Supplementary Information

A Human B Cell Interactome Identifies MYB and FOXM1 as Master Regulators of Proliferation in Germinal Centers

Celine Lefebvre[1,2,3], Presha Rajbhandari[1,2,3], Mariano J. Alvarez[1,2,3], Pradeep Bandaru[1,2,3], Wei Keat Lim[1,4,#], Mai Sato[5,#], Kai Wang[1,4,#], Pavel Sumazin[1,2,3], Manjunath Kustagi[1,2,3], Brygida C. Bisikirska[1,2,3], Katia Basso[5], Pedro Beltrao[6], Nevan Krogan[6], Jean Gautier[5], Riccardo Dalla-Favera[5,7,8], and Andrea Califano[1,2,3,4,5,*]

[1] Center for Computational Biology and Bioinformatics,

[2] Joint Centers for Systems Biology,

[3] Columbia Initiative in Systems Biology,

[4] Department of Biomedical Informatics,

[5] Institute for Cancer Genetics and the Herbert Irving Comprehensive Cancer Center, Columbia University, New York, NY 10032, USA

[6] Department of Cellular and Molecular Pharmacology, University of California at San Francisco, San Francisco, CA, 94158, USA

[7] Department of Pathology & Cell Biology,

[8] Department of Genetics & Development, Columbia University, New York, NY 10032, USA

[#] Present Affiliation: Therasis, Inc., 462 First Avenue, Suite 908, New York, New York 10016, USA (WKL), Kumamoto University 39-1, Kurokami 2-chome, Kumamoto (MS), Pfizer, Inc, San Diego, CA 92121, USA (KW)

* To whom correspondence should be addressed. E-mail: califano@c2b2.columbia.edu

Table of Content

Supplementary Methods

Supplementary Figure 1

Supplementary Figure 2

Supplementary Figure 3

Supplementary Figure 4

Supplementary Figure 5

Supplementary Figure 6

Supplementary Figure 7

Supplementary Figure 8

Supplementary Figure 9

Supplementary Figure 10

Supplementary Figure 11

Supplementary Figure 12

Supplementary Figure 13

Supplementary Figure 14

Supplementary Figure 15

Supplementary Table IV

Supplementary Table VII

Supplementary Table VIII

Supplementary Table IX

Supplementary Table X

Supplementary Methods

The Human B Cell Interactome (HBCI)

Bayesian Evidence Integration Approach (BEIA): We used a Naïve Bayes Classification in which the posterior probability of a specific interaction is computed using the prior of that class of interactions and the product of the Likelihood Ratio (LR) of the individual clues supporting it. Computing priors and likelihoods requires large datasets of both positive and negative examples (i.e. interactions that are respectively known to exist and not to exist). These are called Gold Standard Positive and Negative sets (GSP and GSN respectively) for both Protein-Protein Interactions (PPIs) and Protein-DNA Interactions (PDIs). Each evidence source is represented as categorical data (continuous values are binned as necessary) and used to compute likelihood ratios using GSP and GSN interactions counts. The Bayesian evidence integration model applies the Bayes theorem to compute the posterior odds that a specific interaction exists (O_{post}) as the product of the prior odds (O_{prior}) and of a likelihood ratio (LR):

$$O_{post} = LR \times O_{prior} .$$

The prior odds are defined as the average odds that two random gene products are involved in an interaction and can be calculated as:

$$O_{prior} = \frac{P(I)}{P(\bar{I})}$$

where $P(I)$ is the probability that two random gene products are involved in an interaction and $P(\bar{I})$ is the probability that they are not. The posterior odds of a specific interaction is defined as the ratio of the probabilities that two specific gene products, g_x and g_y, are respectively

involved or not involved in an interaction, conditional to the presence of N different clues, $c_1 ... c_N$:

$$O_{post} = \frac{P\left(I_{xy} \middle| c_1 ... c_N\right)}{P\left(\overline{I_{xy}} \middle| c_1 ... c_N\right)}$$

Similarly, the LR is defined as:

$$LR\left(c_1 ... c_N\right) = \frac{P\left(c_1 ... c_N \middle| I_{xy}\right)}{P\left(c_1 ... c_N \middle| \overline{I_{xy}}\right)}$$

In the Naïve Bayes Classifier (NBC) model, the clues are assumed to be statistically independent. Then, the LR can be computed as the product of individual LRs from the respective datasets:

$$LR\left(c_1 ... c_N\right) = \prod_{i=1}^{i=N} LR\left(c_i\right) = \prod_{i=1}^{i=N} \frac{P\left(c_i \middle| I_{xy}\right)}{P\left(c_i \middle| \overline{I_{xy}}\right)}$$

A useful property of the NBC model is that performance does not significantly deteriorate if weak dependencies among the clues exist. Under this assumption, the posterior odds of a specific interaction can be calculated as:

$$O_{post} = \prod_{i=1}^{i=N} \frac{P\left(c_i \middle| I_{xy}\right)}{P\left(c_i \middle| \overline{I_{xy}}\right)} \times O_{prior}$$

O_{prior} can be estimated from prior knowledge on the number of expected interactions in a cellular context, while the LRs are estimated by counting how many times a specific clue is

observed in a positive and a negative gold standard set. O_{post}, computed as the product of these two values, is related to the probability of an interaction to be true as $P_{post} = O_{post} / (O_{post} + 1)$, then achieving a posterior probability of at least 50% is equivalent to achieve $O_{post} \geq 1$ or $LR \geq 1 / O_{prior}$.

Gold Standard sets: To generate a GSP for PPIs, we extracted 27,568 human PPIs from HPRD (Peri *et al*, 2003), 4,430 from BIND (Bader *et al*, 2003), and 3,522 from IntAct (Hermjakob *et al*, 2004), originating from low-throughput, high quality experiments. This resulted in a GSP set of 28,554 unique PPIs involving 7,826 genes (after homodimers removal). Generating the GSN is somewhat more complicated because negative interaction examples are not easily identified from the literature. Thus, as described previously (Lefebvre *et al*, 2007), the GSN was defined as gene pairs encoding proteins in distinct cell compartment. This resulted in a set of 16,411,614 candidate non interacting gene pairs. In order to keep a realistic proportion between the GSP and the GSN, we extracted the negative pairs involving genes from the GSP, resulting in 5,362,594 negative gene pairs. We split the GSP and GSN in two sets, further used as training and testing sets. The GSP was split in one set of 20,000 interactions used for post-processing datasets and training the evidences and the remaining 8,554 for testing the performance of the classifier. We defined that the prior odds for an interaction is approximately 1 in 800 based on previous estimates of the total number of PPIs in a human cell of ~300,000 among 22,000 proteins (Hart *et al*, 2006; Rual *et al*, 2005). This implies that any protein pair having $LR \geq 800$, after evidence integration, has at least a 50% probability of being involved in a PPI.

To generate the GSP for PDIs, we extracted human interactions from the Transfac® Professional (TRANSFAC) (Matys *et al*, 2003), BIND and Myc (MycDB) databases (Zeller *et*

al, 2003), selecting interactions involving genes expressed in B cells only. This resulted in a GSP PDI set of 1,746 interactions involving 168 TFs and 1,138 targets. The GSP was split in two sets: one set of 1,115 interactions from the TRANSFAC and Myc databases for training the NBC, and an independent set of 631 interactions from the BIND and Myc databases, for testing the performance of the classifier. For defining a GSN, we randomly generated 100,000 gene pairs composed of a TF and a target, excluding pairs where the two genes are involved in a GSP interaction or in the same biological process as defined by Gene Ontology. We created another random set of 50,000 interactions as a testing GSN set, independent of the training GSN set. Instead of using fixed prior odds as for PPIs, we defined TF-specific prior odds. It has been previously demonstrated that the number of targets by TF can be approximated by a power-law distribution (Basso *et al*, 2005; Yu *et al*, 2006), and we argue here that the algorithm ARACNe (Margolin *et al*, 2006a), an information-theoretic method for identifying transcriptional interactions between gene products using microarray expression profile data, identifies a subset of these targets with a false negative rate not dependent on the TF. Then the expected number of targets in B cells for one TF can be approximated by the number of targets identified by ARACNe. LRs are directly computed from the number of targets for each TF in ARACNe with the requirement that the minimum LR is 5.

Evidence Integration: For predicting PPIs, the following evidence sources were used: (a) molecular interactions from the databases IntAct (Hermjakob *et al*, 2004), BIND(Bader *et al*, 2003) and MIPS (Mewes *et al*, 2006) for four eukaryotic organisms (fly, mouse, worm, yeast), (b) human high-throughput screens (Ewing *et al*, 2007; Rual *et al*, 2005; Stelzl *et al*, 2005), (c) the GeneWays literature datamining algorithm (Rzhetsky *et al*, 2004), (d) the Gene Ontology

(GO) biological process annotations (Ashburner *et al*, 2000), (e) co-expression data from a collection of 254 human B cell GEPs (Wang *et al*, 2006), and (f) Interpro protein domain annotations (Mulder *et al*, 2007). A total of 10,404 PPIs (between 2,675 proteins) were inferred, representing either direct physical interactions or same-complex membership. An additional 12,150 PPIs (between 3,433 proteins) were included from the GSP and from the KEGG database (Kanehisa and Goto, 2000). We also included 1,951 pathway-related PPIs predicted by the MINDy algorithm (Wang *et al*, 2006) between 427 proteins (signaling or co-TF) and 249 TFs. MINDy predicts either direct (physical) interactions or indirect ones (pathway-mediated), such as a receptor affecting the turnover of a TF through a signaling pathway.

For predicting PDIs, the following evidence sources were used: (a) mouse interactions from the TRANSFAC and BIND databases (b) human PDIs inferred by the ARACNe (Margolin *et al*, 2006a) and MINDy (Wang *et al*, 2006) algorithms using a collection of 254 Human B cell GEPs, (c) TF binding sites (TFBS) identified in the promoter of target genes (Smith *et al*, 2006), (d) target gene conditional co-expression and (e) GeneWays. We inferred 40,442 PDIs, representing physical TF-target interactions (between 296 TFs and 5,441 targets). Additional interactions were included from the GSP, for a total of 41,728 PDIs.

Gene expression Profiles: Gene expression profiles were collected using the Affymetrix HG-U95Av2 GeneChip® System (approximately 12,600 probe sets). Expression measurements were normalized using MAS5.0, and probe sets with absolute expression mean < 50 and coefficient of variation < 0.3 were considered non-informative and were excluded *a priori* from the analysis. We computed the mutual information (MI), an optimal measure of statistical dependence in a non linear setting, between the 7,349 probes as well defined, corresponding to 6,010 unique

"expressed genes". After applying a threshold (MI >= 0.069), corresponding to a p-value of 10^{-7}, we identified 4,539,340 statistically significant MIs between the 6,010 genes. The highest MI among all the probe-set pairs corresponding to a gene pair was used when multiple probes were present in the set. Among those, 2,561,919 gene pairs showed a positive *Spearman* correlation coefficient. It has been established that some interacting proteins, especially those in stable complexes, tend to be co-expressed (Jansen *et al*, 2003). Thus co-expression in a large expression profile dataset can provide clues about PPIs. We classified the human gene pairs having a positive correlation coefficient, according to their mutual information (MI).

Gene Ontology biological process annotation: It was observed that interacting proteins tend to share the same biological process (Alterovitz *et al*, 2006; Jansen *et al*, 2003). Thus, GO annotations provide additional clues about a PPI. As previously described (Alterovitz *et al*, 2006), we computed the information content of a gene ontology term as follow:

$$I(go_n) = -\log_2 \frac{k(go_n)}{\bigcup_{i=1}^{m} k(go_i)}$$

where go_n represent a GO term, $k(go_n)$ the gene set annotated by go_n and m the number of annotations in the biological process ontology. We classified human gene pairs sharing a biological process annotation using the information content of each GO term, retaining the highest value in case of multiple annotations.

***Orthologous interactions*:** We extracted putative PPIs from IntAct (Hermjakob *et al*, 2004) and BIND (Bader *et al*, 2003) for the three model organisms *Caenorhabditis elegans*, *Drosophila melanogaster* and *Mus musculus* and from IntAct, BIND and MIPS (Mewes *et al*, 2006) for *Saccharomyces cerevisiae*. We defined four different groups of predicted PPIs, one for each

organism, by mapping model organisms' genes to human genes using the Inparanoid database that describes eukaryotic orthologous clusters (O'Brien *et al*, 2005). As these four sources contain redundant information, we chose to combine them in one non-redundant source by classifying each interaction according to the number of evidence sources (organisms) supporting them (from 1 to 4) for computing the LRs.

***Human high-throughput interactions*:** Large-scale human PPIs mapping were recently conducted, including one mass spectrometry scanning (Ewing *et al*, 2007) and two yeast two-hybrid studies (Rual *et al*, 2005; Stelzl *et al*, 2005). We classified each interaction according to the number of studies reporting it (1 or 2, no interaction being common to the three high-throughput studies).

Protein domain annotation: PPIs are mediated by protein domain-domain interactions, therefore, it is possible to infer PPIs from the domain structure of two proteins and the ability of these two domains to interact (Liu *et al*, 2005). We used Interpro to define the protein structures and computed enrichment in domain pairs in known PPIs. Gene pairs were classified using $-log_{10}$ of the p-value computed with a Fisher Exact test.

GeneWays literature datamining algorithm: GeneWays is a computer system designed for automatic analysis of literature data to extract knowledge about molecular interactions. It provides a list of gene pairs associated with a keyword (action), defining the interaction type, and a score (between -1 and 1). By studying the action keyword enrichment for PPIs in the GSP, which were also reported by GeneWays, we identified 19 action keywords associated with PPIs. These include the following: *assemble, associate, bind, coexpress, coimmunoprecipitate, colocalize, connect, coprecipitate, copurify, dephosphorylate, dissociate from, form, form a*

complex, immunoprecipitate, interact, recruit, required for, synergize and ubiquitinate. Enrichment was computed with a fisher exact test ($p \leq 10^{-3}$). The same analysis was performed for PDIs in the GSP (TRANSFAC and BIND). We identified 12 action keywords associated with PDIs, including*: activate, depend on, include, independence, influence, mediate, regulate, repress, transactivate* and *upregulate*. GeneWays interactions were extracted using these lists and further classified in two groups according to their score ($s \leq 0$ and $s > 0$, respectively).

Transcription Factor classification*:* To identify human transcription factors (TFs), we selected the human genes annotated as "transcription factor activity" in Gene Ontology and the list of TFs from TRANSFAC. From this list, we removed general TFs (e.g. stable complexes like polymerases or TATA-box-binding proteins), and added some TFs not annotated by GO, producing a final list of 1,650 TFs, from which 555 were present on the filtered microarray gene set.

ARACNe is an information-theoretic method for identifying transcriptional interactions between gene products using microarray expression profile data (Basso *et al*, 2005; Margolin *et al*, 2006a; Margolin *et al*, 2006b). ARACNe has proven effective in predicting targets of specific TFs (e.g. c-MYC and NOTCH1) that were experimentally validated (Basso *et al*, 2005; Palomero *et al*, 2006). We used the bootstrapping version of ARACNe (Margolin *et al*, 2006b) to predict the targets of the 555 TFs, that were further classified and binned according to the MI shared between the TF and the corresponding target. Note that in the case of interactions predicted by ARACNe and supported by a transcription factor binding sites prediction, we computed the posterior odds based on the extended available biochemical validation ($P_{post} = 67\%$).

MINDy is an algorithm for the prediction of modulators of transcriptional interactions (Wang *et al*, 2006). MINDy predicts post-transcriptional modifications of TFs in the form of 3-ways interactions involving a TF, a target and a modulator of the TF-target interaction. We split these 3-ways interactions into two distinct gene pair interactions: a PDI between the TF and its target that will be further used for PDIs evidence integration and a TF-modulator interaction. These latter interactions can describe a direct interaction between the modulator and the TF, as well as an indirect interaction, when the modulator is indirectly affecting the TF through a cascade of events. TF-target interactions predicted by MINDy were classified with the conditional mutual information between the TF and the targets, keeping the best MI per interaction given the level of expression of the modulator. The TF-modulator interactions were classified with the number of target(s) for each TF a modulator affects and we kept in the HBCI the pairs involving more than 15 targets based on true modulator enrichment for the TF MYC (data not shown).

Mouse interactions*:* We extracted mouse PDIs from the TRANSFAC and BIND databases and used the Inparanoid database to predict human PDIs, selecting the genes associated to a cluster with a score of 1 only.

Transcription factor binding sites (TFBS): Promoters of target genes were defined as sequences from -1,000bp to +200bp around the transcription start site and were retrieved from UCSC Golden Path (hg18) (Karolchik *et al*, 2008). TFBS were identified using motifs represented as position weight matrices that are extracted from the vertebrate subset of Transfac 10.4. Putative binding sites occurrences are scored using a log-likelihood function (Hertz and Stormo, 1999; Smith *et al*, 2006) and TF-target pairs were classified using the log of the sum of the exponential of this score, in order to account for multiple binding sites occurrences in a promoter. When

different matrices mapped to the same TF, we kept the best score for the corresponding TF-target pair.

Target gene conditional co-expression: We used the simple hypothesis that co-expressed genes are likely to be regulated by the same TF, as it has been previously proposed for predicting yeast PDIs (Beyer *et al*, 2006). Therefore, knowing that a particular TF regulates a target gene g_x and that g_x is co-expressed with g_y, we can hypothesize that the same TF regulates g_x and g_y. We created a list of TF-target interactions using the GSP training interactions and the gene expression profiles in B cells, classifying the interactions according to the MI between the known and the predicted target. Note that we excluded interactions also predicted by ARACNe or MINDy from this set to avoid overfitting as these sources are based on the same gene expression profiles.

POU2AF1: POU2AF1 is a transcriptional coactivator and was not present in our list of transcription factors. However, POU2AF1 was experimentally identified as a master regulator of germinal center and could be considered as a transcription factor for this analysis in order to test it. For that, we used information about POU2F1 and POU2F2 transcription factors which are the known partners of POU2AF1 (Lins *et al*, 2003) (POU2AF1 binds to the POU domain of POU2F1 and POU2F2) to infer POU2AF1 targets with the BEIA.

Performance Analysis. For evaluating the performance of the classifiers, we need to train and test the classifier with independent datasets. 10 fold-cross validation is a good choice as we can test our classifier by covering the entire gold standard set. It is possible to do that if the gold standard set is independent from the evidences used. For PDI evidence integration, we use

TRANSFAC and BIND databases as our gold standard set and we also use transcription factor binding site information derived from TRANSFAC interactions. Therefore, we think there is a risk that TFBS and gold-standard TRANSFAC information introduce a bias. Thus, we used gold-standard PDIs extracted from BIND and not present in TRANSFAC to test the algorithm, and leave-one out cross-validation for estimating the performance of the classifier. We report results of leave-one out cross validation for PDI and PPI classification but it should be noted that for PPI, 10-fold cross validation produced the same results. Cross validation analysis confirmed that the integrated analysis outperforms individual methods and could recover 24% (PPI) and 43% (PDI) of interactions in the GSPs, with a corresponding False Positive Rate lower than 0.12% (PPI) and 3.2% (PDI), see Supplementary Figure 1. As expected (Basso *et al*, 2005; Jeong *et al*, 2001), both PDIs and PPIs follow a scale-free distribution (Supplementary Figure 2). When considering a fixed False Discovery Rate (FDR) of 50%, BEIA-inferred PPIs and PDIs achieve 15% and 26% recall respectively. As expected, multiple evidence integration significantly outperforms the use of any evidence in isolation, indirectly validating that evidence sources used in this study are largely statistically independent. Of note, evidence from sources known to contain non-statistically independent data, e.g. target predicted by ARACNe and MINDy using the same gene expression profile data, was eliminated *a priori*. GeneWays infers interactions from literature data. It is thus difficult to effectively separate training and testing data. To evaluate GeneWays contribution, we repeated the analysis without GeneWays evidence. Results were evaluated only for PPI as the number of PDIs inferred by GeneWays was very small. Without GeneWays, fewer PPIs were predicted, achieving 6% recall. However, a significantly

better FPR of 0.06% was also achieved (Supplementary Figure 1). Thus, while GeneWays is an important evidence source, it also introduces some false positive results.

Finally, we compared the performance of the BEIA with the adaptive boosting algorithm adaboost. Adaboost (Schapire and Freund, 1996) is an adaptive boosting algorithm in which a strong classifier is produced by combining complementary weak classifiers, where each weak classifier is trained on the mistakes of the previous one. Individual evidences are used as weak classifiers, defined to have an error rate slightly better than 50%. Let $\langle (x_1, y_1), \ldots, (x_N, y_N) \rangle$ be a set of N labeled examples (NP positive examples and NN negative examples), D the weight distribution over the N examples, $h_t \to \{-1;1\}$ the weak classifier and T the number of boosting steps. The initial example distribution is given by $D_1(i) = \frac{1}{2NP}, \frac{1}{2NN}$. Then for t in 1…T:

1. Find weak classifier with smallest error given D_t
2. Choose α_t
3. Update example weights:

$$D_{t+1} = \frac{D_t(i) \times e^{-\alpha_t u_i}}{Z_t} \text{ where } u_i = y_i h_t(x_i) \text{ and } Z_t = \sum_i D_t(i) \times e^{-\alpha_t u_i}$$

This procedure ensures that the misclassified examples get a higher weight, "forcing" the next weak learner to concentrate on those examples. The output *H* of the strong classifier is given by a majority vote of the weak classifiers:

$$H(x) = sign\left(\sum_{t=1}^{T} \alpha_t h_t(x) \right)$$

As the different evidences concern only a subpopulation of the gene pairs, we used a variant of the original adaboost algorithm in which the weak classifier is given the possibility to abstain

(Schapire and Singer, 1999). In this variant, the range of each weak learner is now $\{-1,0,+1\}$ where 0 correspond to the case where the classifier cannot give a prediction. For a fixed t, we define W_0, W_{-1} (W_-) and W_{+1} (W_+) as $W_b = \sum_{i:u_i=b} D(i)$ for $b \in \{-1,0,+1\}$. Therefore, we can calculate Z_t as:

$$\begin{aligned} Z_t &= \sum_i D(i) e^{-\alpha_t u_i} \\ &= \sum_{b\in\{-1,0,1\}} \sum_{i:u_i=b} D(i) e^{-\alpha_t b} \\ &= W_0 + W_- \mathrm{e}^{\alpha} + W_+ \mathrm{e}^{-\alpha} \end{aligned}$$

Then it becomes possible to find the α_t minimizing Z_t analytically:

$$\alpha_t = \frac{1}{2}\ln\left(\frac{W_+}{W_-}\right) \text{ with } W_+ > W_-$$

For this setting of α_t, Z_t can be written as:

$$Z_t = W_0 + 2\sqrt{W_- W_+}$$

As it has been reported before (Lu *et al*, 2005), we found that boosting the Naïve Bayes classifiers doesn't help for predicting PPIs or PDIs with more accuracy (Supplementary Figure 3). However, it revealed the usefulness of each of the evidence sources for predicting PPIs or PDIs, reflected as the number of boosting iterations the evidence was used. Interestingly, this revealed that evidence sources are used in a more heterogeneous way for inferring PPIs than for PDIs, this later relying mostly on target gene conditional co-expression.

MAster Regulator INference algorithm (MARINa)

For each TF-target interaction in the HCBI, its membership in the positive or negative TF-regulon was determined by computing the TF-target Spearman correlation across all the samples from the gene expression profiles used to assemble the HBCI. When genes were identified by more than one Affymetrix probe in the array, we selected the probe showing the highest difference in expression between centroblast and naïve samples. This is a critical point when two probes mapping to the same gene have anti-correlated expression profiles, especially for the transcription factor. This is not a frequent case but a few TFs have conflicting probes that classify the TF differently (e.g. TP53 and E2F4). Note that we considered GC-activated and GC-repressed targets separately. In the following, TFs showing an increased activity will be referred as to TF_A and TFs with decreased activity as TF_R. GC-activated targets can only be activated targets of a TF_A, based on the hypothesis that the expression of these genes is dependent on a transcriptional activator. On the contrary, GC-repressed targets can be either repressed targets of a TF_A or activated targets of a TF_R, which is by definition no longer able to activate its targets. Therefore, we run MARINa twice. Note than 9 TF_R were not tested (DBP, ERF, FOSB, FOSL2, MLLT7, NRF1, TP53, USF2, YY1) as they were only enriched in their negative regulon. In the following section, we introduce the MARINa algorithm and the definitions of master regulator (MR) and shadow regulator (SR).

MARINa in 4 steps:

1. **Definition of positive and negative regulons**: each TF with more than 20 targets in the HBCI is assigned a positive and a negative regulon by computing the spearman

correlation between every TF-target pair using the B cell GEP used for building the HBCI.

2. **TF enrichment**: GSEA was used to assess the enrichment in GC-genes for each TF's regulon. As a reference, we used a list of genes ranked with the t-statistics obtained by comparing centroblast and naïve samples. This produced a list of 107 enriched TFs at p-value < 0.001 (Supplementary Table II). Results were classified using the DETOR score for Differentially Expressed Target Odds Ratio and computed as:

$$DETOR_{TF_i} = \left(GS_i^{LE} / RS_i^{LE}\right) \Big/ \left(GS_i / RS_i\right),$$

where GS_i^{LE} and RS_i^{LE} are the number of genes before the leading edge in GSEA for the gene set (or regulon) and the reference set while GS_i and RS_i are the sizes of the gene set and the reference set.

3. **Shadow analysis**: If TF_1 and TF_2 regulons, R_1 and R_2, overlap significantly and only TF_1 is enriched, TF_2 may also appear enriched because of common-target enrichment. In this case, TF_2 activity is a shadow of TF_1 activity and we call TF_2 a Shadow Regulator (SR). This can be easily detected, because GSEA enrichment restricted to non-common-targets ($R_2 \setminus R_1 \cap R_2$) will be significantly lower than for the full regulon, R_2.We first compute the GSEA Enrichment Score of the targets of TF_2 (ES_2) that are not targets of TF_1 ($R_2 \setminus R_1 \cap R_2$). We then compute ES scores for 1,000 random subsets of TF_2 of the same size as $R_2 \setminus R_1 \cap R_2$ if the remaining regulon has more than 20 targets. We compute the empirical p-value of observing an ES smaller than ES_2 if $ES_2 > 0$ and greater than ES_2 if $ES_2 < 0$. TF_2 is a shadow of TF_1 if p-value < 0.01 and if TF_1 is not a shadow of TF_2. Note

that we tested each pair in the two directions: TF1 → TF2 and TF2 → TF1, and defined TF2 as a shadow of TF1 only when TF1 was not a shadow of TF2 as well. Ambiguous cases where one direction could not be tested because of the size limitation were ignored. The results of the shadow analysis are reported in Supplementary Table III. GC Master Regulators (MRs) are defined as enriched TFs that are not shadows of any other TF.

4. **Synergy**: Alternatively, two TFs may have a synergistic effect on their common-targets ($R1 \cap R2$) due for instance to multiplicative kinetics. In that case, GSEA enrichment of $R1 \cap R2$ will be significantly higher than for both R_1 and R_2. We first select the pairs having a statistically significant overlap in target genes computed with a fisher exact test. Pairs' regulons were assembled as follow: if the two TFs were positively correlated, we defined a positive and a negative regulon by intersecting the positive and negative regulons of the single TFs ($R^{+}_{TF_1TF_2} = R^{+}_{TF1} \cap R^{+}_{TF2}$ and $R^{-}_{TF_1TF_2} = R^{-}_{TF1} \cap R^{-}_{TF2}$). If the TFs were anti-correlated, we intersected the positive regulon of one TF with the negative regulon of the second TF and vice versa ($R^{+}_{TF_1TF_2} = R^{+}_{TF1} \cap R^{-}_{TF2}$ and $R^{-}_{TF_1TF_2} = R^{-}_{TF1} \cap R^{+}_{TF2}$, positive and negative labels having no meaning for the pair). We then run GSEA with the regulon of pair TF1/TF2 as gene set and the union of the regulons of TF1 and TF2 as reference. The TF pairs tested included all pairs involving 2 MRs and MR-SR pairs from ambiguous cases defined in step 3. Synergistic pairs are defined as the one for which the GSEA enrichment is significant at a p-value threshold of 0.01 (Supplementary Table IV) and the final list of MRs contain all TFs participating in synergistic pairs.

MR inference stability

To show that the MR inference is not dependent on the inclusion of the naïve and centroblast samples for the network reconstruction, we have compared ARACNe and MR inference after using different subset of the gene expression profile. First, ARACNe was run with 240 samples, after exclusion of 5 naïve and 9 centroblast samples. The network obtained shared more than 85% of its predicted interactions with the original network. Moreover, as shown in Supplementary Figure 5, the TF-connectivity and target identity are minimally affected. Similar results were obtained after removing randomly 14 samples, indicating that the difference seen between the 2 networks is not specific to naïve and centroblast samples. We then reconstructed a B cell interactome using the same criteria of naïve and centroblast sample exclusion. MARINa recovers the same top master regulators with very similar rankings (Supplementary Figure 6), illustrating that the network reconstruction and MR inference is stable and not dependent on a small subset of the samples.

Mitotic and pre-replication protein complexes in yeast

Using the ortholog assignments from the Inparanoid database (http://inparanoid.sbc.su.se), 11 yeast orthologs of the 30 mitotic and replication initiation proteins in the module were identified. This set was expanded using their cognate co-complex members obtained from a curated list of *S. cerevisiae* complexes (Wang *et al*, 2009). In this way, 13 mitotic proteins belonging to the Cyclin/Cdc28 protein complexes and 16 proteins from the pre-RC and replication initiation complexes were selected for further analysis (Supplementary Table VII). Genetic interactions for *S. cerevisiae* were compiled from previous E-MAP studies (Collins *et al*, 2007; Schuldiner *et al*,

2005) and from the BioGRID database (www.thebiogrid.org). A total of 11 genetic interactions were detected between the Cyclin/Cdc28 and the replication initiation complex and 9 genetic interactions between the Cyclin/Cdc28 and the pre-RC (Supplementary Table VIII), corresponding to highly significant genetic associations (p-value = 9×10^{-3} and 5×10^{-3} respectively, based on random sampling of equal-size protein sets). Finally, the enrichment of genetic interactions between all combinations of *S. cerevisiae* protein complexes revealed that the mitosis related group was the fourth most significantly associated with the pre-RC complex (Supplementary Figure 14).

References

Alterovitz G, Xiang M, Mohan M, Ramoni MF (2006) GO PaD: the Gene Ontology Partition Database. *Nucleic Acids Res*.

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **25:** 25-29.

Bader GD, Betel D, Hogue CW (2003) BIND: the Biomolecular Interaction Network Database. *Nucleic Acids Res* **31:** 248-250.

Basso K, Margolin AA, Stolovitzky G, Klein U, Dalla-Favera R, Califano A (2005) Reverse engineering of regulatory networks in human B cells. *Nat Genet* **37:** 382-390.

Beyer A, Workman C, Hollunder J, Radke D, Moller U, Wilhelm T, Ideker T (2006) Integrated assessment and prediction of transcription factor binding. *PLoS Comput Biol* **2:** e70.

Collins SR, Miller KM, Maas NL, Roguev A, Fillingham J, Chu CS, Schuldiner M, Gebbia M, Recht J, Shales M, Ding H, Xu H, Han J, Ingvarsdottir K, Cheng B, Andrews B, Boone C, Berger SL, Hieter P, Zhang Z, Brown GW, Ingles CJ, Emili A, Allis CD, Toczyski DP, Weissman JS, Greenblatt JF, Krogan NJ (2007) Functional dissection of protein complexes involved in yeast chromosome biology using a genetic interaction map. *Nature* **446:** 806-810.

Ewing RM, Chu P, Elisma F, Li H, Taylor P, Climie S, McBroom-Cerajewski L, Robinson MD, O'Connor L, Li M, Taylor R, Dharsee M, Ho Y, Heilbut A, Moore L, Zhang S, Ornatsky O, Bukhman YV, Ethier M, Sheng Y, Vasilescu J, Abu-Farha M, Lambert JP, Duewel HS, Stewart, II, Kuehl B, Hogue K, Colwill K, Gladwish K, Muskat B, Kinach R, Adams SL, Moran MF, Morin GB, Topaloglou T, Figeys D (2007) Large-scale mapping of human protein-protein interactions by mass spectrometry. *Mol Syst Biol* **3:** 89.

Hart GT, Ramani AK, Marcotte EM (2006) How complete are current yeast and human protein-interaction networks? *Genome Biol* **7:** 120.

Hermjakob H, Montecchi-Palazzi L, Lewington C, Mudali S, Kerrien S, Orchard S, Vingron M, Roechert B, Roepstorff P, Valencia A, Margalit H, Armstrong J, Bairoch A, Cesareni G, Sherman D, Apweiler R (2004) IntAct: an open source molecular interaction database. *Nucleic Acids Res* **32:** D452-455.

Hertz GZ, Stormo GD (1999) Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics* **15:** 563-577.

Jansen R, Yu H, Greenbaum D, Kluger Y, Krogan NJ, Chung S, Emili A, Snyder M, Greenblatt JF, Gerstein M (2003) A Bayesian networks approach for predicting protein-protein interactions from genomic data. *Science* **302:** 449-453.

Jeong H, Mason SP, Barabasi AL, Oltvai ZN (2001) Lethality and centrality in protein networks. *Nature* **411:** 41-42.

Kanehisa M, Goto S (2000) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* **28:** 27-30.

Karolchik D, Kuhn RM, Baertsch R, Barber GP, Clawson H, Diekhans M, Giardine B, Harte RA, Hinrichs AS, Hsu F, Kober KM, Miller W, Pedersen JS, Pohl A, Raney BJ, Rhead B, Rosenbloom KR, Smith KE, Stanke M, Thakkapallayil A, Trumbower H, Wang T, Zweig AS, Haussler D, Kent WJ (2008) The UCSC Genome Browser Database: 2008 update. *Nucleic Acids Res* **36:** D773-779.

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. *Lecture Notes in Bioinformatics* **4532:** 42-56.

Lins K, Remenyi A, Tomilin A, Massa S, Wilmanns M, Matthias P, Scholer HR (2003) OBF1 enhances transcriptional potential of Oct1. *EMBO J* **22:** 2188-2198.

Liu Y, Liu N, Zhao H (2005) Inferring protein-protein interactions through high-throughput interaction data from diverse organisms. *Bioinformatics* **21:** 3279-3285.

Lu LJ, Xia Y, Paccanaro A, Yu H, Gerstein M (2005) Assessing the limits of genomic data integration for predicting protein networks. *Genome Res* **15:** 945-953.

Margolin AA, Nemenman I, Basso K, Wiggins C, Stolovitzky G, Favera D, Califano A (2006a) ARACNE: An Algorithm for the Reconstruction of Gene Regulatory Networks in a Mammalian Cellular Context. *BMC Bioinformatics* **7 Suppl 1:** S1-7.

Margolin AA, Wang K, Lim WK, Kustagi M, Nemenman I, Califano A (2006b) Reverse engineering cellular networks. *Nat Protocols* **1:** 663-672.

Matys V, Fricke E, Geffers R, Gossling E, Haubrock M, Hehl R, Hornischer K, Karas D, Kel AE, Kel-Margoulis OV, Kloos DU, Land S, Lewicki-Potapov B, Michael H, Munch R, Reuter I, Rotert S, Saxel H, Scheer M, Thiele S, Wingender E (2003) TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res* **31:** 374-378.

Mewes HW, Frishman D, Mayer KF, Munsterkotter M, Noubibou O, Pagel P, Rattei T, Oesterheld M, Ruepp A, Stumpflen V (2006) MIPS: analysis and annotation of proteins from whole genomes in 2005. *Nucleic Acids Res* **34:** D169-172.

Mulder NJ, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, Bork P, Buillard V, Cerutti L, Copley R, Courcelle E, Das U, Daugherty L, Dibley M, Finn R, Fleischmann W, Gough J, Haft D, Hulo N, Hunter S, Kahn D, Kanapin A, Kejariwal A, Labarga A, Langendijk-Genevaux PS, Lonsdale D, Lopez R, Letunic I, Madera M, Maslen J, McAnulla C, McDowall J, Mistry J, Mitchell A, Nikolskaya AN, Orchard S, Orengo C, Petryszak R, Selengut JD, Sigrist CJ, Thomas PD, Valentin F, Wilson D, Wu CH, Yeats C (2007) New developments in the InterPro database. *Nucleic Acids Res* **35:** D224-228.

O'Brien KP, Remm M, Sonnhammer EL (2005) Inparanoid: a comprehensive database of eukaryotic orthologs. *Nucleic Acids Res* **33:** D476-480.

Palomero T, Lim WK, Odom DT, Sulis ML, Real PJ, Margolin A, Barnes KC, O'Neil J, Neuberg D, Weng AP, Aster JC, Sigaux F, Soulier J, Look AT, Young RA, Califano A, Ferrando AA (2006) NOTCH1 directly regulates c-MYC and activates a feed-forward-loop transcriptional network promoting leukemic cell growth. *Proc Natl Acad Sci U S A* **103:** 18261-18266.

Peri S, Navarro JD, Amanchy R, Kristiansen TZ, Jonnalagadda CK, Surendranath V, Niranjan V, Muthusamy B, Gandhi TK, Gronborg M, Ibarrola N, Deshpande N, Shanker K, Shivashankar HN, Rashmi BP, Ramya MA, Zhao Z, Chandrika KN, Padma N, Harsha HC, Yatish AJ, Kavitha MP, Menezes M, Choudhury DR, Suresh S, Ghosh N, Saravana R, Chandran S, Krishna S, Joy M, Anand SK, Madavan V, Joseph A, Wong GW, Schiemann WP, Constantinescu SN, Huang L, Khosravi-Far R, Steen H, Tewari M, Ghaffari S, Blobe GC, Dang CV, Garcia JG, Pevsner J, Jensen ON, Roepstorff P, Deshpande KS, Chinnaiyan AM, Hamosh A, Chakravarti A, Pandey A (2003) Development of human protein reference database as an initial platform for approaching systems biology in humans. *Genome Res* **13:** 2363-2371.

Rual JF, Venkatesan K, Hao T, Hirozane-Kishikawa T, Dricot A, Li N, Berriz GF, Gibbons FD, Dreze M, Ayivi-Guedehoussou N, Klitgord N, Simon C, Boxem M, Milstein S, Rosenberg J, Goldberg DS, Zhang LV, Wong SL, Franklin G, Li S, Albala JS, Lim J, Fraughton C, Llamosas E, Cevik S, Bex C, Lamesch P, Sikorski RS, Vandenhaute J, Zoghbi HY, Smolyar A, Bosak S, Sequerra R, Doucette-Stamm L, Cusick ME, Hill DE, Roth FP, Vidal M (2005) Towards a proteome-scale map of the human protein-protein interaction network. *Nature* **437:** 1173-1178.

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.

Schapire RE, Freund Y (1996) Experiments with a new boosting algorithm. *Machine Learning* **Proceedings of the Thirteenth International Conference:** 148-156.

Schapire RE, Singer Y (1999) Improved Boosting Algorithms using confidence-rate predictions. *Machine Learning* **37:** 297-336.

Schuldiner M, Collins SR, Thompson NJ, Denic V, Bhamidipati A, Punna T, Ihmels J, Andrews B, Boone C, Greenblatt JF, Weissman JS, Krogan NJ (2005) Exploration of the function and organization of the yeast early secretory pathway through an epistatic miniarray profile. *Cell* **123:** 507-519.

Smith AD, Sumazin P, Xuan Z, Zhang MQ (2006) DNA motifs in human and mouse proximal promoters predict tissue-specific expression. *Proc Natl Acad Sci U S A* **103:** 6275-6280.

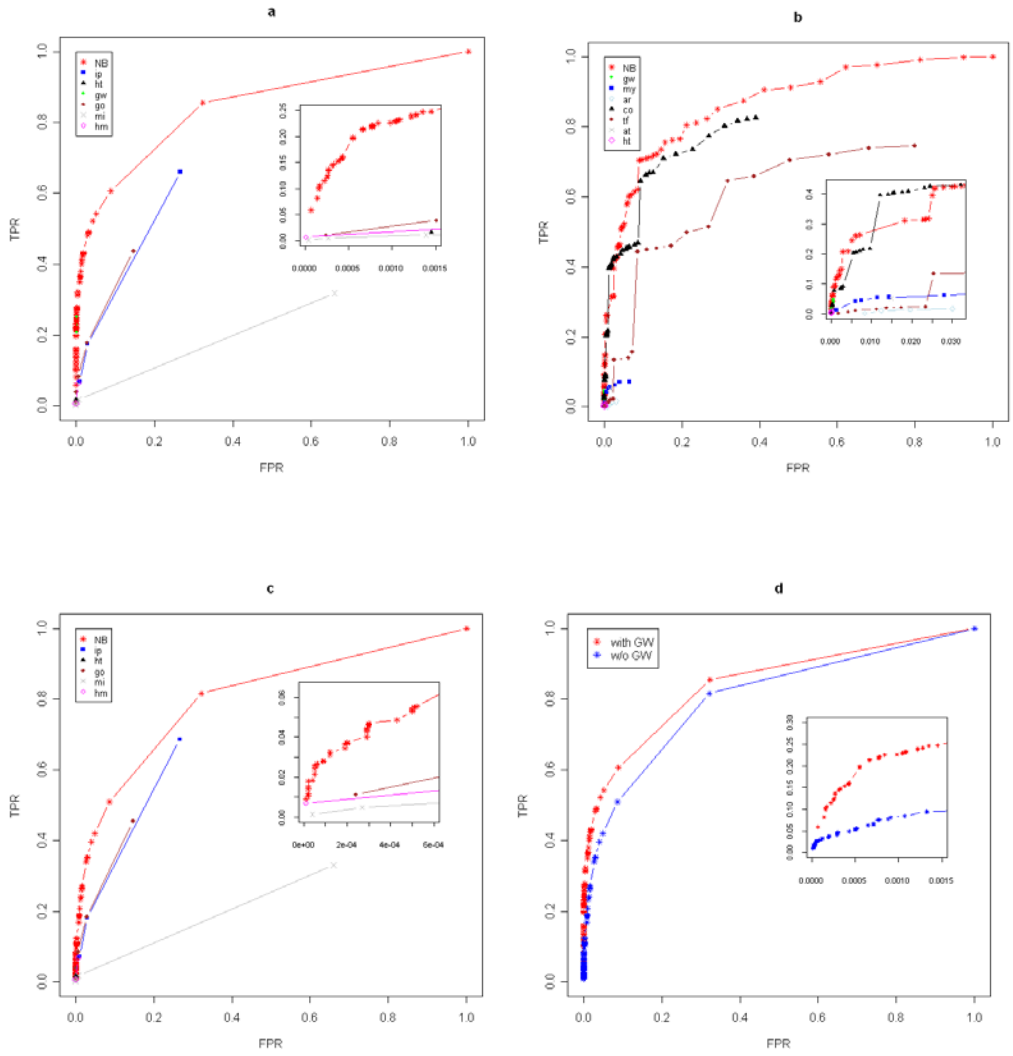
Stelzl U, Worm U, Lalowski M, Haenig C, Brembeck FH, Goehler H, Stroedicke M, Zenkner M, Schoenherr A, Koeppen S, Timm J, Mintzlaff S, Abraham C, Bock N, Kietzmann S, Goedde A, Toksoz E, Droege A, Krobitsch S, Korn B, Birchmeier W, Lehrach H, Wanker EE (2005) A human protein-protein interaction network: a resource for annotating the proteome. *Cell* **122:** 957-968.

Wang H, Kakaradov B, Collins SR, Karotki L, Fiedler D, Shales M, Shokat KM, Walther T, Krogan NJ, Koller D (2009) A complex-based reconstruction of the S. cerevisiae interactome. *Mol Cell Proteomics*.

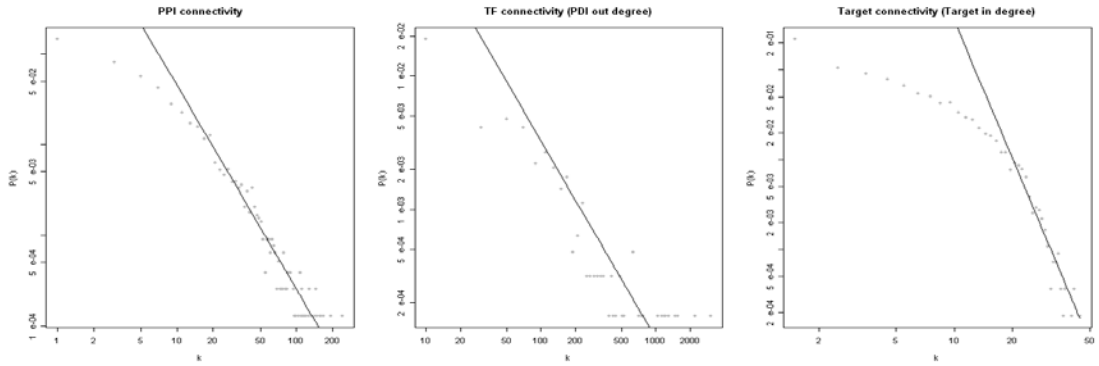
Wang K, Banerjee N, Margolin AA, Nemenman I, Califano A (2006) Genome-wide discovery of modulators of transcriptional interactions in human B lymphocytes. *Lecture Notes in Computer Science* **3909:** 348-362.

Yu H, Xia Y, Trifonov V, Gerstein M (2006) Design principles of molecular networks revealed by global comparisons and composite motifs. *Genome Biol* **7:** R55.

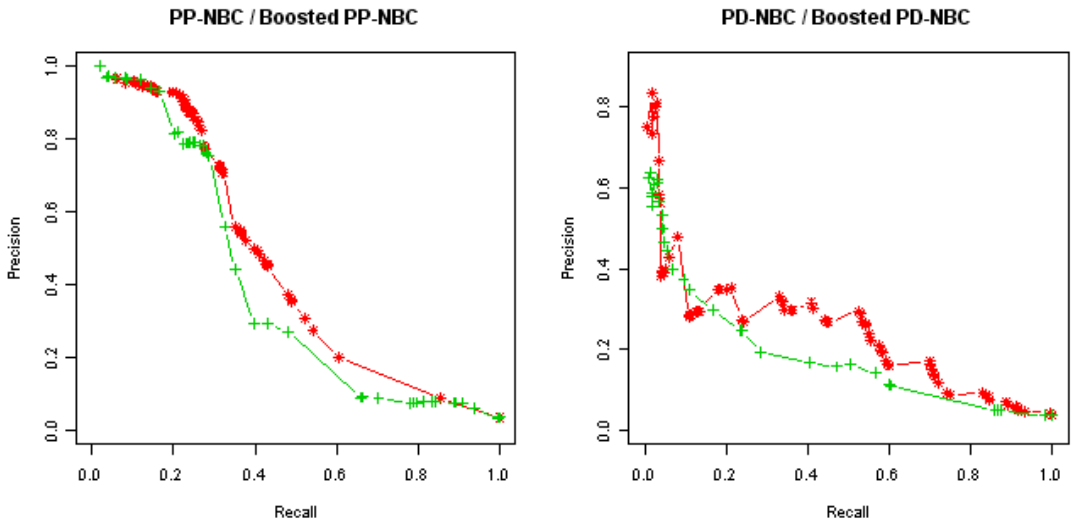
Zeller KI, Jegga AG, Aronow BJ, O'Donnell KA, Dang CV (2003) An integrated database of genes responsive to the Myc oncogenic transcription factor: identification of direct genomic targets. *Genome Biol* **4:** R69.



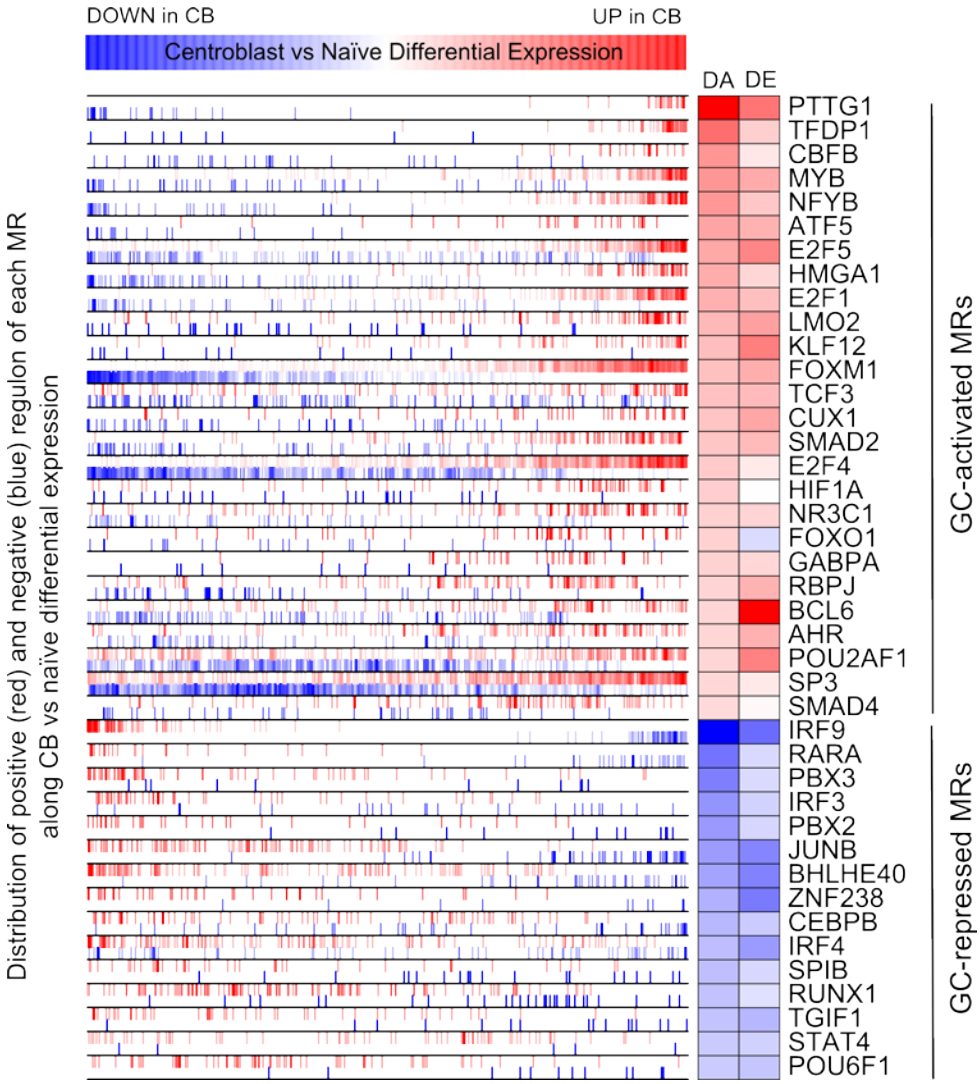
Supplementary Figure 1: ROC curves for: a) the BEIA-inferred PPIs and single evidences including GeneWays; b) BEIA-inferred PDIs and single evidences; c) BEIA-inferred PPIs and single evidences without GeneWays; d) BEIA-inferred PPIs with and without GeneWays (GO: Gene Ontology; GW: GeneWays; HM: Orthologous interactions; HT: High-throughput human interactions; MI: Coexpression; IP: Interpro; TF: Transcription factor binding site; AR: ARACNe; MY: MINDy; AT: ARACNe + TFBS; CO: Conditional target coexpression; MO: Mouse interactions).



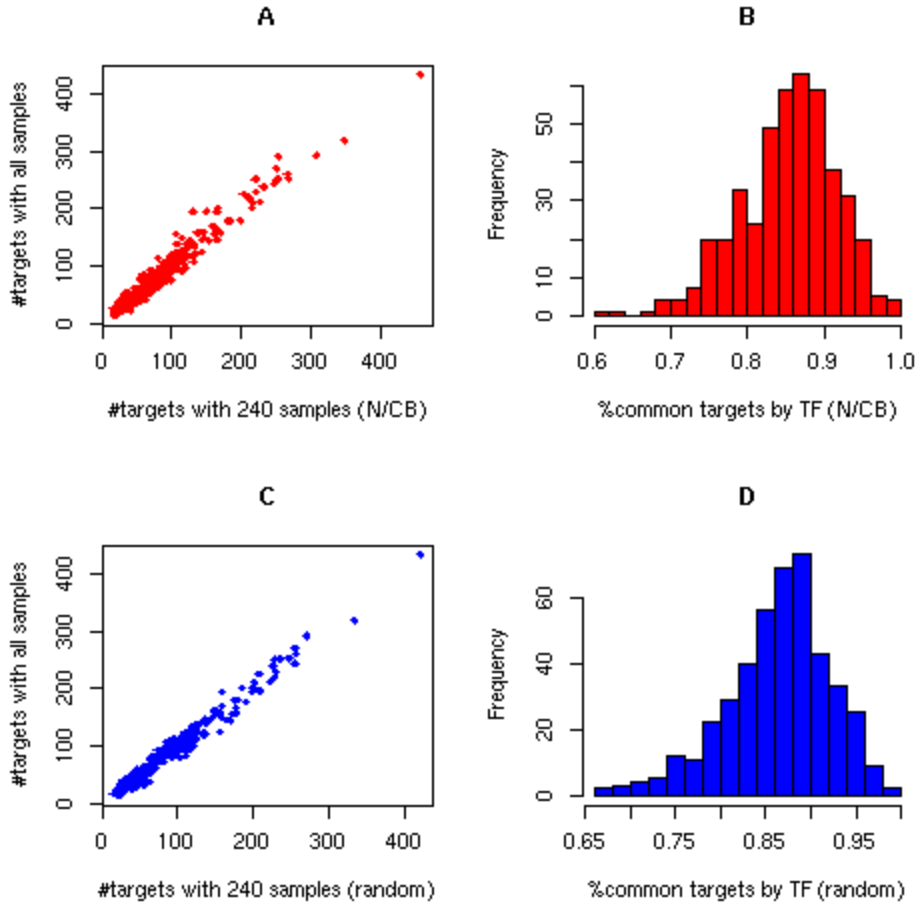
Supplementary Figure 2: HBCI Connectivity. Power law distribution for PPI network (k>10, alpha=2.2), PDI Network out degree (k>20; alpha=1.5) and in degree (k>18; alpha=5.12).



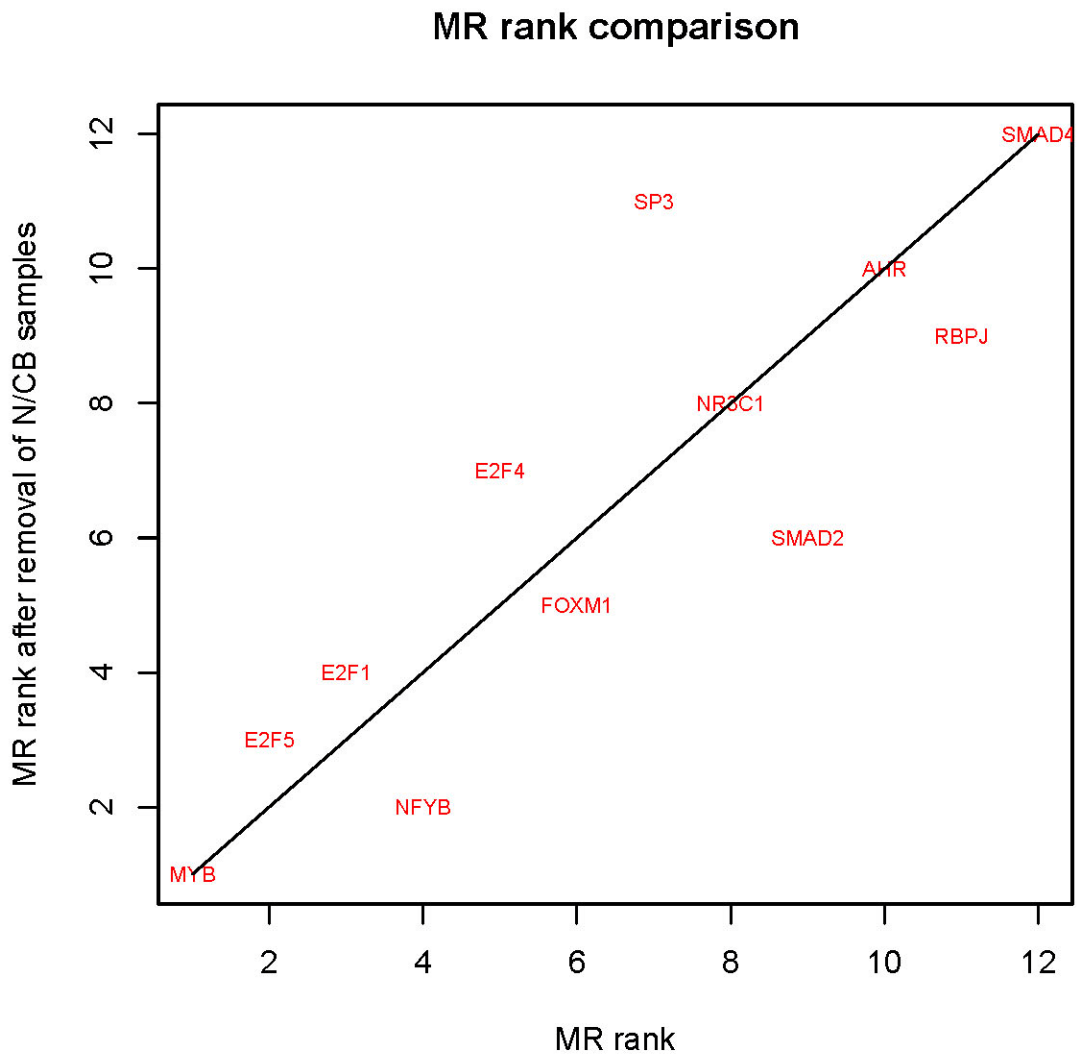
Supplementary Figure 3: Comparison of the performance of the BEIA-inferred PPIs and PDIs (red curves) with the one inferred with adaboost (green curves).



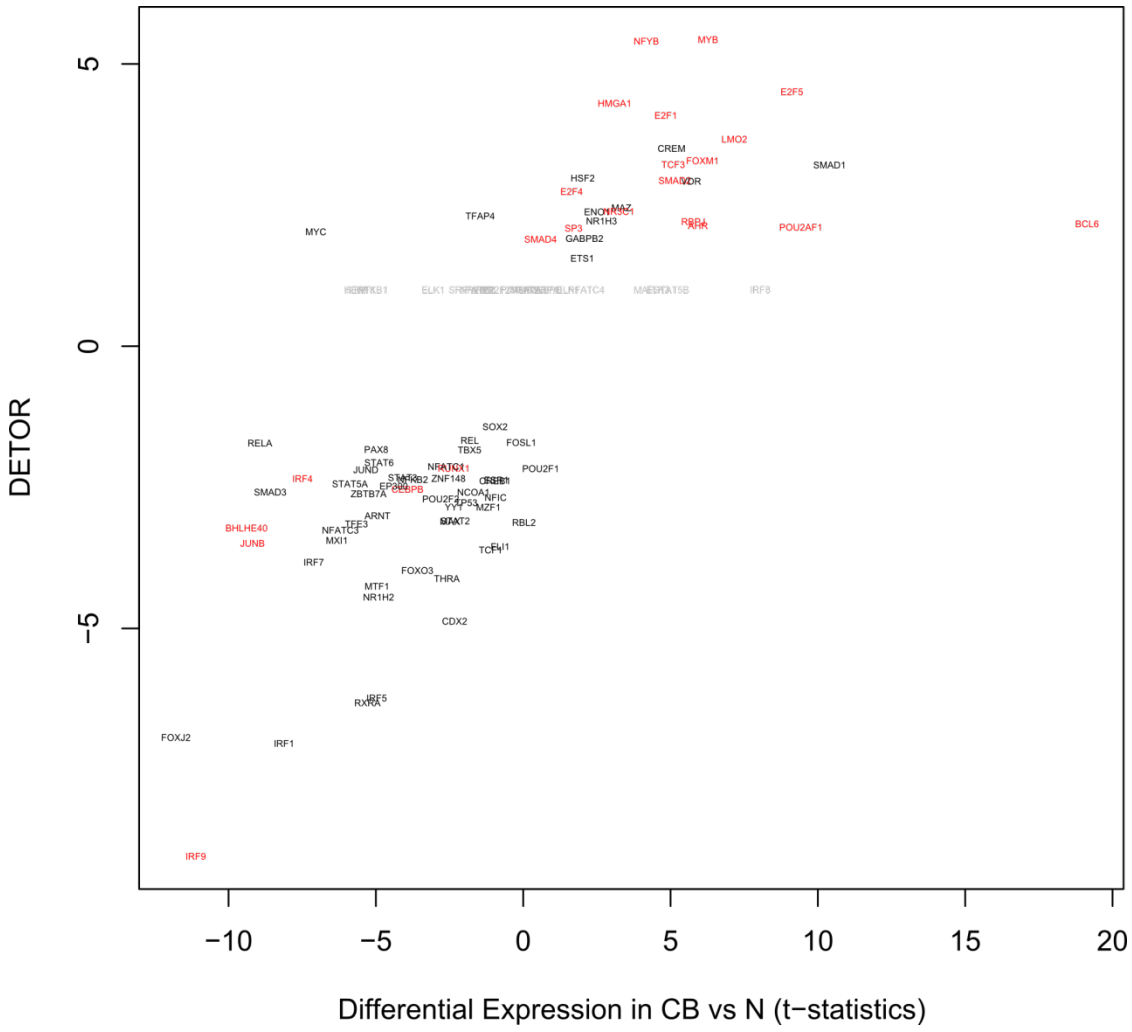
Supplementary Figure 4: Master Regulators of Germinal Center. Left side of the plot shows the distribution of the MR's targets on the gene list ranked by differential expression in centroblast vs naïve samples. The right side of the plot shows a color gradient representing the differential activity of the MR as the DETOR (DA) and its differential expression of the MR in centroblast vs naïve samples as the fold change (DE).



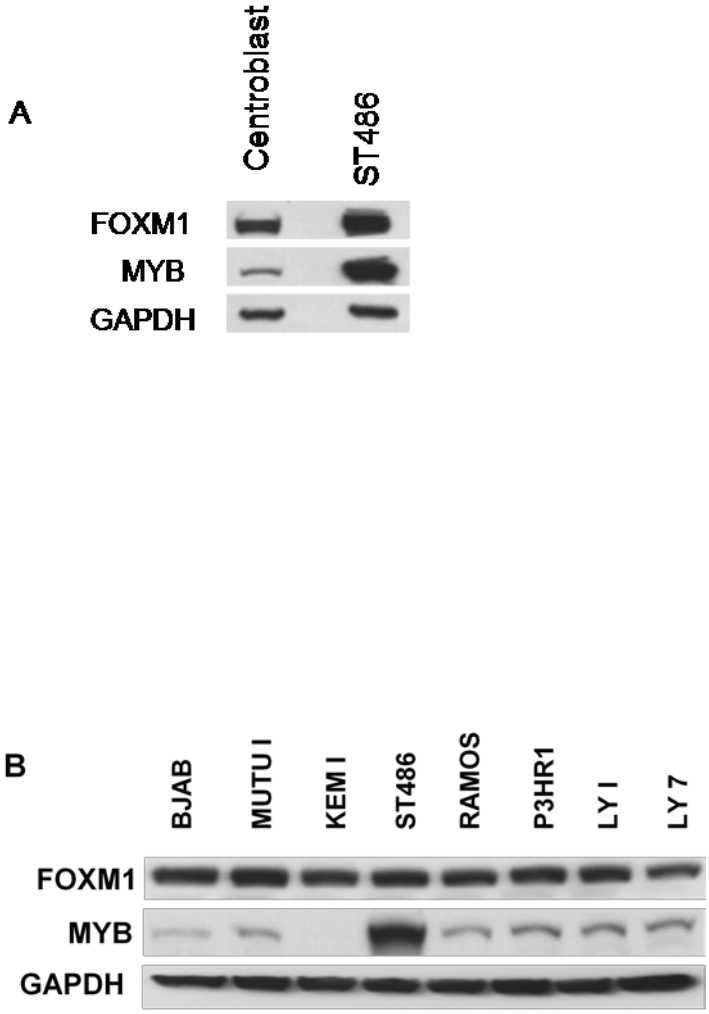
Supplementary Figure 5: (A) Histogram of the percentage of common targets between 2 ARACNe networks obtained with 254 B cell samples and 240 (without naïve and centroblast N/CB) samples; (B) Comparison of the number of targets in the 2 ARACNe networks obtained with 254 B cell samples and 240 (without naïve and centroblast N/CB) samples. Shown are TFs with more than 20 targets in both networks. (C) Histogram of the percentage of common targets between 2 ARACNe networks obtained with 254 B cell samples and 240 (by randomly removing 14 samples) samples; (D) Comparison of the number of targets in the 2 ARACNe networks obtained with 254 B cell samples and 240 (by randomly removing 14 samples) samples. Shown are TFs with more than 20 targets in both networks.



Supplementary Figure 6: comparison of activated MR ranks after using the interactome generated after exclusion of naïve and centroblast samples. Shown are MRs with more than 100 targets in both interactomes (BCL6 and POU2AF1 were not tested again as BCL6 targets were not predicted using the Naïve Bayes classification but with ChIP on chip results and POU2AF1 was tested for illustrating that cofactors could also be recovered as master regulators).

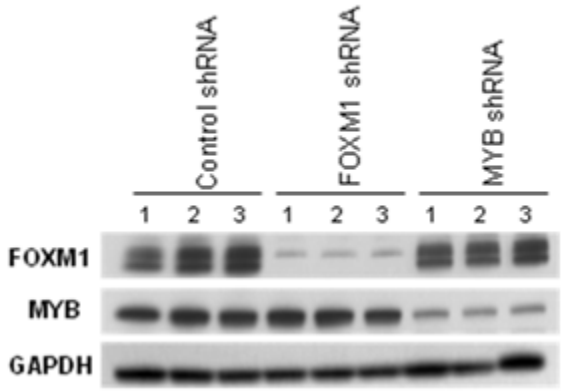


Supplementary Figure 7: Differential expression vs DETOR plot. Differential expression corresponds to the t-statistics computed by comparing the mRNA expression level of the TF in centroblasts (CB) and naïve (N) samples. TFs in the figures have a regulon size of 100 targets or more. Color coding corresponds to: GC-MRs in red, shadow TFs in black and not significant TFs in grey.

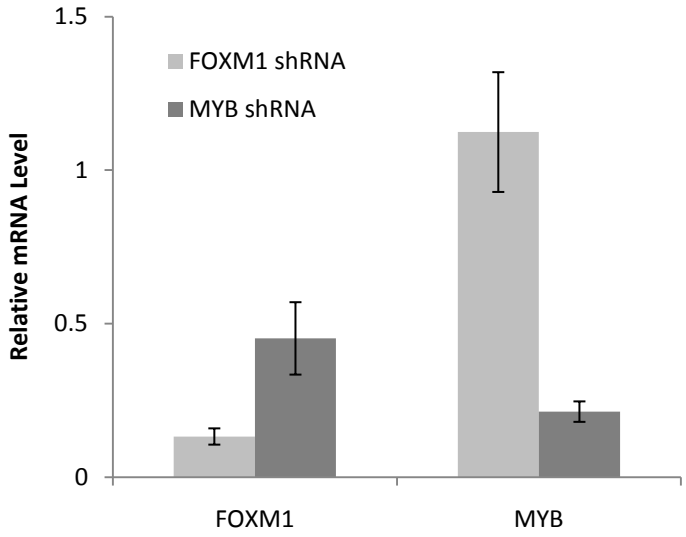


Supplementary Figure 8: FOXM1 and MYB are co-expressed in Germinal Center Centroblasts and Burkitt lymphoma cell lines. Western blot analysis of (A) Centroblasts and ST486 cell line and (B) Burkitt Lymphoma cell lines (BJAB, MUTU I, ST486, Ramos, P3HR1) and GCB-DLBCL cell lines (LY1 and LY7).

A

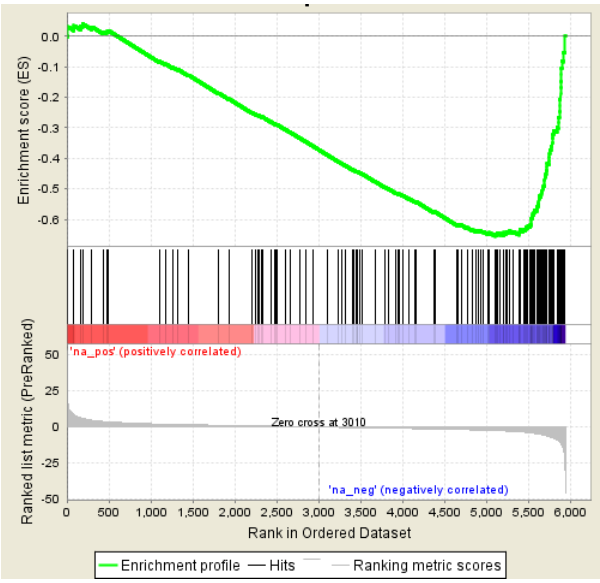


B

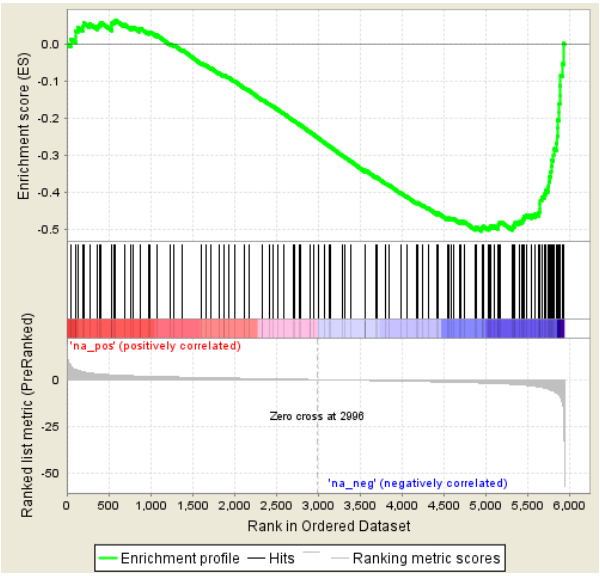


Supplementary Figure 9: Confirmation of FOXM1 and MYB knockdown in the samples used for gene expression profiling. A) Western blot analysis of ST486 total cell lysates 24h after lentiviral-mediated transduction of control, FOXM1 or MYB shRNA confirming knockdown of respective protein. B) Real-time PCR analysis of the mRNA expression of FOXM1 and MYB in the corresponding samples.

A

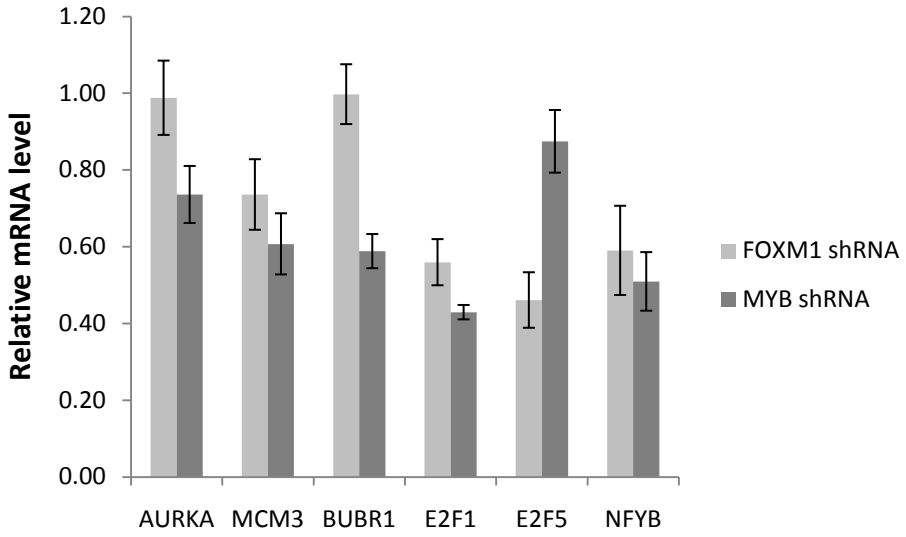


B

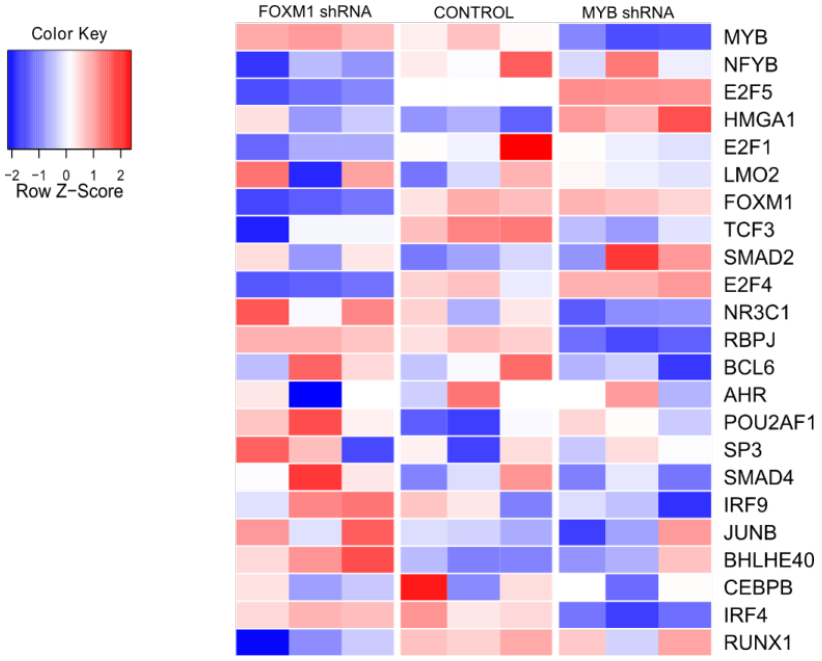


Supplementary Figure 10: Predicted targets enrichment in genes differentially expressed after MYB or FOXM1 silencing. GSEA plots for MYB/FOXM1 positive regulon against a gene list ranked with the t-statistics of (A) MYB (NES = -2.51, p-value < 1E-4) or (B) FOXM1 (NES = -1.92, p-value < 1E-4) silencing compared to control samples.

A

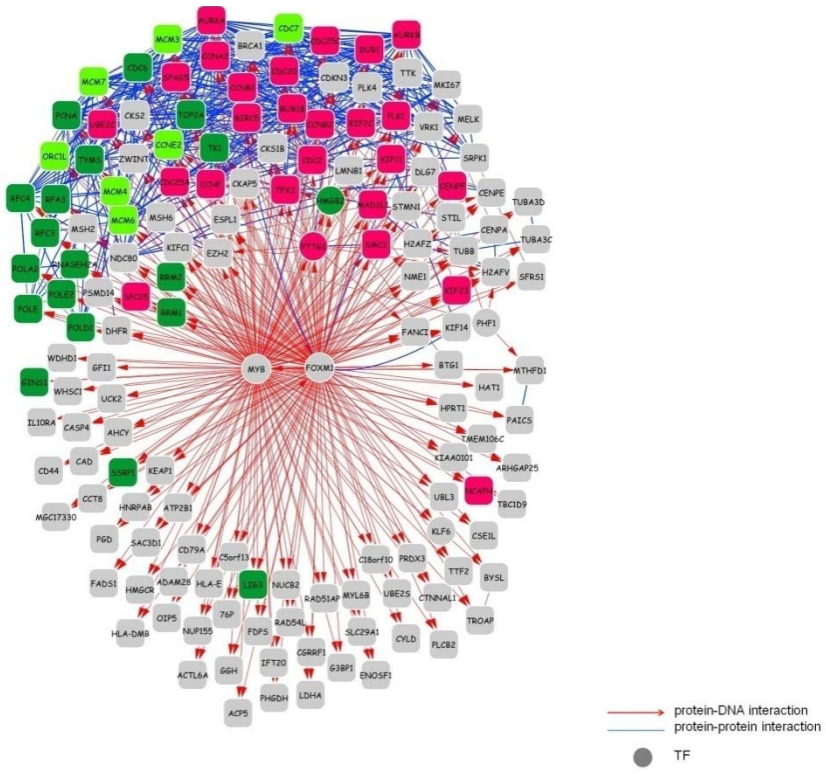


B



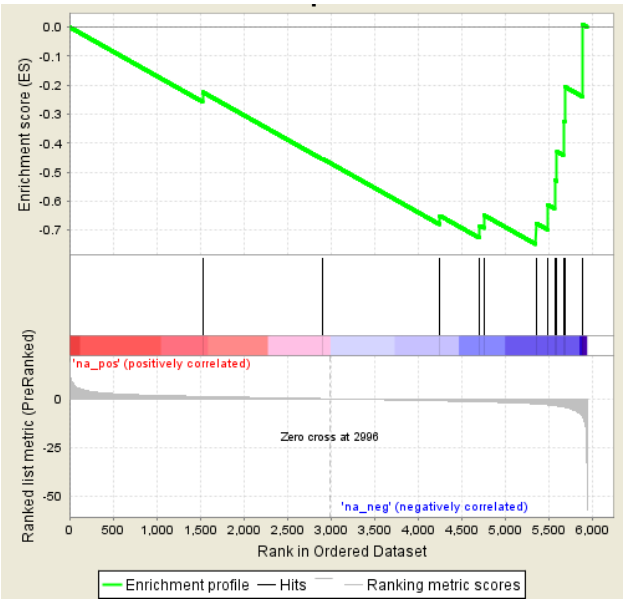
Supplementary Figure 11: Effect of FOXM1/MYB silencing on predicted targets and other MRs mRNA expression. (A) Real-time PCR analysis showing the mRNA expression level of 3 predicted targets of MYB and FOXM1 (AURKA, MCM3 and BUBR1) and 3 predicted MRs (E2F1, E2F5 and NFYB) in ST486 cells transduced with control, FOXM1 and MYB shRNA.

Bars represent the mean +/- SEM between 3 biological replicates. (B) mRNA expression level of the MRs with more than 100 targets in ST486 cells transduced with control, FOXM1 and MYB shRNA.

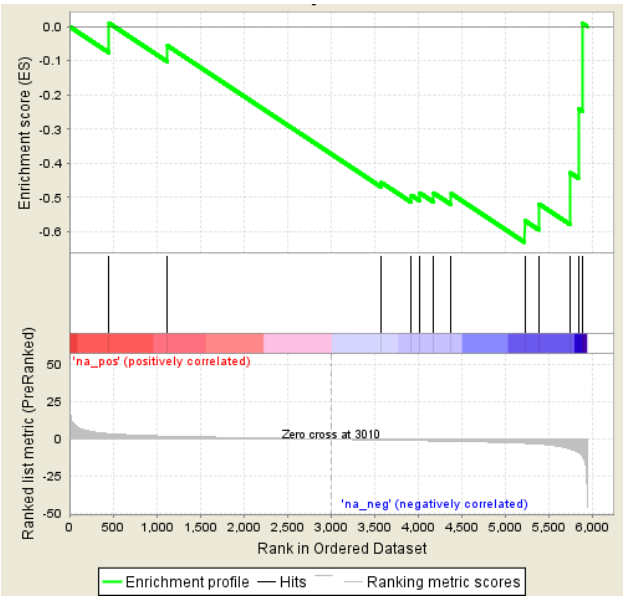


Supplementary Figure 12: FOXM1-MYB network from the HBCI. Node colors correspond to GO annotations mitosis (red), DNA replication (dark green) and DNA replication initiation (light green).

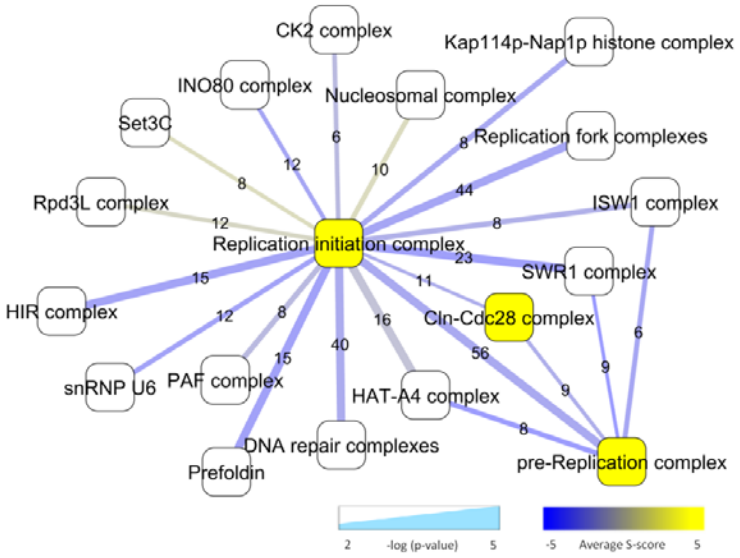
A



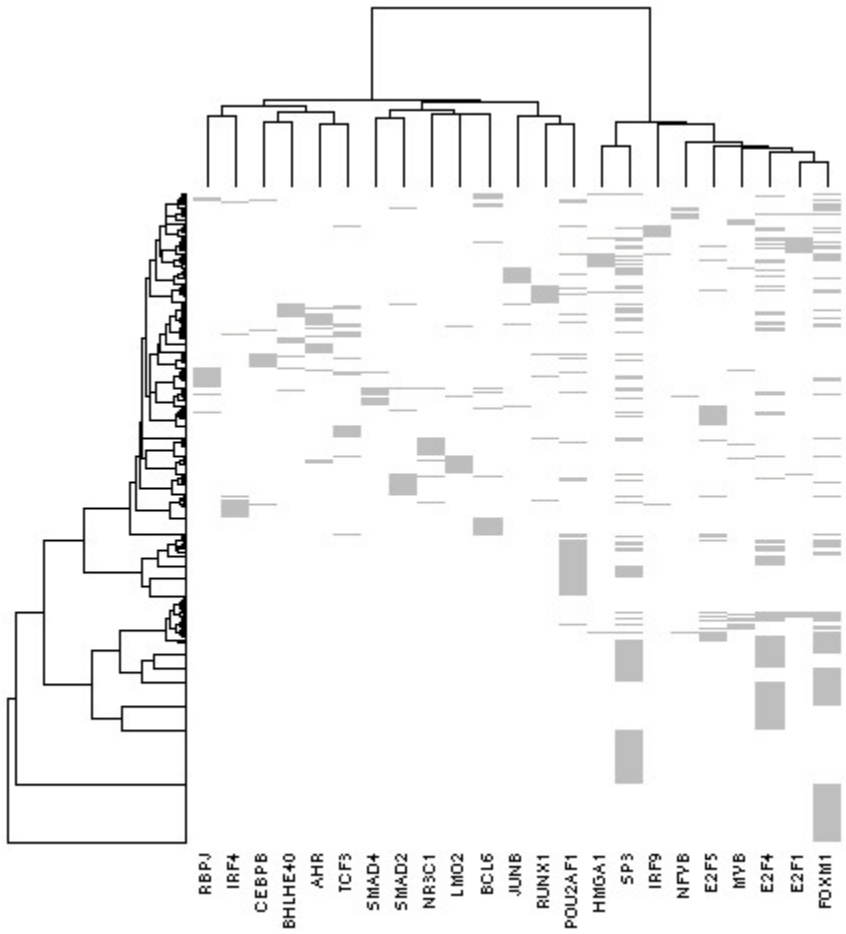
B



Supplementary Figure 13: pre-RC genes differential expression. GSEA plots for 12 pre-RC genes (CDC6, MCM2-7, ORC1L, ORC2L, ORC3L, ORC5L, CDT1) using gene lists ranked with the t-statistics obtained by comparing (A) FOXM1 silencing (p-value = 0.009) and (B) MYB silencing (p = 0.057) to control samples.



Supplementary Figure 14: *Saccharomyces cerevisiae* protein complexes genetically associated with the pre-Replication complex and the Replication initiation complex. A list of curated protein complexes from *S. cerevisiae* was used to search for the groups of proteins most significantly associated with the pre-Replication complex and the Replication Initiation complexes. We determined the number of genetic interactions between any two protein complexes and calculated the likelihood of observing these by chance through random sampling. Complexes with a significant association (p-value<0.01) with either the pre-Replication complex or with Replication Initiation complex are display here in a network diagram. The edge weight connecting two complexes is directly proportional to the significance of the association as defined by –log(p-value) and the number of currently known genetic interactions between these complexes is shown on the edge. The color of the edges corresponds to the average genetic interaction score (S-score). Qualitative genetic interaction data was transformed to approximate quantitative values as follows: synthetic lethal = -5; Phenotypic Enhancement = -3; Phenotypic Suppression=3. The association with the mitotic control proteins (Cln-Cdc28) is the fourth most significant interaction with the pre-Replication complex and the 16th most significant association with the Replication Initiation complex.



Supplementary Figure 15: MRs hierarchical clustering. Hierarchical clustering of the 23 MRs with more than 100 targets based in their regulon overlap. The distance between each MR was computed using their percentage of common targets and we used the ward agglomeration method.

Supplementary Table IV: Synergistic MR pairs with more than 100 targets

MR pair regulon				GSEA for R+				GSEA for R-			
MR pair	**#targets**	**size.pos**	**size.neg**	**ES**	**NES**	**p-value**	**odds ratio**	**ES**	**NES**	**p-value**	**odds ratio**
MYB/E2F4	117	104	13	0.82	3.57	0	6.66	0.00	0.00	1	0.00
MYB/FOXM1	150	129	21	0.82	3.69	0	6.55	-0.76	-2.19	0	0.69
NFYB/FOXM1	118	104	14	0.85	3.66	0	6.33	0.00	0.00	1	0.00
E2F5/E2F4	153	132	21	0.83	3.77	0	5.74	-0.80	-2.32	0	0.81
E2F5/FOXM1	183	157	26	0.82	3.76	0	5.54	-0.82	-2.49	0	1.23
E2F1/FOXM1	174	156	18	0.80	3.72	0	4.87	0.00	0.00	1	0.00
E2F1/E2F4	149	129	20	0.79	3.57	0	4.83	-0.76	-2.21	0	1.01
FOXM1/SP3	340	210	130	0.77	3.64	0	4.47	-0.70	-2.87	0	2.19
FOXM1/POU2AF1	117	66	51	0.76	3.03	0	4.28	-0.71	-2.51	0	2.91
FOXM1/E2F4	573	368	205	0.75	3.80	0	3.74	-0.71	-3.08	0	1.76

Supplementary table VII – Mitotic and replication related *S. cerevisiae* proteins selected for genetic analysis. Using the Inparanoid database we identified 11 yeast orthologs of the human replication/mitosis cluster regulated by MYB/FOXM1. The co-complex members of these yeast genes were obtained from a list of curated *S. cerevisiae* complexes.

Protein Complex Name	Protein Complex Subunits
pre-Replication Complex	TAH11, CDC6, CDC45, DPB11, ORC1, ORC2, ORC3, ORC4, ORC5, ORC6, MCM3, MCM2, MCM5, MCM4, MCM6, CDC47
Replication initiation complex	CDC45, CDC7, CLF1, DBF4, DPB11, MCM1, MCM3, MCM4, MCM5, MCM6, CDC47, MCM10, NOC3, ORC1, ORC2, ORC3, ORC4, ORC5, ORC6, POL1, POL12, PRI1, PRI2, RAD53, SLD3, SUM1
Cln-Cdc28 protein complexes	CLN3 ,CLB2 ,CKS1 ,SRL3 ,CLN2 ,CLB4 ,CLB6 ,CDC28 ,CLB5 ,CLB1 ,SIC1 ,CLN1 ,CLB3

Supplementary table VIII – List of known *S. cerevisiae* genetic interactions between Cln-Cdc28 (mitosis control) complex and the pre-Replication complex or the Replication initiation complex.

Gene from pre-Replication or Replication initiation complexes	Gene from Cln-Cdc28 protein complex	Genetic interaction (S-score)
CDC45	CLB5	Synthetic Lethality (N/A)
CDC6	CLB5	Synthetic Lethality (N/A)
CDC6	SIC1	Synthetic Lethality (N/A)
CDC7	CLB2	Synthetic Lethality (N/A)
CDC7	CDC28	Synthetic Lethality (N/A)
CDC7	CLB5	Phenotypic Suppression (2.2)
DBF4	CLN2	Phenotypic Enhancement (-2.8)
DPB11	CLN3	Phenotypic Enhancement (-2.5)
DPB11	SIC1	Phenotypic Suppression (2.1)
ORC6	CLB5	Synthetic Lethality (N/A)
ORC6	SIC1	Synthetic Lethality (N/A)
ORC6	CLB5	Synthetic Lethality (N/A)
RAD53	CLB5	Synthetic Lethality (N/A)
TAH11	CLB5	Phenotypic Suppression (N/A)
TAH11	SIC1	Synthetic Lethality (N/A)

Supplementary Table IX: Oligonucleotides used for qPCR**.**

Oligonucleotides for qPCR	**Sequence (5'-3')**
GAPDH S	CACCCAGAAGACTGTGGATGGC
GAPDH AS	GTTCAGCTCAGGGATGACCTTGC
FOXM1 S	CACTGGGCCCTGACAACATC
FOXM1 AS	TCACTCAGAGCTTGGGGTG
MYB S	TGGGAGATGTGTGTTGTTGATG
MYB AS	TCCATGCAACAGTTCTGAGACC
E2F1 S	TTCGGCCCTTTTGCTCTG
E2F1 AS	TGCTCTCACCGTCCTACACG
E2F5 S	TTCACTGATTCTGAAGTGTTCTTCC
E2F5 AS	TAGTTACTTTTGGGAGTGGGGAC
NFYB S	AAGGAATTTGAGGCCAGGTATG
NFYB AS	TAAGGGACTACAGGTGTGCACC
AURKA S	GAATGCTGTGTGTCTGTCCG
AURKA AS	CATGGCCTCTTCTGTATCCC
MCM3 S	CTGACCCAAGTCTTTGCCTC
MCM3 AS	TGTCCTCTCCCACTGTCTCC
BUBR1 S	TACAGGTCTTCTGGGATGGG
BUBR1 AS	AACCCCATTCATTTCTGCTG

Supplementary Table X: Oligonucleotides used for qChIP.

Oligonucleotides for qChIP	Sequence (5'-3')
FOXM1 on NFYB S	CATTTTGGCTGCAAGAATCC
FOXM1 on NFYB AS	TCGGTTAGTGGAAGCAGAGG
FOXM1 on BUBR1 S	TGTAGCTTGCCTAAGGTTGC
FOXM1 on BUBR1 AS	GTAGCTTGCCTAAGGTTGCAC
FOXM1 on MCM3 S	CTTTTTCCCCTCTTGAGCTG
FOXM1 on MCM3 AS	CCGCAGAGAAGGATGAAGTG
FOXM1 on AURKA S	AGGACAAGGGCCTTCTTAGG
FOXM1 on AURKA AS	TAGTGGGTGGGGAGACAGAC
FOXM1 on E2F1 S	ATTTTCTCTCCTGGCACTGG
FOXM1 on E2F1 AS	CTCCCTCTGCCTGTCCTTC
FOXM1 on E2F5 S	CTCCTGATCGTCGACTTGC
FOXM1 on E2F5 AS	TGTTATATGTGCAGGGACAGG
FOXM1 on CCNB2 S	TCCTTTGCCGAAAGCTAGAG
FOXM1 on CCNB2 AS	GCAACTGCCAATCTGAAAAAG
FOXM1 on FANCI S	TTCGCTGCTTTTGCCAGG
FOXM1 on FANCI AS	ACCCCTCAGATGTAAGCCCC
FOXM1 on PTTG1 S	GCGGAGTTTGAATGACACAG
FOXM1 on PTTG1 AS	GCCAGGGAGCAAGAGAATATC
MYB on NFYB S	TACGATTTGTGGGTGCTCTG
MYB on NFYB AS	GTTTGCGGTCCCTGTACTTG
MYB on BUBR1 S	GGTTCTGGGGGAATTCAAAG
MYB on BUBR1 AS	TACAAGGAAAAGCCGCAAAC
MYB on AURKA S	CTTTGCCAGACTACCCACTTG
MYB on AURKA AS	TCGTATTTTGTGGGACTCCTG
MYB on E2F1 S	AAAGTCCCGGCCACTTTTAC
MYB on E2F1 AS	GCGTTAAAGCCAATAGGAACC
MYB on E2F5 S	CTCCTGATCGTCGACTTGC
MYB on E2F5 AS	TGTTATATGTGCAGGGACAGG
MYB on CCNB2 S	AAATTCAGAGGCGTCCTACG
MYB on CCNB2 AS	GCACTCTCGCACTCTCATTG
MYB on FANCI S	GTCTACAATGCGAACACAGTCATG
MYB on FANCI AS	AACGACGAAGCAACAGAGCC
MYB on MCM3 S	GTTCGTCAGGCAACGGTATC
MYB on MCM3 AS	TGCAACGACCAAATTCAGAG
MYB on PTTG1 S	GTTTGAGCGTGGTCTCGGAC
MYB on PTTG1 AS	GGGCGTGAGCCAACAAGTAC
MYB on FOXM1 S	CTACCTCAGCGCAAACCTG
MYB on FOXM1 AS	GTATCTTCAGGGCCTAGCGG
β-Actin S	AGCGCGGCTACAGCTTCA
β-Actin AS	CGTAGCACAGCTTCTCCTTAATGTC