A Gene-Coexpression Network for Global Discovery of Conserved Genetic Modules

Joshua M. Stuart,[1*†] Eran Segal,[2*] Daphne Koller,[2‡] Stuart K. Kim[3‡]

To elucidate gene function on a global scale, we identified pairs of genes that are coexpressed over 3182 DNA microarrays from humans, flies, worms, and yeast. We found 22,163 such coexpression relationships, each of which has been conserved across evolution. This conservation implies that the coexpression of these gene pairs confers a selective advantage and therefore that these genes are functionally related. Many of these relationships provide strong evidence for the involvement of new genes in core biological functions such as the cell cycle, secretion, and protein expression. We experimentally confirmed the predictions implied by some of these links and identified cell proliferation functions for several genes. By assembling these links into a gene-coexpression network, we found several components that were animal-specific as well as interrelationships between newly evolved and ancient modules.

The genome sequences of humans and several model organisms have established a nearly complete list of the genes required to enact cellular, developmental, and behavioral processes in these organisms (*1–4*). The next major challenges are to elucidate the functions of the large fraction of genes in the genome whose functions are currently unknown and to discover how the genes interact to perform specific biological processes. DNA microarrays provide us with a first step toward the goal of uncovering gene function on a global scale. Because genes that encode proteins that participate in the same pathway or are part of the same protein complex are often coregulated, clusters of genes with related functions often exhibit expression patterns that are correlated under a large number of diverse conditions in DNA microarray experiments (*5–8*).

However, coregulation does not necessarily imply that genes are functionally related. For example, cis-regulatory DNA motifs are predicted to occur by chance in the genome and might lead to serendipitous transcriptional regulation of nearby genes. In experiments limited to a single species, it would be difficult or even impossible to distinguish accidentally regulated genes from those that are physiologically important. However, evolutionary conservation is a powerful criterion to identify genes that are functionally important from a set of coregulated genes. Coregulation of a pair of genes over large evolutionary distances implies that the coregulation confers a selective advantage, most likely because the genes are functionally related. Because small and subtle changes in fitness can confer selective advantage during evolution, the test for related gene function using evolutionary conservation in the wild is more sensitive than scoring the phenotype resulting from strong loss-of-function mutants in the laboratory.

The recent availability of large sets of DNA microarray data for humans, flies, worms, and yeast makes it possible to measure evolutionarily conserved coexpression on a genomewide scale (*9–11*). We developed a computational method to analyze 3182 DNA microarrays from humans, flies, worms, and yeast (most of which were previously published) to identify gene interactions that are evolutionarily conserved.

Construction of a gene-coexpression network. We selected evolutionarily diverse organisms for which extensive microarray data were available: *Homo sapiens*, *Drosophila melanogaster*, *Caenorhabditis elegans*, and *Saccharomyces cerevisiae*. To identify genes that are coexpressed across multiple organisms, we first associated genes from one organism with their orthologous counterparts in other organisms. We used an approach similar to previous approaches for identifying orthologous sets of genes (*12*, *13*). Orthologs were identified by performing an all-against-all BLAST between every pair of protein sequences from each of the organisms (*14*). We then defined a metagene as a set of genes across multiple organisms whose protein sequences are one another's best reciprocal BLAST hit (*14*). Using this method, we assigned each gene to at most a single metagene. For example, metagene MEG273 refers to the human gene *Psmd4*, the *C. elegans* gene *rpn-10*, the *D. melanogaster* gene *Pros54*, and the *S. cerevisiae* gene *Rpn10*, all

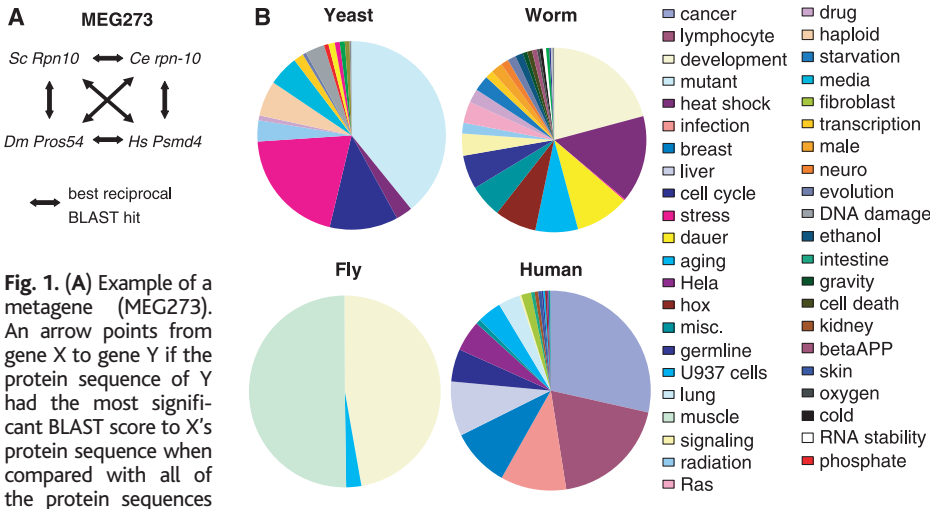


Fig. 1. (**A**) Example of a metagene (MEG273). An arrow points from gene X to gene Y if the protein sequence of Y had the most significant BLAST score to X's protein sequence when compared with all of the protein sequences in Y's database. (**B**) Compendiums of microarray expression data from four organisms included in the analysis. Color shows the type of DNA microarray experiment.

[1]Stanford Medical Informatics, 251 Campus Drive, Medical School Office Building X-215, Stanford, CA 94305–5329, USA. [2]Department of Computer Science, Gates Building 1A, Stanford University, Stanford, CA 94305–9010, USA. [3]Departments of Developmental Biology and Genetics, Stanford University School of Medicine, Stanford, CA 94305–5329, USA.

*These authors contributed equally to this work.
†Present address: Department of Biomolecular Engineering, University of California, Santa Cruz, CA 95064, USA.
‡To whom correspondence should be addressed. E-mail: kim@cmgm.stanford.edu (S.K.K.) and koller@cs.stanford.edu (D.K.)

of which encode a non–adenosine triphosphatase subunit of the 19S proteasome cap (Fig. 1A) (*15*). In total, this construction resulted in 6307 metagenes, consisting of 6591 human genes, 5180 worm genes, 5802 fly genes, and 2434 yeast genes (table S1).

We sought to identify pairs of metagenes that not only were coexpressed in one experiment and in one organism but that also showed correlation in diverse experiments in multiple organisms. We used data from a diverse set of DNA microarray experiments that were obtained from four different organisms: 1202 DNA microarrays from humans, 979 from worms, 155 from flies, and 643 from yeast (Fig. 1B and table S2). These gene-expression databases contain many different expression profiles that show how gene expression is perturbed by developmental stages, different growth conditions, stress, disease, and specific mutations. Correlation of expression profiles for a set of genes across different experimental conditions suggests that the set of genes are functionally related.

We then identified pairs of genes whose expression is significantly correlated in multiple organisms, indicating that their coexpression is conserved across evolution. We computed the Pearson correlation of the expression profiles between every pair of genes in the microarray data sets for each organism and then ranked all other genes according to their Pearson correlations. Next, we used a probabilistic method based on order statistics to evaluate the probability (*P* value) of observing a particular configuration of ranks across the different organisms by chance (*14*, *16*, *17*).

We used this probabilistic model to define a gene-coexpression network. We used $P < 0.05$ as a cutoff to indicate that two metagenes are coexpressed and combined all of the links between pairs of coexpressed metagenes to construct the network. The resulting network contained 3416 metagenes connected by 22,163 expression interactions (available on http://cmgm.stanford.edu/~kimlab/multiplespecies). Under the assumptions of the statistical model we used and our selection criteria ($P < .05$), we expected only 236 interactions by chance, significantly fewer than the 22,163 observed interactions.

We verified the significance of the interactions in the network by means of a variety of statistical tests. First, it was theoretically possible that the set of metagenes exhibited only a few simple types of expression patterns, so that even random pairs of metagenes might appear to have significant coexpression interactions. To rule out this possibility, we generated a set of permuted metagenes, consisting of a random collection of genes from each organism, and constructed a network from these permuted metagenes. We

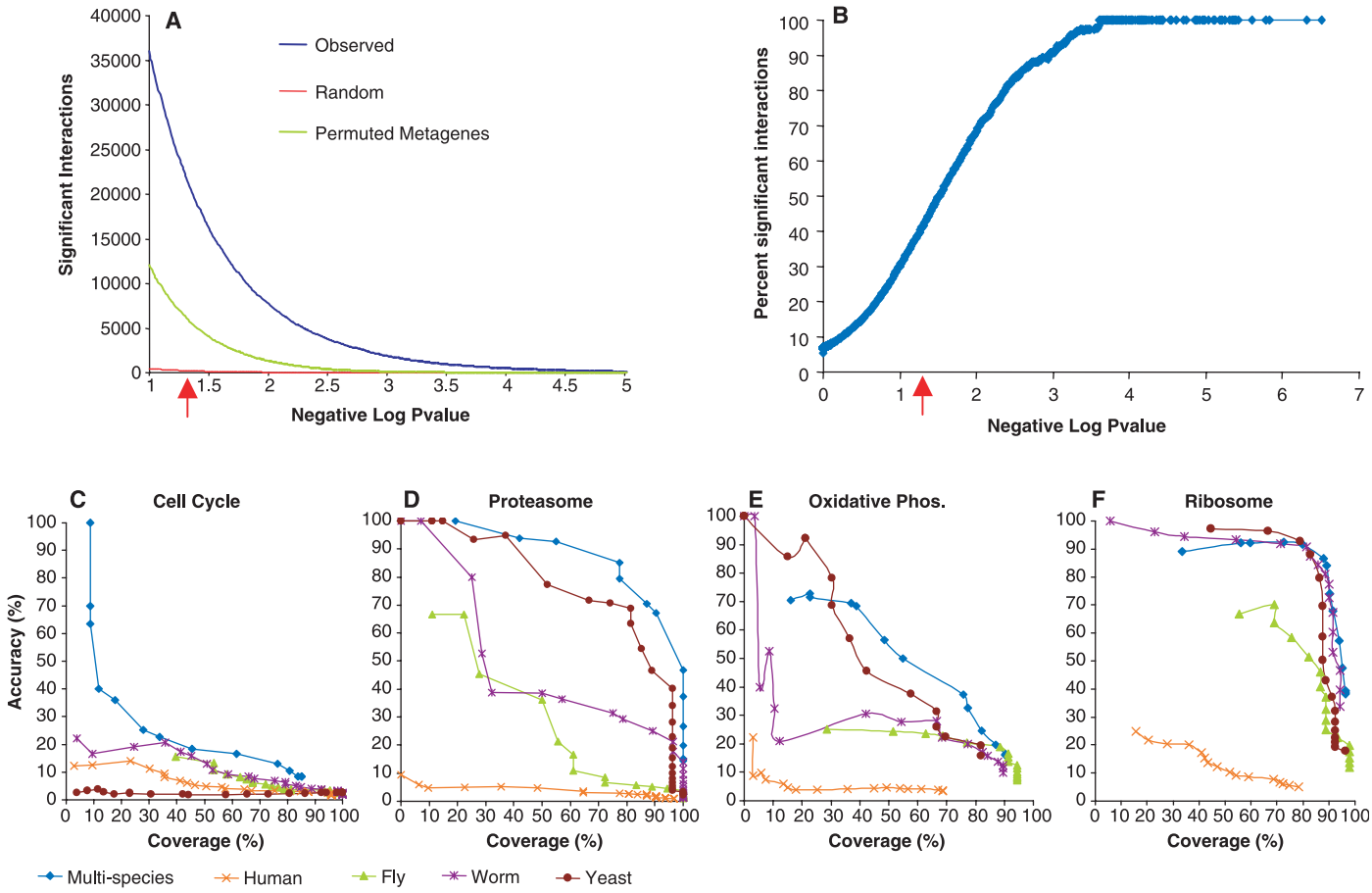


Fig. 2. Statistical validations and comparisons to single-species expression networks. (**A**) The number of metagene interactions (*y* axis) exceeding a *P*-value cutoff (*x* axis) in networks constructed from real metagenes (blue curve), a random distribution (red curve), and randomly permuted metagenes (green curve). *P* values are shown in $\log_{10}$ scale. Red arrow marks $P < .05$, the cutoff used in the gene-coexpression network. (**B**) We randomly divided the databases of each species into two equally sized sets and then generated new networks derived from each half of the data for a series of *P* values. Shown is the percent of metagene pairs with $P < p$ in the first half that have $P < 0.05$ in the second half, for each *P* value *p*. *P* values are shown in logarithmic scale. Three additional randomizations gave identical results. Red arrow marks $P < .05$, the cutoff used in the gene-coexpression network. (**C** to **F**) Comparison of multiple-species and single-species expression networks. We constructed a coexpression network from each species by selecting a Pearson correlation cutoff and linking every pair of genes with a correlation higher than the cutoff. We collected all of the links involving metagenes from a single functional category from KEGG (*22*). We then calculated the percentage of links connecting two members of the category (*y* axis; accuracy) and plotted this against the percentage of metagenes that are connected to at least one other metagene in the category (*x* axis; coverage). We varied the Pearson cutoff for constructing single-species networks and varied the *P*-value cutoff for constructing multiple-species networks to obtain different accuracies and coverages.

compared the number of interactions in the random network with the real network for a wide range of P values (Fig. 2A). We repeated this procedure five times with different random permutations of orthologs. For each permutation and for every P value, we found significantly more links in the real network than the random network; for example, at $P < 0.05$, the real networks contained 3.5 ± 0.03 times as many interactions as the random networks contained.

Second, we wanted to evaluate whether a large fraction of potential gene interactions was represented by the available microarray experiments. The current microarray experiments might reveal only a small fraction of possible gene interactions, and so the particular set of gene interactions found in each organism would be heavily dependent on the specific set of microarray experiments that we conducted. Alternatively, the microarray experiments might be broad and diverse, revealing a large fraction of possible gene-expression interactions. In this case, the coexpression relationships in the network would be robust to the choice of microarray experiments; for example, a significant fraction of the gene-expression links should be present in networks built with only a random half of the data. We randomly split the DNA microarray data in each organism's data set into two halves and then built two coexpression networks, each with only half of the data. We then counted the fraction of interactions that were significant in one network ($P < 0.05$), given that they were significant in the other network at $P < p$ for various values of p. We repeated this entire procedure five times. At $P = 0.05$, we found that 41% of the significant expression interactions in one network were also significant in the other network. These results indicate that the microarray experiments are reasonably broad and diverse, revealing a general set of gene interactions (Fig. 2B).

Finally, because gene-expression measurements contain some inherent variability, we tested whether the network was stable with respect to added noise. For example, *C. elegans* DNA microarray experiments typically vary by a SD of 0.3 to 0.5 ($\log_2$ expression ratios) (*18*, *19*). We added increasing levels of Gaussian noise to the entire data set for each of the organisms and constructed new networks from the perturbed data. Networks constructed following the addition of realistic levels of noise were very similar to the network from the original data (fig. S1).

Biological function of conserved gene-expression links. The multiple-species gene-coexpression network differs from previous gene-expression compendiums (such as the yeast compendium and the worm gene-expression terrain map) (*6*, *7*) in two major ways: (i) the multiple-species network only maps those genes that have orthologs in other species and thus focuses strongly on core, conserved biological processes; (ii) interactions in the multiple-species network imply a functional relationship based on evolutionary conservation, whereas interactions using data from single species only indicate correlated gene expression.

We next visualized the interconnectivity of the network in order to gain insights into the evolutionarily conserved patterns of expression and the coregulation architecture common to all four organisms. We chose to view the network as a terrain map with a three-dimensional (3D) layout program called VxInsight (*20*). In this visualization, metagenes are placed near each other in the x-y plane according to the negative logarithm of their P value, and the density of genes in a region is shown by the altitude in the z direction (Fig. 3). Using VxInsight, one can find highly interconnected areas of the network as peaks in the map as well as specific gene-gene interactions. The network and the VxInsight application are available on http://cmgm.stanford.edu/~kimlab/multiplespecies.

Each link in the terrain map suggests a potential interaction between two genes that has been conserved across evolution, and is therefore likely to be functionally related. We used K-means clustering on the x-y coordinates to define 12 regions of the terrain map that contain a large number of highly interconnected metagenes, and we refer to these regions as components (*14*). Most of the components were enriched for metagenes involved in similar biological processes, such as protein degradation, ribosomal function, cell cycle, metabolic pathways, and neuronal processes (Fig. 3 and Table 1).

Component 5 is an example of a group that is strongly enriched for metagenes involved in a common biological process. This component contains a total of 241 metagenes, 110 of which were previously known to be involved in the cell cycle (out of a total of 202 cell cycle metagenes in the network; 7.7 times as many as were expected using the hypergeometric distribution, $P < 10^{-85}$) (Table 1). Of the cell cycle metagenes, 30 are involved in regulating the cell cycle, such as MEG2742 (encodes cyclin E) and MEG5621 (encodes Wee1), along with 80 that perform terminal cell cycle functions, such as MEG1092 (encodes DNA polymerase-α). The remaining 131 genes were not previously known to be involved in the cell cycle, and so linking these genes to known cell cycle metagenes in the coexpression network suggests new cell cycle functions for these genes.

If a gene is linked in the network to many genes that participate in the same biological process, it is reasonable to hypothesize that it also participates in that process. We experimentally validated some of the gene functions that were predicted by the multiple-species network. We selected five metagenes that showed conserved coexpression with genes known to be involved in cell proliferation and the cell cycle but that were not previously known to be involved in these processes. Specifically, we chose MEG1503 (which encodes an snRNP protein involved in splicing), MEG342 (which encodes a nucleoporin-interacting component), and three

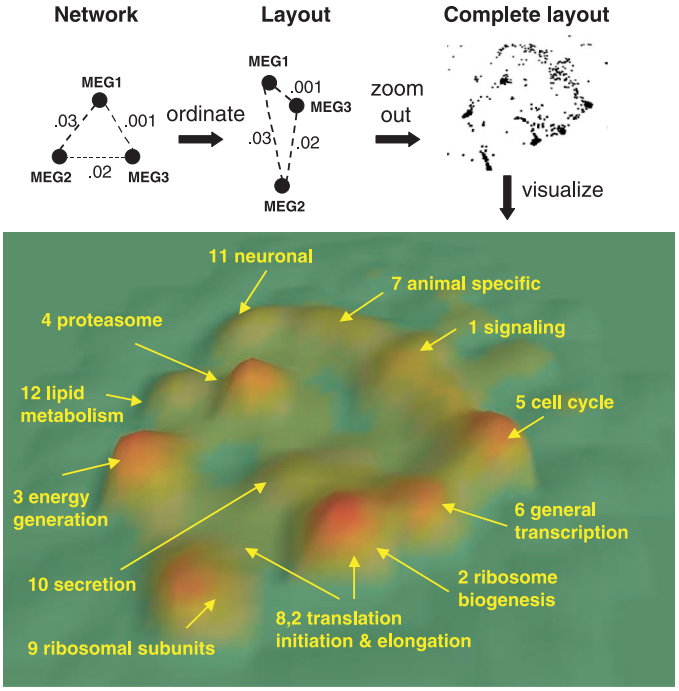


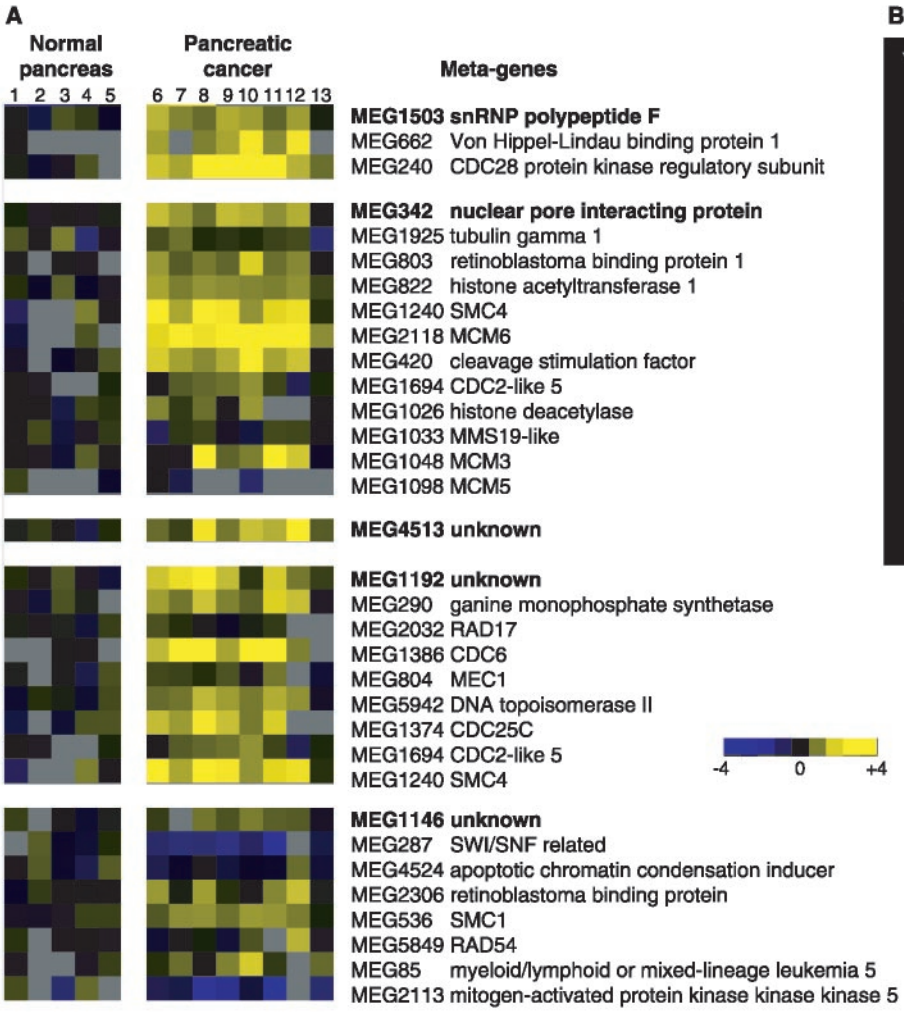
Fig. 3. The negative logarithm of the P values computed for conserved coexpression links were used to position the metagenes on a 2D grid using VxInsight's ordination tool (*20*). Metagenes with smaller P values (indicating a higher significance of conserved coexpression) were placed close to each other, whereas metagenes with larger P values were placed farther apart. The altitude in the final visualization indicates the local density of genes. The bottom panel shows the 3D representation for 3416 metagenes. Twelve components of highly interconnected metagenes are shown along with the main biological functions for which they were enriched. The entire data set can be queried for individual genes using VxInsight, which can be downloaded from http://cmgm.stanford.edu/~kimlab/multiplespecies.

other metagenes (MEG4513, MEG1192, and MEG1146) that encode previously unknown proteins of unknown function (table S1). All five of these metagenes showed a significant number of links in the coexpression networks to known cell proliferation genes (table S3).

We first tested the expression levels of these genes in dividing pancreatic cancer cells and in nondividing normal cells, using recently published data from Iacobuzio-Donahue *et al*. (*21*). (These data were not used to construct the gene-coexpression network.) Figure 4A shows that all five genes are overexpressed in human pancreatic cancers relative to normal tissue, to the same extent as genes known to be involved in cell proliferation. In our second experiment, we tested the loss-of-function mutant phenotype for one of these metagenes, MEG1503, which includes the *C. elegans* gene *ZK652.1*. We induced a loss-of-function mutant phenotype for *ZK652.1* by feeding worms double-stranded ZK652.1 RNA. We found that RNA interference (RNAi) of *ZK652.1* resulted in excess nuclei in the germ line, suggesting that the wild-type function of this gene is to suppress germline proliferation (Fig. 4B). Together, these experiments provide validation for the functional characterization of these genes in two different organisms.

The function of these five genes was much clearer in the multiple-species–coexpression network than it was in a network constructed with data from only a single organism. For each gene, we constructed an organism-specific neighborhood, which consisted of the genes that are most coexpressed with the given gene in that organism. On average, the neighborhoods of these five genes were over four times more enriched for cell proliferation and cell cycle genes in the multiple-species network than they were in the best single-species neighborhood (table S4). This observation supports our hypothesis that the multiple-species network tends to retain coexpression links between functionally related metagenes, whereas it discards spurious gene-expression links.

We evaluated this hypothesis on a more global scale by comparing the ability of the multiple-species and single-species networks to link together genes that were previously known to be involved in a single function, excluding genes not known to participate in that function. For most functional categories [as defined by the Kyoto Encyclopedia of Genes and Genomes (KEGG) database] (*22*), the multiple-species network performed significantly better than the single-species networks. In the case of the cell cycle function, for example, the multiple-species network performed significantly better than any of the networks that were formed using a single species (Fig. 2C). Specifically, for a given *P*-value cutoff, the multiple-species network had a higher percentage of cell cycle genes linked to one another (*y* axis;



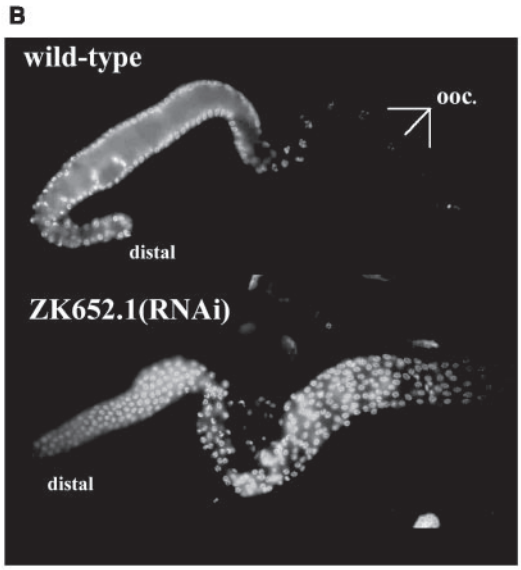


Fig. 4. (**A**) MEG1503, MEG342, MEG4513, MEG1192, and MEG1146 are overexpressed in pancreatic cancers. We plotted the metagenes with the GeneXPress program (http://genexpress.stanford.edu) using data from (*21*). The first five columns correspond to expression data obtained from normal pancreas specimens (pSF2779N, pSF442N, pSF4N, pSF5NT, and pSF768NT), and the remaining eight columns correspond to expression data obtained from pancreatic cancer specimens [a pancreatic cancer cell line (HS766T), five Hopkins/Goggins pancreatic cancer cell cultures (PL2, PL22, PL21, PL1, and PL8), a poorly differentiated pancreas carcinoma (pSF439T), and a pancreas foamy cell adenocarcinoma specimen (pSF1T)]. Each row corresponds to the expression profile of a single metagene across the 13 pancreatic samples. Bold indicates metagenes with unknown functions that are implicated in cell proliferation by the network. Neighbors of each implicated metagene that were previously known to be involved in cell proliferation or cell cycle are also shown. Scale shows $\log_2$ expression ratio. (**B**) RNAi-induced phenotype of *ZK652.1*. Shown are wild-type gonads and gonads from worms that were fed bacteria producing *ZK652.1* double-stranded RNA for 2 days (*29*). Gonads were stained with 4′,6′-diamidino-2-phenylindole to show DNA in the nuclei (*30*). *ZK652.1* (RNAi) gonads have more nuclei than the wild type and lack oocytes (ooc.). *rrf-3(pk1426)* worms were used because they are more sensitive to RNAi (*31*).

accuracy) for any given percentage of total cell cycle genes that are linked (*x* axis; coverage). Genes involved in proteasome function and genes involved in oxidative phosphorylation were also clustered more tightly and more exclusively by the multiple-species network compared with those in any of the single-species networks (Fig. 2C).

For some categories (such as genes involved in ribosomal function), the best network constructed with data from a single organism (such as using only yeast or worm microarray data) was about equal in quality to the network generated using multiple species (Fig. 2C), but the networks in other organisms performed much more poorly.

One possible explanation for the superior performance of the multiple-species network is the trivial fact that the multiple-species network was built from more DNA microarray data. To rule out this possibility, we repeated the functional prediction analysis with a multiple-species network formed from only 979 DNA microarrays (the same number as in the worm data set). We found that the network built from fewer microarrays performed as well as the network built from all of the microarrays in terms of predicting gene functional categories (fig. S2). Thus, we believe that the multiple-species network provides better functional predictions because it uses evolution to filter out gene interactions that are not functionally relevant.

The multiple-species network contains 570 metagenes that encode proteins of unknown function (orthologs that are conserved across evolution but whose function is poorly understood in any organism) (table S1). These metagenes have a total of 3943 connections to other metagenes in the network (many of which have known functions), potentially allowing these metagenes to be characterized.

Conservation and interaction of genetic modules. In addition to learning about the function of individual genes, one can use the network to analyze entire sets of genes to understand the system as a whole. Consider three types of genetic modules: (i) ancient, dedicated modules, (ii) evolving modules, and (iii) modules with interchangeable parts. Ancient modules, such as the group of metagenes involved in ribosomal function, have a main core cellular function that has been conserved from yeast to humans. Metagenes in these modules would be expected to have highly conserved coding regions and to contain gene-expression links that are conserved. Evolving modules, such as those modules involved in neuronal function, show rapid change among the four species. Metagenes in this type of module are expected to lack a yeast ortholog and to show relatively large changes in expression links between invertebrates and humans. Modules with interchangeable parts are composed of metagenes that have different links in different species. For example, *sir-2* encodes a protein that is highly conserved from yeast to humans and is involved in regulating chromatin structure and gene expression, but it has different downstream targets in each species (*23*). Other types of metagenes with adaptable, interchangeable functions would include those encoding transcription factors, signaling molecules, and adaptor proteins. These metagenes have coding sequences that are conserved but are likely to have different sets of gene-expression links in each species.

We tested the extent to which different modules are present in our gene-coexpression network. We split the set of metagenes into a set of 2969 metagenes that contained a yeast ortholog and a set of 3338 metagenes that were animal-specific in that they included genes from worms, flies, or humans but not yeast. Next, we determined the degree to which the gene-expression links have been conserved for each metagene by defining a set theoretic quantity called the expression-conservation index (ECI), in which larger values indicate stronger conservation (*14*).

We examined the degree of conservation of the different biological functions represented by the 12 main components in the gene-coexpression network (Fig. 5). Components 1, 7, and 11 were the most enriched for animal-specific metagenes and also showed the lowest degree of evolutionary conservation of their gene-expression links (ECIs of 0.67, 0.56, and 0.61, respectively) (Table 1 and Fig. 5). Component 1 was enriched for metagenes that were involved in signaling pathways, consistent with the idea that signaling pathways are animal-specific and regulate diverse sets of downstream genes in different organisms. Component 11 was enriched for metagenes involved in neuronal function, consistent with the idea that neuronal functions show large amounts of evolutionary change. Component 7 has yet to be correlated with any biological function. Nevertheless, the low conservation of both the coding region and gene interactions for this component suggests that it may be involved in processes that are evolving. In contrast to components 1, 7, and 11, component 9 is the least enriched for animal-specific metagenes and shows the highest degree of evolutionary conservation (2.4 average conservation degree), consistent with the known biological function associated with this component (ribosomal function).

We used the network to investigate the interconnections between multicellular and core cell-biological processes, and found several examples of processes that were intertwined. Component 3 is enriched for metagenes involved in metabolic pathways, particularly those used to generate energy such as the glycolytic pathway and the tricarboxylic acid cycle (Table 1). Within component 3, there is a small cluster of 76 metagenes that is enriched for

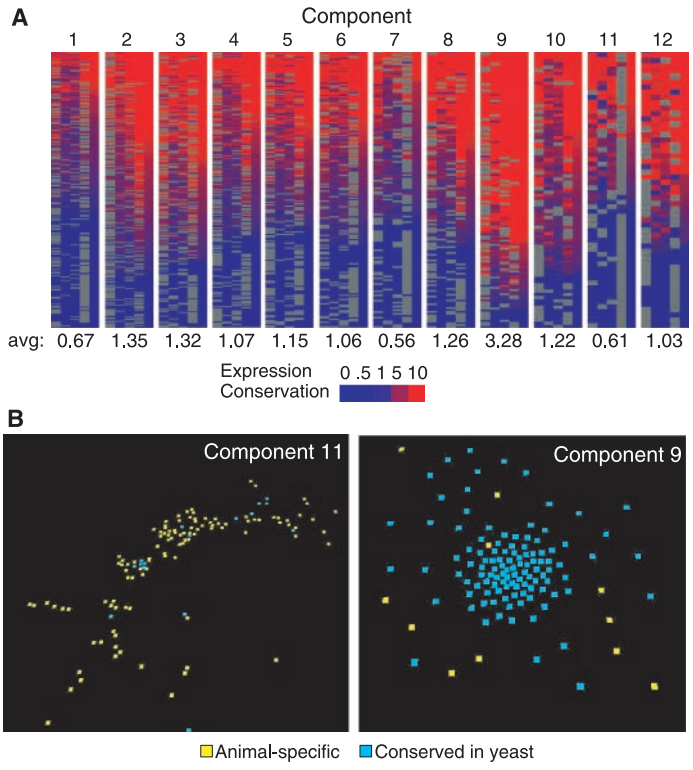


Fig. 5. (**A**) Conservation of components in the multiple-species network. Each box in the heat map represents the percentage of links connected to a metagene in one organism that are also present in the multiple-species–coexpression network (see table S5 for the computed conservation levels). Gray indicates that a metagene lacked an ortholog in a particular organism. The columns correspond to flies, humans, worms, and yeast, respectively. The final column shows the expression conservation index for a single metagene (*14*). The average conservation degree for all metagenes in the component is shown beneath the component's heat map. Scale shows the conservation degree. (**B**) Component 11 is composed primarily of animal-specific metagenes (yellow boxes), and component 9 is composed primarily of metagenes containing a yeast ortholog (blue boxes).

animal-specific metagenes (1.5 times more enriched; $P < 10^{-4}$) and metagenes with a low conservation degree of gene-expression links. Of the 48 animal-specific metagenes, 4 are involved in muscle function (10 times more enriched; $P < 10^{-4.6}$), which is consistent with the very high energy demands of muscle. Component 4 is enriched for metagenes involved in protein degradation, such as metagenes that encode proteasomal subunits (such as MEG1013) or ubiquitin ligases (such as MEG1233) (Table 1). Within this component, there is a cluster of 92 metagenes that are animal-specific. Within this animal-specific portion, three metagenes are involved in apoptosis (2.1 times more enriched; $P < 10^{-1.3}$), indicating a functional link between a core cell-biological process (protein degradation) and an animal-specific process (programmed cell death).

Connectivity properties of genetic networks. Some biological functions require the coordinated effort of a large number of genes. For example, many protein subunits are required for the ribosome to synthesize new polypeptide links, and many enzymes are required to generate energy in the glycosylation pathway. Other biological functions may be comprised of genes that either act alone or that have multiple sets of interaction partners. For example, transcription factor genes act with different regulators to regulate different downstream targets depending on cell context; they thus appear as isolated metagenes in the gene-coexpression network, because they do not have a consistent group of interacting partners. The network of genetic pathways that comprise an organism will be composed of some pathways that are designed to be large and others that are engineered to be small.

To characterize the connectivity properties of the gene-coexpression network, we counted the number of neighbors for each metagene in the network and compared this with neighborhood sizes arising from networks constructed from permuted data (Fig. 6) (*14*). We found that the distribution of gene-expression links in the gene-coexpression network was highly nonrandom, containing significantly more metagenes with a larger number of gene-expression links than the control networks. For example, there were 1290 metagenes with 10 or more links compared with only 168.3 ± 11.1 such links in the control networks.

The connectivity of the network follows a power-law distribution (Fig. 3; linear regression to log-log plot explains 92% of the variation) (*24*). This power-law function has been observed in other analogous natural and social phenomena (such as the distribution of U.S. firm sizes or the neighborhood sizes in the World Wide Web) and also in biological networks (such as protein-protein interaction) (*25–28*). This result suggests the existence of a selective force in the overall design of genetic pathways to maintain a highly connected class of genes.

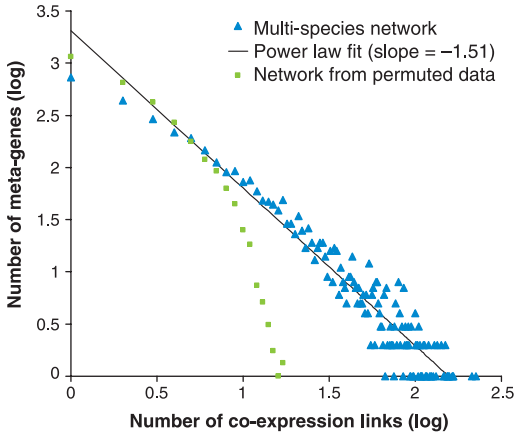


Fig. 6. Distribution of the number of links for each metagene. Shown is the number of links (x axis, $\log_{10}$ scale) compared with the number of metagenes that have that number of links (y axis, $\log_{10}$ scale) in the network (blue triangles) and in the networks constructed from permuted data (green squares) (*14*). The black line (slope of −1.51) depicts the least-squares fit of the data to a linear line in the log-log plot.

Table 1. Network components.

Component	Size*	Biological function†	Genes in component‡	Enrichment; P value§
1	353	Cellular cortex	16/57	2.7; $10^{-6.1}$
		Signaling	44/321	1.3; $10^{-5.8}$
		Animal-specific	195/1441	1.3; $10^{-7.2}$
2	349	Ribosome biogenesis	102/125	8.0; 10^{-83}
3	320	Energy generation	77/147	5.6; 10^{-42}
4	271	Proteasome	31/32	12; 10^{-32}
5	241	Cell cycle	110/202	7.7; 10^{-85}
6	201	General transcription	47/142	5.6; 10^{-24}
7	167	Animal-specific	124/1441	1.8; 10^{-17}
8	156	Translation initiation, elongation, and termination	20/110	4.0; $10^{-7.3}$
		Aminoacyl transfer RNA biosynthesis	14/31	9.9; 10^{-11}
9	139	Ribosomal protein subunits	74/78	23; 10^{-107}
10	92	Secretion	37/85	16; 10^{-38}
11	65	Neuronal	17/42	21; 10^{-19}
		Animal-specific	58/1441	2.1; 10^{-15}
12	57	Lipid metabolism	6/16	22; 10^{-7}
		Peroxisome	14/32	26; 10^{-17}

*The total number of metagenes in the component. †Biological functions were based on edited terms from Gene Ontology (*15*) and the KEGG database (*22*). ‡The number of metagenes in the biological function group and in the component divided by the total number of metagenes in the biological function group that were also in the network. §The ratio between the number of observed metagenes in a category and the number expected by chance. The P value was computed as the probability of obtaining the observed number of overlaps by chance under a hypergeometric distribution.

References and Notes

1. A. Goffeau *et al.*, *Science* **274**, 546 (1996).
2. E. W. Myers *et al.*, *Science* **287**, 2196 (2000).
3. E. S. Lander *et al.*, *Nature* **409**, 860 (2001).
4. J. C. Venter *et al.*, *Science* **291**, 1304 (2001).
5. M. B. Eisen, P. T. Spellman, P. O. Brown, D. Botstein, *Proc. Natl. Acad. Sci. U.S.A.* **95**, 14863 (1998).
6. T. R. Hughes *et al.*, *Cell* **102**, 109 (2000).
7. S. K. Kim *et al.*, *Science* **293**, 2087 (2001).
8. E. Segal *et al.*, *Nature Genet.* (2003).
9. O. Alter, P. O. Brown, D. Botstein, *Proc. Natl. Acad. Sci. U.S.A.* **100**, 3351 (2003).
10. V. van Noort, B. Snel, M. A. Huynen, *Trends Genet.* **19**, 238 (2003).
11. S. A. Teichmann, M. M. Babu, *Trends Biotechnol.* **20**, 407 (2002).
12. R. L. Tatusov, M. Y. Galperin, D. A. Natale, E. V. Koonin, *Nucleic Acids Res.* **28**, 33 (2000).
13. Y. Lee *et al.*, *Genome Res.* **12**, 493 (2002).
14. Materials and methods are available as supporting material on *Science* Online.
15. M. Ashburner *et al.*, *Nature Genet.* **25**, 25 (2000).
16. C. M. Grinstead, J. L. Snell, *Introduction to Probability* (American Mathematical Society, Providence, RI, 1997).
17. K. Siegrist, *Virtual Laboratories in Probability and Statistics* (Univ. of Alabama, Huntsville, AL, 1997), available on www.math.uah.edu/statold/sample/sample7.html.
18. M. Jiang *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **98**, 218 (2001).
19. V. Reinke *et al.*, *Mol. Cell* **6**, 605 (2000).
20. G. S. Davidson, B. Hendrickson, D. K. Johnson, C. E. Meyers, B. N. Wylie, *J. Intelligent Inf. Syst.* **11**, 259 (1998).
21. C. A. Iacobuzio-Donahue *et al.*, *Am. J. Pathol.* **162**, 1151 (2003).
22. H. Ogata *et al.*, *Nucleic Acids Res.* **27**, 29 (1999).
23. L. Guarente, *Genes Dev.* **14**, 1021 (2000).
24. G. K. Zipf, *Human Behavior and the Principle of Least Effort* (Addison Wesley, Cambridge, MA, 1949).
25. S. M. Gomez, S. H. Lo, A. Rzhetsky, *Genetics* **159**, 1291 (2001).
26. R. L. Axtell, *Science* **293**, 1818 (2001).
27. X. Gabaix, P. Gopikrishnan, V. Plerou, H. E. Stanley, *Nature* **423**, 267 (2003).
28. D. Garlaschelli, G. Caldarelli, L. Pietronero, *Nature* **423**, 165 (2003).
29. A. G. Fraser *et al.*, *Nature* **408**, 325 (2000).
30. J. Zalevsky, A. J. MacQueen, J. B. Duffy, K. J. Kemphues, A. M. Villeneuve, *Genetics* **153**, 1271 (1999).
31. F. Simmer *et al.*, *Curr. Biol.* **12**, 1317 (2002).

32. We thank the Stanford Microarray Database for providing data, M. Hansen for providing RNAi feeding

strains, members of the Kim and Villeneuve labs for advice, many *C. elegans* researchers for use of their unpublished microarray data (supporting online material), A. Preble for comments on the manuscript, and K. Reddy for providing help with RNAi. Supported by an NIH genome training grant (J.M.S.), a Stanford Graduate Fellowship (E.S.), grants from the National Institute of General Medical Sciences and National Center for Research Resources (S.K.K.), and NSF grant ACI-0082554 under the Information Technology Research program (D.K.).

Supporting Online Material
www.sciencemag.org/cgi/content/full/1087447/DC1
Materials and Methods
SOM Text
Figs. S1 to S4
Tables S1 to S5

30 May 2003; accepted 13 August 2003
Published online 21 August 2003;
10.1126/science.1087447
Include this information when citing this paper.

Control Mechanism of the Circadian Clock for Timing of Cell Division in Vivo

Takuya Matsuo,[1,2]* Shun Yamaguchi,[1]* Shigeru Mitsui,[1] Aki Emi,[1] Fukuko Shimoda,[1] Hitoshi Okamura[1]†

Cell division in many mammalian tissues is associated with specific times of day, but just how the circadian clock controls this timing has not been clear. Here, we show in the regenerating liver (of mice) that the circadian clock controls the expression of cell cycle–related genes that in turn modulate the expression of active Cyclin B1-Cdc2 kinase, a key regulator of mitosis. Among these genes, expression of *wee1* was directly regulated by the molecular components of the circadian clockwork. In contrast, the circadian clockwork oscillated independently of the cell cycle in single cells. Thus, the intracellular circadian clockwork can control the cell-division cycle directly and unidirectionally in proliferating cells.

Circadian (~24-hour) rhythms and cell division are fundamental biological systems in most organisms. There is substantial evidence that, in mammals, circadian rhythms affect the timing of cell divisions in vivo. Day-night variations in both the mitotic index and DNA synthesis occur in many tissues (e.g., oral mucosa, tongue keratinocytes, intestinal epithelium, skin, and bone marrow) (*1*–*6*), some of which persist even in constant darkness (*7*). However, how the circadian clock controls the timing of cell divisions is not known.

To explore the relationship between cell division and circadian rhythms, we used a mouse model with partial hepatectomy (PH) (*8*–*13*). After a two-thirds partial hepatectomy (*14*), most of the remaining hepatocytes rapidly and simultaneously enter into the cell cycle, resulting in restoration of the liver mass in a few days.

Diurnal control of cell cycle in wild-type mice. PH was performed on mice (*15*) at ZT8 or ZT0 (ZT, Zeitgeber time in a 12 hour light–12 hour dark cycle; ZT0 represents lights on and ZT12, lights off) to compare the kinetics of subsequent cell cycles. The kinetics of S-phase (DNA-synthesizing) hepatocytes for both ZTs were comparable as determined by bromodeoxyuridine (BrdU) incorporation into nuclei, peaking at 36 hours after PH (Fig. 1A). In contrast, subsequent mitotic waves differed (Fig. 1B). When PH was performed at ZT8 (PH/ZT8), a massive entry of hepatocytes into the M phase occurred within 40 hours after PH. In the case of PH/ZT0, however, only a few cells entered the M phase within 44 hours, and a mitotic peak was reached 48 hours after PH. These results suggest that the time of operation has a marked effect on the timing of mitosis controlling the progression of cell cycling itself.

To investigate the molecular mechanism underlying this time of day–dependent regulation of the cell cycle, we examined the kinase activity of Cdc2 (*15*), an initiator of mitosis (*16*). Peaks of Cdc2 activity after PH/ZT8 and PH/ZT0 occurred 40 and 48 hours after PH, respectively, corresponding to the observed mitotic peaks (Fig. 1, B and C). This suggests that the regulation of the expression of the active Cdc2 kinase is an important process for the diurnal control of the cell cycle.

Analysis of the expression profiles of 68 cell cycle–related genes by DNA microarray and Northern blot analysis (fig. S1, A and B; fig. S2; and table S1) revealed that although 11 genes showed moderately different kinetics between PH/ZT8 and PH/ZT0 (differences of 1.5- to 2.2-fold or 0.67- to 0.42-fold; ratios of PH/ZT0 to PH/ZT8) (fig. S2 and table S2), only three genes—*cyclin B1*, *cdc2*, and *wee1*—showed remarkably different expression profiles (a difference of more than 2.7-fold) between 28 and 56 hours after PH (Fig. 1D).

The expression peaks of *cyclin B1* and *cdc2* transcripts, whose products form Cyclin B1-Cdc2-complex kinase, corresponded with Cdc2 kinase activity peaks (Fig. 1, C and D). Both mRNA peaks were delayed by 8 to 12 hours after PH/ZT0 as compared to PH/ZT8. The *wee1* gene product phosphorylates Cdc2 on Tyr-15 [p-Cdc2(Tyr 15)] and keeps it in an inactive form (*17*, *18*); the decrease of its mRNA corresponded with the increase of the Cdc2 kinase activity (Fig. 1, C and D) with the same time delay. These results suggest that the transcript-level regulation of *cyclin B1*, *cdc2*, and *wee1* contributes to the timing of entry into mitosis.

Impaired liver regeneration in arrhythmic *Cry*-deficient mice. Mice that lack the clock regulator *cryptochromes* (*Crys*) completely lack free-running rhythmicity. Yet, embryogenesis and postnatal development in these mice appear normal (*19*), suggesting that clock function is not absolutely required for cell cycling. Although the mean weights of the pre-PH livers of wild-type and *Cry*-deficient animals were the same ($P > 0.05$), the weights of the regenerating livers in *Cry*-deficient mice 72 hours after PH/ZT8 were significantly lower than those of wild-type mice (53.1 ± 3.0% and 64.1 ± 1.7% of pre-PH liver weight, respectively; Student's *t* test, $P < 0.05$) (Fig. 2A). This value for *Cry*-deficient mice was also lower than that of wild-type mice 72 hours after PH/ZT0 (66.4 ± 1.7%; $P < 0.01$), suggesting impairment of hepatocyte proliferation in *Cry*-deficient mice. However, in both genotypes, liver weight returned to pre-PH levels by day 10 (Fig. 2A), and histological analyses (including comparisons of the number and cell size of hepatocytes and average distances separating portal and central hepatic veins), did not reveal any major differences (*20*). This indicates that circadian clock function is required for efficient cell cycling in vivo.

The kinetics of S-phase hepatocytes after PH/ZT8 or PH/ZT0 in *Cry*-deficient and wild-type mice were comparable (Fig. 2B). However, in the subsequent mitotic wave, the maximum value of mitotic hepatocytes was low (less than 4% for both PH/ZT8 and PH/ZT0) (Fig. 2B). Furthermore, Cdc2 kinase activity was reduced (Figs. 2B and 1C). Therefore, the cell cycle progression from S to M phase was impaired during liver regeneration in *Cry*-deficient mice. In

[1]Division of Molecular Brain Science, Department of Brain Sciences, Kobe University Graduate School of Medicine, Chuo-ku, Kobe 650–0017, Japan. [2]Department of Physics, Informatics and Biology, Yamaguchi University, Yamaguchi 753–8512, Japan.

*These authors contributed equally to this work.
†To whom correspondence should be addressed. E-mail: okamurah@kobe-u.ac.jp