nature genetics

Reverse engineering of regulatory networks in human B cells

Katia Basso[1], Adam A Margolin[2], Gustavo Stolovitzky[3], Ulf Klein[1], Riccardo Dalla-Favera[1,4] & Andrea Califano[2]

Cellular phenotypes are determined by the differential activity of networks linking coregulated genes. Available methods for the reverse engineering of such networks from genome-wide expression profiles have been successful only in the analysis of lower eukaryotes with simple genomes. Using a new method called ARACNe (algorithm for the reconstruction of accurate cellular networks), we report the reconstruction of regulatory networks from expression profiles of human B cells. The results are suggestive a hierarchical, scale-free network, where a few highly interconnected genes (hubs) account for most of the interactions. Validation of the network against available data led to the identification of *MYC* as a major hub, which controls a network comprising known target genes as well as new ones, which were biochemically validated. The newly identified *MYC* targets include some major hubs. This approach can be generally useful for the analysis of normal and pathologic networks in mammalian cells.

Cell phenotypes are determined by the concerted activity of thousands of genes and their products. This activity is coordinated by a complex network that regulates the expression of genes controlling common functions, such as the formation of a transcriptional complex or the availability of a signaling pathway. Understanding this organization is crucial to elucidate normal cell physiology as well as to dissect complex pathologic phenotypes. Studies in lower organisms indicate that the structure of both protein-protein interaction and metabolic networks is of a hierarchical scale-free nature[1,2], characterized by an inverse relationship between the number of nodes and their connectivity (scale-free) and by a preferential interaction among highly connected genes, called hubs (hierarchical). Although scale-free networks may represent a common blueprint for all cellular constituents, evidence of scale-free topology in higher-order eukaryotic cells is currently limited to coexpression networks[3,4], which tend to identify entire subpathways rather than individual interactions. Identifying the organizational network of eukaryotic cells is still a key goal in understanding cell physiology and disease.

Genome-wide clustering of gene-expression profiles has provided an initial step towards the elucidation of cellular networks. But the organization of gene-expression profile data into functionally meaningful genetic information has proven difficult and so far has fallen short of uncovering the intricate structure of cellular interactions. This challenge, called network reverse engineering or deconvolution, has led to an entirely new class of methods aimed at producing high-fidelity representations of cellular networks as graphs, where nodes represent genes and edges between them represent interactions, either between the encoded proteins or between the encoded proteins and the genes (we use 'genetic interaction' to refer to both types of mechanisms). Available methods fall into four broad categories: optimization methods[5–7], which maximize a scoring function over alternative network models; regression techniques[8,9], which fit the data to *a priori* models; integrative bioinformatics approaches[10], which combine data from a number of independent experimental clues; and statistical methods[11,12], which rely on a variety of measures of pairwise gene-expression correlation. All these methods suffer from several limitations, including exponential complexity in the local network connectivity, unrealistic assumptions about the network structure, lack of integrative genomic data for higher eukaryotes, overfitting and underconstrained regression analysis. Additionally, coexpression methods identify pairs of genes or proteins that are functionally related rather than involved in direct physical interactions, resulting in exceedingly large false positive rates. As a result, these methods have been successful only in the study of organisms with relatively simple genomes, such as *Saccharomyces cerevisiae*, or have produced networks with only a handful of interactions[13]. No method is currently available for the genome-wide reverse engineering of mammalian cellular networks.

Here we present the successful reverse engineering of gene-expression profile data from human B cells. Our study is based on ARACNe (algorithm for the reconstruction of accurate cellular networks), a new approach for the reverse engineering of cellular networks from microarray expression profiles. ARACNe first identifies statistically significant gene-gene coregulation by mutual information, an information-theoretic measure of relatedness. It then eliminates indirect relationships, in which two genes are coregulated through one or more intermediaries, by applying a well-known staple of data

[1]Institute for Cancer Genetics and [2]Joint Centers for Systems Biology, 1300 St. Nicholas Avenue, Room 912, New York, New York 10032, USA. [3]IBM T.J. Watson Research Center, Yorktown Heights, New York, New York 10598, USA. [4]Departments of Pathology and Genetics & Development, Columbia University, New York, New York 10032, USA. Correspondence should be addressed to A.C. (califano@c2b2.columbia.edu).

Published online 20 March 2005; doi:10.1038/ng1532

transmission theory, the 'data processing inequality' (DPI), which had not been previously applied in this context. Hence, relationships included in the final reconstructed network have a high probability of representing either direct regulatory interactions or interactions mediated by post-transcriptional modifiers that are undetectable from gene-expression profiles.

An essential requirement of any reverse engineering method is the availability of a large set of gene-expression profile data representative of perturbations of the cellular systems, leading to the analysis of a broad range of cellular states and associated gene-expression levels. This is necessary because genetic interactions are best inferred when the genes explore a substantial dynamical range. Traditionally, this has been achieved by systematic perturbations in simple organisms (*e.g.*, by large-scale gene knockouts or exogenous constraints[14]), which are not easily obtained in more complex cellular systems. We show here that an equivalent dynamic richness can be efficiently achieved by assembling a considerable number of naturally occurring and experimentally generated phenotypic variations of a given cell type. To this end, we applied ARACNe to genome-wide expression profiles from a panel of 336 B cell phenotypes representative of a wide selection of normal, transformed and experimentally manipulated human B cells related to the germinal center, a structure in which B cells are selected on the basis of their ability to produce antibodies with high affinity for the antigen and from which many types of B cell lymphoma are derived[15].

ARACNe reconstructed a network suggestive of a hierarchical, scale-free organization, with a power-law relationship between the number of genes and their connectivity[16]. In the inferred network, a relatively small number of highly connected genes interact with most other genes in the cell, either directly or hierarchically, through other highly connected subhubs. The proto-oncogene *MYC* emerged as one of the largest hubs in the network. In-depth analysis of this gene uncovered a hierarchically organized subnetwork, which we used to validate the *in silico* predictions against an extensive number of previously identified targets of the MYC transcription factor. ARACNe recapitulated known *MYC* target genes and identified new candidate targets, which we then validated by biochemical analysis. The success of ARACNe in the genome-wide identification of gene networks indicates that the method can be useful in the dissection of both normal and pathologic mammalian phenotypes.

RESULTS

Validation on a synthetic network

As a first test of ARACNe's performance, we compared it to Bayesian networks[17], which are among the best reverse engineering methods available[5,6,18], by analyzing a synthetic genetic network model[19] that is part of a project aimed at integrating the songbird brain[20,21]. ARACNe correctly identified more interactions than Bayesian networks (13 versus 11 true positives) and incorrectly identified substantially fewer relationships (2 versus 11 false positives; **Fig. 1a**). We then systematically compared the sensitivity (percentage of correctly inferred true interactions) and precision (percentage of correct interactions among all inferred ones) of the two methods as a function of the number of available synthetic samples (ranging between 100 and 2,000; **Fig. 1b**). Overall, ARACNe was comparable to Bayesian networks in sensitivity and largely superior in precision.

The B cell network has hierarchical scale-free behavior

We then used ARACNe to deconvolute cellular networks from a set of 336 expression profiles representative of perturbations of B cell phenotypes, including normal B cell subpopulations, various subtypes of B cell tumors and experimentally manipulated B cells (**Fig. 2** and **Supplementary Table 1** online). Normal cells included resting pre–germinal center naive B cells, proliferating germinal center B cells (centroblasts and centrocytes) and post–germinal center memory B cells. Transformed cells were represented by panels of cell lines and biopsies representative of the more than ten subtypes of B cell malignancies. Experimentally manipulated B cells included cell lines that were treated *in vitro* to induce specific signal transduction pathways or engineered for the conditional expression of several transcription factors.

ARACNe inferred a network with ~129,000 interactions (**Supplementary Table 2** online), which may include those specific to normal and malignant B cells. Given the relatively small number of known phenotypes derived from normal mature B cells, an

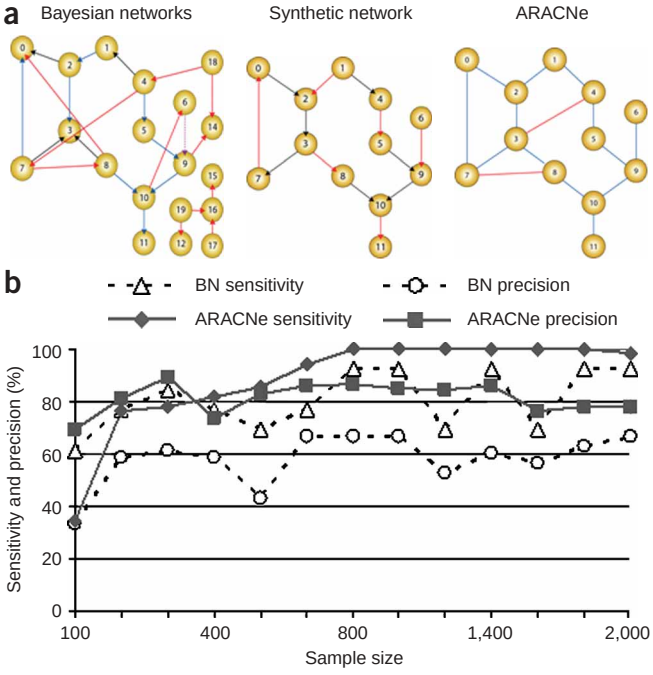


Figure 1 Comparison of the performance of Bayesian networks and ARACNe algorithms in the analysis of a synthetic genetic network. (**a**) Middle panel, the synthetic genetic network proposed to model the songbird brain[19]. The model has 12 interconnected nodes and 7 disconnected nodes (not shown), including a cyclic loop formed by nodes 0, 2, 3 and 7. Interactions between two genes (nodes) are shown as arrows (edges). Black and red arrows represent upregulation and downregulation, respectively. Samples for the synthetic microarrays were obtained from data freely provided by the authors, which was generated from 2,000 simulated samples for each gene in the model. Left panel, results of the reverse engineering using Bayesian networks on the full data set. Correct edges with correct or incorrect direction are shown by blue and black arrows, respectively. Incorrect and missing edges are represented by red and dotted purple arrows, respectively. Although edge direction is shown, it was not used for the method comparison. Right panel, results of the reverse engineering on the full data set using ARACNe. ARACNe identifies only a slightly more correct edges than Bayesian networks (13 versus 11) but performs substantially better in the assignment of incorrect edges. Correct, incorrect and missing edges are represented by blue, red and dotted purple lines, respectively. (**b**) Performance of ARACNe and Bayesian networks (BN) as a function of the sample size. Sensitivity and precision are plotted as a function of the number of samples used for the analysis. Results for ARACNe were produced using a Gaussian Kernel mutual information estimator with an error tolerance $\varepsilon = 0.1$ for the DPI and a *P*-value threshold $P_0 = 0.1$.

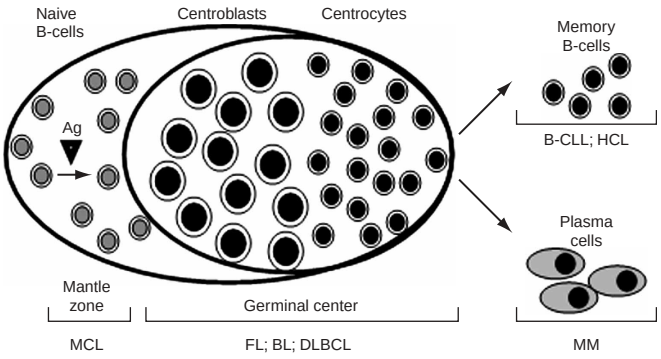


Figure 2 Schematic representation of the germinal center reaction and related normal and malignant B cell phenotypes. The scheme provides a biologic framework to identify the various B cell populations used in this study. When naive B cells encounter the antigen (Ag) in the secondary lymphoid organs, they are stimulated to proliferate and form specific histological structures called germinal centers[15]. The germinal center includes two B cell populations: centroblasts and centrocytes. At the end of the germinal center reaction, the subset of centrocytes, which acquired the ability to express high-affinity immunoglobulin receptors, is positively selected and further differentiates to memory or plasma cells. Malignant transformation can affect each of the normal populations, leading to different types of lymphoma and leukemia. B-CLL, B cell chronic lymphocytic leukemia; BL, Burkitt lymphoma; DLBCL, diffuse large B cell lymphoma; FL, follicular lymphoma; HCL, hairy cell leukemia; MCL, mantle cell lymphoma; MM, multiple myeloma.

equivalent network for normal B cells could not be produced with acceptable accuracy. Because the inferred network is too complex to be shown in its entirety, we summarized its global connectivity properties (**Fig. 3**). The results show a power-law tail in the relationship between the number of genes, n, in the network and their number of interactions, k. This tail extends over slightly more than one order of magnitude. This is suggestive of a scale-free underlying network structure[16]. The deviation of the curve at low values of connectivity ($k < 12$) from the power-law extrapolated by linear fit from larger connectivity values ($k \geq 12$) is probably a consequence of the limited number of available genes. In fact, only ~6,000 genes on the microarray have enough dynamic range in our expression profiles to allow the inference of at least one interaction. According to the theoretical curve, however, ~100,000 genes would be needed to assess the power-law behavior down to $k = 1$.

Construction of the *MYC* subnetwork

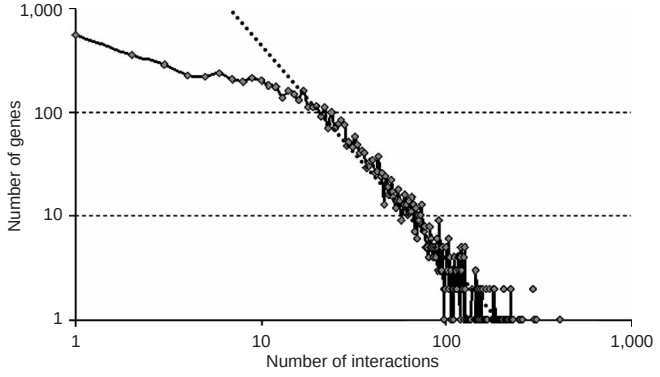
As expected given the scale-free nature of the network, a small percentage of genes accounts for most of the connections. We defined as major hubs the largest 5% of hubs in the network; they collectively participate in ~50,000 interactions, almost as many as the remaining 95% of genes combined. Using the Gene Ontology classification and the GOMiner tool[22], we analyzed the biological processes affected by the top 5% of hubs and identified substantial enrichment for genes involved in essential cellular processes such as cell cycle regulation, protein synthesis and catabolism, RNA processing and metabolism, and transcription (**Supplementary Table 3** online).

The ARACNe-generated network could be used to identify the subnetworks associated with any gene of interest. We chose to study the proto-oncogene *MYC*, which emerged as one of the top 5% of largest cellular hubs. Because *MYC* is extensively characterized as a transcription factor, we could compare the ARACNe-inferred interactions with those previously identified by biochemical methods. Because ARACNe should be able to identify genes both upstream and downstream of *MYC*, we expected the subset of genes directly connected to *MYC* by an edge in the network (first neighbors) to be significantly enriched in transcriptionally regulated targets. To identify the *MYC* subnetwork, we selected from the complete network only those genes that had statistically significant mutual information with any of the available probe sets for *MYC* ($P < 10^{-7}$). The inferred subnetwork structure included 2,063 genes, 56 of which were directly connected to *MYC* (**Fig. 4** and **Supplementary Table 4** online). Previously reported *MYC* target genes were identified in the subnetwork using as reference the MYC database[23]. The network has several limitations: (i) edges lack directionality (*i.e.*, they do not indicate which gene is 'upstream' or 'downstream'); (ii) some direct connections may involve unknown intermediates, as not all biochemical species participating in cellular interactions are represented on the microarray (*e.g.*, missing probes or post-transcriptionally modified intermediates); (iii) some direct interactions may have been incorrectly removed by the DPI (*e.g.*, any regulatory loop formed by three interacting genes would result in the weakest of the three interactions being removed, unless its strength is within 15% of that of the intermediate interaction).

The *in silico* network is enriched in known *MYC* targets

Twenty-nine of the 56 (51.8%) predicted first neighbors were previously reported (or biochemically validated here) as *MYC* targets. This represents a highly significant enrichment ($P = 1.8 \times 10^{-22}$ by

Figure 3 Identification of key regulatory hubs in the human B cell network. Three hundred and thirty-six expression profiles from different B cell phenotypes were used by ARACNe to deconvolute cellular networks. The reconstructed network includes ~129,000 interactions, and its connectivity properties are graphically represented. The plot shows the distribution of nodes with a specific number of incident edges (degree of connectivity) in log-log scale using a tolerance $\varepsilon = 0.15$. The dotted line indicates a theoretical power-law extrapolated from the experimental data for connectivity greater than 12 (*i.e.*, genes with more than 12 interactions). The difference between the experimental and the theoretical plots is due to a saturation effect for low connectivity values: the theoretical curve would require ~100,000 genes with a single interaction. Because the experimental data have a power-law–like behavior from a connectivity of ~12, an assessment over two orders of magnitudes would require either significant experimental data for genes with up to 1,200 interactions or a much larger number of genes to avoid the saturation effect for low connectivity values.



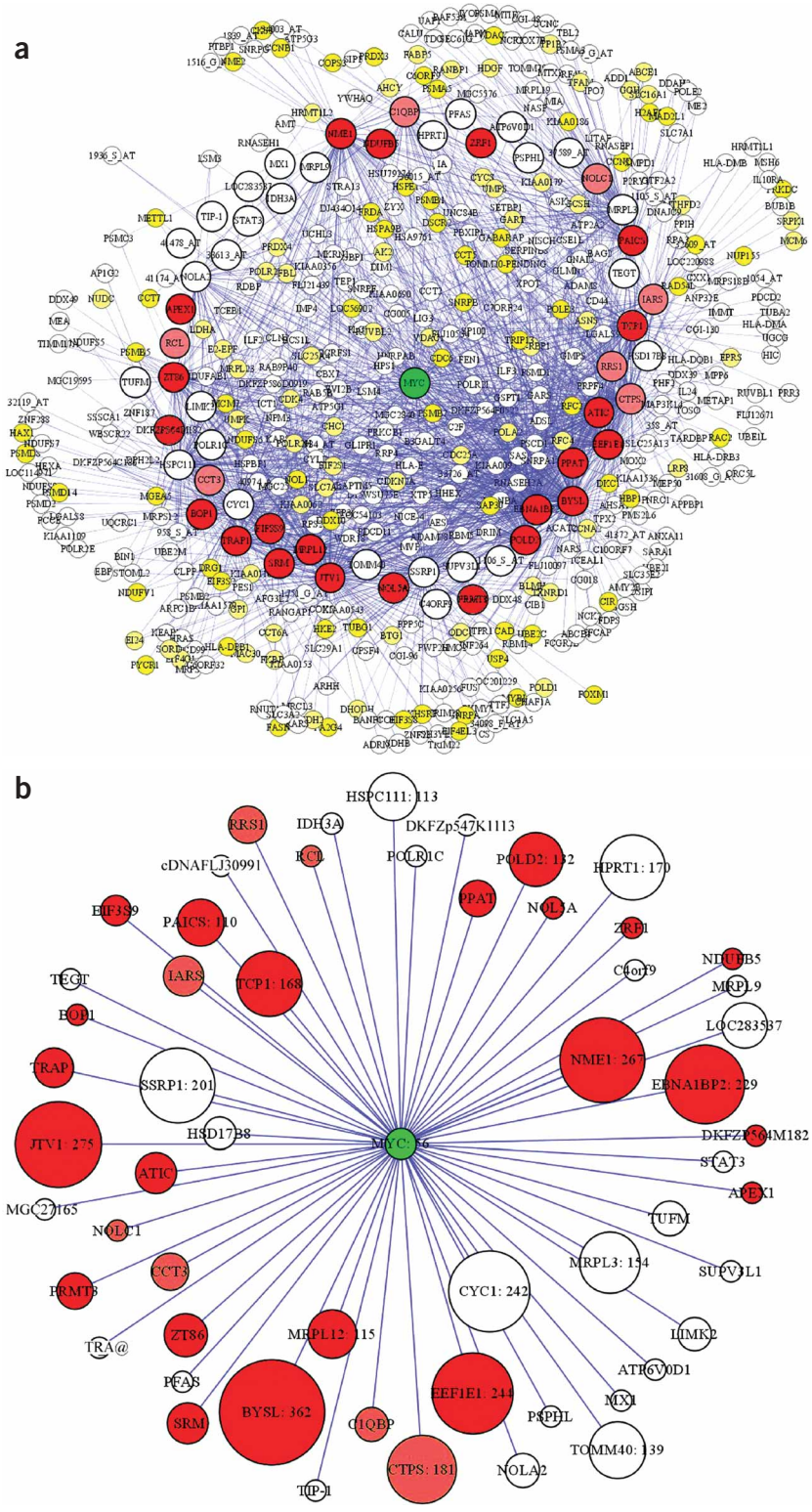


Figure 4 The *MYC* subnetwork. (**a**) A *MYC*-specific subnetwork was obtained by including all the genes that have $P < 10^{-7}$ based on their pairwise mutual information with *MYC*. The faster bin-counting estimator was used with an error tolerance $\varepsilon = 0.15$. The *MYC* subnetwork includes 56 genes directly connected to *MYC* (first neighbors; represented by larger circles) and 2,007 genes connected through an intermediate (second neighbors). For representation purposes, only the first 500 genes are shown, including all 56 first neighbors and the 444 most statistically significant second neighbors. Red or pink nodes represent first neighbor target genes for which ChIP data is available or not available, respectively; yellow and light yellow nodes represent second neighbor target genes for which ChIP data is available or not available, respectively; *MYC* is shown in green; white nodes represent genes for which no *MYC*-related information is available. The complete list of genes, including gene symbol, Affymetrix ID and LocusLink ID, is given in **Supplementary Table 4** online. (**b**) The first neighbors of the *MYC* subnetwork. The size of each circle is proportional to the number of the gene interactions. For hubs with more than 100 interactions, the exact number of first neighbors is shown beside the gene symbol.

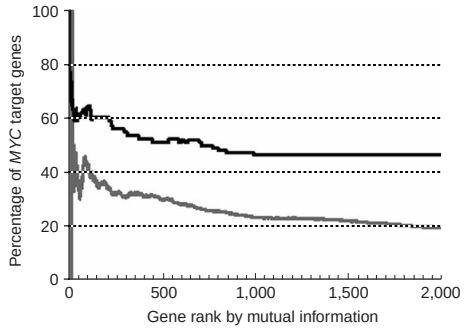


Figure 5 Distribution of known *MYC* target genes among first and second neighbors in the *MYC* subnetwork. The genes identified by ARACNe as first or second neighbors of *MYC* were ranked according to their pairwise mutual information with *MYC*. The plot shows the percentage of known *MYC* target genes among first (black line) and second (gray line) neighbors in the first *n* top genes, ranked by mutual information value, where *n* is shown on the *x* axis. For example, for a value of $n = 100$ on the *x* axis, the percentage of targets among the 100 genes with highest mutual information with *MYC* is shown. Results include the ChIP validation assays reported here.

χ^2 test) with respect to the expected 11% of *MYC* targets among randomly selected genes[24]. In addition, known *MYC* target genes were significantly more enriched among first neighbors than second neighbors (51.8% versus 19.4%; $P = 1.4 \times 10^{-9}$), indicating that ARACNe is effective at separating direct regulatory interactions from indirect ones (**Fig. 5**). Moreover, 37.5% of first neighbors (as opposed to 12% of second neighbors) were also validated by chromatin immunoprecipitation (ChIP) assay, either previously or as part of this study. In total, 419 of the 2,063 (20%) genes in the predicted *MYC* subnetwork were known *MYC* targets. Some *bona fide* direct interactions may be identified by ARACNe as second neighbors for two reasons: (i) if the gene is involved in a three-gene loop and the edge representing the interaction with *MYC* is the weakest, it is removed by the DPI; and (ii) if the interaction between the two genes requires a cofactor expressed only in a subset of the data, mutual information may be too low to be detected. Overall, we identified ~40% of the genes reported in the MYC database and validated in human B cells.

Candidate new *MYC* targets are biochemically validated

We validated new candidate *MYC* targets biochemically by ChIP assay to show direct binding of *MYC* to their promoter regions *in vivo*. We identified the 34 first neighbors present among the top 100 genes ranked based on mutual information and examined 2 kb of genomic DNA upstream and downstream from the transcription initiation site for the presence of canonical MYC binding sites (E boxes, CACGTG; **Supplementary Table 5** online). After eliminating 9 genes as previously reported *MYC* direct targets, we selected a representative set of 12 genes for validation by ChIP in the Ramos B cell lymphoma line. A large fraction (11 of 12) of the tested genomic sequences could be immunoprecipitated by specific MYC antibodies but not by control antibodies (**Fig. 6**), indicating that MYC was bound to these regions *in vivo*. Among them there are genes associated with programs known to be affected by MYC[23,25,26], such as purine biosynthesis (ATIC), folding of newly translated protein (TCP1), mitochondrial ribosomal protein (MRPL12), ribosomal function and regulation of protein synthesis (HRMT1L3), and genes whose function has not yet been fully investigated (ZRF1, BYSL, RSL1D1). Taken together, these results indicate that ARACNe can identify *MYC* targets with >90% accuracy.

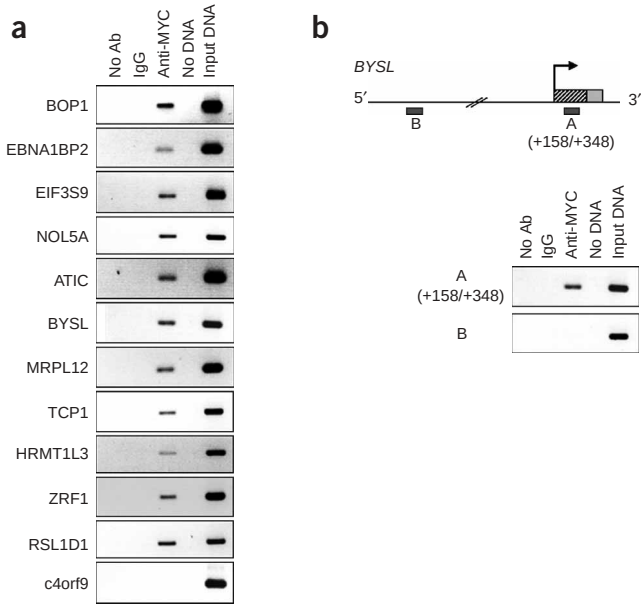


Figure 6 Identification of new direct *MYC* targets by ChIP analysis. (**a**) ChIP was done on the regulatory regions of 12 genes of the *MYC* subnetwork. Eleven of the 12 tested genomic sequences were immunoprecipitated in the Ramos cell line using antibodies to MYC but not using irrelevant antibodies (IgG). Total chromatin before immunoprecipitation (input DNA) was used as positive control for PCR. (**b**) Schematic representation of the localization of the regions analyzed by ChIP in *BYSL*. The investigated E box is located in the noncoding region of the first exon. A set of primers was designed to detect a control region (B) ~19 kb upstream of the initiation site.

MYC has a hierarchical control structure

The identification of the *MYC* subnetwork allowed further investigation of its extended connections. Seventeen of 56 *MYC* first neighbors (30%) are themselves very large genetic hubs, with more than 100 first neighbors in the complete network (range 106–362; **Fig. 4b**). By comparison, only a small minority (<1.4%) of the genes represented in the HG-U95Av2 array have more than 100 first neighbors. This is suggestive of a hierarchical structure of the network, such that *MYC* can modulate a substantial percentage of all genes in the cell through a relatively small number of highly connected subhubs.

Analysis of the genes representing *MYC* larger subhubs using the Gene Ontology categories confirmed the involvement of *MYC* in multiple cellular processes (**Supplementary Table 6** online). Notably, *MYC* directly controls *BYSL*, a gene whose function is poorly characterized but which is the largest hub in the whole B cell network.

DISCUSSION

A key result of this study is the ability to infer genetic interactions on a genome-wide scale from gene-expression profiles of mammalian cells. This objective has so far been elusive except for few instances involving limited number of genes, as shown by the study on the early development of spinal cord in mice[27], which elucidated interactions between 112 genes.

In general, mammalian networks have presented a formidable algorithmic challenge for most optimization- and regression-based algorithms, and integrative approaches are not yet fully applicable

given the very scattered nature of the available mammalian cell information. For instance, although the analysis of conservation of coexpression patterns across human, fly, worm and yeast expression profiles[28] identified key genetic modules common to all of these organisms with reasonable accuracy, it produced $\leq$10% accuracy in the dissection of human cellular networks in isolation. Similarly, a genome-wide study of *cis*-acting elements across coexpressed genes in mammalian cell lines produced only nine interactions among six genes[13]. The success of ARACNe, on the other hand, is probably a direct result of its inherent simplicity, low algorithmic complexity and lack of empirical rules ('*ad hoc* heuristics') such as adding penalties to more complex networks. Consistent with these observations, the comparative analysis (**Fig. 1**) of ARACNe versus Bayesian networks shows that ARACNe offers substantially higher precision at an equal or better sensitivity, even in the analysis of relatively simple networks. ARACNe differs fundamentally from coexpression networks[28] because it explicitly discriminates between direct and indirect interactions (the latter being the majority) resulting in very few false positives. Coexpression networks create an interaction for any pair of genes whose expression is related in a statistical significant way. ARACNe, instead, analyzes all possible alternative paths between two genes and then introduces a direct connection in the network only if this is the most likely interaction among all the paths. Furthermore, as opposed to other methods[28], ARACNe is useful for the dissection of networks that are organism- and tissue-specific, thus producing highly specific interactions.

The results from the synthetic data analysis (**Fig. 1**) were further confirmed by the high success rate for the validation of *MYC* targets in human B lymphocytes using real gene-expression profile data. In particular, the reported precision on the synthetic networks (80–90% at $\sim$340 data points) is comparable to the success rate of the biochemical validation step on the *MYC* putative targets ($>$90%). Moreover, the percentage of *bona fide* interactions among first neighbors will probably increase as additional upstream regulators and downstream targets of *MYC* are identified among the untested genes. Such high precision is a particularly desirable feature of a reverse engineering algorithm, suggesting that the method could be applied to identify candidate interactions without the need for an extensive set of biological validations, which are time- and resource-consuming.

A second result of this study is that gene-expression profile data for systematic experimental perturbations, such as gene inactivations, can be successfully substituted by expression profiles for a wide variety of naturally occurring cellular phenotypes, such as the ones represented by normal and transformed B cells. Using such data implies that the expression of many genes changes over relatively broad dynamic ranges, allowing many regulatory constraints to be identified by algorithmic means. This approach is particularly relevant for studies involving mammalian cells, in which experimental gene perturbations are more technically challenging and time-consuming than in lower organisms. Its systematic application could make reverse engineering techniques broadly applicable to a variety of existing expression profile data for human and mouse tissue.

Substantial evidence indicates that the structure of both protein-protein interactions and metabolic networks in lower organisms is of a scale-free nature[1,2], meaning that the network topology is dominated by a few highly connected hubs and an increasing number of less-connected nodes. Although hierarchical, scale-free networks may represent a common blueprint for all cellular constituents, evidence for this has not been obtained for higher-order eukaryotic cells. ARACNe's ability to reconstruct networks with a broad range of connectivity (more than 400 interactions) could be used to further investigate this hypothesis. **Figure 3** shows a power-law tail in the log-log connectivity plot, spanning slightly more than one order of magnitude, with a saturation effect for connectivity ranges below 12. Similar saturation effects are reported for other scale-free networks when the maximum connectivity range is below 1,000 (ref. 29). The power-law should ideally extend over two or more orders of magnitudes, but these results nevertheless support the hypothesis that mammalian networks are scale-free. Furthermore, the marked enrichment of large genetic hubs among first neighbors of other large hubs and the partial overlap of their first neighbors, as observed for *MYC*, support the notion of a hierarchical control mechanism, providing both redundancy and finegrain combinatorial control of the cellular genetic programs by a handful of genes. The main functional features of hierarchical scale-free networks are their robustness (*i.e.*, the high redundancy of their pathways), their modularity and their high degree of error tolerance. These properties have important implications for the ability of cells to react to physiologic stimuli and to resist mutations. In addition, the understanding that cellular phenotypes are controlled modularly, by a relatively small number of key genetic hubs, should lead to efforts to identify them as key biological determinants and, possibly, therapeutic targets in disease.

We chose to investigate the *MYC* subnetwork because the large amount of available data might allow a validation of ARACNe's power and limitations. The generated network includes $\sim$40% of previously identified *MYC* target genes in human B cells. There are several reasons for the absence in the network of previously reported *MYC* target genes. First, ARACNe cannot produce reliable predictions for genes that are expressed at very low levels, given the substantial impact of measurement noise. Second, the stringency used in our analysis may hinder the detection of genes that have lower mutual information, including those genes for which *MYC* is not the principal regulator. Third, a substantial number of genes are not represented in the microarray used. Nonetheless, ARACNe was able both to reproduce a large fraction of the data on *MYC* targets that were collected over a decade of research using traditional molecular biology methods and to identify a substantial number of new targets.

The results generated by ARACNe provide a wealth of information on *MYC* target genes that require further analysis exceeding the scope of this study. Although the idea that *MYC* is a major hub was perhaps expected, on the basis of evidence that *MYC* can bind to the promoter region of a large number of genes[24], the hierarchical structure of the *MYC*-dependent network has new and important conceptual implications for the function of this gene. This structure suggests that the large number of targets may increase further and be 'modularized' by the use of subhubs. Most of these major subhubs do not seem to encode DNA binding transcription factors, suggesting that *MYC* may function primarily by inducing or suppressing the expression of molecules, such as transcriptional cofactors or kinases, that can indirectly control transcription. Overall, the structure of the network indicates that the subhubs are crucial mediators of *MYC* function. Therefore, future studies of *MYC* should consider these subhubs as priority targets.

One of the most notable aspects of the *MYC* network is that one of the *MYC* subhubs and a new biochemically validated *MYC* target is *BYSL*, which is the most connected gene in the entire cellular network generated by ARACNe. *BYSL* is a highly evolutionarily conserved gene whose function is poorly understood[30]. Studies in human cells have suggested that it has a cytoplasmic localization and a role in cell adhesion[31]; the yeast *BYSL* homolog, *ENP1*, is nuclear and has a putative role in ribosomal RNA splicing and ribosome biogenesis[32]. *Drosophila BYSL* (*bys*) has nuclear localization, with a pattern of

expression mirroring that of a number of *MYC* target genes, and a suggested role in cell growth[33]. A preliminary analysis of the *BYSL* subnetwork suggests that many genes involved in nucleic acid metabolism, cell proliferation and ribosome biogenesis are represented (**Supplementary Table 6**). Although additional studies are needed to elucidate its function, the high number of connections of *BYSL* indicates that the encoded protein may be an important cellular molecule and a critical effector of MYC function.

The approach and results shown here have broad general applicability. First, the data set generated in B cells can be used to identify the network of any gene whose range of expression is sufficiently broad in these cells, leading to accumulation of reliable data that usually requires extensive experimental studies. Second, the network can be used as a basis to identify changes in its structure in other tissues and in specific subtypes of B cell malignancy. Third, the efficiency of ARACNe in identifying cellular connections suggests that it can be readily applied to available gene-expression profile data sets, provided that they have adequate size and complexity.

METHODS

Degree of connectivity. We define the degree of connectivity of a node (gene), k, as the number of nearest neighbors of that gene in the adirected graph that represent the genetic interaction network. This is equal to the sum of the in-degrees (inbound edges) and out-degrees (outbound edges) of a node. The connectivity structure of a network can be summarized by plotting the number of genes having degree of connectivity k against k in log-log scale. A decreasing linear dependency in this plot indicates that the network has a scale-free structure, associated with a corresponding power-law: $n(k) \propto k^{-\gamma}$. This corresponds to a linear relationship in log-log space.

Mutual information. Mutual information for a pair of discrete random variables, x and y, is defined as $I(x,y) = S(x) + S(y) - S(x,y)$, where $S(t)$ is the entropy of an arbitrary variable t. Entropy for a discrete variable is defined as the average of the log probability of its states:

$$S(t) = \langle \log p(t_i) \rangle = \sum_i p(t_i) \log p(t_i),$$

where $p(t_i) = \Pr(t = t_i)$ is the probability associated with each discrete state or value of the variable. If the variable is continuous, the entropy is replaced by the differential entropy, which has the same definition as $S(t)$ in the preceding equation but where the summation is replaced by an integral and the discrete distribution is replaced by a probability density. To estimate the entropy, we use the property that mutual information is invariant under any invertible reparameterization of either x or y. This can be expressed as $I(x' = f_1(x), y' = f_2(y)) = I(x, y)$, with both f_1 and f_2 being invertible. Here, we reparameterize the data using a rank transformation that projects the N_m measurements for each gene into equally spaced real numbers in the interval [0,1], preserving their original order. This transformation is also called copula[34], and it has the advantage of transforming the probability density of the individual variables into a constant, $p(x') = p(y') = 1$. Under this transformation, both $S(x')$ and $S(y')$ become constant and equal to zero. As a result, only $S(x',y')$ must be estimated. For the synthetic analysis, this is done using a Gaussian Kernel estimator[35] where

$$I(x', y') = \frac{1}{N_m} \sum_i \log \left[\frac{p(x_i, y_i)}{p(x_i)p(y_i)} \right].$$

Here, $p(x_i)$, $p(y_i)$ and $p(x_i,y_i)$ are defined as

$$p(x_i) = \frac{1}{\sqrt{2\pi} N d_1} \sum_j e^{-\frac{[x_i - x_j]^2}{2d_1^2}}; p(x_i, y_i) = \frac{1}{2\pi N d_2^2} \sum_j e^{-\frac{[x_i - x_j]^2 + [y_i - y_j]^2}{2d_1^2}}.$$

We obtained the optimal values of the smoothing parameters d_1 and d_2 from Monte Carlo simulations, using a wide range of bivariate normal probability densities. For the large set of B cell expression profiles, we used a slightly less accurate but much more computationally efficient approximation, which estimates $p(x',y')$ by averaging data counts over overlapping windows of size $2n \times 2n$ points located at each discrete position (x',y') that are integer multiples of n. We used a value of $n = N_m/6$ on the basis of the optimization of the error on the mutual information estimate from analytical probability densities. The faster method, albeit slightly less accurate than the Gaussian Kernel, still compares favorably with Bayesian Networks in terms of sensitivity and precision.

Statistical threshold for mutual information. For each value of N_m in the synthetic data analysis, we obtained the P value associated with a given value of mutual information in the null hypothesis by Monte Carlo simulation using 10,000 iterations. The null hypothesis corresponds to pairs of nodes that are disconnected from the network and from each other, such as nodes 12 to 19 in the model. These follow a random-walk dynamic, in the range [1,100], with a noise term drawn from a uniform probability density over the interval [−10,10], as previously reported[19]. For the *MYC* analysis, we computed the P value by Monte Carlo simulations using one million iterations. Because a null-hypothesis dynamical model is not available, it is defined as a pair of existing genes whose values are randomly shuffled at each iteration with respect to the microarray profile in which they were observed.

DPI. First we define two genes, x and y, as indirectly interacting through a third gene, z, if the conditional mutual information $I(x,y|z)$ is equal to zero. The two genes are directly interacting if no such third gene exists, implying that there is direct transfer of information between them. The DPI asserts that if both (x,y) and (y,z) are directly interacting, and (x,z) are indirectly interacting through y, then $I(x,z) \leq I(x,y)$ and $I(x,z) \leq I(y,z)$. This inequality is not symmetric, meaning that there may be situations where the triangle inequality is satisfied but x and z may be directly interacting. As a result, by applying the DPI to discard indirect interactions (*i.e.*, (x, z) relationships for which the inequality is satisfied), we may be discarding some direct interactions as well. These are of two kinds: (i) cyclic or acyclic loops with exactly three genes and (ii) sets of three genes whose information exchange is not completely captured by the pairwise marginals. A typical example of the latter would be the Boolean operator XOR, for which the mutual information between any subpair of the three variables is zero. This does not prevent us from discovering loops of size four and above, as our results show. We introduce a percent tolerance for the DPI to account for inaccurate estimates of the difference between two close mutual information values. This is implemented by rewriting the DPI using a percent tolerance threshold ε: $I(x, z) \leq I(x, y)[1 - \varepsilon]$ and $I(x, z) \leq I(y, z)[1 - \varepsilon]$. This has the advantage of avoiding rejection of some borderline edges, resulting in some loops of size three to occur in the predicted topology.

This is the case for the *MYC* network, in which several three-node loops are present. We used $\varepsilon = 0.1$ (10%) for the synthetic network and $\varepsilon = 0.15$ (15%) for the B cell network. The higher value used for B cells is associated with the use of the less-accurate bin counting estimator. These values were determined by Monte Carlo analysis so as to minimize the effect of mutual information estimation errors, using a wide range of bivariate Gaussian distributions, for which the correct mutual information could be estimated analytically. Although the estimate error on the mutual information can be substantial, the error on the difference of close values of mutual information is much smaller, as the estimate bias tends to cancel out. The quality of the reconstruction, both on the synthetic data and the real expression profiles, justifies this choice of parameters.

Synthetic model. We tested ARACNe on a simulated genetic network[19] to study dynamic Bayesian network inference. We used a 20-gene network containing 14 gene regulatory interactions with one negative feedback loop. Regulatory interactions are defined to affect the transcriptional rate of the target gene linearly, as modeled in a discrete time step simulation by the formula $Y_{t+1} = f(Y_t) = A(Y_t - T) + \varepsilon$, where Y_t is a vector representing the expression levels of all genes at time t, with expression levels ranging from 0 to 100; A is a matrix of gene regulatory interactions; T is a vector of threshold regulating values, which causes the influence of each gene on its target to be proportional to its deviation from this threshold value; and ε is a noise term drawn uniformly from the interval [−10,10]. We studied performance by choosing precision,

over specificity, as a metric for two reasons: (i) precision translates directly into the expected success rate of a subsequent biological validation step; and (ii) the large number of potential interactions in a network, proportional to the square of the number of genes, makes the specificity a less relevant metric.

Bayesian networks. A Bayesian network is a representation of a joint probability distribution as a directed acyclic graph whose vertices correspond to random variables $\{X_1,\ldots,X_n\}$ and whose edges correspond to dependencies between variables. The most likely graph G for a given data set D can be inferred by searching for the optimal graph based on a statistically motivated scoring metric. In this study, we used the Bayesian Scoring Metric[36], defined as $S(G:D) = \log P(G|D) = \log P(D|G) + \log P(G) - \log P(D)$, where $\log P(D)$ is independent of G and can be treated as a constant. $P(G)$ is the prior over graphs, for which we use a uniform prior, following a previously described method[6,19,20]. The results of the analysis (**Fig. 1**) were produced with the Bayesian Networks software LibB2.1 (ref. 17), which is one of the best implementations of the method. The graph space was explored using the greedy hill-climbing algorithm with random restarts (other search methods were tested with similar results). We used a Dirichlet prior of 1, which is suggested for the method. The full Dirichlet parameter space was explored without any performance improvement over ARACNe (data not shown).

Generation of gene-expression profiles. The gene-expression profiles data set (336 samples) includes normal purified cord blood (5 samples)[37], germinal center (10), memory (5) and naive (5) B cells[38]; 34 samples of B cell chronic lymphocytic leukemia[37]; 68 samples of diffuse large B cell lymphomas, including cases further classified as immunoblastic or centroblastic; 27 samples of Burkitt lymphoma; 6 samples of follicular lymphoma; 9 samples of primary effusion lymphoma[39]; 8 samples of mantle cell lymphoma; 16 samples of hairy cell leukemia[40]; 4 cell lines derived from Hodgkin disease[41]; 5 B cell lymphoma cell lines; and 5 lymphoblastic cell lines. The data set includes a Burkitt lymphoma cell line (Ramos) treated *in vitro* to activate CD40 or BCR signaling[42] and cell lines engineered to stably express BCL6 and BCL6(ΔPEST) mutant[43] or to conditionally express BCL6 or MYC[44]. BCL6 conditional expression was obtained in EREB cells[45] using a metallothionein responsive promoter. A detailed description of the samples is given in **Supplementary Table 1.**

We generated gene-expression profiles by following the protocol recommended by Affymetrix, Inc. We extracted total RNA using the Trizol reagent (Invitrogen Life Technologies) and purified it using the RNeasy Kit (Qiagen). We generated double-stranded cDNA from 5 μg of total RNA using the SuperScript Choice System (Invitrogen Life Technologies) and a poly-dT oligonucleotide that contains a T7 RNA polymerase initiation site. We used double-stranded cDNA as template to generate biotinylated cRNA by *in vitro* transcription using MEGAscript T7 High Yield Transcription kit (Ambion), Biotin-11-CTP and Biotin-11-UTP (PerkinElmer Life Sciences). We purified biotinylated cRNA using the Rneasy Kit (Qiagen) and fragmented it in accordance with the Affymetrix protocol. We hybridized 15 μg of fragmented cRNA to HG-U95Av2 microarrays (Affymetrix). We determined gene-expression values by Affymetrix Microarray Suite 5.0 software, using the Global Scaling option.

ChIP. We identified the E-box sequences using the MatInspector software (Genomatix Software GmbH). We carried out ChIP analysis of a Burkitt lymphoma cell line (Ramos) using a previously reported protocol[46]. We cross-linked proteins to DNA and carried out immunoprecipitation using N262 MYC antibodies (Santa Cruz) as well as species- and isotype-matched control antibodies (rabbit IgG; Sigma) as control for the specificity. After reverse cross-linking, we purified DNA by phenol-chloroform extraction and resuspended it in 30 μl of Tris-EDTA (150 μl for the total input sample, representing 3% of the soluble chromatin before immunoprecipitation). We detected E-box sequences by PCR amplification using 2 μl of the above DNA preparations. Oligonucleotide sequences, annealing temperatures and number of cycles of amplification are given in **Supplementary Table 7** online. We resolved PCR products on 2% agarose gels and visualized them by ethidium-bromide staining.

URLs. The results shown in **Figure 1** were produced with the Bayesian Network software LibB 2.1 (http://www.cs.huji.ac.il/labs/compbio/LibB/). The *MYC* target gene database is available at http://www.myc-cancer-gene.org/site/mycTargetDB.asp. The ARACNe platform is available at http://amdec-bioinfo.cu-genome.org/html/caWorkBench.htm.

GEO accession number. Gene-expression profiles, GSE2350.

Note: Supplementary information is available on the Nature Genetics website.

ACKNOWLEDGMENTS

We thank I. Nemenman for his expertise in Information Theory, M. Mattioli for contributing to the generation of the B-cell gene expression database and V. Miljkovic for help with the microarray hybridizations. K.B. is supported by a fellowship from the American-Italian Cancer Foundation, A.M. by the National Library of Medicine Medical Informatics Research Training Program at Columbia and U.K. by a fellowship from the Human Frontiers Science Program. This study was supported by a US National Institutes of Health grant to R.D.-F. and by the computational resources of the AMDeC Bioinformatics Core at Columbia University.

COMPETING INTERESTS STATEMENT

The authors declare that they have no competing financial interests.

Received 14 December 2004; accepted 8 February 2005
Published online at http://www.nature.com/naturegenetics/

1. Han, J.D. *et al.* Evidence for dynamically organized modularity in the yeast protein-protein interaction network. *Nature* **430**, 88–93 (2004).
2. Jeong, H., Tombor, B., Albert, R., Oltvai, Z.N. & Barabasi, A.L. The large-scale organization of metabolic networks. *Nature* **407**, 651–654 (2000).
3. Jordan, I.K., Marino-Ramirez, L., Wolf, Y.I. & Koonin, E.V. Conservation and coevolution in the scale-free human gene coexpression network. *Mol. Biol. Evol.* **21**, 2058–2070 (2004).
4. Lukashin, A.V., Lukashev, M.E. & Fuchs, R. Topology of gene expression networks as revealed by data mining and modeling. *Bioinformatics* **19**, 1909–1916 (2003).
5. Friedman, N., Linial, M., Nachman, I. & Pe'er, D. Using Bayesian networks to analyze expression data. *J. Comput. Biol.* **7**, 601–620 (2000).
6. Hartemink, A.J., Gifford, D.K., Jaakkola, T.S. & Young, R.A. Using graphical models and genomic expression data to statistically validate models of genetic regulatory networks. *Pac. Symp. Biocomput.* **2001**, 422–433 (2001).
7. Gat-Viks, I. & Shamir, R. Chain functions and scoring functions in genetic networks. *Bioinformatics* **19** Suppl 1, i108–i117 (2003).
8. Gardner, T.S., di Bernardo, D., Lorenz, D. & Collins, J.J. Inferring genetic networks and identifying compound mode of action via expression profiling. *Science* **301**, 102–105 (2003).
9. Yeung, M.K., Tegner, J. & Collins, J.J. Reverse engineering gene networks using singular value decomposition and robust regression. *Proc. Natl. Acad. Sci. USA* **99**, 6163–6168 (2002).
10. Ideker, T. *et al.* Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science* **292**, 929–934 (2001).
11. Butte, A.J. & Kohane, I.S. Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. *Pac. Symp. Biocomput.* **2000**, 418–429 (2000).
12. Butte, A.J., Tamayo, P., Slonim, D., Golub, T.R. & Kohane, I.S. Discovering functional relationships between RNA expression and chemotherapeutic susceptibility using relevance networks. *Proc. Natl. Acad. Sci. USA* **97**, 12182–12186 (2000).
13. Elkon, R., Linhart, C., Sharan, R., Shamir, R. & Shiloh, Y. Genome-wide in silico identification of transcriptional regulators controlling the cell cycle in human cells. *Genome Res.* **13**, 773–780 (2003).
14. Hughes, T.R. *et al.* Functional discovery via a compendium of expression profiles. *Cell* **102**, 109–126 (2000).
15. MacLennan, I.C. Germinal centers. *Annu. Rev. Immunol.* **12**, 117–139 (1994).
16. Barabasi, A.L. & Bonabeau, E. Scale-free networks. *Sci. Am.* **288**, 60–69 (2003).
17. Friedman, N. Inferring cellular networks using probabilistic graphical models. *Science* **303**, 799–805 (2004).
18. Smith, V.A., Jarvis, E.D. & Hartemink, A.J. Influence of network topology and data collection on network inference. *Pac. Symp. Biocomput.* **2003**, 164–175 (2003).
19. Yu, J., Smith, A.V., Wang, P.P., Hartemink, A.J. & Jarvis, E.D. Using Bayesian network inference algorithms to recover molecular genetic regulatory networks. in *3rd International Conference on Systems Biology* (Karolinska Institute, Stockholm, Sweden, 2002).
20. Smith, V.A., Jarvis, E.D. & Hartemink, A.J. Evaluating functional network inference using simulations of complex biological systems. *Bioinformatics* **18**, S216–S224 (2002).
21. Jarvis, E.D. *et al.* A framework for integrating the songbird brain. *J. Comp. Physiol. A. Neuroethol. Sens. Neural. Behav. Physiol.* **188**, 961–980 (2002).
22. Zeeberg, B.R. *et al.* GoMiner: a resource for biological interpretation of genomic and proteomic data. *Genome Biol.* **4**, R28 (2003).
23. Zeller, K.I., Jegga, A.G., Aronow, B.J., O'Donnell, K.A. & Dang, C.V. An integrated database of genes responsive to the Myc oncogenic transcription factor: identification of direct genomic targets. *Genome Biol.* **4**, R69 (2003).
24. Fernandez, P.C. *et al.* Genomic targets of the human c-Myc protein. *Genes Dev.* **17**, 1115–1129 (2003).

25. Dang, C.V. c-Myc target genes involved in cell growth, apoptosis, and metabolism. *Mol. Cell. Biol.* **19**, 1–11 (1999).
26. O'Connell, B.C. *et al.* A large scale genetic analysis of c-Myc-regulated gene expression patterns. *J. Biol. Chem.* **278**, 12563–12573 (2003).
27. D'Haeseleer, P., Wen, X., Fuhrman, S. & Somogyi, R. Linear modeling of mRNA expression levels during CNS development and injury. *Pac. Symp. Biocomput.* **1999**, 41–52 (1999).
28. Stuart, J.M., Segal, E., Koller, D. & Kim, S.K. A gene-coexpression network for global discovery of conserved genetic modules. *Science* **302**, 249–255 (2003).
29. Barabasi, A.L. & Albert, R. Emergence of scaling in random networks. *Science* **286**, 509–512 (1999).
30. Roos, J., Luz, J.M., Centoducati, S., Sternglanz, R. & Lennarz, W.J. ENP1, an essential gene encoding a nuclear protein that is highly conserved from yeast to humans. *Gene* **185**, 137–146 (1997).
31. Suzuki, N. *et al.* A cytoplasmic protein, bystin, interacts with trophinin, tastin, and cytokeratin and may be involved in trophinin-mediated cell adhesion between trophoblast and endometrial epithelial cells. *Proc. Natl. Acad. Sci. USA* **95**, 5027–5032 (1998).
32. Chen, W., Bucaria, J., Band, D.A., Sutton, A. & Sternglanz, R. Enp1, a yeast protein associated with U3 and U14 snoRNAs, is required for pre-rRNA processing and 40S subunit synthesis. *Nucleic Acids Res.* **31**, 690–699 (2003).
33. Stewart, M.J. & Nordquist, E.K. Drosophila Bys is nuclear and shows dynamic tissue-specific expression during development. *Dev. Genes Evol.* **215**, 97–102 (2005).
34. Joe, H. *Multivariate Models and Dependence Concepts* (Chapman & Hall, Boca Raton, Florida, 1997).
35. Steuer, R., Kurths, J., Daub, C.O., Weise, J. & Selbig, J. The mutual information: Detecting and evaluating dependencies between variables. *Bioinformatics* **18** Suppl 2: S231–S240 (2002).
36. Cooper, G.F. & Herskovits, E. A Bayesian method for the induction of probabilistic networks from data. *Mach. Learn.* **9**, 309–347 (1992).
37. Klein, U. *et al.* Gene expression profiling of B cell chronic lymphocytic leukemia reveals a homogeneous phenotype related to memory B cells. *J. Exp. Med.* **194**, 1625–1638 (2001).
38. Klein, U. *et al.* Transcriptional analysis of the B cell germinal center reaction. *Proc. Natl. Acad. Sci. USA* **100**, 2639–2644 (2003).
39. Klein, U. *et al.* Gene expression profile analysis of AIDS-related primary effusion lymphoma (PEL) suggests a plasmablastic derivation and identifies PEL-specific transcripts. *Blood* **101**, 4115–4121 (2003).
40. Basso, K. *et al.* Gene expression profiling of hairy cell leukemia reveals a phenotype related to memory B cells with altered expression of chemokine and adhesion receptors. *J. Exp. Med.* **199**, 59–68 (2004).
41. Kuppers, R. *et al.* Identification of Hodgkin and Reed-Sternberg cell-specific genes by gene expression profiling. *J. Clin. Invest.* **111**, 529–537 (2003).
42. Basso, K. *et al.* Tracking CD40 signaling during germinal center development. *Blood* **104**, 4088–4096 (2004).
43. Niu, H., Cattoretti, G. & Dalla-Favera, R. BCL6 controls the expression of the B7-1/CD80 costimulatory receptor in germinal center B cells. *J. Exp. Med* . **198**, 211–221 (2003).
44. Wu, K.J., Polack, A. & Dalla-Favera, R. Coordinated regulation of iron-controlling genes, H-ferritin and IRP2, by c-MYC. *Science* **283**, 676–679 (1999).
45. Kempkes, B. *et al.* B-cell proliferation and induction of early G1-regulating proteins by Epstein-Barr virus mutants conditional for EBNA2. *EMBO J.* **14**, 88–96 (1995).
46. Frank, S.R., Schroeder, M., Fernandez, P., Taubert, S. & Amati, B. Binding of c-Myc to chromatin mediates mitogen-induced acetylation of histone H4 and gene activation. *Genes Dev.* **15**, 2069–2082 (2001).