Molecular Systems Biology 4; Article number 169; doi:10.1038/msb.2008.2
**Citation:** ***Molecular Systems Biology*** **4:169**

www.molecularsystemsbiology.com

molecular systems biology

REPORT

# A systems biology approach to prediction of oncogenes and molecular perturbation targets in B-cell lymphomas

**Kartik M Mani[1,*], Celine Lefebvre[2], Kai Wang[1], Wei Keat Lim[1], Katia Basso[3], Riccardo Dalla-Favera[3] and Andrea Califano[1,2,3]**

[1] Department of Biomedical Informatics (DBMI), Columbia University, New York, NY, USA, [2] Center for Computational Biology and Bioinformatics (C2B2), Columbia University, New York, NY, USA and [3] Institute for Cancer Genetics and Herbert Irving Comprehensive Cancer Center, Columbia University, New York, NY, USA

* Corresponding author. Department of Biomedical Informatics, Columbia University, 1130 St. Nicholas Ave, New York, NY 10032, USA. Tel.: +212-851-5183; Fax: +212-851-5149; E-mails: kartik.mani@dbmi.columbia.edu or califano@c2b2.columbia.edu

Received 2.10.07; accepted 14.12.07

**The computational identification of oncogenic lesions is still a key open problem in cancer biology. Although several methods have been proposed, they fail to model how such events are mediated by the network of molecular interactions in the cell. In this paper, we introduce a systems biology approach, based on the analysis of molecular interactions that become dysregulated in specific tumor phenotypes. Such a strategy provides important insights into tumorigenesis, effectively extending and complementing existing methods. Furthermore, we show that the same approach is highly effective in identifying the targets of molecular perturbations in a human cellular context, a task virtually unaddressed by existing computational methods. To identify interactions that are dysregulated in three distinct non-Hodgkin's lymphomas and in samples perturbed with CD40 ligand, we use the B-cell interactome (BCI), a genome-wide compendium of human B-cell molecular interactions, in combination with a large set of microarray expression profiles. The method consistently ranked the known gene in the top 20 (0.3%), outperforming conventional approaches in 3 of 4 cases.**

*Molecular Systems Biology* 12 February 2008; doi:10.1038/msb.2008.2
*Subject Categories:* molecular biology of disease; metabolic & regulatory networks
*Keywords:* B-cell lymphoma; drug mechanism-of-action (MOA); gene network; interactome; oncogene



## Introduction

Cancer is a complex and highly heterogeneous disease that is mediated by a myriad of distinct cellular pathways, according to tissue of origin, specific set of chromosomal aberrations/mutations, and environmental conditions. In leukemia, for instance, there are several documented oncogenic lesions that work cooperatively to drive the cell to tumorigenesis (Mullighan *et al*, 2007). As a result, cancer phenotypes can exhibit a great range of genetic variability. With analytical methods still in their relative infancy, it is thus not surprising that we are only in the very preliminary stages of assembling a complete repertoire of germ-line and somatic oncogenic lesions for each cancer phenotype.

Such knowledge, albeit still partial, has already proven useful as a guide for therapeutic intervention (Downward, 2006) and is expected to become a key driver in the development of new personalized, diagnostic, and therapeutic strategies. Therefore, the computational inference of oncogenic events, as well as their specific impact on pathway dysregulation, has become the subject of intense focus in molecular biology.

High-throughput technologies are now producing vast amounts of biological data representing the availability of specific molecular species in a cellular population. These include, among many others, gene expression and genotypic profiles (Schena *et al*, 1995), DNA-binding profiles from chromatin immunoprecipitation (Ren *et al*, 2000), genomic sequences, and protein abundance from mass spectrometry (Perez and Nolan, 2002). These data have been used extensively to characterize the differences between cancer cells and their normal counterpart. Gene expression profiling,

in particular, has been successful in classifying tumors or patient prognosis based on specific molecular signatures. These have been applied to several phenotypes, including leukemia (Golub *et al*, 1999) and breast cancer (van 't Veer *et al*, 2002). In a similar context, expression profiling has also been used to characterize the molecular signatures arising from specific pharmacological interventions in the cell (Lamb *et al*, 2006).

Recently, using these data, a number of computational methods have been proposed for the identification of oncogenes, tumor-suppressor genes, and even entire pathways that are dysregulated in cancer. A highly recurrent gene fusion event, for instance, was identified in prostate cancer from expression profiles using an 'outlier' analysis approach (Tomlins *et al*, 2005). Additionally, genome-wide SNP profiling and array-based comparative genomic hybridization were applied to the identification of germ-line and somatic lesions in several cancers, including leukemia (Mullighan *et al*, 2007) and breast cancer (Yao *et al*, 2006). Integrative approaches were also proposed: copy-number and expression profile data, for instance, were successfully used in the identification of specific chromosomal amplifications in breast cancer (Adler *et al*, 2006). Other context-dependent methods have been proposed such as those that use reference signatures of specific activated pathways to characterize tumors and establish drug sensitivity (Bild *et al*, 2006).

These methods, while partially successful, still focus primarily on characteristics of individual genes or gene products. It is not possible, therefore, to infer any details on how a protein's behavior has changed, nor the specific mechanisms that led to the pathologic transition.

In this paper, we introduce the interactome dysregulation enrichment analysis (IDEA) algorithm, which uses a genome-wide molecular interaction map as a systematic framework for the identification of genes playing a role in oncogenesis. Furthermore, we show that the same approach is also effective in identifying both targets and effectors of specific biochemical perturbations, a problem also known as the 'drug mechanism-of-action' (MOA). Interestingly, while highly related, there are no available computational algorithms to address the MOA problem in a human cellular context; although interesting solutions have been proposed in bacteria (Gardner *et al*, 2003) and yeast (di Bernardo *et al*, 2005). We suggest that studying dysregulation patterns at a cellular network level, rather than in a 'gene-centric' manner, can provide a highly efficient method for addressing both problems. Furthermore, the use of cellular networks provides a much-needed molecular interaction context to further characterize any gene predictions emerging from the analysis.

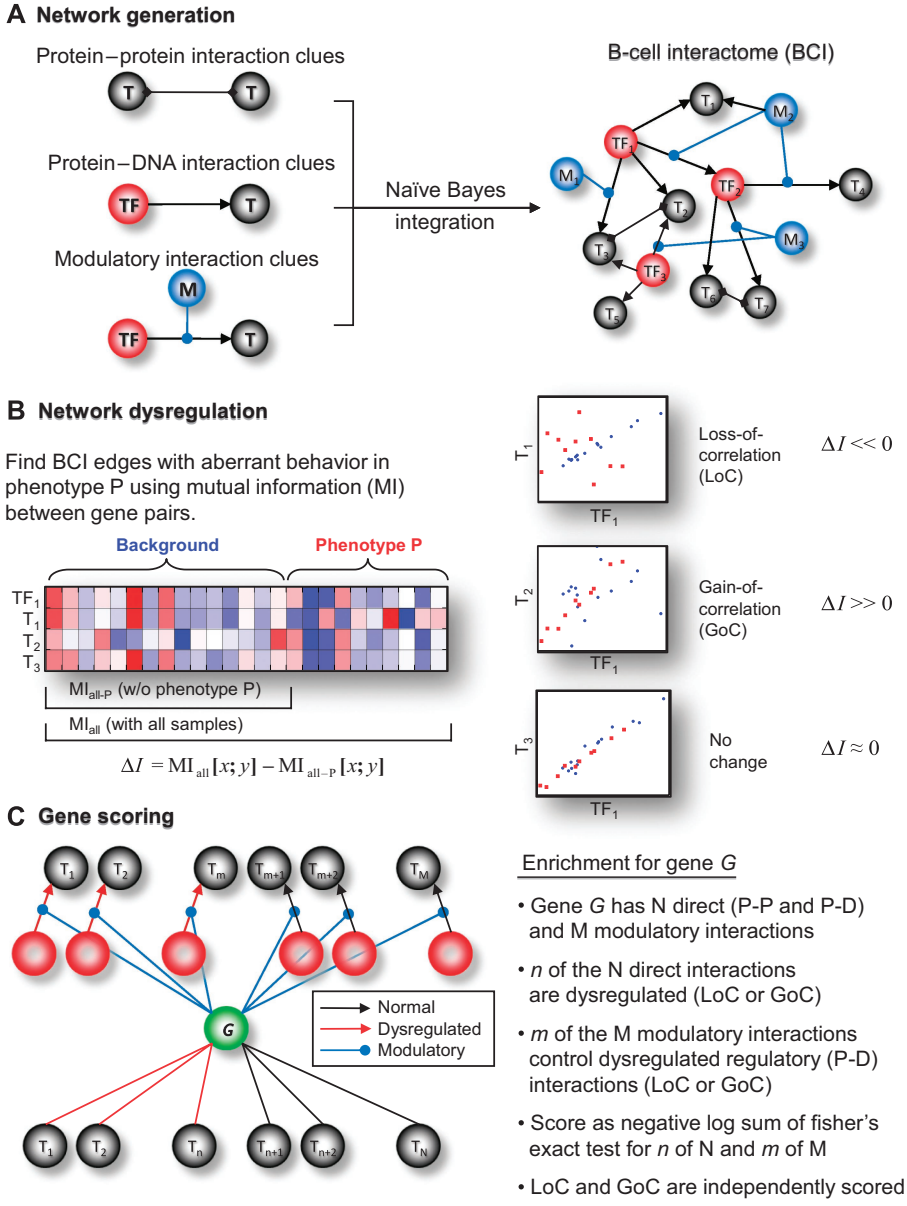
The use of an interaction network for gene–disease association is not novel *per se*. A few recent studies have leveraged the growing repertoire of interaction data for this purpose. In one example (Lage *et al*, 2007), protein–protein interaction networks were combined with Online Mendelian Inheritance in Man (OMIM) (Hamosh *et al*, 2000) annotation data to identify complexes implicated in disease progression. In another study specific to prostate cancer (Ergun *et al*, 2007), a regulatory network was inferred from microarray data and used as a filter to infer genetic mediators of disease progression. The approach was successful in identifying the androgen-receptor-signaling pathway, whose role in prostate cancer is already well documented. Both methods however, like others in this category, still adopt a gene-centric approach, using the underlying network essentially as a filter to identify clusters of significant genes. Furthermore, only individual interaction layers, such as the transcriptional layer or the protein complex layer, were modeled by these methods. Finally, no explicit biochemical validation is provided to support their prediction accuracy.

In this paper, we use an existing genome-wide cellular network, the B-cell interactome (BCI), originally assembled by our laboratory (Lefebvre *et al*, 2007) and further enhanced by including post-translational modulation events (C Lefebvre *et al*, in preparation). The BCI is a mixed-interaction network, representing several key molecular interaction types in a human B cell, including transcriptional, signaling, and complex formation. The proposed analysis works in two steps. We first use a large compendium of microarray expression profiles from normal, tumor-related, and experimentally manipulated B cells to identify BCI interactions showing either a gain of correlation (GoC) or a loss of correlation (LoC) pattern in the phenotype of interest. These interactions are either lost (LoC) or gained (GoC) in the specific phenotype compared with the background, based on an information-theoretic test. We then rank genes according to the statistical significance of the LoC/GoC enrichment among the interactions in which they directly participate (see Box 1 for method overview).

The study introduces four key innovations as follows: (1) by adopting a genome-wide, mixed-interaction network, instead of the individual interaction layers of previous studies, we cover a far greater range of processes within the cell; (2) rather than analyzing the differential properties of individual genes (e.g., expression profile or genotypic data), we identify molecular *interactions* that are significantly dysregulated in a particular phenotype of interest. We hypothesize that genes implicated in cancer initiation and progression (as well as those targeted by specific biochemical perturbations) will show dysregulated interactions with their molecular partners. Biologically, this is quite plausible, since biochemical perturbations as well as a wide variety of oncogenic events (gene fusion or translocation, post-translational protein modification, structural mutation) will manifest through gains or losses of regulatory, signaling, and protein–complex interaction capability; (3) we validate on three distinct tumor models (follicular (FL), Burkitt's (BL), and mantle cell lymphoma (MCL)), whose oncogenic lesions are both known and completely different. In each case, we show that the known gene is identified in the 20 most significant by the analysis; (4) finally, we biochemically validate the approach by perturbing B-cell lines (using the *CD40* ligand/antibody) and by showing that the method is successful in identifying the perturbation targets (*CD40* pathway genes).

A key advantage of such a network-centric approach is that it can identify relatively small, yet tightly connected areas of the network (modules) that are dysregulated, providing a window over the mechanistic and possibly synergistic processes underlying oncogenesis and biochemical perturbation.

**Box 1 Interactome Dysregulation Enrichment Analysis (IDEA)**




An overview of the proposed network-based analysis to characterize oncogenic mechanisms and pharmacological interventions. (**A**) In step 1, a comprehensive network of interactions is generated for B cells using a Bayesian evidence integration approach, including predictions of post-translational modifications. In this diagram, transcription factors are shown in red, non-transcription factors in gray, and modulators are shown in blue. Directed arrows indicate protein–DNA (P-D) interactions, and undirected indicate protein–protein (P-P) interactions or modulation events. Evidences, or clues, include curated databases, literate mining, orthologous interactions from model organisms, and reverse engineering algorithms. (**B**) In step 2, each interaction is analyzed to determine which show aberrant behavior in a specific phenotype (P); that is, interactions that show correlation in all samples except P (TF1 and T1), or interactions that are not correlated in any samples except P (TF1 and T2). These dysregulated interactions are classified as LoC or GoC, respectively, for every edge in the BCI. (**C**) In step 3, these dysregulated interactions are pooled together and a statistical enrichment is calculated which identifies genes having an unusually high number of these interactions in its neighborhood, either through direct or modulated links.

## Results

The enhanced version of the BCI (http://amdec-bioinfo.cu-genome.org/html/BCellInteractome.html) includes 64 649 unique pairwise interactions (160 730 non-unique interactions between probes). This network represents an 'average' set of molecular interactions, supported by the majority of B-cell samples from several stages of normal development—naïve (N), memory (M) and germinal center (GC)—as well as from several tumor phenotypes. Interactions that are present only in a small phenotypic subset are not represented. For each phenotype, Table I shows the number of dysregulated interactions detected by IDEA divided by LoC and GoC category. Figure 1 shows a comprehensive view of all the

dysregulated interactions in each represented phenotype, using a 'barcode' like representation. Two findings are intriguing from this global analysis. First, a large percentage of the network interactions are not dysregulated in any of the phenotypes (80.5%), implying that many of the interactions represent a cellular network 'backbone' that behaves consistently across phenotypes. Second, as shown, cancer barcodes for different phenotypes appear highly distinctive. See Materials and methods section for a clear definition of LoC and GoC interactions.

**Table I** Distribution of phenotypes and LoC and GoC signatures

| Phenotype | No. of samples | LoC | GoC |
|---|---|---|---|
| B-CLL | 34 | 1813 | 10 815 |
| B-CLL-mut | 18 | 121 | 3417 |
| B-CLL-unmut | 16 | 92 | 1430 |
| BL | 26 | 383 | 701 |
| pDLCL | 15 | 596 | 17 |
| pFL | 6 | 183 | 9 |
| HCL | 16 | 3399 | 824 |
| pMCL | 8 | 488 | 16 |
| PEL | 9 | 1839 | 1204 |

Abbreviations: BL, Burkitt's lymphoma; CLL-mut, chronic lymphocytic leukemia from mutated; CLL-unmut, chronic lymphocytic leukemia from unmutated; DLCL, diffuse large B-cell lymphoma; FL, follicular lymphoma; GoC, gain of correlation; LoC, loss of correlation; MCL, mantle cell lymphoma; PEL, primary effusion lymphoma.

The method's performance was benchmarked using three extensively characterized B-cell tumor phenotypes and a set of biochemical perturbation assays. In all four assays, the method correctly identified the known gene in the top 20 candidates out of approximately 7900 probes on the chip, after filtering non-informative genes based on the coefficient of variation. These tests are discussed below.

## FL benchmark

FL is one of the most common B-cell non-Hodgkin's lymphomas (NHLs), the key genetic lesion (found in ~90% of FL samples) is the t(14;18) rearrangement. This translocation causes the constitutive expression of the antiapoptotic *BCL2* oncogene (Bende *et al*, 2007). FL shows a relatively small network dysregulation signature, with only 192 LoC/GoC interactions. *BCL2*, which supports eight of those interactions, is ranked first by our enrichment analysis method. By comparison, differential expression analysis between FL samples and GC samples (the normal FL counterpart) ranks *BCL2* in the fifty-ninth position. Furthermore, the analysis identified the *SMAD1* gene, ranked sixth. This gene, although not detectable by differential expression analysis in our data set, has been shown to have an aberrant pathway activation in FL and other NHL phenotypes, mediated by tumor-transforming growth factor-β (Munoz *et al*, 2004).

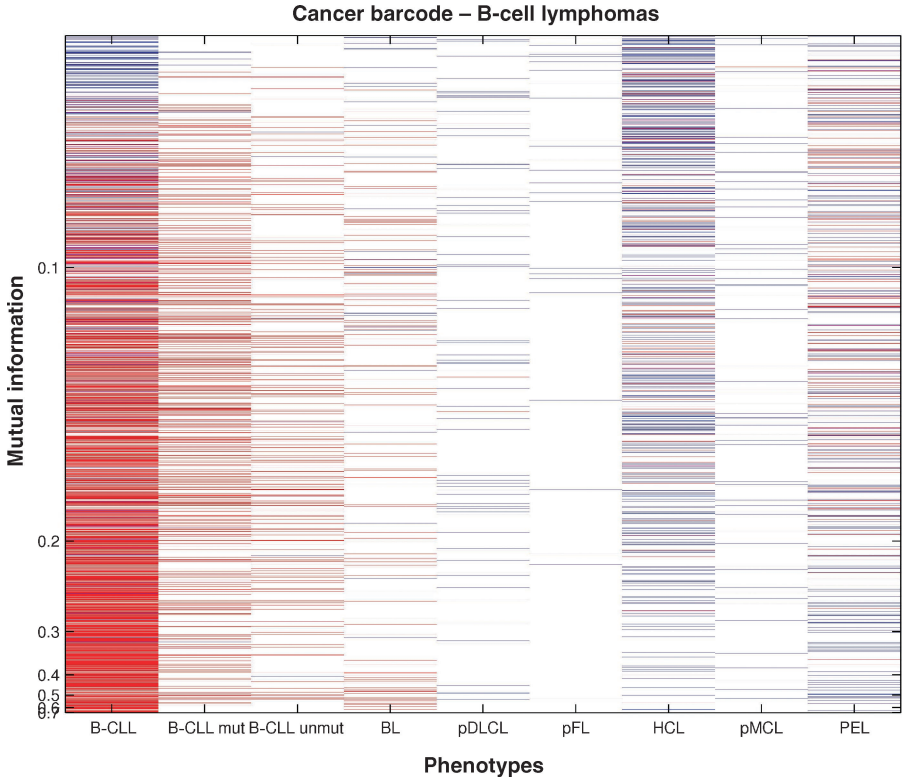



**Figure 1** Cancer barcode: In this figure we show the complete set of affected BCI interactions for each analyzed phenotype. The rows represent these BCI interactions sorted in ascending order (from top to bottom) by their MI computed over the complete set of BCGEP samples. Each column is one analyzed phenotype. These phenotypes shown include CLL-mut and CLL-unmut subsets, BL, DLCL, FL, MCL, and PEL. A 'p' preceding a phenotype name indicates those samples were purified. Interactions are color coded in blue for LoC and red for GoC. Clearly visible from this figure is that these phenotypes all appear to have very distinct areas of the network, which define their pathologic activity.

**Table II** Comparative ranks of GoC, LoC, and combined enrichments for B-cell lymphoma phenotypes as well as *CD40*-stimulated Ramos cells

| Phenotype | Gene | LoC | GoC | Combined | *t*-Test |
|---|---|---|---|---|---|
| BL | *MYC* | 308 | 7 | 15 | 32 |
| FL | *BCL2* | 1 | NA | 1 | 59 |
| MCL | *CCND1* | 28 | 21 | 18 | 6 |
| Ramos/*CD40* | *CD40* | NA | 7 | 9 | 24 |

Abbreviations: *CCND1*, cyclin D1/*BCL1*; FL, follicular lymphoma; GoC, gain of correlation; LoC, loss of correlation; MCL, mantle cell lymphoma; NA, not available.
Last column indicates ranking by differential expression analysis.

## BL benchmark

BL is endemic among children in equatorial Africa and occurs sporadically in other geographic areas, where it also affects adults (Bellan *et al*, 2003). In these malignancies, a key oncogenic lesion is the translocation of the proto-oncogene *MYC* from chromosome 8 to either the immunoglobulin heavy-chain region on chromosome 14, or one of the light-chain regions on chromosome 2 or chromosome 22. *MYC* has been shown to have a global regulatory role in BL (Li *et al*, 2003). *MYC* is also one of the most connected hubs in the BCI, having 4079 probe-based interactions. Sixty of these interactions were dysregulated, giving this gene the fifteenth most significant enrichment score. By differential expression analysis between BL and GC cells (BL's normal counterpart), *MYC* has a rank of thirty-two (see Table II). While this result is encouraging *per se*, our method was also successful in identifying other key effectors of *MYC* in BL. In particular, *MTA1*, an established target of *MYC*, was ranked third, even though it is not even ranked in the top 1000 genes by differential expression. *MTA1* was recently identified as a primary downstream effector of *MYC* function. Specifically, its silencing blocks the ability of *MYC* to produce a pathologic transformation (Zhang *et al*, 2005).

## MCL benchmark

MCL is an aggressive type of NHL that generally occurs in middle-aged and elderly people. Cyclin D1/*BCL1* (*CCND1*) is a cell-cycle protein that is overexpressed in MCL as a result of the translocation t(11;14) involving the immunoglobulin heavy-chain gene on chromosome 14 and a region on chromosome 11 harboring *CCND1*. (Miranda *et al*, 2000). In the BCI, cyclin D1 is connected to six dysregulated interactions, ranking it eighteenth in our list. By differential expression analysis with non-GC samples (MCL's normal counterpart) *CCND1* has a rank of six (see Table II). In addition, our analysis ranked *HDAC1* third among all candidates. Histone deacetylases inhibitors have recently been suggested as potentially useful in the therapy of MCL (Heider *et al*, 2006), so this finding is another piece of supporting evidence that our method identifies the correct patterns. *HDAC1* is also highly differentially expressed, and ranked fourteenth. These results indicate that in some cases conventional analysis do indeed capture the correct gene(s). However, as shown, our method seems to consistently identify these key genes as well as effectors, which may be undetectable by differential expression.

In these three cases, it is important to note that we expect the translocated gene to be differentially expressed. It is significant therefore, that against a benchmark where differential expression should be very useful, our method still outperforms it in two out of three cases, and consistently ranks these genes at the very top throughout.

Interestingly, when the scores for these phenotypes are shown distinctly for LoC and GoC interactions (see Table II), *MYC* appears heavily weighted toward GoC, *BCL2* toward LoC, and *CCND1* shows a mixed mode of both. These results may indicate that the progression of these lymphomas is marked by distinct types of changes in the network.

## Biochemical validation

Although the above examples provide some evidence that our method can correctly identify key regulators and effectors in three separate tumors, a more robust form of validation can be provided by a biochemical perturbation of a specific pathway. We proceeded to analyze a set of samples from Ramos (BL) cell lines stimulated with *CD40* ligand or antibody against a non-stimulated set. To quantitatively measure the performance of the method, we considered an established signature of 41 genes in the *CD40* pathway and used the gene set enrichment analysis (GSEA) (Subramanian *et al*, 2005) to compare our method to differential expression analysis.

Our method ranked 379 probes as having a non-zero score. Using GSEA, this ranked list produced a nominal enrichment *P*-value of 0 ($P < 1e{-}3$ given 1000 permutations), with 13 of the *CD40* pathway genes appearing in the list, many of them clustered at the very top. Remarkably, of the top 10 genes five are in the *CD40* pathway set, including *CD40* itself, which is ranked ninth. The other four *CD40* pathway genes include *NFKB1* (second), *NFKBIA* (third), *NFKBIE* (fifth), and *NFKB2* (tenth), all known to be key effectors of *CD40* signaling. Since our method produces a score of zero for all genes that do not participate in any dysregulated interactions, it is not possible to analyze enrichment beyond these 379 probes. When compared with differential expression using the same cutoff of 379 probes, GSEA produces a nominal *P*-value of 0.12, showing no statistically significant enrichment of the *CD40* pathway gene list. *CD40* itself is ranked twenty-fourth. Furthermore, in our analysis, we find eight *CD40* pathway genes in the top 25 (*P*-value=0 by Fisher's exact test, below machine precision), compared with only 4 of 25 by differential expression analysis (*P*-value $< 2e{-}5$). Although both approaches show significant enrichment, the new method captures twice as many relevant genes within the top 25, while finding the actual perturbation target within the top 10. This further supports the use of our method for the identification of targets of compounds of unknown activity. When looking at these results, the extreme enrichment of the *CD40* pathway members, both in the top 10 and 25 genes is likely to make the difference between identifying and missing the perturbation MOA. Note that, similar to the other benchmarks, *CD40* itself is upregulated upon binding the *CD40* ligand. Thus, as expected, differential expression analysis appears partially effective. However, as shown for *MTA1*, *SMAD1*, and other effectors (see Figure 2), IDEA does not

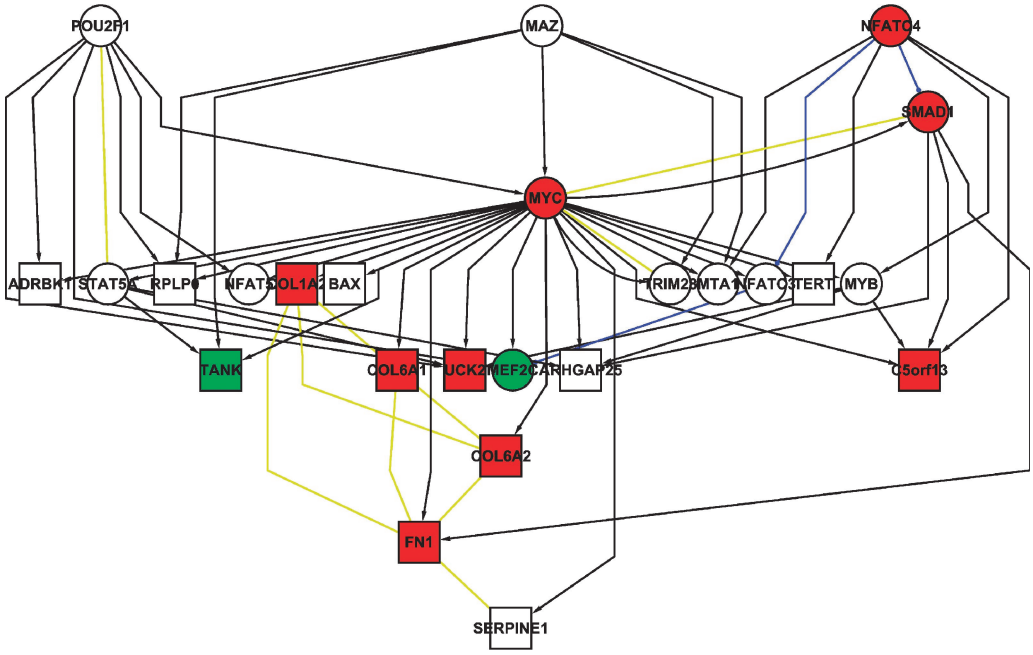



**Figure 2** BL module: A network visualization of the top 25 scoring genes in BL. Transcription factors are shown as circles, whereas other proteins are shown as squares. Protein–protein interactions are also shown in beige, protein–DNA interactions are black with an arrowhead, and transcription factor-modulated interactions are shown in blue with a circular endpoint. Red/green indicates overexpression or underexpression ($P < 1e{-}8$), respectively in BL versus GC cells. There are some notable characteristics of this figure. First, all 25 genes form a connected module, which would not occur by chance. Second, *MYC* appears to be a central regulator of this module, as a full 21 out of the 25 members are *MYC* targets. *MYC* also appears regulated by *MYC*-associated zinc-finger protein (*MAZ*), which is also not differentially expressed. Third, there are interesting sets of genes that emerge, such as *SMAD1*, which is known to be associated with some NHL, and members of the *NFAT* family, including *NFATC3*, *NFATC4*, and *NFAT5* (these proteins are members of the Wnt-signaling pathway). There also appears to be a protein complex of *COL1A2*, *COL6A1 COL6A2*, and *FN1*, which are all upregulated (and members of the cell signaling and ECM–receptor interaction pathway). These module diagrams can serve as a useful platform for further hypothesis generation and biochemical investigation.

require the gene to be differentially regulated in order to be identified as a likely candidate.

## Visualization and interpretation

One benefit of a network-based approach candidate is that gene lists can be viewed in a network context. When we map the top scoring genes from the phenotypes listed above across the network, they tend to tightly cluster in specific areas. Figure 2 shows a visualization of the top 25 genes predicted in BL, which form a connected module. Of interest is the fact that *MYC* is a key regulator of this module (with 21 of 25 genes being its target, including *MTA1*). These 'cancer module' diagrams provide more context than a ranked list of genes, and as shown, can effectively complement existing methods such as differential expression.

IDEA is useful for generating testable hypotheses in a number of different contexts. In the first case, ranked genes can be viewed in a network module to identify key regulators. As discussed in Figure 2, this approach would identify *MYC*, which upon visualization clearly controls the vast majority of top ranked genes. These candidate driver genes could be experimentally validated using siRNA knockdowns or other perturbation assays. Second, these lists can be analyzed for enrichment in specific pathways. We compared the ranked output to a set of Kyoto Encyclopedia of Genes and Genomes, or KEGG (Kanehisa *et al*, 2006), pathway annotations. For BL, this method identified focal adhesion ($P$=0) and the ECM–receptor interaction pathway ($P$=0), which contain similar sets of genes, which are more commonly associated with solid tumors. Also identified were the B-cell receptor-signaling pathway ($P$=0.006) and the Jak-Stat-signaling pathway ($P$=0.057), which has been associated with several different cancer phenotypes. Lastly, genes that score high across multiple phenotypes could be identified pertaining to common mechanisms. When the scores across all phenotypes are averaged, the top scoring genes contain several key oncogenic regulators. Included in the top of this list are *MYC*, the tumor repressor *PRDM2*, *JAK3*, the transcriptional repressor *DRAP1*, and the estrogen receptor *ESR1*. Ranked second was the transcription factor *POU6F1*, which is known to have a role in several eukaryotic development processes, but has not been previously associated with lymphoma, and may warrant further investigation.

We applied this approach to the analysis of chronic lymphocytic leukemia (CLL), a complex tumor phenotype, for which oncogenic lesions have not been identified. The top-ranked genes include *PRDM2*, *MYC*, and *MLL*, which are known to be translocated in different subtypes of leukemia, and *SMAD3*, which is active in several NHL phenotypes. The top 25 genes also form a tightly connected cluster, with almost half the connections being modulated interactions. Pathway enrichment identified the cell-cycle ($P$=0), B-cell receptor ($P$=0.0007), TGFβ ($P$=0.038), and P53-signaling pathways ($P$=0.05). These pathways are commonly associated with B-cell lymphomas and this is not surprising, but the presence

of *MYC*, *MLL*, and *PRDM2*, all strong oncogenic effectors, may be worthy of inquiry in CLL, as they have not previously been associated with this malignant phenotype. *MYC* shows a high level of connectivity in the module diagram, connecting to 18 out of the 24 other genes. It is also predicted to be a regulator and modulator of *PRDM2*. As translocations of *MYC* and *MLL* are exceedingly rare in CLL (Reddy *et al*, 2006), it is unclear what role they have in this specific cancer.

### Discussion

We have proposed IDEA, a systems biology approach to the identification of mechanisms associated with the presentation of a specific tumor phenotype or biochemical perturbation. We have shown that this approach identifies known oncogenic lesions and downstream effectors for 3 malignant B-cell phenotypes. We have also shown its applicability to artificially perturbed cellular systems using Ramos cell line samples where the *CD40* pathway was specifically stimulated.

IDEA gains coverage by generating a network from multiple sources. In our approach, we chose to use a hybrid interactome containing protein–protein, protein–DNA and post-translational interactions inferred by the MINDy algorithm. This decision allows the method to capture several different mechanisms of action associated with oncogenic lesions and biochemical perturbations. As indicated from the results, two of the known lesions correctly identified were not transcription factors (*BCL1*/cyclin D1 and *BCL2*), indicating that we can capture oncogenic candidates that fall outside of typical regulatory network models (and more so that the method is not inherently biased to only find transcriptional regulators). Furthermore, post-translational interactions have not been integrated into other network-based analyses. Although this more inclusive approach may add noise to the analysis, the conservative threshold we apply, along with the fact that incorrect edges would be distributed randomly through the network, leads us to have strong confidence in the tolerance of this approach to false positive and false negative interactions.

A key difference from other network-based methods is that we identify dysregulated network edges (interactions) instead of dysregulated nodes (genes) to assemble disease-related signatures. By focusing on the behavior of gene pairs, as opposed to their individual expression or genetic characteristic, this analysis is capable of identifying patterns other methods may not.

Although we observed results consistent with published data on specific oncogenes, IDEA also identified secondary effectors that were associated with the phenotypic transition. *SMAD1* was identified in FL, and it is known that this pathway is affected in FL and other NHLs. Perhaps the best example of this trend is with BL, where the third-ranked gene was *MTA1*. *MTA1* is a known target of *MYC*, but its higher rank reflects the observation that *MYC* loses its transforming capability in cells without *MTA1*. It is remarkable that both *SMAD1* and *MTA1* are not detected by differential expression analysis and would likely be missed by conventional analysis. Thus, our method not only identifies oncogenic candidates, but also key effectors of the phenotypic transition, where gene expression alone would not support their association.

The ability to visualize these disease modules is also a potential platform for further investigation. It provides advantages beyond simple gene lists, especially with respect to producing a systems level representation of the molecular mechanisms supporting the phenotype. These findings can lead to testable hypotheses and rational models. As noted, when combined with specific pathway enrichment statistics, novel mechanisms may emerge, such as *MYC* as a regulator of proteins involved in the ECM–receptor interaction in BL.

One drawback of this methodology is the large background population that is necessary for comparison. As dependency metrics like mutual information (MI) require a certain sample size to establish significance, this may pose a difficulty in situations where sample sizes are limited. We encountered this very problem in analyzing our B-cell phenotypes, and chose to use our entire set as a background instead. Although this tactic may dilute signals in the data, the positive evidence suggests that we can still detect highly specific details, even among a noisy background. As more data becomes available, this problem will become less apparent.

A second problem deals with the thresholding we apply to classify interactions as GoC and LoC. By being conservative, we may improve accuracy, but the undesired effect is that interactions not meeting this threshold are not used in enrichment, causing the majority of probes to have a zero value. This limitation creates shorter ranked lists of genes that are potentially adding a number of false negatives. We are currently investigating non-threshold-based enrichment statistics, which can allow us to score all the probes accurately.

Next steps in developing this methodology include more fully leveraging the underlying network to infer affected mechanisms. Currently a gene's enrichment is only calculated based on its immediate neighborhood, which is potentially eliminating secondary effects that propagate from one area of the network. If propagation through regulatory and signaling interaction were used, for example, *MYC*'s position as a key regulator of highly ranked genes in BL would further increase its already significant score/rank.

## Materials and methods

The procedure is split into three distinct parts, as described in Box 1. The first part is the generation of the integrated BCI network. The second part is a phenotype analysis to identify dysregulated interactions. The third part is enrichment analysis and gene scoring. Benchmarking was performed against three B-cell lymphomas with known oncogenic lesions, and against CD40-stimulated Ramos cell line samples. The three steps are summarized below. A much more detailed description is available in the Supplementary Information.

### Network assembly

The BCI is a mixed-interaction network composed of protein–protein (PP) and protein–DNA (PD) interactions in a human B-cell context (Lefebvre *et al*, 2007). The former include both same-complex protein interactions and transient ones, such as those supporting signaling pathways. This network has since been enhanced (C Lefebvre *et al*, in preparation) to include additional post-translational interactions predicted by the MINDy algorithm (Wang *et al*, 2006). These interactions include those cases where the ability of a transcription factor (TF) to regulate its target(s) (T) is modulated by a third protein (M) (e.g., an activating kinase). The BCI is generated using 'gold-standard' evidences from curated databases, by applying a Naïve

Bayes classifier to integrate a large number of experimental and computational evidence. Evidence is drawn from several sources, including literature mining from GeneWays (Rzhetsky *et al*, 2004), transcription factor-binding motif enrichment, orthologous interactions from model organisms, and reverse engineering algorithms, including ARACNe (Basso *et al*., 2005; Margolin *et al*, 2006) and MINDy for regulatory and post-translational interactions, respectively. A likelihood ratio (LR) for each evidence source was generated using the positive and negative gold-standard sets. Individual LRs are then combined into a global LR for each interaction. A threshold corresponding to a posterior probability $P \geqslant 0.5$ was used to qualify interactions as present or absent. See the Supplementary Information for full details of the method.

### Dysregulation analysis

Analysis was performed using a large compendium of over 200 microarray expression profiles in B cells (BCGEP), including primary tissue as well as cell line samples, available in the NIH Gene Expression Omnibus (GSE2350). Samples in this set were hybridized to the Affymetrix HG-U95Av2 GeneChip®. After filtering for uninformative probes (those having less than a mean of 50 and a coefficient of variation less than 0.3 in the BCGEP), 7907 remained for analysis. Hierarchical clustering was performed to identify relatively homogeneous phenotype groups suitable for this analysis. The three benchmarking phenotypes used included BL (26 purified and unpurified samples), FL (six purified samples), and MCL (eight purified samples). Other phenotypes represented in this data set included germinal center (GC), naïve (N), memory (M), CLL from mutated (CLL-mut) and unmutated (CLL-unmut) subsets, diffuse large B-cell lymphoma (DLCL), and primary effusion lymphoma (PEL). A list of the analyzed cancerous phenotypes can be seen in Table II. For the CD40 perturbation analysis, a set of 24 CD40-stimulated Ramos cell line samples was used against a background of 43 Ramos samples.

For each phenotype, each BCI interaction was analyzed in sequence to determine if it could be classified as either a GoC, LoC, or no change (NC). The test was based on the estimate of the MI between the expression profiles of the two genes in the interaction. MI is an information theoretic measure of statistical dependence, which is zero if and only if two variables are statistically independent. It was calculated using Gaussian kernel estimation (Margolin *et al*, 2006). Specifically, we tested whether the MI increased (LoC) or decreased (GoC) when the samples corresponding to the specific phenotype were removed from the entire compendium (used to compute the background MI). A null distribution was computed to assess the statistical significance of an MI change as a function of the background MI and of the number of removed samples. More detailed interpretations of LoC and GoC events are shown in Box 1. See full details in the Supplementary Information.

### Scoring

Genes were scored by the enrichment of their direct network neighborhood in GoC/LoC interactions, using a Fisher' exact test. Specifically, for both LoC and GoC, two partial *P*-values were separately computed, based on the number of dysregulated interactions a gene was directly involved in or it was modulating within its direct neighborhood. A global *P*-value was computed as the product of all four partial *P*-values. All scoring totals can be seen in the Supplementary Information, where the score is the negative log of the global *P*-value.

### Benchmarking

We benchmarked the performance of this approach using three well-annotated lymphoma phenotypes, where the oncogenic lesion is reported in the literature. These are BL (*MYC*), FL (*BCL2*), and MCL (*BCL1/CCND1*). The results of our analysis were compared with conventional differential expression analysis using a *t*-test. Each *t*-test was computed using log2-transformed data and taking each phenotype against its normal counterpart (BL/GC, FL/GC, and MCL/N + M), applying Welch correction for sample sets of different size.

This approach was also run against Ramos cell line samples, where the *CD40*-signaling pathway had been biochemically perturbed (either by co-culturing with *CD40* ligand-producing fibroblasts, or using a CD40-specific antibody). Enrichment was calculated for the top scoring genes against a reference set of 41 CD40-signaling pathway genes using GSEA (Subramanian *et al*, 2005). This reference set was generated using two *CD40* sets available at the Molecular Signatures Database, or MSigDB, available with GSEA (http://www.broad.mit.edu/gsea/msigdb/). These results were also compared with differential expression analysis (same procedure as above, with *CD40*-stimulated against unstimulated). Enrichment of the top 25 genes in both cases was calculated via a Fisher's exact test.

Network visualization was also performed to create disease modules based on the top scoring genes in each phenotype. These visualizations were produced using the Cytoscape software package (http://www.cytoscape.org/) (Shannon *et al*, 2003). Enrichment of specific cellular pathways was computed using GSEA on the top-ranked list of probes in each phenotype, and compared with these visualizations.

### Supplementary information

Supplementary information is available at the *Molecular Systems Biology* website (www.nature.com/msb).

## Acknowledgements

This work was supported by the NCI (R01CA109755), the NIAID (R01AI066116), and the National Centers for Biomedical Computing NIH Roadmap initiative (U54CA121852). KMM is supported by the NLM Informatics Research Training Program (5 T15 LM007079-15).

## References

Adler AS, Lin M, Horlings H, Nuyten DS, van de Vijver MJ, Chang HY (2006) Genetic regulators of large-scale transcriptional signatures in cancer. *Nat Genet* **38:** 421–430

Bellan C, Lazzi S, De Falco G, Nyongo A, Giordano A, Leoncini L (2003) Burkitt's lymphoma: new insights into molecular pathogenesis. *J Clin Pathol* **56:** 188–192

Bende RJ, Smit LA, van Noesel CJ (2007) Molecular pathways in follicular lymphoma. *Leukemia* **21:** 18–29

Bild AH, Yao G, Chang JT, Wang Q, Potti A, Chasse D, Joshi MB, Harpole D, Lancaster JM, Berchuck A, Olson Jr JA, Marks JR, Dressman HK, West M, Nevins JR (2006) Oncogenic pathway signatures in human cancers as a guide to targeted therapies. *Nature* **439:** 353–357

di Bernardo D, Thompson MJ, Gardner TS, Chobot SE, Eastwood EL, Wojtovich AP, Elliott SJ, Schaus SE, Collins JJ (2005) Chemogenomic profiling on a genome-wide scale using reverse-engineered gene networks. *Nat Biotechnol* **23:** 377–383

Downward J (2006) Cancer biology: signatures guide drug choice. *Nature* **439:** 274–275

Ergun A, Lawrence CA, Kohanski MA, Brennan TA, Collins JJ (2007) A network biology approach to prostate cancer. *Mol Syst Biol* **3:** 82

Gardner TS, di Bernardo D, Lorenz D, Collins JJ (2003) Inferring genetic networks and identifying compound mode of action via expression profiling. *Science* **301:** 102–105

Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD, Lander ES (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* **286:** 531–537

Hamosh A, Scott AF, Amberger J, Valle D, McKusick VA (2000) Online Mendelian Inheritance In Man (OMIM). *Hum Mutat* **15:** 57–61

Heider U, Kaiser M, Sterz J, Zavrski I, Jakob C, Fleissner C, Eucker J, Possinger K, Sezer O (2006) Histone deacetylase inhibitors reduce VEGF production and induce growth suppression and apoptosis in human mantle cell lymphoma. *Eur J Haematol* **76:** 42–50

Kanehisa M, Goto S, Hattori M, Aoki-Kinoshita KF, Itoh M, Kawashima S, Katayama T, Araki M, Hirakawa M (2006) From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res* **34:** D354–D357

Lage K, Karlberg EO, Storling ZM, Olason PI, Pedersen AG, Rigina O, Hinsby AM, Tumer Z, Pociot F, Tommerup N, Moreau Y, Brunak S (2007) A human phenome–interactome network of protein complexes implicated in genetic disorders. *Nat Biotechnol* **25:** 309–316

Lamb J, Crawford ED, Peck D, Modell JW, Blat IC, Wrobel MJ, Lerner J, Brunet JP, Subramanian A, Ross KN, Reich M, Hieronymus H, Wei G, Armstrong SA, Haggarty SJ, Clemons PA, Wei R, Carr SA, Lander ES, Golub TR (2006) The connectivity map: using gene-expression signatures to connect small molecules, genes, and disease. *Science* **313:** 1929–1935

Lefebvre C, Lim WK, Basso K, Dalla-Favera R, Califano A (2007) A context-specific network of protein–DNA and protein–protein interactions reveals new regulatory motifs in human B cells. *Lect Notes Bioinform (LNCS)* **4532:** 42–56

Li Z, Van Calcar S, Qu C, Cavenee WK, Zhang MQ, Ren B (2003) A global transcriptional regulatory role for c-Myc in Burkitt's lymphoma cells. *Proc Natl Acad Sci USA* **100:** 8164–8169

Margolin AA, Nemenman I, Basso K, Wiggins C, Stolovitzky G, Dalla Favera R, Califano A (2006) ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics* **7** (Suppl 1)**:** S7

Miranda RN, Briggs RC, Kinney MC, Veno PA, Hammer RD, Cousar JB (2000) Immunohistochemical detection of cyclin D1 using optimized conditions is highly specific for mantle cell lymphoma and hairy cell leukemia. *Mod Pathol* **13:** 1308–1314

Mullighan CG, Goorha S, Radtke I, Miller CB, Coustan-Smith E, Dalton JD, Girtman K, Mathew S, Ma J, Pounds SB, Su X, Pui CH, Relling MV, Evans WE, Shurtleff SA, Downing JR (2007) Genome-wide analysis of genetic alterations in acute lymphoblastic leukaemia. *Nature* **446:** 758–764

Munoz O, Fend F, de Beaumont R, Husson H, Astier A, Freedman AS (2004) TGFbeta-mediated activation of Smad1 in B-cell non-Hodgkin's lymphoma and effect on cell proliferation. *Leukemia* **18:** 2015–2025

Perez OD, Nolan GP (2002) Simultaneous measurement of multiple active kinase states using polychromatic flow cytometry. *Nat Biotechnol* **20:** 155–162

Reddy K, Satyadev R, Bouman D, Hibbard MK, Lu G, Paolo R (2006) Burkitt t(8;14)(q24;q32) and cryptic deletion in a CLL patient: report of a case and review of literature. *Cancer Genet Cytogenet* **166:** 12–21

Ren B, Robert F, Wyrick JJ, Aparicio O, Jennings EG, Simon I, Zeitlinger J, Schreiber J, Hannett N, Kanin E, Volkert TL, Wilson CJ, Bell SP, Young RA (2000) Genome-wide location and function of DNA binding proteins. *Science* **290:** 2306–2309

Rzhetsky A, Iossifov I, Koike T, Krauthammer M, Kra P, Morris M, Yu H, Duboue PA, Weng W, Wilbur WJ, Hatzivassiloglou V, Friedman C (2004) GeneWays: a system for extracting, analyzing, visualizing, and integrating molecular pathway data. *J Biomed Inform* **37:** 43–53

Schena M, Shalon D, Davis RW, Brown PO (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* **270:** 467–470

Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* **13:** 2498–2504

Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci USA* **102:** 15545–15550

Tomlins SA, Rhodes DR, Perner S, Dhanasekaran SM, Mehra R, Sun XW, Varambally S, Cao X, Tchinda J, Kuefer R, Lee C, Montie JE, Shah RB, Pienta KJ, Rubin MA, Chinnaiyan AM (2005) Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. *Science* **310:** 644–648

van 't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AA, Mao M, Peterse HL, van der Kooy K, Marton MJ, Witteveen AT, Schreiber GJ, Kerkhoven RM, Roberts C, Linsley PS, Bernards R, Friend SH (2002) Gene expression profiling predicts clinical outcome of breast cancer. *Nature* **415:** 530–536

Wang K, Banerjee N, Margolin AA, Nemenman I, Califano A (2006) Genome-wide discovery of modulators of transcriptional interactions in human B lymphocytes. *Lect Notes Comput Sci* **3909:** 348–362

Yao J, Weremowicz S, Feng B, Gentleman RC, Marks JR, Gelman R, Brennan C, Polyak K (2006) Combined cDNA array comparative genomic hybridization and serial analysis of gene expression analysis of breast tumor progression. *Cancer Res* **66:** 4065–4078

Zhang XY, DeSalle LM, Patel JH, Capobianco AJ, Yu D, Thomas-Tikhonenko A, McMahon SB (2005) Metastasis-associated protein 1 (MTA1) is an essential downstream effector of the c-MYC oncoprotein. *Proc Natl Acad Sci USA* **102:** 13968–13973