
<td>ACO2</td>
<td>DBI</td>
</tr>
<tr>
<td></td>
<td>AHCY1L</td>
<td>DGFZP564B167</td>
</tr>
<tr>
<td></td>
<td>ANK3</td>
<td>DLL</td>
</tr>
<tr>
<td></td>
<td>APP</td>
<td>ECH1</td>
</tr>
<tr>
<td></td>
<td>ATR1A</td>
<td>ECHS1</td>
</tr>
<tr>
<td></td>
<td>ATR1A</td>
<td>EGF</td>
</tr>
<tr>
<td></td>
<td>ATR1A</td>
<td>ESG</td>
</tr>
<tr>
<td></td>
<td>ATR5C1</td>
<td>FEZI</td>
</tr>
<tr>
<td></td>
<td>ATR5L</td>
<td>GAD1</td>
</tr>
<tr>
<td></td>
<td>ATR6V1A</td>
<td>GJA1</td>
</tr>
<tr>
<td></td>
<td>ATR6V1D</td>
<td>GK</td>
</tr>
<tr>
<td></td>
<td>C5orf18</td>
<td>GPM6B</td>
</tr>
<tr>
<td></td>
<td>CALB1</td>
<td>HSPA8</td>
</tr>
<tr>
<td></td>
<td>CDH16</td>
<td>HSPA9B</td>
</tr>
</tbody>
</table>

---

**APPENDIX 1. DETAILED INFORMATION OF CGEMS**

**Mode** Human genes Mouse genes

### Human Genes

- **ABCC2**
- **ACADM**
- **ACOX1**
- **ACS1**
- **ADH1C**
- **ADH5**
- **AGT**
- **AHCY**
- **AK2**
- **ALB**
- **ALDH9A1**

### Mouse Genes

- **ABCC2**
- **ACADM**
- **ACOX1**
- **ACS1**
- **ADH1C**
- **ADH5**
- **AGT**
- **AHCY**
- **AK2**
- **ALB**
- **ALDH9A1**

---

**H4M3**

### Human Genes

- **ALDOB**
- **AMBP**
- **APOC4**
- **APOE**
- **APOH**
- **APOM**
- **ARG1**
- **ASGR1**
- **ASL**
- **AZGP1**
- **C4A**
- **C4BPA**

### Mouse Genes

- **ALDOB**
- **AMBP**
- **APOC4**
- **APOE**
- **APOH**
- **APOM**
- **ARG1**
- **ASGR1**
- **ASL**
- **AZGP1**
- **C4A**
- **C4BPA**

---

**H1M1**

### Human Genes

- **ATRX**
- **BDNF**
- **BNC1**
- **BUB1**
- **CALB1**
- **CCNE2**
- **CCR5**
- **CDH11**
- **CDH8**
- **COP2S**
- **CLCL10**
- **CYP23A1**
- **CYP7A1**
- **DKK1**
- **DSC2**
- **ACADM**
- **ACO2**
- **AHCY1L**
- **ANK3**
- **APP**
- **ATP1A1**
- **ATP1A2**
- **ATP5A1**
- **ATP5C1**
- **ATP5L**
- **ATP6V1A**
- **ATP6V1D**
- **C5orf18**
- **CALB1**
- **CDH16**
### Appendix 1. Detailed Information of CGEMS (Cont’d)

#### A. Representative Genes

<table>
<thead>
<tr>
<th>Mode</th>
<th>Human genes</th>
<th>Mouse genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>CD1</td>
<td>HSPD1</td>
<td>UCHL1</td>
</tr>
<tr>
<td>CDG</td>
<td>HSEI</td>
<td>PDCD8</td>
</tr>
<tr>
<td>COX5B</td>
<td>IGFBP5</td>
<td>CDHC8</td>
</tr>
<tr>
<td>COX7A</td>
<td>ICA</td>
<td>UMOD</td>
</tr>
<tr>
<td>COX7A2</td>
<td>IFAP5</td>
<td>VSNL1</td>
</tr>
<tr>
<td>CRYZ</td>
<td>KLK1</td>
<td>PFG3</td>
</tr>
<tr>
<td>AKA4</td>
<td>DNAJ9B</td>
<td>PFKM</td>
</tr>
<tr>
<td>ANXA4</td>
<td>EPHB4</td>
<td>LUM</td>
</tr>
<tr>
<td>ATP6VOE</td>
<td>FN1</td>
<td>SEC61B</td>
</tr>
<tr>
<td>Clorf8</td>
<td>FCAI</td>
<td>SERPINE1</td>
</tr>
<tr>
<td>CALU</td>
<td>GADD45G</td>
<td>SLC39A6</td>
</tr>
<tr>
<td>CCT5</td>
<td>GM2A</td>
<td>TRACSTD1</td>
</tr>
<tr>
<td>CD151</td>
<td>GNG5</td>
<td>PABPC1</td>
</tr>
<tr>
<td>CLDN3</td>
<td>GRP58</td>
<td>PRM1</td>
</tr>
<tr>
<td>COL3A1</td>
<td>HMGN1</td>
<td>TAPF9</td>
</tr>
<tr>
<td>COL4A1</td>
<td>HSPA5</td>
<td>Calu</td>
</tr>
<tr>
<td>COL4A2</td>
<td>HSPG4</td>
<td>DAPPR8</td>
</tr>
<tr>
<td>C3SRP1</td>
<td>KRT18</td>
<td>RNAS4</td>
</tr>
<tr>
<td>CTSI2</td>
<td>LAPTM4A</td>
<td>RNP2</td>
</tr>
<tr>
<td>CYR61</td>
<td>LDHC</td>
<td>RPN2</td>
</tr>
<tr>
<td>DBI</td>
<td>LGALS3</td>
<td>VIL2</td>
</tr>
<tr>
<td>DLK1</td>
<td>LGALS3BP</td>
<td>WE1</td>
</tr>
<tr>
<td>CD74</td>
<td>HLA-DQ</td>
<td>PSMB8</td>
</tr>
<tr>
<td>APOE</td>
<td>CD8A</td>
<td>JAK1</td>
</tr>
<tr>
<td>ARHGEF1</td>
<td>CEL</td>
<td>KLK1</td>
</tr>
<tr>
<td>BIRC2</td>
<td>CR2</td>
<td>LCP1</td>
</tr>
<tr>
<td>C1OB</td>
<td>CSK</td>
<td>LCP2</td>
</tr>
<tr>
<td>C4A</td>
<td>CTS</td>
<td>LGMN</td>
</tr>
<tr>
<td>C7orf32</td>
<td>CXCL9</td>
<td>LOC28428</td>
</tr>
<tr>
<td>CCL5</td>
<td>CCR4</td>
<td>LTB</td>
</tr>
<tr>
<td>CCR7</td>
<td>CYP1B1</td>
<td>LY86</td>
</tr>
<tr>
<td>CD19</td>
<td>DITD4</td>
<td>MARCKS</td>
</tr>
<tr>
<td>CD2</td>
<td>DNTT</td>
<td>MARCKS1L</td>
</tr>
<tr>
<td>CD28</td>
<td>DUSP2</td>
<td>MS4A1</td>
</tr>
<tr>
<td>CD3D</td>
<td>FCER2</td>
<td>MX1</td>
</tr>
<tr>
<td>CD3E</td>
<td>FKBP5</td>
<td>MFD4</td>
</tr>
<tr>
<td>CD3Z</td>
<td>GMFG</td>
<td>PPC</td>
</tr>
<tr>
<td>CD48</td>
<td>GN2B</td>
<td>PLA2G7</td>
</tr>
<tr>
<td>CD53</td>
<td>HA-1</td>
<td>PSCD8</td>
</tr>
<tr>
<td>ABCC2</td>
<td>CSH1</td>
<td>ICAM1</td>
</tr>
<tr>
<td>AHCHY</td>
<td>CHSL1</td>
<td>IGBP5</td>
</tr>
<tr>
<td>AHCHYI</td>
<td>DAO</td>
<td>IGFBP5</td>
</tr>
<tr>
<td>ALDOB</td>
<td>DDX17</td>
<td>KIBRA</td>
</tr>
<tr>
<td>ANK3</td>
<td>EGF</td>
<td>KNG1</td>
</tr>
<tr>
<td>AQP4</td>
<td>ENPEP</td>
<td>LGMN</td>
</tr>
<tr>
<td>ATP1B3</td>
<td>FPB1</td>
<td>LIPA</td>
</tr>
<tr>
<td>ATP6V1A</td>
<td>FTHI</td>
<td>LPLA</td>
</tr>
<tr>
<td>ATP6V1B</td>
<td>FT1</td>
<td>LSCA2A</td>
</tr>
<tr>
<td>ATP6V1B2</td>
<td>GA21</td>
<td>SLCA2</td>
</tr>
<tr>
<td>C7orf2</td>
<td>GH1</td>
<td>NAT8</td>
</tr>
<tr>
<td>CALB1</td>
<td>GKI</td>
<td>PDZK1</td>
</tr>
<tr>
<td>CD48</td>
<td>GM2A</td>
<td>PPIA</td>
</tr>
</tbody>
</table>

(continued)
## Appendix 1. Detailed Information of CGEMS (Cont’d)

### Mode

<table>
<thead>
<tr>
<th>Human genes</th>
<th>Mouse genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>CD53</td>
<td>GPC4</td>
</tr>
<tr>
<td>CD74</td>
<td>GPM6B</td>
</tr>
<tr>
<td>CDH16</td>
<td>GPX3</td>
</tr>
<tr>
<td>CRYZ</td>
<td>H2AFZ</td>
</tr>
<tr>
<td>H10M8 ACADM</td>
<td>CRYR1</td>
</tr>
<tr>
<td>ALDOB DIFZP564B167</td>
<td>DLD</td>
</tr>
<tr>
<td>ATP6V03</td>
<td>CYR61</td>
</tr>
<tr>
<td>CA4</td>
<td>ECH1</td>
</tr>
<tr>
<td>CD74</td>
<td>ENG</td>
</tr>
<tr>
<td>CDH13</td>
<td>ENPEP</td>
</tr>
<tr>
<td>H10M8</td>
<td>CRYZ</td>
</tr>
<tr>
<td>H3M4</td>
<td>H2AFZ</td>
</tr>
<tr>
<td>H8M10</td>
<td>H2AFZ</td>
</tr>
<tr>
<td>H8M10 ALDOB</td>
<td>EPB42</td>
</tr>
<tr>
<td>H8M10 ARFGEF1</td>
<td>EZH2</td>
</tr>
<tr>
<td>H8M10 ATP6V0C</td>
<td>FBPI</td>
</tr>
<tr>
<td>H8M10 BPGM</td>
<td>FLT1</td>
</tr>
<tr>
<td>H8M10 CALB1</td>
<td>GALC</td>
</tr>
<tr>
<td>H8M10 CCNB2</td>
<td>GK</td>
</tr>
<tr>
<td>H8M10 CCNB2</td>
<td>GK</td>
</tr>
<tr>
<td>H8M10 CCR2</td>
<td>GNS</td>
</tr>
<tr>
<td>H8M10 CCR5</td>
<td>GPR56</td>
</tr>
<tr>
<td>H8M10 CD59</td>
<td>GPX3</td>
</tr>
<tr>
<td>H8M10 CDH16</td>
<td>HBA1</td>
</tr>
<tr>
<td>H8M10 CDPN</td>
<td>HMOX1</td>
</tr>
<tr>
<td>H8M10 CTS2L</td>
<td>HSD17B2</td>
</tr>
<tr>
<td>H8M10 CYP17A1</td>
<td>IGBF1</td>
</tr>
<tr>
<td>H8M10 DAO</td>
<td>KNG1</td>
</tr>
<tr>
<td>H8M10 DLD1</td>
<td>KRT18</td>
</tr>
<tr>
<td>H8M10 DNAJB1</td>
<td>LGMN</td>
</tr>
</tbody>
</table>
## Appendix 1. Detailed Information of CGEMs (Cont’d)

### A. Representative Genes

<table>
<thead>
<tr>
<th>Mode</th>
<th>Human genes</th>
<th>Mouse genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>AGER</td>
<td>CYP2A13</td>
<td>ITGA6</td>
</tr>
<tr>
<td>ALDH3A1</td>
<td>CYP4B1</td>
<td>KLK1</td>
</tr>
<tr>
<td>AMY2A</td>
<td>DOCK9</td>
<td>LAMB3</td>
</tr>
<tr>
<td>ANSN</td>
<td>EIF2AK2</td>
<td>LM07</td>
</tr>
<tr>
<td>ATP11A</td>
<td>ELF3</td>
<td>LOC29482</td>
</tr>
<tr>
<td>ATP2A1</td>
<td>ELL2</td>
<td>LYZ</td>
</tr>
<tr>
<td>CAMK4G</td>
<td>EMP2</td>
<td>MAPK11</td>
</tr>
<tr>
<td>H12M13</td>
<td>CASP4</td>
<td>F2R</td>
</tr>
<tr>
<td>H7M7</td>
<td>ADD1</td>
<td>FGA</td>
</tr>
<tr>
<td>ALB</td>
<td>FGB</td>
<td>PGDFRA</td>
</tr>
<tr>
<td>AMBP</td>
<td>FGG</td>
<td>PEA15</td>
</tr>
<tr>
<td>APOE</td>
<td>FLJ13052</td>
<td>PEG3</td>
</tr>
<tr>
<td>APOH</td>
<td>FNI</td>
<td>PENK</td>
</tr>
<tr>
<td>APP</td>
<td>FTH1</td>
<td>PLG</td>
</tr>
<tr>
<td>AQP4</td>
<td>FYN</td>
<td>PL53</td>
</tr>
<tr>
<td>ARG1</td>
<td>GC</td>
<td>PTGDS</td>
</tr>
<tr>
<td>ATP6V1B2</td>
<td>GLUL</td>
<td>RAB31</td>
</tr>
<tr>
<td>C4BPA</td>
<td>HRG</td>
<td>RAB6IP1</td>
</tr>
<tr>
<td>C9</td>
<td>HSPCA</td>
<td>RANBP1</td>
</tr>
<tr>
<td>CDC2L1</td>
<td>KIFAP3</td>
<td>RBP4</td>
</tr>
<tr>
<td>CREM1</td>
<td>LCAT</td>
<td>RND2</td>
</tr>
<tr>
<td>DIA1</td>
<td>LOC112714</td>
<td>SERPIN1A</td>
</tr>
<tr>
<td>DPP6</td>
<td>LUM</td>
<td>SERPIN1</td>
</tr>
<tr>
<td>FABP1</td>
<td>MACF1</td>
<td>SERPIN1</td>
</tr>
<tr>
<td>FEZ1</td>
<td>MMP23B</td>
<td>SERPIN1</td>
</tr>
<tr>
<td>ACTG1</td>
<td>CYP1B1</td>
<td>RBP1</td>
</tr>
<tr>
<td>ACT3</td>
<td>EPHB4</td>
<td>RNASE4</td>
</tr>
<tr>
<td>ADRBK2</td>
<td>ESF</td>
<td>RPL10</td>
</tr>
<tr>
<td>ALDH1A2</td>
<td>EIR3</td>
<td>RPL13</td>
</tr>
<tr>
<td>AMID3</td>
<td>IF</td>
<td>RPL19</td>
</tr>
<tr>
<td>AMY2A</td>
<td>IGFBP5</td>
<td>RPL23A</td>
</tr>
<tr>
<td>ARG1</td>
<td>IGI</td>
<td>RPL30</td>
</tr>
<tr>
<td>ASS</td>
<td>IMPDH2</td>
<td>RPL5</td>
</tr>
<tr>
<td>CALB1</td>
<td>LTF</td>
<td>RPL6</td>
</tr>
<tr>
<td>H11M11</td>
<td>CCL11</td>
<td>LUM</td>
</tr>
<tr>
<td>CCL5</td>
<td>MAF</td>
<td>RPS10</td>
</tr>
<tr>
<td>CCR2</td>
<td>MFAP5</td>
<td>RPS11</td>
</tr>
<tr>
<td>CCR5</td>
<td>MYY11</td>
<td>RPS12</td>
</tr>
<tr>
<td>CD2</td>
<td>NAT8</td>
<td>RPS19</td>
</tr>
<tr>
<td>CD3D</td>
<td>PADD2</td>
<td>RPS3</td>
</tr>
<tr>
<td>CDH16</td>
<td>PDZK1</td>
<td>RPS3A</td>
</tr>
<tr>
<td>CSG2</td>
<td>PLAT</td>
<td>RPS7</td>
</tr>
<tr>
<td>CSRP1</td>
<td>PTHR1</td>
<td>RPSA</td>
</tr>
<tr>
<td>CTSL2</td>
<td>PTMA</td>
<td>SERPIN1</td>
</tr>
</tbody>
</table>
### APPENDIX 1. Detailed Information of CGEMS (Cont’d)

#### B. Functional Annotation and Transcription Factor Binding Information

<table>
<thead>
<tr>
<th>Mode</th>
<th>Common over-represented GO terms</th>
<th>Common over-represented pathways</th>
<th>Common over-represented words</th>
<th>Common over-represented TFs</th>
<th>Correlated GSEA motifs</th>
<th>Over-represented COG functional classes</th>
</tr>
</thead>
<tbody>
<tr>
<td>H4M3</td>
<td>Blood coagulation, acute-phase response</td>
<td>Complement and coagulation cascades, fatty acid metabolism, glycolysis/gluconeogenesis, caprolactam degradation</td>
<td>Alcohol, nitrogenous, ikk-alpha, apolipoprotein, apoL, lact, none, apoE, dehydrogenase, fatty, metabolism, lipoprotein</td>
<td>HNF4, HNF1, ER</td>
<td>ERR1:ERR1 motif, HNF1:HNFn1 motif</td>
<td>Lipid transport and metabolism</td>
</tr>
<tr>
<td>H1M1</td>
<td>Homophilic cell adhesion</td>
<td>bdnf, phex, cells, c-rel, tyrosinase, leptin, bile, blast, none, unigene</td>
<td>OCT1_B</td>
<td>SF1:Splicing factor 1</td>
<td></td>
<td>Transcription</td>
</tr>
<tr>
<td>H5M5</td>
<td>Electron transport, ATP biosynthesis, energy metabolism</td>
<td>Oxidative phosphorylation, citrate cycle (TCA cycle), ATP synthesis, propanoate metabolism</td>
<td>Epidermal, n-cadherin, beta-amyloid, precursor, amloid, abeta, connexin, communication</td>
<td>SF1, CREB</td>
<td>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</td>
<td>Energy production and conversion</td>
</tr>
<tr>
<td>H6M6</td>
<td>ECM-receptor interaction</td>
<td>Ezrin, plasminogen, activator</td>
<td>PAX6, USF, E2F1</td>
<td>SOX9:SRY (sex determining region Y)-box 9 (campomelic dysplasia, autosomal sex-reversal)</td>
<td></td>
<td>Intracellular trafficking, secretion, vesicular transport</td>
</tr>
<tr>
<td>H13M12</td>
<td>Immune response</td>
<td>Toll-like receptor signaling pathway, cytokine-cytokine receptor interaction</td>
<td>Cells, l-selectin, chemokine, apolipoprotein, antigen, costimulatory, blast rantes, adhesion, none</td>
<td>MYC</td>
<td>AML:RUNX1—Runt-related transcription factor 1 (acute myeloid leukemia 1; aml1 oncogene)</td>
<td></td>
</tr>
<tr>
<td>H9M9</td>
<td>Organic acid transport, proton transport</td>
<td>Carbon fixation, pentose phosphate pathway</td>
<td>Epidermal, prkc z, ezrin, hormone, growth</td>
<td>GATA4</td>
<td>PITX2:Paired-like homeodomain transcription factor 2</td>
<td></td>
</tr>
<tr>
<td>-------</td>
<td>---------------------------------------</td>
<td>---------------------------------------------</td>
<td>-------------------------------------------</td>
<td>-------</td>
<td>--------------------------------------------------</td>
<td></td>
</tr>
<tr>
<td>H10M8</td>
<td>Cell adhesion, generation of energy</td>
<td>Oxidative phosphorylation, pyruvate metabolism, glycolysis/gluconeogenesis</td>
<td>n-cadherin, blast, adhesion, superoxide, intercellular, integrin, manganese, connexin,</td>
<td></td>
<td>ETS2:V-ets erythroblastosis virus, E26 oncogene homolog 2 (avian), MEF2: myocyte enhancer factor 2</td>
<td></td>
</tr>
<tr>
<td>H3M4</td>
<td>Mitotic cell cycle</td>
<td>Cell cycle</td>
<td>Cyclin, cycle</td>
<td>NFY</td>
<td>ELK1:ELK1, member of ETS oncogene family, NFY:nuclear factor Y, NRF1:Nuclear respiratory factor 1, SP1:Sp1 transcription factor</td>
<td></td>
</tr>
<tr>
<td>H8M10</td>
<td>Cation homeostasis, cellular macromolecule catabolism, Glycolysis/gluconeogenesis</td>
<td>Gamma-secretase, matrix, blast, none, ezrin, presenilin, plasminogen, unigene, activator</td>
<td></td>
<td>ZTA</td>
<td>GATA:GATA motif</td>
<td></td>
</tr>
<tr>
<td>H12M13</td>
<td>Gamma-Hexachlorocyclohexane degradation, toll-like receptor signaling pathway</td>
<td>ikappabalpha, thrombin, occludin, inflammatory, rage, g-csf, blast, none, unigene</td>
<td>SREBP1, CREL, NFKB, HNF1</td>
<td>AP4:NA</td>
<td>E4F1:E4F transcription factor 1, NFKB:NFKB motif</td>
<td></td>
</tr>
</tbody>
</table>

(continued)
## Appendix 1. Detailed Information of CGEMs (Cont’d)

### B. Functional Annotation and Transcription Factor Binding Information

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