Dynamics of cellular level function and regulation derived from murine expression array data

Benjamin de Bivort*[†], Sui Huang[‡], and Yaneer Bar-Yam*[§]

*Department of Molecular and Cellular Biology, Harvard University, Cambridge, MA 02138; [‡]Vascular Biology Program, Departments of Pathology and Surgery, Children's Hospital and Harvard Medical School, Boston, MA 02115; and [§]New England Complex Systems Institute, Cambridge, MA 02138

Edited by Charles R. Cantor, Sequenom, Inc., San Diego, CA, and approved November 4, 2004 (received for review September 10, 2004)

A major open question of systems biology is how genetic and molecular components interact to create phenotypes at the cellular level. Although much recent effort has been dedicated to inferring effective regulatory influences within small networks of genes, the power of microarray bioinformatics has yet to be used to determine functional influences at the cellular level. In all cases of data-driven parameter estimation, the number of model parameters estimable from a set of data is strictly limited by the size of that set. Rather than infer parameters describing the detailed interactions of just a few genes, we chose a larger-scale investigation so that the cumulative effects of all gene interactions could be analyzed to identify the dynamics of cellular-level function. By aggregating genes into large groups with related behaviors (megamodules), we were able to determine the effective aggregate regulatory influences among 12 major gene groups in murine B lymphocytes over a variety of time steps. Intriguing observations about the behavior of cells at this high level of abstraction include: (*i*) a medium-term critical global transcriptional dependence on ATP-generating genes in the mitochondria, (*ii*) a longer-term dependence on glycolytic genes, (*iii*) the dual role of chromatin-reorganizing genes in transcriptional activation and repression, (*iv*) homeostasis-favoring influences, (*v*) the indication that, as a group, G protein-mediated signals are not concentration-dependent in their influence on target gene expression, and (*vi*) short-term-activating/long-term-repressing behavior of the cell-cycle system that reflects its oscillatory behavior.

module | network

Since the advent of DNA microarray technology (1), various efforts have been made to infer molecular network function from array data. First, genes were clustered by similar responses to perturbation (2–4), and genes in these expression clusters were found to share cis-regulatory elements, providing a molecular basis for their similarity in expression behavior (5). To determine the dynamics of regulatory networks, several reverse-engineering approaches have been suggested: discrete networks, linear models, Bayesian networks of dependencies, etc. (6–10). Some successful inferences have been demonstrated for small circuits (relatively low-dimensional subsystems) (6, 11, 12). Additionally, genomic data have been integrated into models based on experimental genetics (13). However, in organisms not amenable to genetic manipulation, the possibility of obtaining a phenomenological model of a genome-scale influence network appears remote, due to noise in microarray studies (14) and the large number of variables involved. Even assuming a linear influence model, in which a vector of changes in n gene expression levels $\mathbf{Y}$ at time t_2 is determined by expression level changes $\mathbf{X}$ at time t_1 and a transition matrix $\mathbf{M}$ by $\mathbf{Y} = \mathbf{M}\cdot\mathbf{X}$, one must solve for $n \times n$ influence variables in $\mathbf{M}$. Thus for 10^4 genes (and $n^2 = 10^8$ influences), one would need 10,000 transitions or perturbations for mathematical solubility, which is currently beyond experimental capacity.

One way around this problem is to analyze the cell at a higher level of abstraction, thereby reducing the number of variables. If the modular hypothesis of network organization is at all valid, then to the extent that genes combine to form minimodules of insulated function (which in turn aggregate into meso- and megamodules) (15), they form higher-level regulatory networks with fewer interactions. The level of study can therefore be tuned so that the set of effective interactions is mathematically soluble given the data available (16). The effective regulatory influences inferred based upon sequential observations do not represent physical interactions and do not account for indirect effects, e.g., prior, intermediate, or parallel transitions that are always present under the conditions observed. This study is no different from others in that regard, although it uses a larger scale of analysis than others, inferring regulatory networks from expression data.

Materials and Methods

Data from the Alliance for Cell Signaling (AfCS) splenic B lymphocyte ligand screen (19) track the expression levels of ≈16,000 cDNAs at time intervals of no delay and 0.5, 1, 2, and 4 h after perturbation by the addition of 1 of 32 ligands and without perturbation as a control. Thus each gene is associated with 33 expression-level time courses. To group the genes by similar activity profiles, the 33 time courses were concatenated to form an expression profile for each gene with 132 values, which were allocated into 12 bins by profile similarity using the self-organizing map (SOM) algorithm (17), as implemented in the GEDI (16) software add-on to MATLAB (18). The centroid profile of each bin was then used as the expression profile for each of the gene groups.

To assign functions onto each of the gene groups, we used the AfCS probe identifications to map Gene Ontology annotations onto the gene names within each group. To identify large-scale processes associated with gene groups, we sorted the 943 Process (P) annotations into 39 categories. χ^2 analysis was used to determine which P categories were overrepresented in which gene groups. For greater resolution, we also performed the analogous χ^2 test with the 943 P annotations uncategorized.

The 132-value profile data set was parsed into six different profiles of 66 values, each containing an initial and final time point for the 33 different ligand conditions. Using all combinations of initial and final time points, we analyzed the following transitions: 0.5 h (0.5 h initial to 1 h final), 1 h (1–2 h), 1.5 h (0.5–2 h), 2 h (2–4 h), 3 h (1–4 h), and 3.5 h (0.5–4 h).

The effective influence of gene group a on gene group b (α_{ab}) was determined for each of the 12 groups over the six different time intervals by calculating the least-squares fit of the parameters $\alpha_{0a}, \alpha_{1a}, \ldots, \alpha_{11\text{a}}$ to the equations

$$x_{a,i,t+k} = \alpha_{0a}\, x_{0,i,t} + \alpha_{1a}\, x_{1,i,t} + \ldots + \alpha_{aa}\, x_{a,i,t} + \ldots + \alpha_{11a}\, x_{11,i,t}$$

for all i, where $x_{a,i,t+k}$ is the expression level of group a at time $t+k$ (k is one of the six time intervals), and i is one of the 33 ligand experiments. Systems of these 33 equations in the 12 parameters

This paper was submitted directly (Track II) to the PNAS office.

Abbreviations: AfCS, Alliance for Cellular Signaling; SOM, self-organizing map.

[†] To whom correspondence should be addressed. E-mail: bivort@fas.harvard.edu.

www.pnas.org/cgi/doi/10.1073/pnas.0406707102

influencing a particular gene group are mutually independent and were solved sequentially for each gene group to determine 144 influence coefficients.

These values were alternatively calculated by using a bootstrapping methodology. For each bootstrap replicate, 12 of 33 experiments were chosen randomly, and the expression values for each gene group at times t, and $t+k$, were used to solve 144 equations (12 equations, one for each gene group, are derived from each of 12 experiments), for the 144 influence coefficients simultaneously. Because the estimates of both the signs and magnitudes of the coefficients were sensitive to which 12 experiments were chosen in the bootstrap replicate (probably due to a high degree of noise in the array data), the coefficients were first normalized by the variance across the 144 coefficients and then averaged over 2,800 bootstrap replicates. Therefore, using the bootstrap analysis, each coefficient was represented by a distribution of 2,800 estimates. t tests between these distributions were performed in a pairwise manner for all coefficients to calculate the P values shown in Fig. 3. The number of bootstrap replicates was chosen for convergence in the P values.

Gene groups were grouped by similarity of either inputs (rows of influence tables in Fig. 2) or outputs (columns) across all time intervals by using distance-based clustering with correlation coefficient as a distance metric. To gauge the relationships between expression profile and input or output, the Euclidean distance between each gene group's vector of input (or output) coefficients was calculated in a pairwise fashion. Similarly, the Euclidean distance between the positions of all pairs of tiles in the SOM analysis (equivalent to an *a priori* estimate of their similarity) was calculated, as was the correlation coefficient between all pairs of expression profiles, and the correlation coefficient between these sets of distances was determined.

Results

We analyzed the gene expression data made available by the AfCS, in which the response of murine B cells to 32 perturbations was monitored over time (19). We present the effective influences among 12 megamodules and find that, even at this level of aggregation, modules are significantly enriched for specific gene functions, as annotated in the Gene Ontology database. At this level of abstraction, cellular transcription behavior appears to be dominated by three major influences: ATP generation and consumption, the cell cycle, and the need to impose bounds on the level of transcriptional activity.

Logarithmic changes in expression level of ≈15,000 genes at 0.5, 1, 2, and 4 h after the addition of a signaling ligand molecule comprise the AfCS data. We used the SOM algorithm (17) to reduce the number of unknown variables in the transition matrix $\mathbf{M}$ from ≈225,000,000 (for $n = 15{,}000$) to 144 ($n = 12$). The SOM uses the set of observed expression profile time courses to generate an array of "representative time courses" that are spatially related to each other (for example, a logarithmically increasing time course would tend to occur adjacent to other monotonically increasing profiles, and far from decreasing profiles). Other techniques have been used to reduce the number of model parameters, in particular, singular value decomposition (8, 10, 20) and principle component analysis (21). Our application of SOM differs from these techniques in that it uses the full set of available expression data to partition the genes into similarly behaving but still mutually influencing aggregates, rather than identifying the principal linear modes of variation.

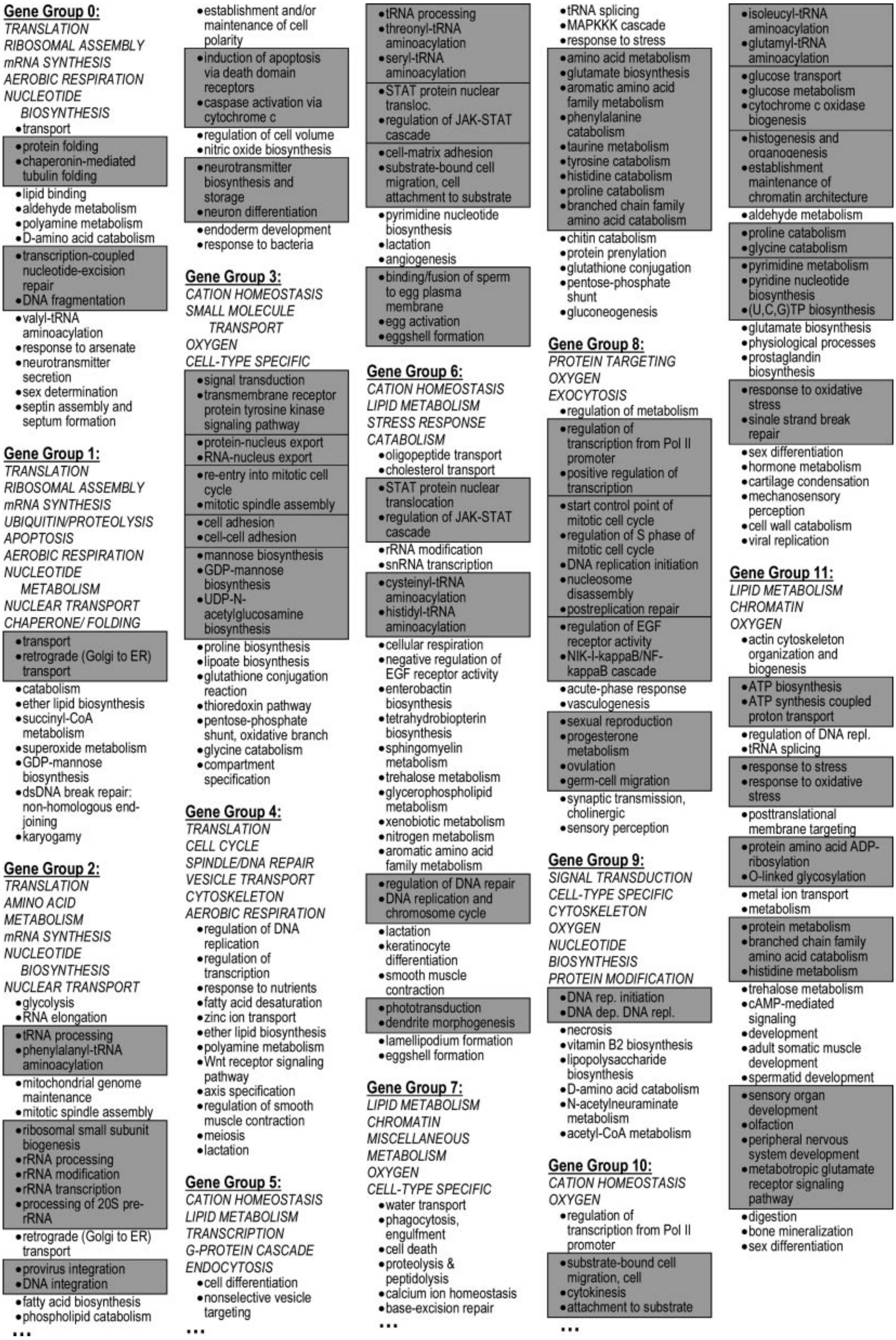
Applying the SOM algorithm to the AfCS data yielded 12 groups (in a 4 × 3 array) with characteristic gene expression time courses across all time points of the 33 ligand experiments (see Tables 1 and 2, which are published as supporting information on the PNAS web site). Associated with ≈5,000 genes in the AfCS database are functional ontology labels that identify the cellular process(es) in which each gene is involved. To uncover higher-level trends in the function of each gene group, we sorted the 943 unique ontology labels into 39 categories (see Table 3, which is published as supporting information on the PNAS web site). Functional labels associated with statistical significance (P value <0.0016) in genes of a particular SOM group are shown in Fig. 1. Total categories associated with each gene group are shown in italics, functions overrepresented individually are in lowercase, and individual functions contributing to a process category that was not overrepresented as a whole appear in shaded boxes.

The gene expression levels that characterized each of the 12 SOM gene groups at a particular time step in the ligand experiments were used to solve (using least-squared fitting and bootstrap analysis; see *Materials and Methods*) for the effective regulatory influences ($\mathbf{M}$) that each gene group has over every other gene group over that time step. These influences are presented in Fig. 2. For a given time step, the value in the ith row and jth column represents the effective influence of j on i, the extent to which the presence of gene group j at the initial time activates (green) or inhibits (red) the presence of gene group i after the time step. Columns are sorted by the similarity of outputs across all time intervals; rows are sorted by the similarity of inputs (see *Supporting Text*, which is published as supporting information on the PNAS web site).

To control for artifacts of the computational methods, we randomized the order of expression values for each gene group across all experiments and recalculated the influence coefficients (Fig. 2, R). The clear qualitative differences between this table and the actual data indicate that the analysis reveals actual patterns in the data. We also performed two stability analyses, convergence and bootstrap resampling (see *Supporting Text*), to verify that our results were not being overly influenced by a few dominant observations.

The data generated from the bootstrapping analysis allowed us to evaluate our confidence in each of the coefficient values; high levels of confidence can be assigned to coefficients that are calculated consistently across all bootstrap replicates. A pairwise comparison among all influences reveals which are statistically distinct. Fig. 3 shows the P values associated with the comparison of the means of every pair of casual coefficients: red, $P < 0.05$; yellow, $0.05 < P < 0.1$; gray, $P > 0.1$. Rows and columns are given in the ranked order of the influence's average value over the 2,800 bootstrap replicates, arranged from most inhibiting to most activating. The presence of gray block features along the diagonal indicates sets of influences that are not statistically distinct, i.e., types of effectively equivalent influence, such as "weak activator" or "weak repressor."

Fig. 2 provides a direct understanding of the coupling between change in gene expression at one time and a later time, across a wide variety of conditions. The coupling is summarized as a set of effective influences. Among the interesting observations is the possibility that a single gene group can effectively act as both an activator and a repressor of another gene group, depending on which time interval is analyzed. For example, an increase in gene group 11 implies an increase in gene group 7 after 1.5 h but a decrease after 2 h. The presence of activation and inhibition over different times illustrates that, whereas the discrete time-step influences are assumed to be linear, the separate treatment of each time interval can reveal underlying biological nonlinearity. Several biological mechanisms can explain this nonlinearity. Time delays in influence may arise from the accumulation of transcription products to regulatory thresholds; if, for example, multiple other groups mediate the effect of 11 on 7, then the earliest of these groups could be activators and the later groups inhibitors. Alternatively, group 11's activation of 7 could follow chemical equilibrium dynamics where the accumulation of the "product," gene group 7, eventually acts to inhibit its own production; or the inhibition subsequent to activation could reflect negative feedback serving to dampen global transcriptional activity.



Gene Group 0:
TRANSLATION
RIBOSOMAL ASSEMBLY
mRNA SYNTHESIS
AEROBIC RESPIRATION
NUCLEOTIDE BIOSYNTHESIS
- transport
- protein folding
- chaperonin-mediated tubulin folding
- lipid binding
- aldehyde metabolism
- polyamine metabolism
- D-amino acid catabolism
- transcription-coupled nucleotide-excision repair
- DNA fragmentation
- valyl-tRNA aminoacylation
- response to arsenate
- neurotransmitter secretion
- sex determination
- septin assembly and septum formation

Gene Group 1:
TRANSLATION
RIBOSOMAL ASSEMBLY
mRNA SYNTHESIS
UBIQUITIN/PROTEOLYSIS
APOPTOSIS
AEROBIC RESPIRATION
NUCLEOTIDE METABOLISM
NUCLEAR TRANSPORT
CHAPERONE/ FOLDING
- transport
- retrograde (Golgi to ER) transport
- catabolism
- ether lipid biosynthesis
- succinyl-CoA metabolism
- superoxide metabolism
- GDP-mannose biosynthesis
- dsDNA break repair: non-homologous end-joining
- karyogamy

Gene Group 2:
TRANSLATION
AMINO ACID METABOLISM
mRNA SYNTHESIS
NUCLEOTIDE BIOSYNTHESIS
NUCLEAR TRANSPORT
- glycolysis
- RNA elongation
- tRNA processing
- phenylalanyl-tRNA aminoacylation
- mitochondrial genome maintenance
- mitotic spindle assembly
- ribosomal small subunit biogenesis
- rRNA processing
- rRNA modification
- rRNA transcription
- processing of 20S pre-rRNA
- retrograde (Golgi to ER) transport
- provirus integration
- DNA integration
- fatty acid biosynthesis
- phospholipid catabolism

…

- establishment and/or maintenance of cell polarity
- induction of apoptosis via death domain receptors
- caspase activation via cytochrome c
- regulation of cell volume
- nitric oxide biosynthesis
- neurotransmitter biosynthesis and storage
- neuron differentiation
- endoderm development
- response to bacteria

Gene Group 3:
CATION HOMEOSTASIS
SMALL MOLECULE TRANSPORT
OXYGEN
CELL-TYPE SPECIFIC
- signal transduction
- transmembrane receptor protein tyrosine kinase signaling pathway
- protein-nucleus export
- RNA-nucleus export
- re-entry into mitotic cell cycle
- mitotic spindle assembly
- cell adhesion
- cell-cell adhesion
- mannose biosynthesis
- GDP-mannose biosynthesis
- UDP-N-acetylglucosamine biosynthesis
- proline biosynthesis
- lipoate biosynthesis
- glutathione conjugation reaction
- thioredoxin pathway
- pentose-phosphate shunt, oxidative branch
- glycine catabolism
- compartment specification

Gene Group 4:
TRANSLATION
CELL CYCLE
SPINDLE/DNA REPAIR
VESICLE TRANSPORT
CYTOSKELETON
AEROBIC RESPIRATION
- regulation of DNA replication
- regulation of transcription
- response to nutrients
- fatty acid desaturation
- zinc ion transport
- ether lipid biosynthesis
- polyamine metabolism
- Wnt receptor signaling pathway
- axis specification
- regulation of smooth muscle contraction
- meiosis
- lactation

Gene Group 5:
CATION HOMEOSTASIS
LIPID METABOLISM
TRANSCRIPTION
G-PROTEIN CASCADE
ENDOCYTOSIS
- cell differentiation
- nonselective vesicle targeting

…

- tRNA processing
- threonyl-tRNA aminoacylation
- seryl-tRNA aminoacylation
- STAT protein nuclear transloc.
- regulation of JAK-STAT cascade
- cell-matrix adhesion
- substrate-bound cell migration, cell attachment to substrate
- pyrimidine nucleotide biosynthesis
- lactation
- angiogenesis
- binding/fusion of sperm to egg plasma membrane
- egg activation
- eggshell formation

Gene Group 6:
CATION HOMEOSTASIS
LIPID METABOLISM
STRESS RESPONSE
CATABOLISM
- oligopeptide transport
- cholesterol transport
- STAT protein nuclear translocation
- regulation of JAK-STAT cascade
- rRNA modification
- snRNA transcription
- cysteinyl-tRNA aminoacylation
- histidyl-tRNA aminoacylation
- cellular respiration
- negative regulation of EGF receptor activity
- enterobactin biosynthesis
- tetrahydrobiopterin biosynthesis
- sphingomyelin metabolism
- trehalose metabolism
- glycerophospholipid metabolism
- xenobiotic metabolism
- nitrogen metabolism
- aromatic amino acid family metabolism
- regulation of DNA repair
- DNA replication and chromosome cycle
- lactation
- keratinocyte differentiation
- smooth muscle contraction
- phototransduction
- dendrite morphogenesis
- lamellipodium formation
- eggshell formation

Gene Group 7:
LIPID METABOLISM
CHROMATIN
MISCELLANEOUS METABOLISM
OXYGEN
CELL-TYPE SPECIFIC
- water transport
- phagocytosis, engulfment
- cell death
- proteolysis & peptidolysis
- calcium ion homeostasis
- base-excision repair

…

- tRNA splicing
- MAPKKK cascade
- response to stress
- amino acid metabolism
- glutamate biosynthesis
- aromatic amino acid family metabolism
- phenylalanine catabolism
- taurine metabolism
- tyrosine catabolism
- histidine catabolism
- proline catabolism
- branched chain family amino acid catabolism
- chitin catabolism
- protein prenylation
- glutathione conjugation
- pentose-phosphate shunt
- gluconeogenesis

Gene Group 8:
PROTEIN TARGETING
OXYGEN
EXOCYTOSIS
- regulation of metabolism
- regulation of transcription from Pol II promoter
- positive regulation of transcription
- start control point of mitotic cell cycle
- regulation of S phase of mitotic cell cycle
- DNA replication initiation
- nucleosome disassembly
- postreplication repair
- regulation of EGF receptor activity
- NIK-I-kappaB/NF-kappaB cascade
- acute-phase response
- vasculogenesis
- sexual reproduction
- progesterone metabolism
- ovulation
- germ-cell migration
- synaptic transmission, cholinergic
- sensory perception

Gene Group 9:
SIGNAL TRANSDUCTION
CELL-TYPE SPECIFIC
CYTOSKELETON
OXYGEN
NUCLEOTIDE BIOSYNTHESIS
PROTEIN MODIFICATION
- DNA rep. initiation
- DNA dep. DNA repl.
- necrosis
- vitamin B2 biosynthesis
- lipopolysaccharide biosynthesis
- D-amino acid catabolism
- N-acetylneuraminate metabolism
- acetyl-CoA metabolism

Gene Group 10:
CATION HOMEOSTASIS
OXYGEN
- regulation of transcription from Pol II promoter
- substrate-bound cell migration, cell
- cytokinesis
- attachment to substrate

…

- isoleucyl-tRNA aminoacylation
- glutamyl-tRNA aminoacylation
- glucose transport
- glucose metabolism
- cytochrome c oxidase biogenesis
- histogenesis and organogenesis
- establishment maintenance of chromatin architecture
- aldehyde metabolism
- proline catabolism
- glycine catabolism
- pyrimidine metabolism
- pyridine nucleotide biosynthesis
- (U,C,G)TP biosynthesis
- glutamate biosynthesis
- physiological processes
- prostaglandin biosynthesis
- response to oxidative stress
- single strand break repair
- sex differentiation
- hormone metabolism
- cartilage condensation
- mechanosensory perception
- cell wall catabolism
- viral replication

Gene Group 11:
LIPID METABOLISM
CHROMATIN
OXYGEN
- actin cytoskeleton organization and biogenesis
- ATP biosynthesis
- ATP synthesis coupled proton transport
- regulation of DNA repl.
- tRNA splicing
- response to stress
- response to oxidative stress
- posttranslational membrane targeting
- protein amino acid ADP-ribosylation
- O-linked glycosylation
- metal ion transport
- metabolism
- protein metabolism
- branched chain family amino acid catabolism
- histidine metabolism
- trehalose metabolism
- cAMP-mediated signaling
- development
- adult somatic muscle development
- spermatid development
- sensory organ development
- olfaction
- peripheral nervous system development
- metabotropic glutamate receptor signaling pathway
- digestion
- bone mineralization
- sex differentiation

Fig. 1. Gene ontology labels found in the SOM gene groups with strong statistical significance abundance ($P < 0.0016$). Uppercase italic letters indicate entire overrepresented categories. Individual items are bulleted. Shaded boxes mark individually related processes whose category was not as a whole overrepresented.

The task of understanding the many effective influences is simplified by the observation that, at a particular time, a gene group generally up- or down-regulates transcription globally. For example, group 11 is a global activator after 1.5 h and a global inhibitor after 2 h. This feature is manifest in the red and green columns of Fig. 2. At the longest time interval (3.5 h), however, groups 4 and 11 are moderate activators of some gene groups while inhibiting others. Despite these exceptions, the general tendency is striking.

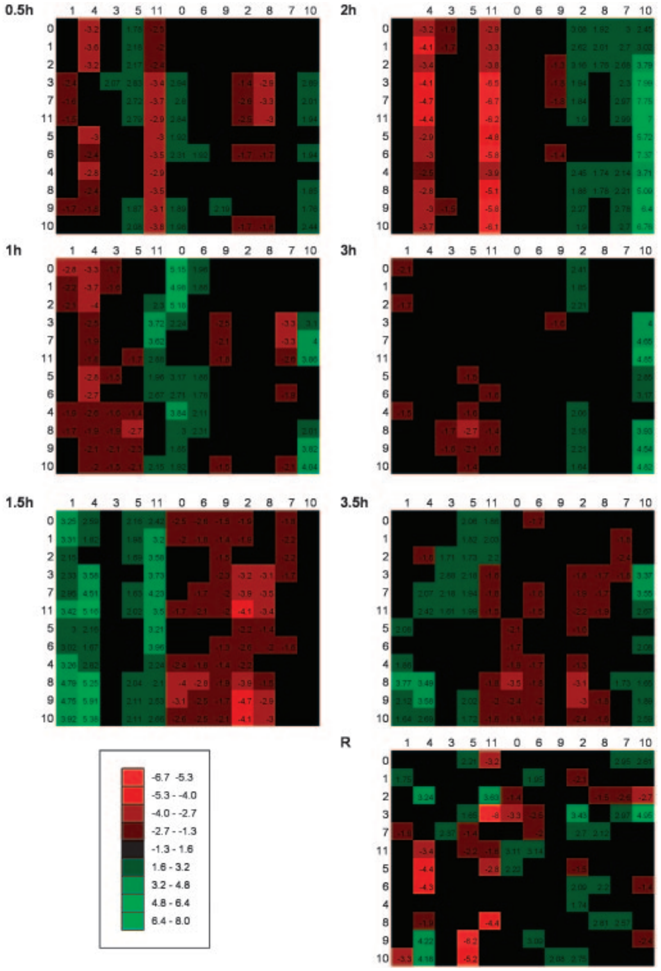


Fig. 2. Effective regulatory influences among the 12 gene groups arranged by time step. The influence of gene group A on gene group B is given in the Bth row, Ath column of each table. Red indicates an inhibitory influence, green an activating one, with brighter values indicating stronger effects. Influence tables are given for 0.5-, 1-, 1.5-, 2-, 3-, and 3.5-h transitions. The order of rows is determined by clustering the gene groups by similarity of input (row values) across all time steps; column order is clustered by output similarity. R indicates a control analysis in which the order of gene expression levels at the second time point was randomized.

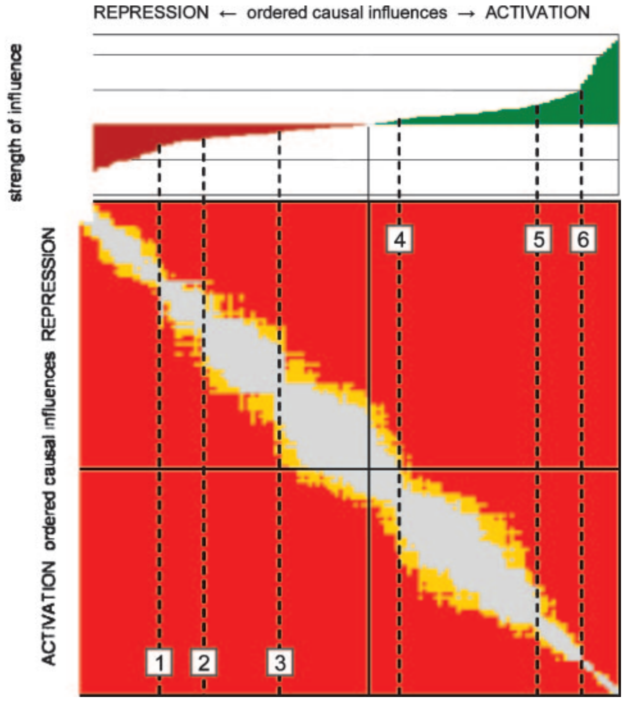


Fig. 3. Pair-wise comparison of gene group regulatory influences. Plotted are the *P* values associated with a pair-wise *t* test comparison of bootstrap distributions for each coefficient across all 144 coefficients: red, $P < 0.05$; yellow, $P < 0.1$; gray, $P > 0.1$. Coefficients are ordered from most inhibitory to most activating. The top plot shows the value of each coefficient; red, inhibition; green, activation; solid line indicates the position of ''0'' influence. Dashed lines demarcate the boundaries of ''influence types,'' groups of coefficients that are mutually statistically indistinguishable, e.g., lines 3 and 4 bound a type characterizable as ''minimal effect.''

A possible explanation for the tendency of groups to have uniform influences on all other groups is that there exist processes in the cell that regulate transcriptional levels globally, and the effect of each gene group on all other groups is determined primarily by its effect on these upstream critical processes. Several categories stand out as possible upstream determinants of global transcription, in particular: transcription, mRNA synthesis, nucleotide metabolism, nucleotide biosynthesis, and aerobic respiration. If any of these act as global activators of transcription, then their category presence will correlate with a global transcription activating influence.

We calculated the mean global activating/inhibiting influence of each gene group over each time interval (see Fig. 4*b*). Gene group 10 attained the greatest mean activating influence (5.59) over 2 h, followed by groups 4 (3.45), 1 (3.35), 11 (3.11) over 1.5 h (see Fig. 4*a*), and 0 (3.05) over 1 h. The correlation between the fraction of genes within a particular group dedicated to a process category and these activating levels was highest for aerobic respiration (correlation $r = 0.471$), followed by spindle/DNA repair ($r = 0.424$), and chromatin ($r = 0.420$). Two of these processes are consistent with the hypothesis of critical processes controlling global transcription; aerobic respiration generates the ATP required for transcription, and chromatin unpacking acetylases will open up major regions of the DNA to transcriptional activation (22, 23).

DNA replication ($r = 0.617$) and chromatin ($r = 0.419$) were the processes most correlated to inhibiting influences. It is not surprising that these two groups were inhibitory in nature. For example, DNA replication is a major ATP sink and is associated with regulated repression of transcription (24), whereas the chromatin category contains deacetylases, which are inhibitory in nature (22, 25).

Interestingly, the presence of the G protein cascade category in gene groups was anticorrelated to both strong activation ($r = -0.371$) and repression ($r = -0.209$); generally, groups containing many G protein-signaling genes never attained a strong regulatory effect. This is reasonable, considering that as critical cell-surface signaling molecules, G proteins are unlikely to function in a concentration-limited manner. Thus, changes in their concentration should not greatly alter the expression of their target genes.

Discussion

The purpose of the megamodule level of description is not microcausal relationships but rather an understanding of effective regulatory influences at the aggregate level. It is possible that a ''master-regulator'' gene, or genes, with the ability to exert genome-wide transcriptional influence (a sort of supertranscription factor) could be present in each of the gene groups in low numbers and thus not identified by our statistical tests. Although we cannot strictly exclude the possibility that the above hypotheses, e.g., ATP-limited transcriptional control, may be *ad hoc* explanations, the sequential regulatory patterns we observe are

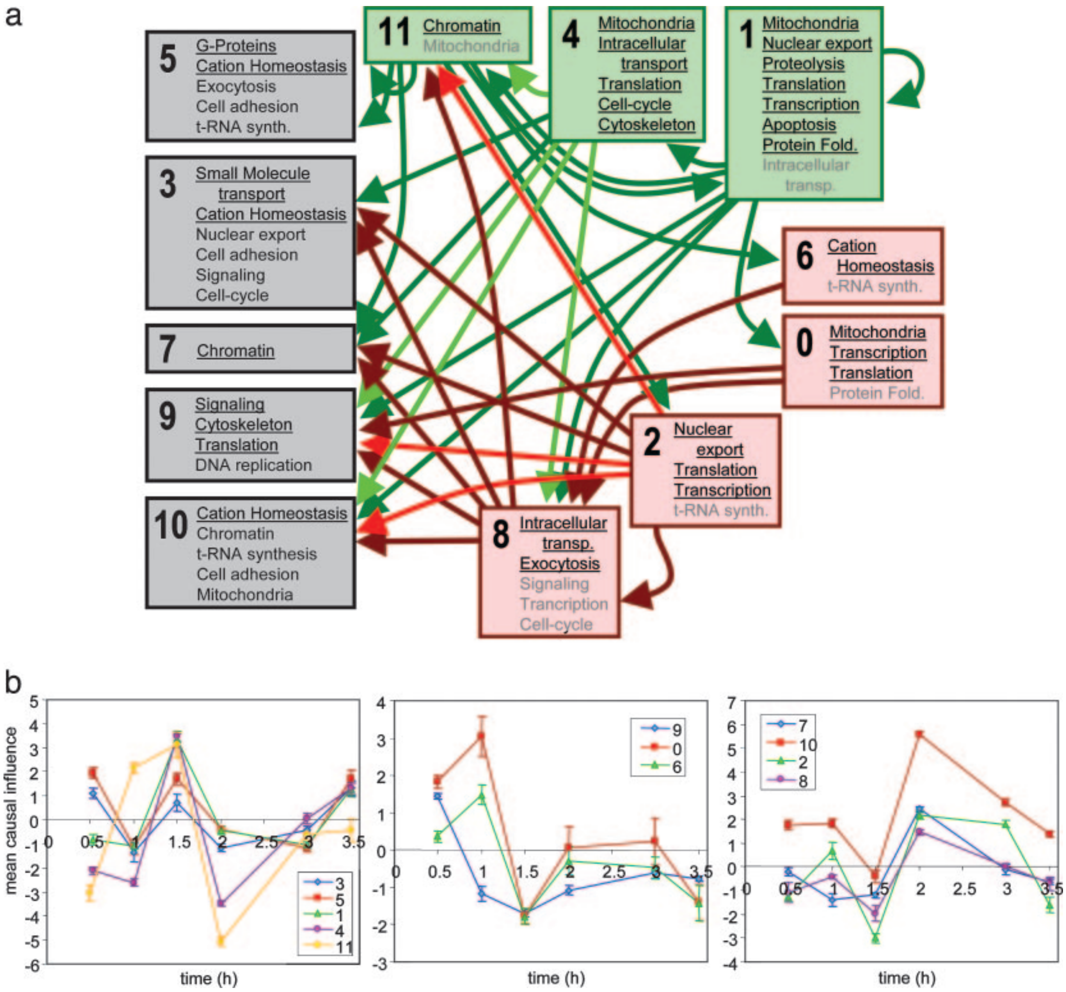


Fig. 4. Cellular regulatory influence diagrams. (*a*) A global comprehensive cellular influence network for the 1.5-h transition. Each gene group is represented by its number, along with those ontology labels that are overrepresented in that gene group. Underlined labels are given when an entire function category is overrepresented, and grayed labels when several individually related functions are each overrepresented, but the category is not. A green arrow connecting one gene group to another indicates activation over 1.5 h, red arrows indicates repression, with brighter arrows indicating stronger effects. (*b*) The mean level of gene group output influences across all 12 target groups is plotted versus the time step. Error bars are the standard error of the mean. Gene groups 9, 0, and 6 can be characterized as ''very early activating'' gene groups. Gene groups 3, 5, 1, 4, and 11 are ''early activating;'' groups 4 and 11 are also ''long-term repressors.'' Groups 7, 10, 2, and 8 are ''long-term activating.''

at play in mouse B cells, either as a direct or a secondary result of the ontology content of each gene group. Moreover, it is unlikely that a group of genes small enough to be statistically anonymous could exert genome-wide influence dominant over the summed influence of the ≈1,000 other genes in its coexpression group. Furthermore, for the hypothesis of ATP production as a critical global determinant of transcriptional levels, two complementary observations provide strong corroboration. First, ATP-consuming behaviors that also inhibit transcription (DNA replication and chromatin remodeling) are correlated to strong inhibitory influence of a group. Second, aerobic respiration, the ATP-generating process, is correlated with ubiquitous transcriptional activation.

Analysis of the time courses of influence of each gene group reveals that the majority of strong activating and inhibiting influences occur over 1.5 or 2 h (see Fig. 4*b*). This observation imposes a limit of 12 to 16 major transitions, or functional steps, that can be made in 24 h (a cell-cycle duration typical of eukaryotes).

It is notable that the strongest activating influences (>2.5) occur almost exclusively in the 1.5-h time interval (see Fig. 4). That there were no immediate (0.5-h) strong activating effects is consistent with the hypothesis that activation is controlled by an intermediate ATP-controlling step (e.g., aerobic respiration). Gene group 10 is the strongest global activator of transcription; yet the aerobic respiration category is not overrepresented in it. However, two upstream processes are ''glucose metabolism'' and ''glucose transport.'' Thus, it is not surprising that the activating influence of gene group 10 is delayed by 0.5 h.

Only one gene group (4) has a statistical overabundance of cell-cycle-regulating genes (see Fig. 1). This group additionally has a differential between strongest activating and inhibiting influences second only to gene group 11 (see Fig. 4*b*). It is possible that the inhibiting-activating-inhibiting temporal profile of group 4 underlies an oscillatory behavior associated with the cell cycle.

The data also supported the notion that when two gene groups exhibit related expression profiles, they are also likely to be under similar input regulation from the other gene groups. Similarity in input vectors was correlated (r = 0.403) to the pair-wise distance between two gene groups in the SOM array (an *a priori* measure of expression profile relatedness) and even more highly correlated (r = 0.650) to similarity in expression profiles (rows in Fig. 2) (see *Materials and Methods*), suggesting quite reasonably that similar regulatory inputs lead to similar expression profiles.

Similarity in regulatory outputs is slightly anticorrelated to similarity in expression profiles (for the expression profile and SOM distance comparisons, $r = -0.103$ and $r = -0.080$, respectively). This result can be explained by a cellular need to maintain transcriptional homeostasis. If all gene groups that are coexpressed similarly activate or inhibit global transcription, the level of transcription would fluctuate wildly in an all-or-none manner. Anticorrelation between expression profiles and functional output means that whenever a particular activating force is turned on, an antithetical inhibiting force will arise to maintain relative homeostasis in global transcriptional levels.

An important aspect of obtaining the set of influences is the possibility of predicting the outcomes of perturbations. We found that the matrix of influences converges after fitting with 26 of 33 available perturbations (see Fig. 5, which is published as supporting information on the PNAS web site). The ability to predict the rest is demonstrated by the correlation of outputs calculated from the matrix and the actual experimental observations which, e.g., for the 1.5-h transition, have a correlation of 0.993 (see Fig. 6, which is published as supporting information on the PNAS web site).

Combining ontology and casual influence analysis allows one to visualize the interaction of functions at the cellular level (see Fig. 4*a*). This model is unique, because it is comprehensive in describing all major cellular regulatory influences that occur over a 1.5-h time step, and because it sets forth dozens of experimentally falsifiable hypotheses. Although this analysis has been performed at the megamodule level, this technique can conceptually be used to infer all transcriptional interactions, because its resolution is limited only by the number of time-step experiments. Practically, however, to infer strictly linear effects of every gene, given the levels of experimental noise in microarray data would require ≈90,000 observations in mammalian systems or 18,000 observations in yeast. Although these numbers appear daunting, rapidly developing high-throughput automation techniques (26, 27) suggest that such an effort might soon be feasible and economical.

We are indebted to E. Perlstein and J. Strader for helpful discussions and critiques and to V. Vishwanath, G. Eichler, and T. Smith for help with the Gene Ontology databases.

1. Lockhart, D. J. & Winzeler, E. A. (2000) *Nature* **405,** 827–836.
2. Iyer, V. R., Eisen, M. B., Ross, D. T., Schuler, G., Moore, T., Lee, J. C., Trent, J. M., Staudt, L. M., Hudson, J., Jr., Boguski, M. S., *et al.* (1999) *Science* **283,** 83–87.
3. Eisen, M. B., Spellman, P. T., Brown, P. O. & Botstein, D. (1998) *Proc. Natl. Acad. Sci. USA* **95,** 14863–14868.
4. Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E., Lander, E. S. & Golub, T. R. (1999) *Proc. Natl. Acad. Sci. USA* **96,** 2907–2912.
5. Tavazoie, S., Hughes, J. D., Campbell, M. J., Cho, R. J. & Church, G. M. (1999) *Nat. Genet.* **22,** 281–285.
6. D'Haeseleer, P., Wen, X., Fuhrman, S. & Somogyi, R. (1999) *Pac. Symp. Biocomput*., 41–52.
7. Friedman, N., Linial, M., Nachman, I. & Pe'er, D. (2000) *J. Comput. Biol.* **7,** 601–620.
8. Yeung, M. K., Tegner, J. & Collins, J. J. (2002) *Proc. Natl. Acad. Sci. USA* **99,** 6163–6168.
9. Tamada, Y., Kim, S., Bannai, H., Imoto, S., Tashiro, K., Kuhara, S. & Miyano, S. (2003) *Bioinformatics* **19**, II227–II236.
10. Holter, N. S., Maritan, A., Cieplak, M., Fedoroff, N. V. & Banavar, J. R. (2001) *Proc. Natl. Acad. Sci. USA* **98,** 1693–1698.
11. Tegner, J., Yeung, M. K., Hasty, J. & Collins, J. J. (2003) *Proc. Natl. Acad. Sci. USA* **100,** 5944–5949.
12. Gardner, T. S., di Bernardo, D., Lorenz, D. & Collins, J. J. (2003) *Science* **301,** 102–105.
13. Reed, J. L., Vo, T. D., Schilling, C. H. & Palsson, B. O. (2003) *Genome Biol.* **4,** R54.1–R54.12.
14. Chen, Y., Kamat, V., Dougherty, E. R., Bittner, M. L., Meltzer, P. S. & Trent, J. M. (2002) *Bioinformatics* **18,** 1207–1215.
15. Stuart, J. M., Segal, E., Koller, D. & Kim, S. K. (2003) *Science* **302,** 249–255.
16. Eichler, G. S., Huang, S. & Ingber, D. E. (2003) *Bioinformatics* **19,** 2321–2322.
17. Kohonen, T. (2001) *Self-Organizing Maps* (Springer, Berlin).
18. Little, J. & Moler, C. (2003) MATLAB (Mathworks, Natick, MA).
19. Gilman, A. G., Simon, M. I., Bourne, H. R., Harris, B. A., Long, R., Ross, E. M., Stull, J. T., Taussig, R., Arkin, A. P., Cobb, M. H., *et al.* (2002) *Nature* **420,** 703–706.
20. Alter, O., Brown, P. O. & Botstein, D. (2000) *Proc. Natl. Acad. Sci. USA* **97,** 10101–10106.
21. Hornquist, M., Hertz, J. & Wahde, M. (2002) *Biosystems* **65,** 147–156.
22. Davie, J. R. & Spencer, V. A. (1999) *J. Cell Biochem.* **Suppl. 32–33,** 141–148.
23. Berger, S. L. (1999) *Curr. Opin. Cell Biol.* **11,** 336–341.
24. Voorma, H. O. (1983) *Horiz. Biochem. Biophys.* **7,** 139–153.
25. Taunton, J., Hassig, C. A. & Schreiber, S. L. (1996) *Science* **272,** 408–411.
26. Heller, M. J. (2002) *Annu. Rev. Biomed. Eng.* **4,** 129–153.
27. King, R. D., Whelan, K. E., Jones, F. M., Reiser, P. G., Bryant, C. H., Muggleton, S. H., Kell, D. B. & Oliver, S. G. (2004) *Nature* **427,** 247–252.