Supplementary Information

A Systems Biology Approach to Prediction of Oncogenes and Molecular Perturbation Targets in B Cell Lymphomas

Kartik M. Mani, Celine Lefebvre, Kai Wang, Wei Keat Lim, Katia Basso, Riccardo Dalla-Favera and Andrea Califano

1. Network-Based Approach

The approach taken in this paper is comprised of three distinct steps: (1) a comprehensive network of interactions in human B cells (the B Cell Interactome, or BCI) is generated using an integrative framework. (2) Individual B Cell phenotypes represented in our dataset are analyzed to identify interactions in the network that show a specific gain-of-correlation (GoC) or loss-of-correlation (LoC) pattern. (3) Statistical enrichment for each gene based on its proximity to these affected interactions is computed. Section 1.1 is a condensed description of the BCI formation, which is fully detailed in (Lefebvre et al., 2007) and (Lefebvre et al., in preparation).

1.1 The B Cell Interactome

The BCI comprised a mixed interaction network of protein-protein (PP), protein-DNA (PD) and modulated interactions between a transcription factor and its target. These were predicted using a Naïve Bayes classification (NBC) algorithm using evidences from a variety of sources and gold-standard positive (GSP) and gold-standard negative (GSN) sets. These sources are outlined here.

1.1.1 Protein-Protein Interactions: A GSP for PP interactions was generated using 27,568 human PP interactions from HPRD (Peri et al., 2003), 4,430 from BIND (Bader et al., 2003), and 3,522 from IntAct (Hermjakob et al., 2004), all originating from low-throughput, high quality experiments. The resultant GSP had 28,554 unique PP interactions involving 7,826 genes (after homodimers removal). The GSN was defined as gene pairs involving proteins in different cellular compartments, resulting in a set of 16,411,614 candidate non interacting gene pairs. The negative pairs involving genes from the GSP were extracted, leaving 5,362,594 negative gene pairs.

Evidences for PP interactions were integrated from the following sources:

- four eukaryotic organisms (fly, mouse, worm, yeast) from the databases HPRD (Peri et al., 2003), IntAct (Hermjakob et al., 2004), BIND (Bader et al., 2003) and MIPS (Mewes et al., 2006)
- human high-throughput screens (Ewing et al., 2007; Rual et al., 2005; Stelzl et al., 2005),
- GeneWays literature data mining algorithm (Rzhetsky et al., 2004)
- Gene Ontology (GO) biological process annotations (Ashburner et al., 2000)
- gene co-expression data from B cell expression profiles (Basso et al., 2005)
- Interpro protein domain annotations (Mulder et al., 2007).

Each evidence source was represented as categorical data (continuous values were binned as necessary) and used to compute a likelihood ratio (LR) based on comparison with the GSP and GSN sets. The NBC was trained with all the genes and only the output was filtered for genes expressed in B cells (using B cell expression data listed above).

The prior odds for a PP interaction was approximately 1 in 800 based on previous estimates of the total number of PP interactions in a human cell of ~300,000 among 22,000 proteins (Hart et al., 2006; Rual et al., 2005). From this value, any protein pair having an $LR \geq 800$, after evidence integration, has at least a 50% probability of being involved in a PP interaction. Based on this threshold, the final set had 10,405 PP interactions (2,677 genes) with a posterior probability $P \geq 50\%$ of being true interactions. All missing interactions in the GSPs (10,765 interactions and 3,926 genes) were also re-introduced.

1.1.2 Protein-DNA Interactions: To generate the GSP for PD interactions, human interactions were extracted from the TRANSFAC Professional (Matys et al., 2003), BIND and Myc (MycDB) databases (Zeller et al., 2003), selecting interactions involving genes expressed in B cells only. The resultant GSP PD interaction set had 1,752 interactions involving 197 transcription factors (TFs) and 972 targets. For the GSN, a set of 100,000 random gene pairs was used, 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 in Gene Ontology. The GSP was split in two sets: one set of 1,116 interactions from the TRANSFAC Professional and Myc databases was used for

training the NBC, and the remaining 636 interactions from the BIND and Myc databases were used for testing the performance of the classifier. Another random set of 24,000 interactions was created as a testing GSN set as described above and did not contain any interactions from the training GSN set. A TF-specific prior odds was used, as it has been previously demonstrated that the number of targets regulated by a TF can be approximated by a power-law distribution (Basso et al., 2005; Yu et al., 2006). Predictions by the ARACNe algorithm (Margolin et al., 2006), an information-theoretic method for identifying transcriptional interactions between genes using microarray data, were used to approximate the expected number of targets for a single TF and compute the TF-specific prior odds.

Information on PD interactions from different sources including

- mouse interactions from the databases TRANSFAC Professional and BIND
- human PD interactions inferred by the algorithms ARACNe and MINDy (Wang et al., 2006)
- transcription factor binding sites identified in the promoter of target genes (Smith et al., 2006)
- target gene conditional co-expression based on the B cell expression profiles and GSP interactions.

The data from each evidence source was binned and tested against the GSP and GSN to compute a LR, reflecting the ability of individual evidence sources to predict transcriptional interactions. The NBC produced a final set of 40,798 PD interactions (303 TFs and 5,448 putative targets) with a posterior probability $P \geq 50\%$ of being true interactions. As with PP interactions, all missing interactions from TRANSFAC Professional, BIND, and B cell Myc targets from the MycDB verified by a Chromatin Immunoprecipitation experiment were re-introduced (927 PD interactions).

1.1.3 Post-translational modification: The MINDy algorithm predicts post-translational modulation events, where a TF and target appear to only have an interaction in the presence or absence of a third modulator gene (M). These 3-way interactions were split into two distinct pairwise interactions: a PD interaction between the TF and its target and a TF-modulator interaction that could be either a P-TF or a TF-TF interaction, depending on if the modulator was

a TF as well. These interactions were classified with the number of target(s) a modulator affects for a single TF, and only modulators affecting 15 or more targets per TF were included (based on evidence from known modulator enrichment for *MYC*). This resultant set included 1,925 PP interactions (of which 13 are supported by a direct PP interaction as previously defined) involving 246 TFs and 430 modulators.

1.2 Dysregulation Analysis

Analysis was performed using a large compendium of microarray expression profiles in B cells (BCGEP), including primary tissue as well as cell line samples (NIH Gene Expression Omnibus, record GSE2350) (Basso et al., 2005). Samples were hybridized to Affymetrix HGU95Av2 chips, and normalized using the MAS5.0 algorithm available in the bioconductor package of System R. This set houses data from over 15 distinct phenotypes, including Germinal Center (GC), Naïve (N), Memory (M), Chronic Lymphocytic Leukemia (CLL) both mutated (CLL-mut) and unmutated (CLL-unmut), Diffuse Large B Cell Lymphoma (DLBCL), Primary Effusion Lymphoma (PEL), Follicular Lymphoma (FL), Mantle Cell Lymphoma (MCL), Burkitt Lymphoma (BL) and various tumor cell line samples. For the primary tumor analysis, all cell line samples were removed *a priori* to restrict the data to primary tissue. This resulted in an overall set of over 200 samples, and the whole set was used as background for each specific phenotype. Hierarchical clustering, using Pearson correlation and average linkage, was also performed to verify phenotype groups of interest were relatively homogeneous, making them suitable for this analysis. For the *CD40* enrichment analysis, 67 Burkitt Lymphoma Ramos cell lines were used, with 28 unstimulated samples and 15 stimulated with the anti-IgM antibody used as a background population. 24 samples which were *CD40*-stimulated - 12 co-cultured with *CD40*-ligand expressing fibroblasts (at 8 and 24 hours exposures), 6 stimulated using a *CD40* antibody, and 6 exposed to CD40L and anti-IgM treatment – were used as the test phenotype to characterize the *CD40* perturbation.

The BCI, comprising 64,649 interactions, was split into all possible probes pairs represented on these chips, resulting in a non-unique interaction index of 160,730 interactions. For each phenotype, all non-unique BCI interactions were classified as either a gain-of-correlation (GoC), loss-of-correlation (LoC) or no change (NC). The mutual information (MI) between each gene

pair represented in the BCI was calculated for all samples, and for all samples other than the specific phenotype of interest. The MI is computed using Gaussian kernel estimation (Margolin, et al., 2006). We define a BCI interaction between genes x and y to be affected in the phenotype P, if and only if the MI difference below is statistically significant.

$$\Delta I = I_{All}[x;y] - I_{All-P}[x;y]$$

$I_{All}[x;y]$ is the MI between x and y estimated from all the BCGEP samples, while $I_{All-P}[x;y]$ is estimated from all the BCGEP samples except those in P.

The threshold that defines whether ΔI is statistically significant was calculated by sampling a subset of interactions across 100 equally sized MI bins covering the full MI range in the network. For each bin of 100 interactions, sample sets of various sizes (representing the size of each phenotype group) were randomly removed from the BCGEP and the ΔI was calculated. A total of 10,000 values were computed for each bin and fit with a Gaussian distribution. A bonferroni-corrected p-value of 0.05 was used to threshold a test for a given sample set size and original MI value.

Note that the ΔI value will be negative in the LoC cases (as the MI increases after removal), and positive in the GoC cases (vice-versa). All interactions that passed the threshold were labeled as -1 or 1 respectively.

1.3 Enrichment

Enrichment for each gene was calculated using a set of hypergeometric tests. For each phenotype, all affected interactions were split into LoC or GoC categories. A p-value for each case was computed, based on the total interactions (N), the number of LoC or GoC interactions the gene is directly connected to (D), its' natural connectivity in the BCI (H), and the size of the overall LoC/GoC signature for that particular phenotype (S). The sizes of these signatures are shown in Table 1. As shown below, the p-value is equivalent to a Fisher Exact Test, and is computed for LoC and GoC cases separately.

$$p-value(G)=1-\int_{i=1}^{D-1}\frac{\binom{H}{i}\binom{N-H}{S-i}}{\binom{N}{S}}$$

An additional set of p-values was computed based on modulating effects from each gene as well. As described in section 1.1, we incorporate predictions from the MINDy algorithm about three-way interactions between a transcription factor, its target, and a third modulator gene. Thus, we also include an enrichment based on the number of interactions a gene is predicted to modulate that fall into the LoC or GoC category.

In total, these 4 p-values are combined in a negative log sum operation. The reason for this decision is a simplifying assumption that LoC and GoC cases can be treated independently, as can direct effects and modulatory effects. Although this type of enrichment may bias the analysis against hubs, we find that we still identify those hubs when they are, in fact, related to the phenotype being analyzed. *MYC*, for instance, is one of the most widely connected hubs in our network and still emerges at the top in our analysis of Burkitt Lymphoma.

2. Benchmarking

2.1 Differential Expression

Benchmarking was performed against differential expression analysis in three phenotypes where the causal gene was known, and in *CD40*-stimulated versus unstimulated sets. In the primary tumor analysis, each phenotype of interest (BL, FL, and MCL) was compared with its normal counterpart in the dataset (GC, GC, and non-GC). In the *CD40* set, the 24 samples stimulated with *CD40* were compared against the 43 additional Ramos samples. Microarray data was log2-transformed and a t-test was performed using Welch correction, adjusting for varying degrees of freedom from different sample sizes. The rank was established by taking the first occurrence of a probe representing the gene of interest (*MYC* in BL, *BCL2* in FL, and *CCND1* in MCL). Analysis was conducted in MATLAB.

2.2 Gene Set Enrichment Analysis

Gene Set Enrichment Analysis (Subramanian et al., 2005) was also conducted to establish enrichment of *CD40*-related genes using our method and differential expression. The *CD40* reference set was obtained from the MSigDB available through the GSEA website (http://www.broad.mit.edu/gsea/msigdb/index.jsp). The unique union of two curated sets, one from Biocarta (http://www.biocarta.com/pathfiles/h_CD40PATHWAY.asp) and the other from SIGNALINGAlliance (http://www.broad.mit.edu/gsea/msigdb/genesetCard.jsp?geneset=SIG_CD40PATHWAYMAP) was used for reference, and the *CD40* gene itself was manually added as it is known to be self-regulated. This list was also manually reviewed with collaborators to ensure accuracy. GSEA was conducted using the unweighted classic scoring analysis, so as to be comparable with differential expression, which has a different rank statistic. The minimum set overlap threshold was set to 10.

All genes with a non-zero score from our method were ranked and analyzed using GSEA. The same set size of 379 was used as a cutoff for differential expression, representing a group approximately equivalent to a Bonferroni-corrected p-value of 0.05.

2.3 Visualization

Images of disease modules used in this study were produced using the Cytoscape software package (http://www.cytoscape.org) (Shannon et al., 2003)

References

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.
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, et al. (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.
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 (LNCS)* **4532:** 42-56.
Margolin AA, Nemenman I, Basso K, Wiggins C, Stolovitzky G, Favera D, Califano A (2006) ARACNE: An Algorithm for the Reconstruction of Gene Regulatory Networks in a Mammalian Cellular Context. *BMC Bioinformatics* **7 Suppl 1:** S1-7.
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, et al. (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, et al. (2007) New developments in the InterPro database. *Nucleic Acids Res* **35:** D224-228.
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, et al. (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, et al. (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.
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.
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, et al. (2005) A human protein-protein interaction network: a resource for annotating the proteome. *Cell* **122:** 957-968.

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 U S A* **102:** 15545-15550.
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.