DOI 10.1002/pmic.201100491 *Proteomics* 2012, *12*, 2084–2093

RESEARCH ARTICLE

Network analysis of reverse phase protein expression data: Characterizing protein signatures in acute myeloid leukemia cytogenetic categories t(8;21) and inv(16)

Heather York[1], *Steven M. Kornblau*[2] *and Amina Ann Qutub*[1]

[1] Department of Bioengineering, Rice University, Houston, TX, USA
[2] Departments of Leukemia and Stem Cell Transplantation and Cellular Therapy, The University of Texas M.D. Anderson Cancer Center, Houston, TX, USA

Received: September 15, 2011
Revised: March 5, 2012
Accepted: March 26, 2012

Acute myeloid leukemia (AML) patients present with cancerous cells originating from bone marrow. Proteomic data on AML patient cells provides critical information on the key molecules associated with the disease. Here, we introduce a new computational approach to identify complex patterns in protein signaling from reverse phase protein array data. We analyzed the expression of 203 proteins in cells taken from AML patients. Dominant overlapping protein networks between subtypes of AML patients were characterized computationally, through a paired *t*-test approach looking at relative protein expression. In the first application of this method, we compared recurrent cytogenetic abnormalities inv(16) and t(8;21), both affecting core-binding factor (CBFβ), to normal $CD34^+$ cells and to each other. Six hundred seventy-eight sets of proteins were identified as significantly different in both inv(16) and t(8;21) compared to controls, at the Bonferroni number, $\alpha < 2.44 \times 10^{-6}$. We strengthened our predictions by comparing results to those obtained using lasso regression analysis. Signaling networks were constructed from the protein pairs that were significantly different in the *t*-test and lasso regression analysis. Predicted networks were also compared to known networks from public protein–protein interaction and signaling databases. By characterizing unique "protein signatures" through this rapid computational analysis, and placing them in the context of canonical biological networks, we identify signaling pathways distinct to subcategories of AML patients.

Keywords:
AML / Cytogenetics / Network analysis / Protein signatures / RPPA / Systems biology

1 Introduction

Acute myeloid leukemia (AML) will be diagnosed in over 12 000 American adults this year. Despite extensive research, 5-year overall survival remains at 25%. AML is markedly heterogeneous, with the diagnosis representing a collection of diseases that share a common clinical presentation despite arising from diverse mutations and genetic events. As such, the response of AML patients to similar therapies varies widely. Understanding how to classify and characterize AML on the basis of biologically functional differences is a critical step in developing more efficacious targeted therapies on an individualized basis.

Classification methods for AML have evolved from the French–American–British (FAB) [1, 2] system, which was based on cell morphology and differentiation stage, to the

Correspondence: Dr. Amina Ann Qutub, Department of Bioengineering, BioSciences Research Collaborative, Rm 613, Rice University, 6500 Main St., Houston, TX 77030, USA.
E-mail: aminaq@rice.edu
Fax: (713) 348-5877

Abbreviations: AML, acute myelogenous/myeloid leukemia; **CBF**, core-binding factor; **FDR**, false-discovery rate; **GEP**, gene expression profile; **inv(16)**, inversion 16; **PPI**, protein-protein interactions; **RPPA**, reverse phase protein array

Colour Online: See the article online to view Figs. 1 and 3 in colour.

current WHO system which incorporates cytogenetic and mutation states as well as context (prior chemotherapy) into the classification criteria. Newer data from whole-genome sequencing, gene expression profiling (GEP), micro-RNA profiling, and proteomics are emerging [3–5], but how these data should be incorporated into a classification scheme remains unclear. The desired classification would help explain the heterogeneity of AML in a way that provides guidance toward the selection of the appropriate targeted therapy. In this context, proteomics has an advantage over gene expression profiling because it can measure the protein expression and activation state (phosphorylation, cleavage, etc.) of proteins, features that are unknown from GEP. However, proteomics is currently limited by lower throughput relative to GEP. We have previously used reverse phase protein arrays (RPPA) to show that cases of AML can be classified on the basis of protein expression signatures [1]. In this prior analysis, we looked at proteins individually and considered absolute expression values to be of primary interest. However, since proteins function in networks and interact with many partners, we felt that a computational systems biology approach that evaluated the RPPA-based proteomic data and deduced connectivity and utilization would be superior. Furthermore, we wanted to analyze the relative levels of protein pairs. We therefore set out to develop the means to use RPPA data to build interaction networks. Our goals were to determine what networks were present and whether these networks followed known canonical pathways. Furthermore, we sought to determine whether we could identify previously unknown connections in AML.

In order to develop this ability, we utilized a subset of cases of AML known as "core-binding factor" (CBF) leukemias which arise from inversion of chromosome 16, or inv(16), and translocation of chromosomes 8 and 21, or t(8;21). Leukemias with these two molecular events have a favorable prognosis in response to current therapy with anthracycline and cytosine arabinoside (ara-C). Despite the functional similarity of both affecting CBF, and the similarity in clinical responsiveness, these two cytogenetic groups also have unique characteristics—patients with inv(16) have dysplastic esoinophils; they are also more likely to relapse and to suffer from central nervous system relapses than t(8;21) patients. These distinct commonalities and differences make inv(16) and t(8;21) cytogenetic patients a good test case to try our computational approach for analyzing protein array expression data from patient cells.

Our two main hypotheses cannot be readily tested by previously established proteomics methods alone. These hypotheses are: (i) different subcategories of AML use different molecular signaling pathways (or routes) to obtain similar phenotypic results; and (ii) different subcategories of AML share intracellular pathways that define their cancerous phenotype, however utilize them to different degrees. The two hypotheses are not mutually exclusive. A goal of this research is to determine whether both situations occur, and to identify dominant signaling pathways as a function of AML subcategories.

To that end, we leveraged existing statistical techniques to develop novel computational methods that identify shared and distinct signaling pathways across AML patients, using protein expression data obtained from patient cells. Here, we applied the methods to inv(16) and t(8;21) patients, and bolstered our approach by a second computational analysis using lasso regression. Protein sets resulting from both methods were then compared to known protein interactions obtained from queries to public databases. By this means, we identified the proteins that were significantly changed in AML patients versus control, related these proteins to each other through a network representation, and found novel interactions where no known direct protein connections had been previously documented.

2 Materials and methods

2.1 RPPA protein dataset

A dataset consisting of expression levels of 203 proteins from AML patient samples was collected by researchers at M.D. Anderson Cancer Center from patient blood, marrow, and plasma. The protein expression levels were obtained using RPPA. As previously fully described [2], for RPPA, whole cell lysates are spotted onto nitrocellulose slides in serial dilution and each slide is probed with a highly validated antibody against a protein (total, phopsho or cleaved) of interest. Slides are scanned and digitized using MicroVigene® software. Established protocols were used to normalize the data [1, 2]. To account for variations in staining, background and loading, a "pooled control" lysate from a mixture of 11 AML cell lines served as an overall positive control; lysate buffer was the negative control. Values collected using RPPA are representative of the protein expression level in each sample, rather than absolute protein concentrations. Of the total 539 AML patient samples taken, 21 were from patients who have the cytogenetic abnormality of inversion 16, while 17 were from those with the chromosome abnormality translocation (8;21). These 38 samples from inversion 16 and t(8;21) patients, and 11 bone marrow derived $CD34^+$ normal samples, were used for the following analysis.

2.2 Network building through recognition of protein–protein interactions (PPI)

In order to build networks, we started by identifying those pairs of proteins where the relative expression level in the diseased setting was distinct from that in the normal setting. Because of the nature of the data normalization, taking the

difference between two data points was equivalent to taking the ratio of the expression levels of those proteins. To consider the full dataset, we used an iterative procedure to generate an $m \times n$ matrix, where m was the number of patients included in the study and n was the total number of possible pairs of proteins.

2.3 Paired *t*-test analysis

To analyze the matrix of relative expression levels, we employed a standard *t*-test. We wanted to compare samples from the two cytogenetic categories of AML patients, inv(16) and t(8;21), with the set of bone marrow derived CD34$^+$ normal cells. For the *t*-test analysis of the data, all possible pairwise combinations of the 203 proteins in the dataset were considered, for a total of 20 503 protein pairs. The relative expression level difference between the proteins in each pair was calculated for the 21 inv(16) patients, 17 t(8:21) patients, and 21 normal controls. The averages and standard deviations (SDs) of each pair in each of these groups were used in the *t*-test. From this, the *p*-value of each protein pair for each cytogenetic group was calculated relative to the set of normal patients using the standard *t*-distribution (Fig. 1A).

$$t_p = \frac{\overline{X}_p - \overline{Y}_p}{\sqrt{\frac{s_{xp}^2}{n_x} + \frac{s_{yp}^2}{n_y}}}$$

where t_p is the test statistic for protein pair p, $\overline{X}_p$ is the average of the relative expression levels of pair p for a cytogenetic category, $\overline{Y}_p$ is the average of the relative expression levels of pair p for the controls, s_{kp} is the SD for pair p in group k, and n_k is the number of patients in group k.

$$(p - value)_p = 2\left(1 - tcdf\left(\left|t_p\right|, df_p\right)\right)$$

where $(p - value)_p$ is the *p*-value for protein pair p when comparing a cytogenetic group to the controls and $tcdf$ computes the Student's *t*-cumulative distribution function at the given test statistic with the given degrees of freedom, df_p, as defined in the Supporting Information Material. This *p*-value tells us how similar or different a protein pair belonging to

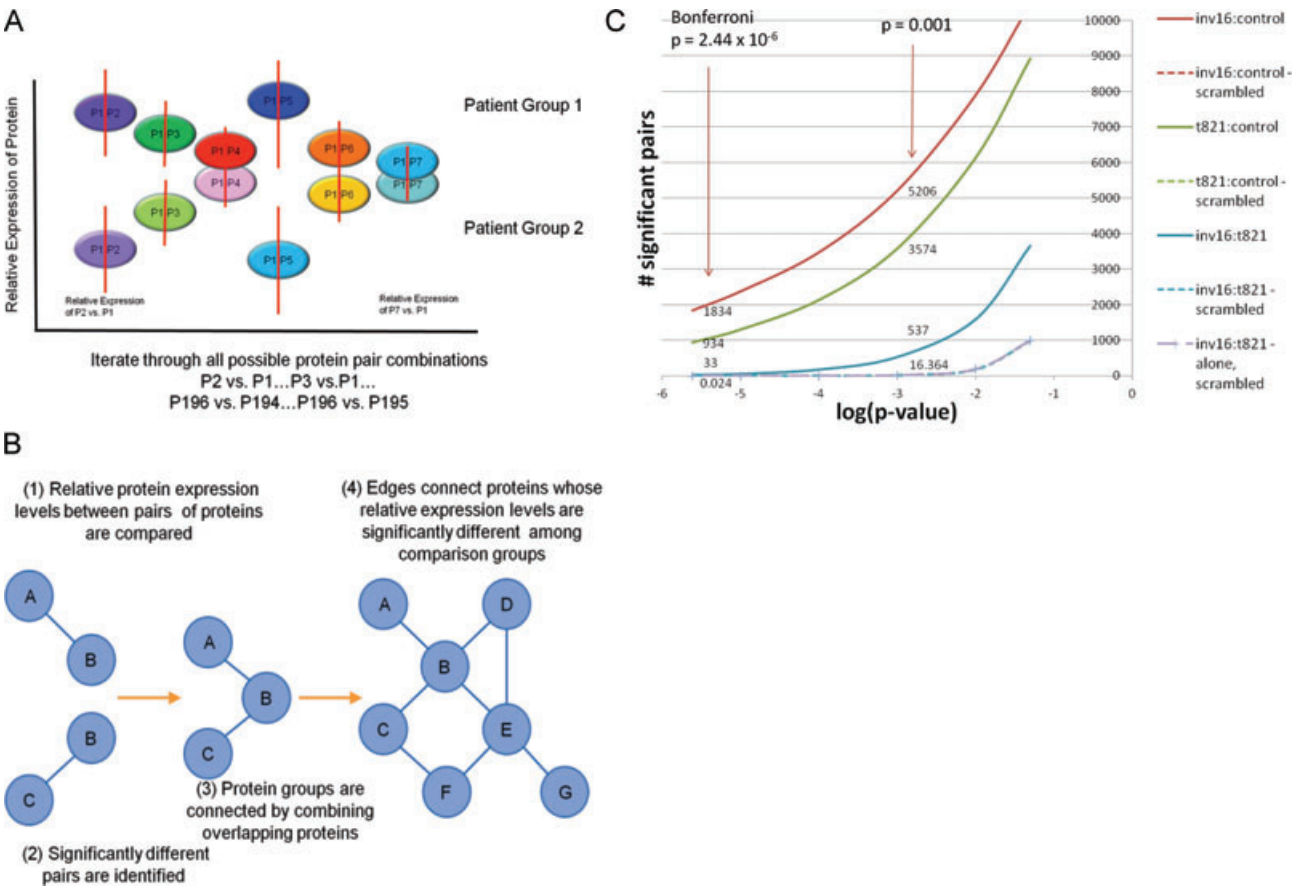


Figure 1. Development of the *t*-test algorithm. Schematics illustrating the paired *t*-test algorithm (A) and the networks developed from combining the pairs of proteins identified as significantly different across patient subcategories (B). In (B), possible interactions are compared using the *t*-test and lasso regression; significantly different pairs are identified; and only these pairs are used to build the graphical network. (C) The false discovery rate for the paired *t*-test algorithm, across inv(16) and t(8;21) patient subcategories.

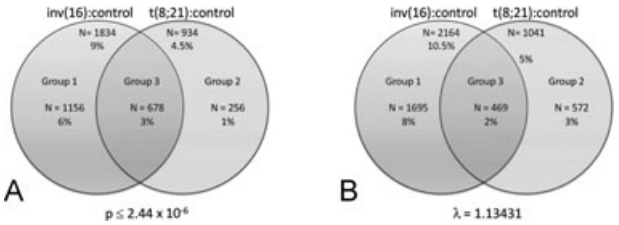


Figure 2. Number of significant pairs found for inv(16) alone, t(8;21) alone and the overlap between inv(16) and t(8;21) using: (A) the *t*-test approach at $\alpha = 2.44 \times 10^{-6}$ and (B) the lasso method at a penalty $\lambda = 1.13$.

the diseased group is when compared to the control group, with a high *p*-value indicating that the relative expression level of the proteins in the pair is not statistically distinguishable between the diseased group and the normal group. By setting a threshold *p*-value, α, a subset of protein pairs that have significantly different relative expression levels between the diseased set and the normal set could then be constructed. From the identified sets of statistically different protein pairs between patients and controls, we build a network representation (Fig. 2B). Protein pairs are connected by joining overlapping proteins (nodes). Edges are our initial hypothesis of probabilistic interactions between the identified proteins.

2.4 False discovery rate and α

In order to determine what value of α would be appropriate for use with this dataset, a study was conducted to determine the false-discovery rate (FDR) associated with the data. First, the selected matrix of data (all protein expression levels for all patients in the inv(16), t(8;21), and normal groups) was scrambled so that each data point was given a new, randomly selected location within the matrix. We also produced a scrambled matrix of inv(16) and t(8;21) protein expression levels alone, without including normals. Using these scrambled matrices of data, the entire *t*-test procedure was repeated, starting with the calculation of relative expression levels of all possible protein pairs, and dividing the data into groups that corresponded to what were previously the two cytogenetic groups and the control group. This entire procedure, beginning with the scrambling of the data and concluding with the *t*-test, was repeated 1000 times so that a variety of possible combinations of the data were considered. Any statistically significant protein pairs identified in the scrambled datasets were considered false discoveries and were an indication of the number of protein pairs found in the real dataset that should be considered as α-type errors. The number of statistically significant protein pairs identified in the scrambled dataset was compared to the number identified in the real dataset for several values of α in order to determine a maximum α, above which the number of false discoveries became a significant portion of the total number of pairs discovered.

2.5 Lasso regression analysis

To supplement our method, we also used an alternate computational approach to analyze the dataset. We then compared its results to the paired *t*-test method. The second approach we employed is the lasso (least absolute shrinkage and selection operator) technique. Lasso is a regression shrinkage method that has been previously used to analyze large protein expression datasets [6–8]. The lasso technique first calculates the covariance matrix between sets of variables, each with multiple observations. For this study, variables were all the possible 203 proteins. Observations were divided into the set of patient samples for the two cytogenetic categories and the bone marrow derived $CD34^+$ normals. We illustrate this using the example of the inv(16) patient set:

$$S_{inv16} = \begin{matrix} Actin_{P1} & \dots & ZNF342_{P1} \\ \vdots & & \vdots \\ Actin_{P21} & \dots & ZNF342_{P21} \end{matrix},$$

where S_{inv16} is the set of 21 observations, i.e. patients (P1 through P21), for the inv(16) samples, measuring 203 variables, i.e. proteins listed alphabetically from Actin to ZNF342. We then describe this data by its covariance matrix, as follows. On the main diagonal of the covariance matrix, variance between observations of each protein expression level is computed. In all other matrix entries, the covariance between pairs of proteins is calculated. Covariance is calculated by:

$$COV = \frac{\sum_{i=1}^{n}(X_i - \overline{x})(Y_i - \overline{y})}{n-1},$$

where X and Y are the variables to compare over a set of i observations, and $\overline{x}$ and $\overline{y}$ are the means of those variables across all observations. Writing this for the inv(16) example, where Actin expression levels are tested for correlation to ZNF342 levels, yields:

$$COV = \frac{\sum_{i=1}^{21}(Actin_{Pi} - \overline{Actin})(ZNF342_{Pi} - \overline{ZNF342})}{21-1}$$

The covariance matrix calculates this value for all possible pairs of proteins.

The power of the lasso method comes into play once we have calculated the covariance matrix. The technique employs a L^1-norm absolute value penalty on the matrix [9, 10]. This selectively sets entries in the covariance matrix to zero, depending on the strength of the penalty imposed. By iteratively imposing an increasingly rigorous penalty, entries which correspond to protein pairs that are not well correlated are eliminated, so that only entries where protein levels are statistically dependent on each other remain. We want to identify protein–protein pairs whose relative expression levels are statistically different from normal. To find these pairs, we take the following approach—for the lasso method, three penalized $n_p x\, n_p$ covariance matrices were calculated, where n_p is the number of proteins in the dataset; a penalized $n_p x\, n_p$ covariance matrix was calculated for each of the two

cytogenetic groups considered and for the $CD34^+$ bone marrow derived normals. These covariance matrices were calculated using several different imposed penalty values. Then each value in the covariance matrix for each cytogenetic group was compared to the corresponding value in the covariance matrix for the control group. Any values in the covariance matrices that were set to zero indicate that the relative difference in expression levels between the two proteins corresponding to that value are uncorrelated. Therefore, if a value was zero in either the cytogenetic covariance matrix or the control covariance matrix, but not in the other, that means that the relative difference in expression levels between the two proteins corresponding to that value were correlated in one group but not the other, indicating a characteristic difference between the two groups. Hence, for comparison to our paired *t*-test analysis, we are interested in matrix values that satisfy the following:

$$\left(X'_{i,j} = 0 \; and \; Y'_{i,j} \neq 0\right) or \left(X'_{i,j} \neq 0 \; and \; Y'_{i,j} = 0\right)$$

where $X'_{i,j}$ is the covariance value corresponding to protein i and protein j in the cytogenetic matrix and $Y'_{i,j}$ is the covariance value corresponding to protein i and protein j in the control matrix.

To employ the lasso method with our dataset, we applied a previously developed lasso regression program [9]. Additionally, we wrote a script to prepare the matrices containing values of protein expression levels for the inv(16), t(8;21), and control groups. The covariance matrix was then calculated for each cytogenetic group and for the control group at a range of L^1-norm penalty values. The entries remaining in the covariance matrix after the L^1-norm penalty is imposed correspond to values where the protein levels are statistically related to each other. By setting a relaxed L^1-norm penalty, the number of values remaining in the covariance matrix will be large, corresponding to a relatively small quantity of entries that have been set to zero. We explored a range of L^1-norm penalties to identify a small subset of relative protein expression levels for further study. We compare the resulting "statistically different" pairs to our *t*-test analysis.

2.6 Proteins appearing in many significant pairs

Another approach to examining the data was to identify the proteins that appear in a large number of significant pairs. The prevalence of these proteins in significantly different pairs between cancerous and control groups, make them potential chemotherapeutic targets that could broadly affect signaling in patient cells. Table 2 shows a list of these proteins found with the *t*-test. The α value used was the Bonferroni number, which is calculated by the following equation:

$$\alpha = \frac{0.05}{n_s} = \frac{0.05}{\frac{n_p!}{2!*(n_p-2)!}},$$

where n_s is the number of outcomes, and n_p is the number of proteins studied. The Bonferroni correction identifies the maximum *p*-value needed to maintain significance when comparing multiple hypotheses or outcomes for the given dataset. n_s in our case is the number of possible protein pairs that could be significantly different between patient groups.

In order to determine how many significant pairs a protein should be involved in to be considered a large number, first, the average and SD of number of pairs each protein was involved in for each group was determined. Any protein involved in a number of significant pairs greater than two SDs above the mean was considered to be significantly large for the *t*-test results.

2.7 Comparisons to known protein interactions

After identifying protein pairs and potential networks of interest, we wanted to compare our findings from the *t*-test and lasso studies to information that is already well known and documented in literature. To do so, we built combined PPI and signaling networks from available public databases and graphed them in the open source program Cytoscape [11–13]. Interactions between two proteins are represented by edges in the graph. Nodes represent proteins. In the case of the PPI databases, protein interactions are direct molecular interactions (e.g. binding, phosphorylation, transport) documented in published experimental data. In the signaling databases, interactions can include transcriptional interactions or unknown mechanisms where one protein is known to affect another protein's activity or level. As our goal is to determine all existing known relationships between protein pairs, we include both PPI and signaling databases in our analysis. Henceforth, we use the term interaction to refer to both signaling and molecular binding events. In this way, bionetworks are built which capture known relationships between proteins in a query group. In addition, we can query not just direct interactions among proteins in a set, but also interactions among their nearest neighbors (Fig. 3D). This means that all known protein interactions will also be displayed for each protein in the set, even if the interaction involves a protein not included in the original dataset. For our purposes, we allowed two proteins in a query set to be linked by interactions through no more than two intermediate, additional proteins.

Within Cytoscape, we queried public databases to establish known networks for the proteins identified as significant by the *t*-test and lasso method. Additionally, in order to expand the results of the queries, we performed secondary queries that included nearest neighbors of the proteins contained in the subsets. The database sources used in these queries were all those available through the MiMI Plugin 3.0.1 (including PPI and signaling networks: BIND, CCSB, DIP, GRID, HPRD, IntAct, MDC, MINT, KEGG, PubMed, and reactome) [11,12], restricted to human protein data. The results of these queries showed us how the proteins identified by our

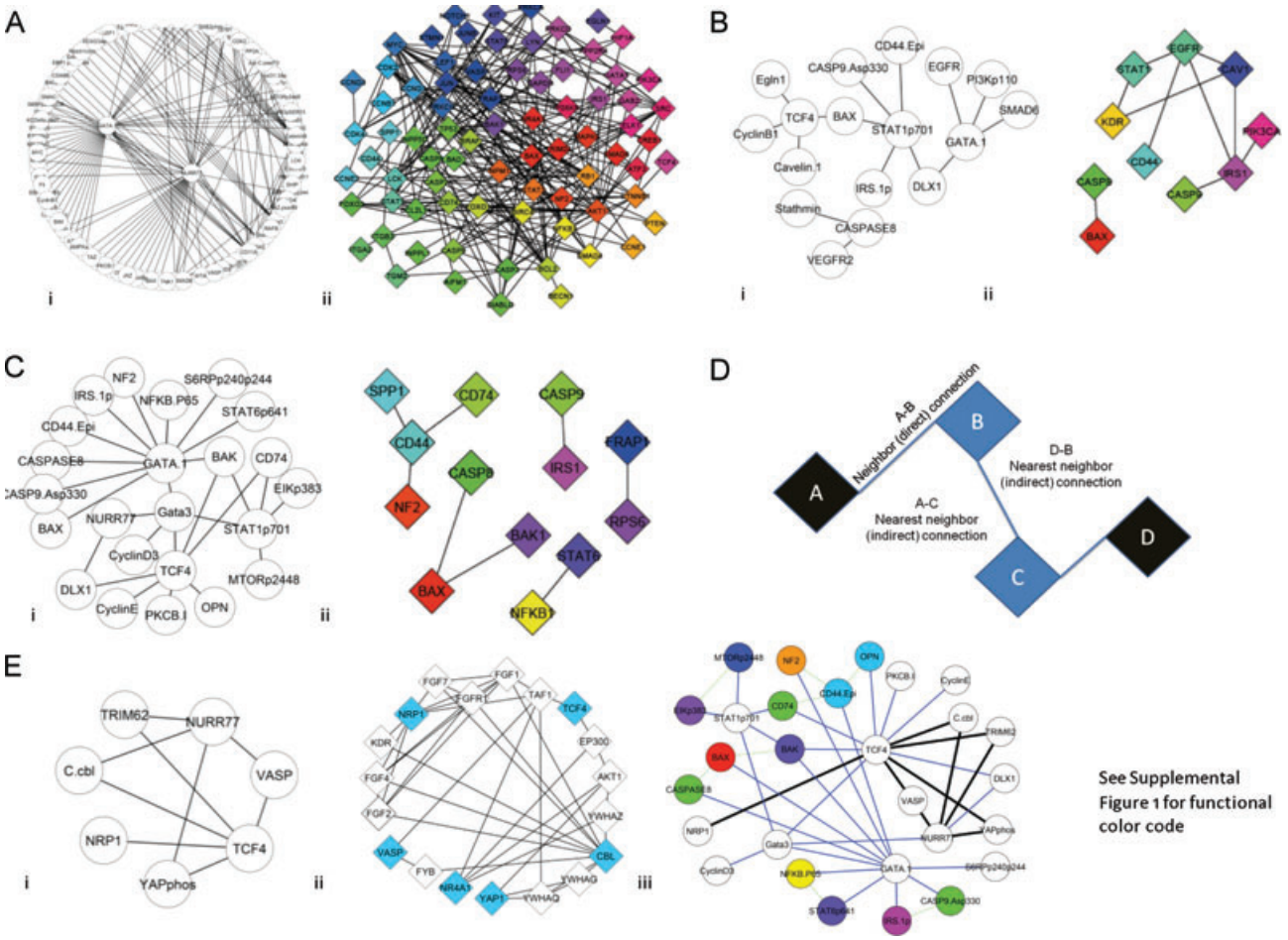


Figure 3. Interactions between proteins identified as part of a highly different pair or pairs graphed in Cytoscape. (A–C) Circular nodes represent proteins from the dataset analyses alone, while diamonds indicate proteins queried from public databases. (A) inv(16) patients compared to normals. Group 1. Results from the paired *t*-test alone at *p*-value = 10^{-10} (i) and direct interactions known from public databases (ii). (B) t(8;21) patients compared to normals. Group 2. Paired *t*-test alone at *p*-value = 10^{-10} (i) and direct interactions from databases (ii). (C) Protein pairs in both the inv(16) and t(8;21) patients (shared proteins) compared to normals. Group 3. Paired *t*-test at *p*-value = 10^{-10} (i) and direct interactions from databases (ii). (D) Illustration of direct and nearest neighbor (indirect) connectivity where A and D are hypothetical proteins from the RPPA dataset, and B and C are their neighbors. (E) Interactions for Group 3 predicted by the *t*-test at *p*-value = 2.44×10^{-6} and by lasso regression at $\lambda = 1.42$ but not yet known to be direct connections, from public databases. No direct connections between proteins shared in the *t*-test and lasso methods were found by public databases. Nearest neighbors shared by at least two query proteins (blue) are shown (ii). The combined and weighted network for the *t*-test alone (blue edges, *p*-value = 10^{-10}), *t*-test overlapping with the lasso regression (black edges, *p*-value = 2.44×10^{-6} and lasso at $\lambda = 1.42$), and public databases (green edges), showing overlap for the inv(16) and t(8;21) patients. Public database connections are for the full protein.

statistical techniques were known to interact either directly (PPI) or through transcriptional signaling (signaling databases). By highlighting the proteins from our dataset, we could see where any previously known interactions may occur. Any edges already known served to confirm the results of our study, while any edges predicted by our methods that did not appear in the Cytoscape query may represent newly discovered edges that could be pursued by further in vitro studies to determine if these are previously unrecognized edges that exist in normal cells or, preferentially, disease-specific edges.

2.8 Implementation

The *t*-test code and iterative script were written in Matlab (MathWorks). We adapted the lasso regression analysis from a program written in FORTRAN [9] and run from Matlab. Network representations were graphed in Cytoscape, Versions 2.6.3 or 3.0.1 [11–13]. Programs were run on a Linux server (ThinkServer, 2.66 GHZ, 500 GB Harddrive, 24 GB DIMM). The paired *t*-test runs on average took 0.96 s. The graphical lasso (glasso method) approach converged on three

covariance matrices of 203 proteins in 0.97–3.01 s, depending on the penalty value used. Smaller penalty values resulted in longer runtimes.

3 Results

As we analyzed the dataset, we obtained three main sets of results: (i) protein pairs in AML patients that were identified as significantly different from control using the paired *t*-test analysis; (ii) protein pairs identified as significantly different using the lasso analysis; and (iii) predicted signaling network relationships between proteins in the dataset based on our computational analysis as well as queries to public databases.

Table 1 and Fig. 2 show the number of protein pairs that were significantly different between each AML subcategory and the control group for several different tolerance levels using the *t*-test and the number of protein pairs which satisfy the logical condition for each AML group at a range of penalty values using the lasso method. The results were split into three groups: protein pairs that were unique when comparing inv16 patients to the control group (Group 1), protein pairs that were unique when comparing t(8:21) patients to the control group (Group 2), and protein pairs that appeared to be significantly different between both inv16 and t(8:21) patients when compared to the control group (Group 3).

The results of the FDR study for the *t*-test are shown in Fig. 1C. Three categories were considered for this study: Group 1, Group 2, and protein pairs that were significantly different when comparing inv(16) patients to t(8;21) patients. The last group is considered to be a secondary test of the validity of the dataset, because any protein pairs appearing in this group would be significantly different between two cytogenetic groups, but not between the cytogenetic groups and the control group. That is, this group represents protein pairs that are more different between two leukemia groups than the control group. This is unlikely to be a real occurrence and also explains why this category reports the smallest number of significant protein pairs in the unscrambled dataset. For $\alpha < 0.001$, the number of false discoveries are insignificant, and the number of protein pairs in the group that compares inv(16) patients to t(8;21) patients is also very low. Therefore, we considered 0.001 as the largest α at which reliable results for this analysis would be expected.

In order to produce results that can be more easily digested and interpreted, an even lower α than the maximum value of 0.001 was used to generate even smaller subsets of protein pairs that are very significantly different between the cytogenetic groups and the control group. Supporting Information Table 1 in the appendix shows a list of all significant protein pairings with $\alpha < 10^{-10}$. Likewise, Supporting Information Table 2 in the appendix shows a list of all significant protein pairings after an imposed lasso penalty of 1.89. Once again, the results were divided into three sections: Group 1, Group 2, and Group 3, as before. The stringent α value of 10^{-10} and the penalty of 1.89 were selected to make the lists of resulting significant proteins pairs of lengths that are manageable and readable. The protein pairs identified in Supporting Information Tables 1 and 2 should be considered to be highly significant. Results from building networks of the *t*-test pairs and comparisons to public databases are shown in Fig. 3 and Supporting Information Fig. 2. Protein interactions hypothesized as a result of both the *t*-test and lasso analysis—and not found to be known direct connections from public databases—are presented in Fig. 3E. For the network in Fig. 3E(iii), we weighted the edges based on whether they were identified by the *t*-test alone, by the *t*-test and lasso regression, or by the public databases. The calculation for edge weights is presented in Supporting Information Fig. 3.

We also list the proteins that appear in a large number of significant pairs when using the *t*-test in Table 2. Proteins appearing in the overlap group belong in both cytogenetic categories. For example, NURR77 is involved in a total of 199 protein pairs in the inv(16) cytogenetic group. Forty-six of the pairs are unique to inv(16) when compared to controls, while 153 of these pairs are also significantly different between t(8;21) patients and controls.

4 Discussion

The presented research is the first of its kind to employ a series of computational modeling techniques to: (i) predict the dominant signaling pathways used by specific AML patient cytogenetic subcategories; (ii) identify and rank highly different protein interactions compared to normals; and (iii) map identified key proteins onto known signaling pathways from public databases.

Through our paired *t*-test approach and network development, we identified pathway utilization common—as well as distinct—to inv(16) and t(8;21) cytogenetic categories. These pathways are shown in Fig. 3A–C. Our analysis shows inv16 and t(8;21) share a number of protein pairs that are significantly different from normals—39% of protein pairs found significantly different at $p \leq 2.44 \times 10^6$ and 32% at $p \leq 0.001$ are shared, and they share thousands of protein pairs that were not found different from controls (Table 2). These results are consistent with a large degree of overlap observed between the inv16 and t(8;21) patient phenotype, and the knowledge that both chromosome abnormalities involve CBF mutations. Our analysis also identified characteristics that are unique to each patient category (Table 3). Those protein pairs found significantly different between inv16 and t(8;21) could underlie observed clinical differences such as inv16 patients presenting with dysplastic eosinophils and inv16 patients' higher risk for CNS relapse.

When our predicted pathways are compared to a public database query, it is clear that, while some previously documented direct interactions are identified by our analysis, there are also many connections which had not been

Table 1. Number of significant pairs found using the *t*-test approach as a function of α values and number of significant pairs found using the lasso method as a function of penalty values

		Group 1	Group 2	Group 3
t-test	α	Unique to inv(16)	Unique to t(8;21)	Overlap between inv(16) and t(8;21)
	0.05	3946	2404	6529
	0.01	3570	1805	4323
	0.001	2751	1119	2455
	10^{-4}	2004	663	1465
	10^{-5}	1450	375	920
	2.44×10^{-6}	1156	256	678
	10^{-10}	184	15	26
Lasso method	λ	Unique to inv(16)	Unique to t(8;21)	Overlap between inv(16) and t(8;21)
	0.567154	5614	5170	4860
	0.708942	5316	2979	2538
	0.945256	3455	1211	1053
	1.13431	1695	572	469
	1.41788	755	183	104
	1.89051	254	25	18
	2.83577	36	0	1

For the *t*-test, values at the Bonferroni number are shaded. Likewise, for the lasso method, values at a penalty of 1.13 are shaded. Shaded values are illustrated by the Venn diagrams in Fig. 2.

documented or were previously only documented through indirect protein interactions, i.e. through interactions with up to two nonqueried neighbors, as shown in Fig. 3E. Therefore, any protein pairs identified by our computational techniques as being statistically different between AML cytogenetic categories and controls and not identified as a known direct interaction through a database query represent potential PPI that were previously unknown, and are of interest for future in vitro study. Furthermore, when we also applied lasso regression analysis to the data, overlap in significantly different protein pairs was found between the results of the *t*-test and the lasso regression, as shown in Table 3 and Fig. 3E(i). These protein pairs may be of particularly strong interest for in vitro verification of interaction, due to the fact that they are identified as being unique by two different statistical techniques. No direct interactions between the protein pairs predicted using both the *t*-test and lasso analysis were previously documented. Shared nearest neighbor interactions are identified by public databases for all proteins but TRIM62 (Fig. 3E(ii)).

We see several advantages of employing the presented approach. Previous approaches to analyzing protein array data used the absolute expression levels given in the postnormalized RPPA data [1]. However, it is possible that there could be instances where the relative expression between two proteins is of greater importance. For example, there could be one protein that is at the lower end of the normal range and a second protein that is at the higher end of the normal range, yet that patient may be in the diseased state. The relative expression between the two proteins would reveal the abnormality of the situation and may be indicative of that

Table 2. List of proteins that have the highest number of pairings whose relative expression level is significantly different from controls

Group 1 (inv(16)) Group mean = 12.094 Group SD = 15.628		Group 2 (t8;21) Group mean = 4.063 Group SD = 4.805		Group 3 (overlap) Group mean = 7.0895 Group SD = 17.825	
Protein name	PP	Protein name	PP	Protein name	PP
CateninB	108	BAX	31	NURR77	153
CDK4	96	STAT1p701	22	TCF4	152
SHIP2	72	IRS.1p	18	GATA.1	96
GATA.1	65	XIAP	17	STAT1p701	73
CD11A	63	GAPDH	16		
BAD	57	TCF4	16		
GATA3	52	TNK1	15		
S6RPp240p244	47				
NURR77	46				
ZNF342	45				

Proteins with a pairing number greater than two standard deviations from the group mean are shown; $\alpha < 2.44 \times 10^{-6}$ (Bonferroni number). PP, number of significantly different protein pairs that include the indicated protein. The protein pairings enumerated in Group 3, the overlap of inv16 and t(8;21), are distinct from those listed in Groups 1 and 2.

Table 3. Protein pairs that are identified as being significantly different between cytogenetic categories and the control group by both the *t*-test, at $\alpha < 2.44 \times 10^{-6}$, and the lasso method, with $\lambda = 1.42$

Group 1 (inv(16))		Group 2 (t8;21)		Group 3 (overlap)	
NURR77	AKTp308	NURR77	CASP3clvd	NURR77	C.cbl
NURR77	ARC	NURR77	Caspase3	NURR77	TRIM62
NURR77	BADp112	NURR77	CDC2	NURR77	VASP
NURR77	HIF1a	NURR77	CDK2	TCF4	C.cbl
NURR77	HSP27	NURR77	CDK4	TCF4	NRP1
NURR77	Jun.C.pser73	NURR77	JAB1	TCF4	TRIM62
NURR77	FOXO3Ap	NURR77	MCL1	TCF4	VASP
NURR77	RAC123	NURR77	MSI2	TCF4	YAPphos
NURR77	XIAP	NURR77	NRP1		
TCF4	BAK	NURR77	P62		
TCF4	Beclin.1	NURR77	PKCBII		
TCF4	CateninA	NURR77	PPARgam		
TCF4	CateninBp	NURR77	X14.3.3Sigma		
TCF4	CIAP	NURR77	YAPphos		
TCF4	Cox2	TCF4	CASP3clvd		
TCF4	CyclinD1	TCF4	Caspase3		
TCF4	DLX1	TCF4	CDC2		
TCF4	EGFR	TCF4	CDK2		
TCF4	EGFRp992	TCF4	CDK4		
TCF4	FoxO1.3Ap	TCF4	JAB1		
TCF4	FOXO3A	TCF4	MCL1		
TCF4	HER3	TCF4	MSI2		
TCF4	IGF1	TCF4	P62		
TCF4	IGFBP2	TCF4	PKCBII		
TCF4	MEK	TCF4	PPARgam		
TCF4	MEKp217p221	TCF4	X14.3.3Sigma		
TCF4	NF2p	VASP	XIAP		
TCF4	NPM				
TCF4	ODC				
TCF4	OPN				
TCF4	P70S6K				
TCF4	PI3Kp110				
TCF4	PKCB.I				
TCF4	PLAC1				
TCF4	RAC123				
TCF4	SMAD6				
TCF4	STAT3				
TCF4	Survivin				
TCF4	TAZ				
TCF4	TNK1				
TCF4	VHL				
TCF4	XIAP				
TCF4	YAP				
BAK	CDK2				
BAK	FOXO3A				
BAK	IGFBP2				
BAK	P27				
BAK	P70S6K				
BAK	Survivin				
Beclin.1	EIF2				
CDK2	Cox2				
CDK2	CyclinD1				
CDK2	EGFRp992				
CDK2	FoxO1.3Ap				
CDK2	OPN				
CDK2	PKCB.I				
CDK2	SMAD6				
CDK2	TAZ				
CDK2	TNK1				
Cox2	Survivin				
FOXO3Ap	VASP				
Jun.C.pser73	VASP				

particular disease. It is our hypothesis that by comparing relative, rather than absolute, protein expression levels, we will be able to discover relationships between proteins that were previously unknown and may prove to be diagnostically and, potentially, therapeutically significant. A second, and major, benefit of this analysis is the ability to characterize complex network differences across patient groups. We have started to build hypotheses as to the structure and differential utilization of the underlying signaling network, and we compare these to documented protein interactions. Applying this technique to RPPA data, also allows us to include changes in protein states due to phosphorylation or cleavage.

It is also important to note current limitations of this approach. Edges in the networks we develop from the paired *t*-test method indicate that the relative level between those two proteins is significantly different in one patient group compared to normals. They may—but do not necessarily—indicate direct network connections, or interactions, between those proteins. Nor, in the presented analysis, do they provide directionality to the difference. Future improvements to this method would consider directionality. Additionally, the comparison database queries are limited by the accuracy and sources for the databases—we have currently included both signaling and PPI networks from public sources, and as databases become more refined and comprehensive we will be able to include phosphorylation comparisons.

The technique we presented using the lasso method gives a different but synergistic result: protein pairs that are correlated in one group but not another. The overlapping proteins identified using both the *t*-test and the lasso method, i.e., those are both correlated differently and whose relative expression levels are significantly different across groups, may prove the most powerful in characterizing specific AML groups by unique protein signatures (Fig. 3E). By querying public databases to find where these protein pairs fit in known networks, we also are able to identify potential new protein signaling relationships. We recently applied this approach to study signaling through the friend leukemia virus integration 1 protein (FLI1) pathway, and characterized newly identified interactions with SMAD4 across AML patients [14]. Experimental functional studies of the FLI1–SMAD4 interaction and of all subsequent predictions from the presented approach will further enhance confidence in its utility.

Now that we have developed the computational methods to analyze the RPPA data, build networks, and compare predicted protein interactions to known pathways, we will begin to analyze the entire RPPA dataset. The goal of this work is to assess if we can classify AML on the basis of network utilization. In combination with existing methods such as protein clustering and Bayesian network analysis, the presented approach has the promise to become more prognostic, or therapy directing. Furthermore, we can apply our techniques to other RPPA datasets and presumably, as newer high throughput techniques for proteomics become available, to those as well.

The authors have declared no conflict of interest.

5 References

[1] Kornblau, S. M., Tibes, R., Qiu, Y. H., Chen, W. et al., Functional proteomic profiling of AML predicts response and survival. *Blood* 2009, *113*, 154–164.

[2] Tibes, R., Qiu, Y., Lu, Y., Hennessy, B. et al., Reverse phase protein array: validation of a novel proteomic technology and utility for analysis of primary leukemia specimens and hematopoietic stem cells. *Mol. Cancer Ther.* 2006, *5*, 2512–2521.

[3] Yasar, D., Karadogan, I., Alanoglu, G., Akkaya, B. et al., Array comparative genomic hybridization analysis of adult acute leukemia patients. *Cancer Genet. Cytogenet.* 2010, *197*, 122–129.

[4] Deneberg, S., Grovdal, M., Karimi, M., Jansson, M. et al., Gene-specific and global methylation patterns predict outcome in patients with acute myeloid leukemia. *Leukemia* 2010, *24*, 932–941.

[5] Marcucci, G., Maharry, K., Radmacher, M. D., Mrozek, K. et al., Prognostic significance of, and gene and microRNA expression signatures associated with, CEBPA mutations in cytogenetically normal acute myeloid leukemia with high-risk molecular features: a Cancer and Leukemia Group B Study. *J. Clin. Oncol.* 2008, *26*, 5078–5087.

[6] Friedman, J., Hastie, T., Tibshirani, R., Applications of the lasso and grouped lasso to the estimation of sparse graphical models. *Technical Reports, Available Online: Stanford University,* 2010.

[7] Meinshausen, N., Buhlmann, P., High dimensional graphs and variable selection with the lasso. *Ann. Stat.* 2006, *34*, 1–32.

[8] Tibshirani, R., Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B* 1996, *58*, 267–288.

[9] Friedman, J., Hastie, T., Tibshirani, R., Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* 2008, *9*, 432–441.

[10] Gradshteyn, I. S., Ryzhik, I. M., *Tables of Integrals, Series, and Products*, 6th Ed., Academic Press, San Diego 2000, pp. 1114–1125.

[11] Tarcea, V. G., Weymouth, T., Ade, A., Bookvich, A. et al., Michigan molecular interactions r2: from interacting proteins to pathways. *Nucleic Acids Res.* 2009, *37*, D642–D646.

[12] Gao, J., Ade, A. S., Tarcea, V. G., Weymouth, T. E. et al., Integrating and annotating the interactome using the MiMI plugin for cytoscape. *Bioinformatics* 2009, *25*, 137–138.

[13] Assenov, Y., Ramirez, F., Schelhorn, S. E., Lengauer, T. et al., Computing topological parameters of biological networks. *Bioinformatics* 2008, *24*, 282–284.

[14] Kornblau, S. M., Qiu, Y. H., Zhang, N., Singh, N. et al., Abnormal expression of FLI1 protein is an adverse prognostic factor in acute myeloid leukemia. *Blood* 2011, *118*, 5604–5612.

Supplemental Material

Calculation of Degrees of Freedom

$$df_p = \frac{\left(\frac{s^2_{xp}}{n_x} + \frac{s^2_{yp}}{n_y}\right)^2}{\frac{\left(\frac{s^2_{xp}}{n_x}\right)^2}{n_x - 1} + \frac{\left(\frac{s^2_{yp}}{n_y}\right)^2}{n_y - 1}}$$

where df_p is the degrees of freedom for protein pair p when comparing a cytogenetic group to the controls.

Table S1. List of protein pairs with $\alpha < 10^{-10}$.

Unique to inv(16)		Unique to t(8;21)		Overlap	
AIF	GATA.1	BAX	STAT1p701	BAK	GATA.1
AIF	NURR77	BAX	TCF4	BAK	STAT1p701
AKT	CD11A	CASP9.Asp330	STAT1p701	BAK	TCF4
AMPKa	GATA.1	CASPASE8	Stathmin	BAX	GATA.1
ATF3	GATA.1	CASPASE8	VEGFR2	CASP9.Asp330	GATA.1
BAD	BAK	Cavelin.1	TCF4	CASPASE8	GATA.1
BAD	CD11A	CD44.Epi	STAT1p701	CD44.Epi	GATA.1
BAD	Gata3	CyclinB1	TCF4	CD74	STAT1p701
BAD	S6RPp240p244	DLX1	GATA.1	CD74	TCF4
BAD	SHIP2	DLX1	STAT1p701	CyclinD3	Gata3
BADp112	CD11A	Egln1	TCF4	CyclinE	TCF4
BADp136	NURR77	EGFR	GATA.1	DLX1	NURR77
BADp155	GATA.1	GATA.1	PI3Kp110	DLX1	TCF4
BADp155	Jun.C.pser73	GATA.1	SMAD6	EIKp383	STAT1p701
BADp155	NURR77	IRS.1p	STAT1p701	GATA.1	Gata3
BADp155	STAT1p701			GATA.1	IRS.1p
BADp155	TCF4			GATA.1	NF2
BAK	CateninB			GATA.1	NFKB.P65
BAK	CDK2			GATA.1	S6RPp240p244
BAK	CDK4			GATA.1	STAT6p641
BAK	CyclinD3			Gata3	NURR77
BAK	Jun.C.pser73			Gata3	STAT1p701
BAK	NURR77			Gata3	TCF4
BAK	P70S6K			MTORp2448	STAT1p701
BAK	PP2A			OPN	TCF4
BAX	NURR77			PKCB.I	TCF4
BCL2	GATA.1				
BCL2	NURR77				
Beclin.1	GATA.1				
Beclin.1	NURR77				
BIM	GATA.1				
CASP3clvd	GATA.1				
CASP7Clvd	NURR77				
CASP7Clvd	TCF4				
CASP9.Asp330	CateninB				
CASPASE8	CateninB				
CASPASE8	CDK4				

CateninB	CD11A				
CateninB	CD44.Epi				
CateninB	CD74				
CateninB	CREBps133				
CateninB	CyclinE				
CateninB	EBP1.pthr70				
CateninB	EIKp383				
CateninB	FoxO1.3Ap				
CateninB	Gata3				
CateninB	IRS.1p				
CateninB	JUNB				
CateninB	LCK				
CateninB	MTORp2448				
CateninB	NF2				
CateninB	P53pSER15				
CateninB	PKCBII				
CateninB	PKCg645				
CateninB	S6RPp240p244				
CateninB	SHIP2				
CateninB	SRCp416				
CateninB	STAT6p641				
CateninBp	NURR77				
CateninBp	TCF4				
CD11A	CD49B				
CD11A	CDK4				
CD11A	CyclinD3				
CD11A	GATA.1				
CD11A	IntegrinB3				
CD11A	Jun.C.pser73				
CD11A	LEF1				
CD11A	NURR77				
CD11A	PKCap657				
CD11A	RBp807p811				
CD11A	SHIP				
CD11A	SMAD4				
CD11A	STAT1p701				
CD11A	TCF4				
CD11A	TG2				
CD11A	XIAP				
CD11A	ZNF342				
CD44.Epi	NURR77				

CD44.Epi	TCF4				
CD74	CDK4				
CD74	GATA.1				
CD74	NURR77				
CDK2	Gata3				
CDK4	EBP1.pthr70				
CDK4	Egln1				
CDK4	EIKp383				
CDK4	Gata3				
CDK4	IRS.1p				
CDK4	JNK2				
CDK4	JUNB				
CDK4	LCK				
CDK4	MTORp2448				
CDK4	NFKB.P65				
CDK4	P53pSER15				
CDK4	S6RPp240p244				
CDK4	SHIP2				
CDK4	STAT6p641				
CREB	GATA.1				
CREBps133	GATA.1				
CyclinB1	GATA.1				
CyclinD1	STAT1p701				
CyclinD3	IRS.1p				
CyclinD3	LCK				
CyclinD3	S6RPp240p244				
CyclinE	GATA.1				
CyclinE	NURR77				
CyclinE2	Gata3				
EBP1.pthr37.46	SHIP2				
EBP1.pser65	GATA.1				
EBP1.pthr70	GATA.1				
EIKp383	GATA.1				
EIKp383	Jun.C.pser73				
EIKp383	NURR77				
EIKp383	TCF4				
Fli	GATA.1				
FOXO3Ap	Gata3				
GAB2phos	NURR77				
GAB2phos	STAT1p701				
GAB2phos	TCF4				

GAPDH	SHIP2				
GATA.1	JNK2				
GATA.1	JUNB				
GATA.1	Kit.C				
GATA.1	LCK				
GATA.1	LYN				
GATA.1	MTOR				
GATA.1	MTORp2448				
GATA.1	MYC				
GATA.1	NPM3542				
GATA.1	P53pSER15				
GATA.1	P70S6Kp				
GATA.1	PI3Kp85				
GATA.1	PKCBII				
GATA.1	PKCDelta.p507				
GATA.1	PKCg645				
GATA.1	PTENp				
GATA.1	S6RPp235p236				
GATA.1	SHIP2				
GATA.1	SMAC				
GATA.1	SRCp416				
GATA.1	STAT3p727				
GATA.1	TRIM24				
GATA.1	VASP				
GATA.1	WTAP				
Gata3	HIF1a				
Gata3	Jun.C.pser73				
Gata3	LEF1				
Gata3	PKCap657				
Gata3	RAFB				
Gata3	SHIP				
Gata3	SMAD4				
Gata3	TAZ.pser89				
Gata3	XIAP				
Gata3	ZNF342				
IntegrinB3	SRCp416				
JAZ	NURR77				
Jun.C.pser73	NURR77				
Kit.C	NURR77				
MTOR	NURR77				
MTORp2448	NURR77				

Notch1clvd	SHIP2				
NPM3542	NURR77				
NURR77	OPN				
NURR77	PI3Kp110				
NURR77	PKCB.I				
NURR77	PTENp				
NURR77	S6RPp235p236				
NURR77	S6RPp240p244				
NURR77	SHIP2				
NURR77	SMAC				
NURR77	SMAD6				
NURR77	SSBP2				
NURR77	TAZ				
NURR77	WTAP				
RAFB	SHIP2				
RBp807p811	S6RPp240p244				
S6RPp240p244	SHIP				
S6RPp240p244	SMAD4				
S6RPp240p244	TCF4				
SHIP2	Stathmin				
SHIP2	TCF4				
SHIP2	ZNF342				
SMAD6	TCF4				
TCF4	TNK1				

Table S2. List of significant protein pairs found using $\lambda = 1.89$

Unique to inv(16)		Unique to t(8;21)		Overlap	
CASP9	CateninA	CASP9	HIF1a	AKTp308	CASP9
CASP9	CateninBp	CASP9	JAB1	AKTp308	C.cbl
CASP9	Cox2	CASP9	MCL1	AKTp308	Jun.C.pser73
CASP9	CyclinD1	CASP9	MSI2	AKTp308	NURR77
CASP9	DLX1	CASP9	PPARgam	AKTp308	TCF4
CASP9	EGFR	HIF1a	JAB1	AKTp308	TRIM62
CASP9	HER3	HIF1a	MCL1	CASP9	C.cbl
CASP9	IGF.1	HIF1a	MSI2	CASP9	Jun.C.pser73
CASP9	IGFBP2	HIF1a	NURR77	CASP9	TRIM62
CASP9	NF2p	HIF1a	PPARgam	C.cbl	Jun.C.pser73
CASP9	NRP1	HIF1a	TCF4	C.cbl	NURR77
CASP9	NURR77	JAB1	MCL1	C.cbl	TCF4
CASP9	OPN	JAB1	MSI2	C.cbl	TRIM62
CASP9	PI3Kp110	JAB1	NURR77	Jun.C.pser73	NURR77
CASP9	PKCB.I	JAB1	PPARgam	Jun.C.pser73	TCF4
CASP9	PLAC1	JAB1	TCF4	Jun.C.pser73	TRIM62
CASP9	SMAD6	MCL1	MSI2	NURR77	TRIM62
CASP9	STAT3	MCL1	NURR77	TCF4	TRIM62
CASP9	Survivin	MCL1	PPARgam		
CASP9	TAZ	MCL1	TCF4		
CASP9	VHL	MSI2	NURR77		
CASP9	YAP	MSI2	PPARgam		
CateninA	CateninBp	MSI2	TCF4		
CateninA	Cox2	NURR77	PPARgam		
CateninA	CyclinD1	PPARgam	TCF4		
CateninA	DLX1				
CateninA	EGFR				
CateninA	HER3				
CateninA	IGF.1				
CateninA	IGFBP2				
CateninA	NF2p				
CateninA	NRP1				
CateninA	OPN				
CateninA	PI3Kp110				
CateninA	PKCB.I				
CateninA	PLAC1				
CateninA	SMAD6				
CateninA	STAT3				

CateninA	Survivin				
CateninA	TAZ				
CateninA	TCF4				
CateninA	VHL				
CateninA	YAP				
CateninBp	Cox2				
CateninBp	CyclinD1				
CateninBp	DLX1				
CateninBp	EGFR				
CateninBp	HER3				
CateninBp	IGF.1				
CateninBp	IGFBP2				
CateninBp	NF2p				
CateninBp	NRP1				
CateninBp	OPN				
CateninBp	PI3Kp110				
CateninBp	PKCB.I				
CateninBp	PLAC1				
CateninBp	SMAD6				
CateninBp	STAT3				
CateninBp	Survivin				
CateninBp	TAZ				
CateninBp	TCF4				
CateninBp	VHL				
CateninBp	YAP				
Cox2	CyclinD1				
Cox2	DLX1				
Cox2	EGFR				
Cox2	HER3				
Cox2	IGF.1				
Cox2	IGFBP2				
Cox2	NF2p				
Cox2	NRP1				
Cox2	OPN				
Cox2	PI3Kp110				
Cox2	PKCB.I				
Cox2	PLAC1				
Cox2	SMAD6				
Cox2	STAT3				
Cox2	Survivin				
Cox2	TAZ				

Cox2	TCF4				
Cox2	VHL				
Cox2	YAP				
CyclinD1	DLX1				
CyclinD1	EGFR				
CyclinD1	HER3				
CyclinD1	IGF.1				
CyclinD1	IGFBP2				
CyclinD1	NF2p				
CyclinD1	NRP1				
CyclinD1	OPN				
CyclinD1	PI3Kp110				
CyclinD1	PKCB.I				
CyclinD1	PLAC1				
CyclinD1	SMAD6				
CyclinD1	STAT3				
CyclinD1	Survivin				
CyclinD1	TAZ				
CyclinD1	TCF4				
CyclinD1	VHL				
CyclinD1	YAP				
DLX1	EGFR				
DLX1	HER3				
DLX1	IGF.1				
DLX1	IGFBP2				
DLX1	NF2p				
DLX1	NRP1				
DLX1	OPN				
DLX1	PI3Kp110				
DLX1	PKCB.I				
DLX1	PLAC1				
DLX1	SMAD6				
DLX1	STAT3				
DLX1	Survivin				
DLX1	TAZ				
DLX1	TCF4				
DLX1	VHL				
DLX1	YAP				
EGFR	HER3				
EGFR	IGF.1				
EGFR	IGFBP2				

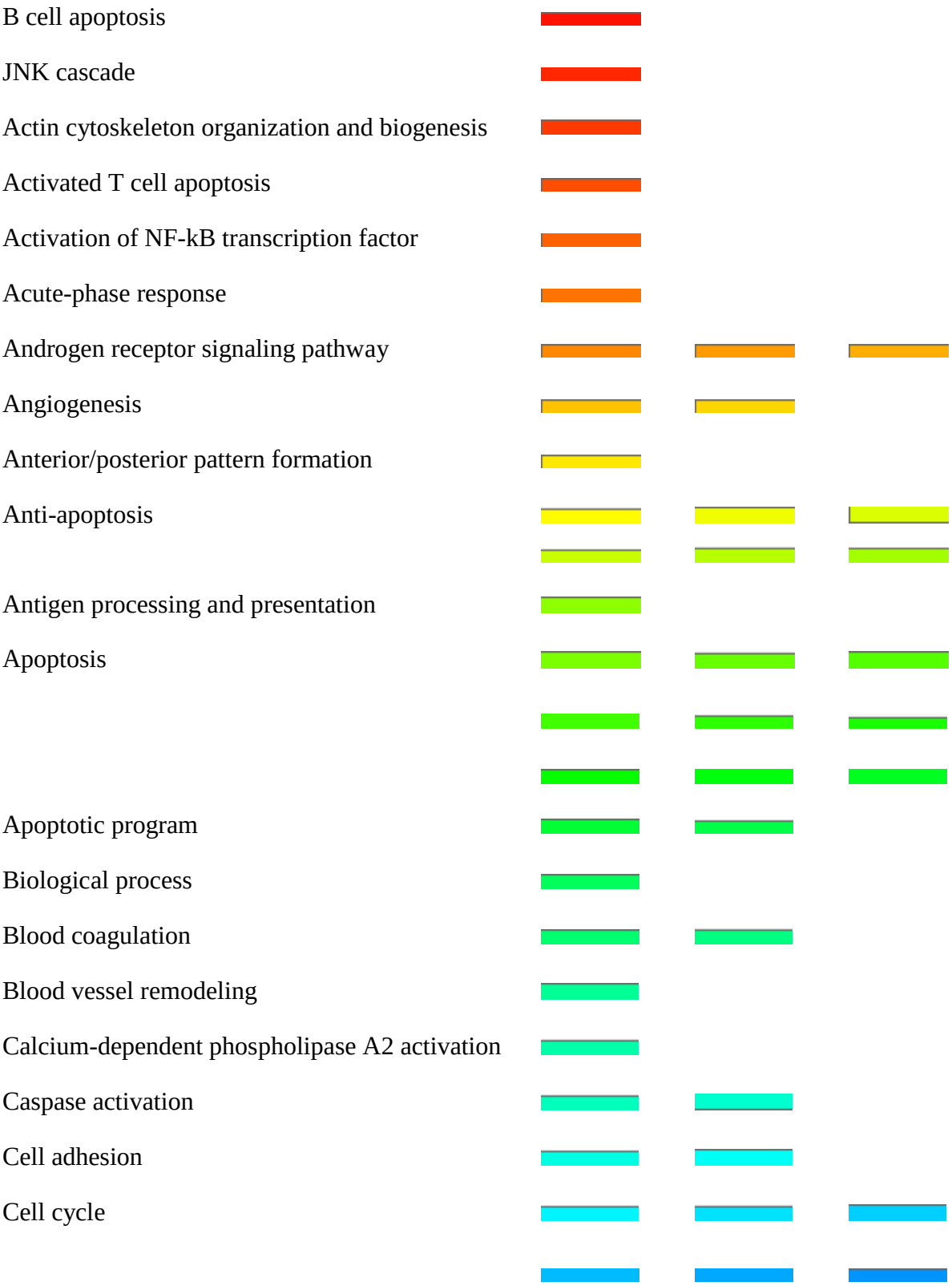
EGFR	NF2p				
EGFR	NRP1				
EGFR	OPN				
EGFR	PI3Kp110				
EGFR	PKCB.I				
EGFR	PLAC1				
EGFR	SMAD6				
EGFR	STAT3				
EGFR	Survivin				
EGFR	TAZ				
EGFR	TCF4				
EGFR	VHL				
EGFR	YAP				
HER3	IGF.1				
HER3	IGFBP2				
HER3	NF2p				
HER3	NRP1				
HER3	OPN				
HER3	PI3Kp110				
HER3	PKCB.I				
HER3	PLAC1				
HER3	SMAD6				
HER3	STAT3				
HER3	Survivin				
HER3	TAZ				
HER3	TCF4				
HER3	VHL				
HER3	YAP				
IGF.1	IGFBP2				
IGF.1	NF2p				
IGF.1	NRP1				
IGF.1	OPN				
IGF.1	PI3Kp110				
IGF.1	PKCB.I				
IGF.1	PLAC1				
IGF.1	SMAD6				
IGF.1	STAT3				
IGF.1	Survivin				
IGF.1	TAZ				
IGF.1	TCF4				
IGF.1	VHL				

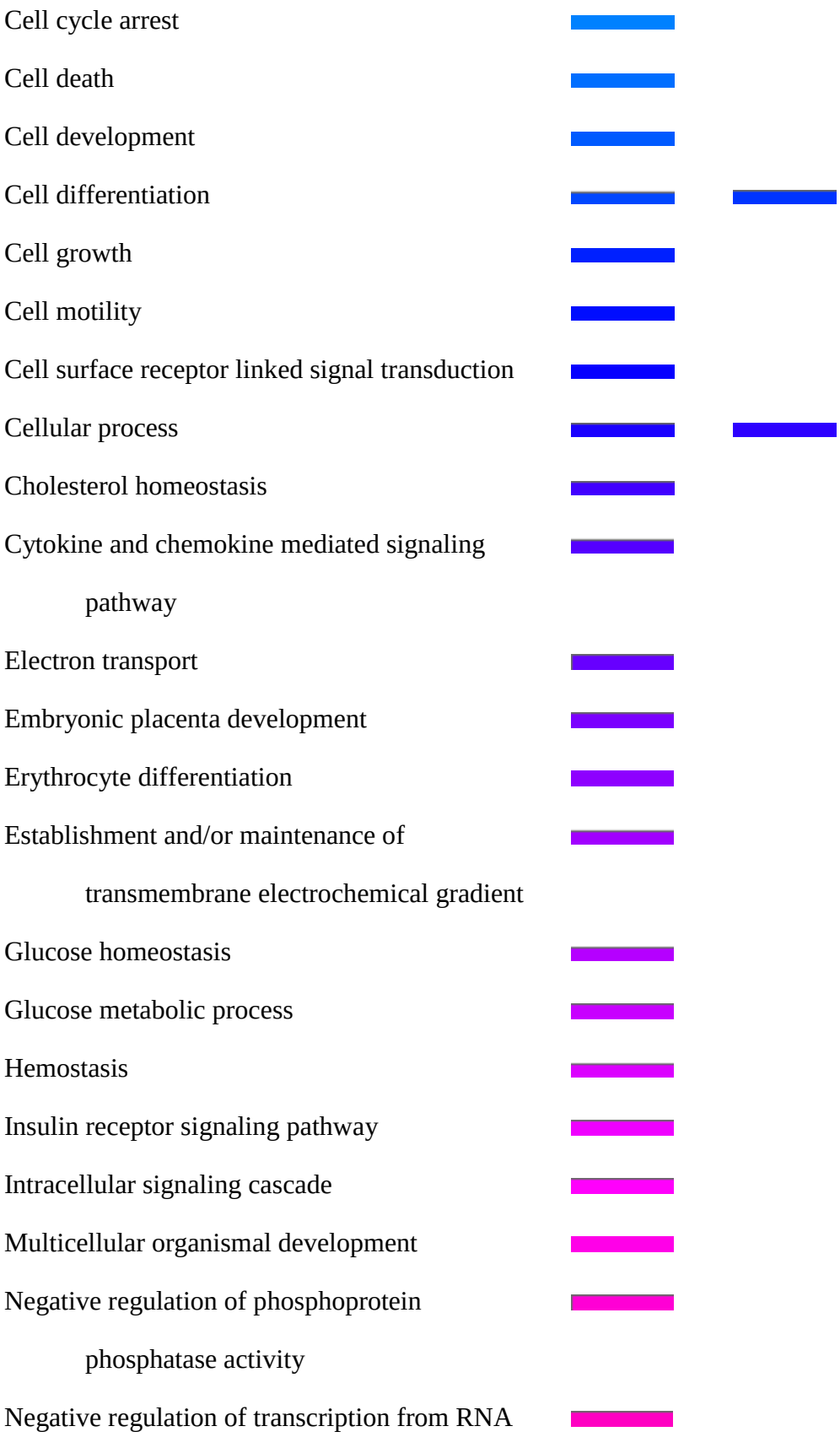
IGF.1	YAP				
IGFBP2	NF2p				
IGFBP2	NRP1				
IGFBP2	OPN				
IGFBP2	PI3Kp110				
IGFBP2	PKCB.I				
IGFBP2	PLAC1				
IGFBP2	SMAD6				
IGFBP2	STAT3				
IGFBP2	Survivin				
IGFBP2	TAZ				
IGFBP2	TCF4				
IGFBP2	VHL				
IGFBP2	YAP				
NF2p	NRP1				
NF2p	OPN				
NF2p	PI3Kp110				
NF2p	PKCB.I				
NF2p	PLAC1				
NF2p	SMAD6				
NF2p	STAT3				
NF2p	Survivin				
NF2p	TAZ				
NF2p	TCF4				
NF2p	VHL				
NF2p	YAP				
NRP1	OPN				
NRP1	PI3Kp110				
NRP1	PKCB.I				
NRP1	PLAC1				
NRP1	SMAD6				
NRP1	STAT3				
NRP1	Survivin				
NRP1	TAZ				
NRP1	TCF4				
NRP1	VHL				
NRP1	YAP				
NURR77	TCF4				
OPN	PI3Kp110				
OPN	PKCB.I				
OPN	PLAC1				

OPN	SMAD6				
OPN	STAT3				
OPN	Survivin				
OPN	TAZ				
OPN	TCF4				
OPN	VHL				
OPN	YAP				
PI3Kp110	PKCB.I				
PI3Kp110	PLAC1				
PI3Kp110	SMAD6				
PI3Kp110	STAT3				
PI3Kp110	Survivin				
PI3Kp110	TAZ				
PI3Kp110	TCF4				
PI3Kp110	VHL				
PI3Kp110	YAP				
PKCB.I	PLAC1				
PKCB.I	SMAD6				
PKCB.I	STAT3				
PKCB.I	Survivin				
PKCB.I	TAZ				
PKCB.I	TCF4				
PKCB.I	VHL				
PKCB.I	YAP				
PLAC1	SMAD6				
PLAC1	STAT3				
PLAC1	Survivin				
PLAC1	TAZ				
PLAC1	TCF4				
PLAC1	VHL				
PLAC1	YAP				
SMAD6	STAT3				
SMAD6	Survivin				
SMAD6	TAZ				
SMAD6	TCF4				
SMAD6	VHL				
SMAD6	YAP				
STAT3	Survivin				
STAT3	TAZ				
STAT3	TCF4				
STAT3	VHL				

STAT3	YAP				
Survivin	TAZ				
Survivin	TCF4				
Survivin	VHL				
Survivin	YAP				
TAZ	TCF4				
TAZ	VHL				
TAZ	YAP				
TCF4	VHL				
TCF4	YAP				
VHL	YAP				

Supplemental Figure 1.




Cell cycle arrest
Cell death
Cell development
Cell differentiation
Cell growth
Cell motility
Cell surface receptor linked signal transduction
Cellular process
Cholesterol homeostasis
Cytokine and chemokine mediated signaling pathway
Electron transport
Embryonic placenta development
Erythrocyte differentiation
Establishment and/or maintenance of transmembrane electrochemical gradient
Glucose homeostasis
Glucose metabolic process
Hemostasis
Insulin receptor signaling pathway
Intracellular signaling cascade
Multicellular organismal development
Negative regulation of phosphoprotein phosphatase activity
Negative regulation of transcription from RNA

polymerase II promoter

Oxygen homoeostasis

Phosphoinositide phosphorylation

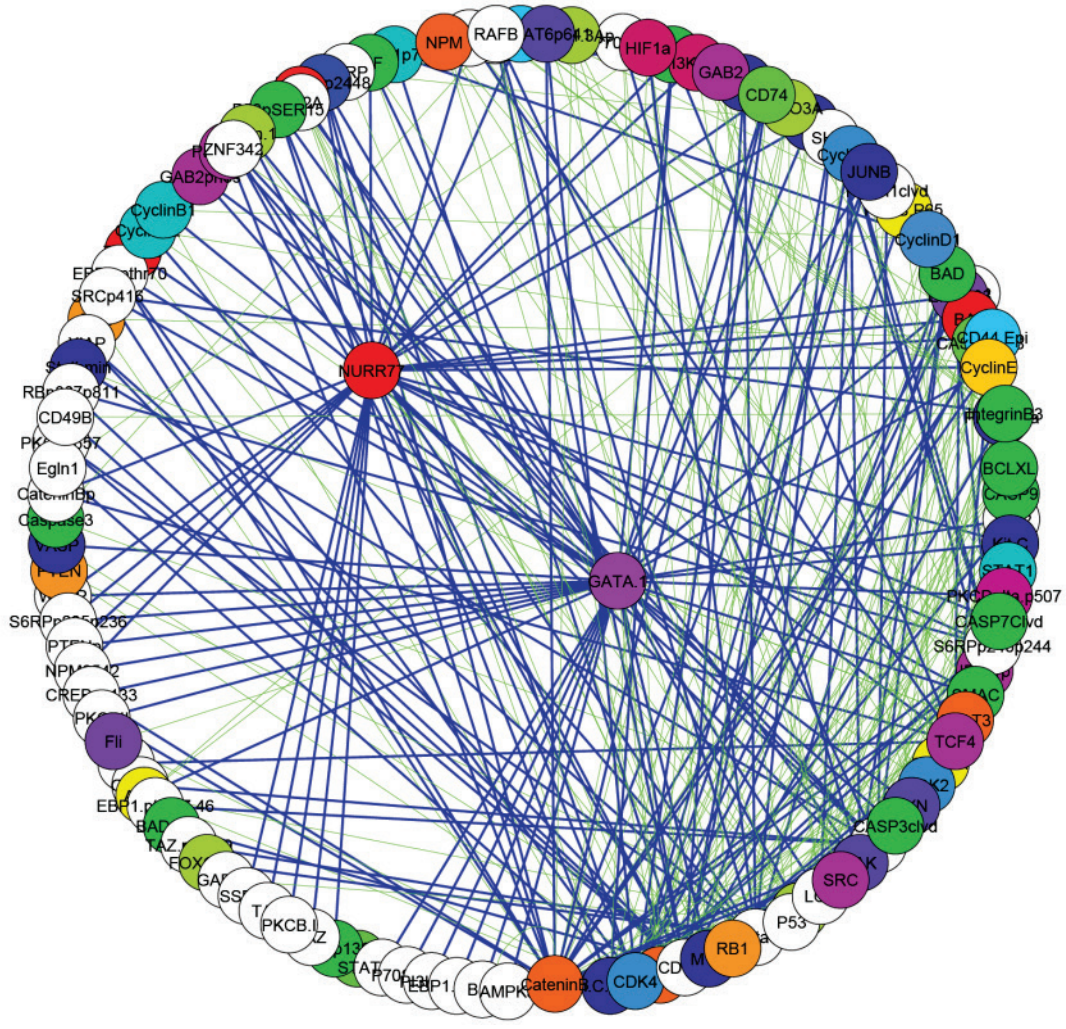
Positive regulation of transcription from RNA polymerase II promoter

Protein amino acid phosphorylation

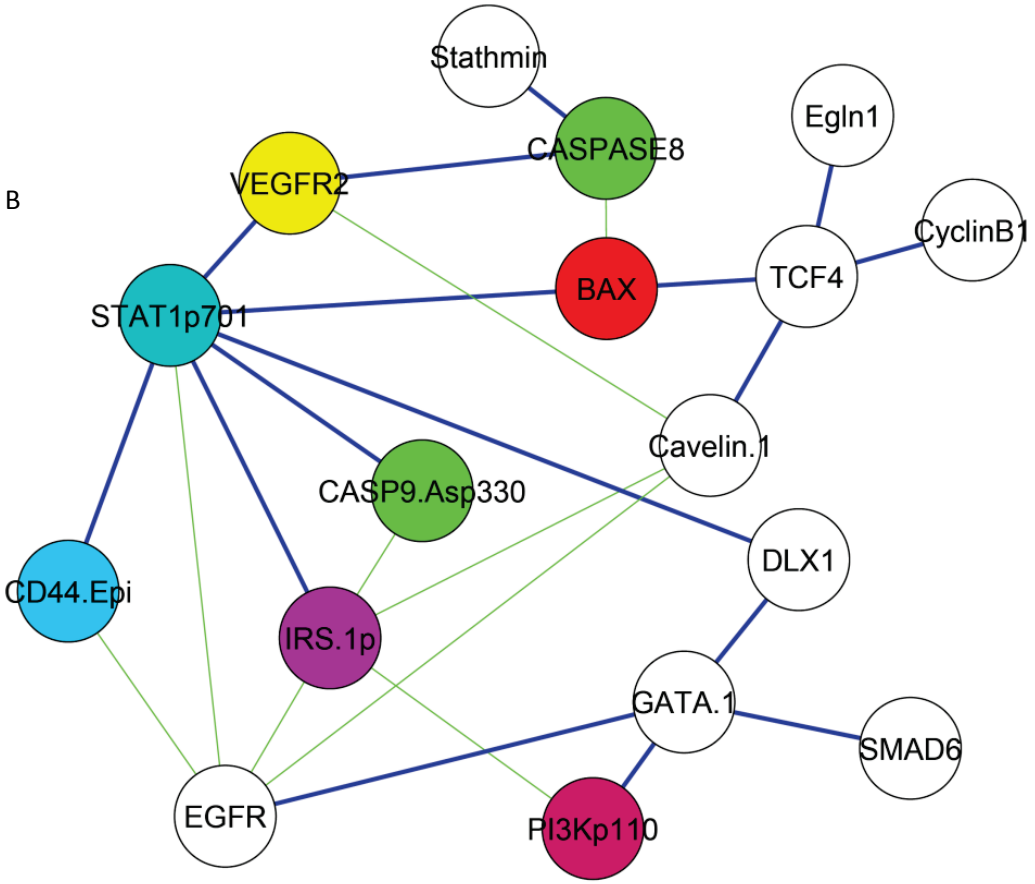
Regulation of transcription, DNA-dependent

Supplemental Figure 2. Interactions between proteins identified as part of a highly different pair or pairs, at p-value = 10^{-10}, graphed in Cytoscape combined with direct interactions between the proteins found from public databases. (A) inv(16) patients compared to normals. Group 1. (B) t(8;21) patients compared to normals. Group 2. (C) Overlap between inv(16) and t(8;21) patients compared to normals. Group 3. Note: public database information was available for the full protein and did not include post-translational changes like phosphorylation, which the patient dataset does. Green edges = public database. Blue edges = patient data.

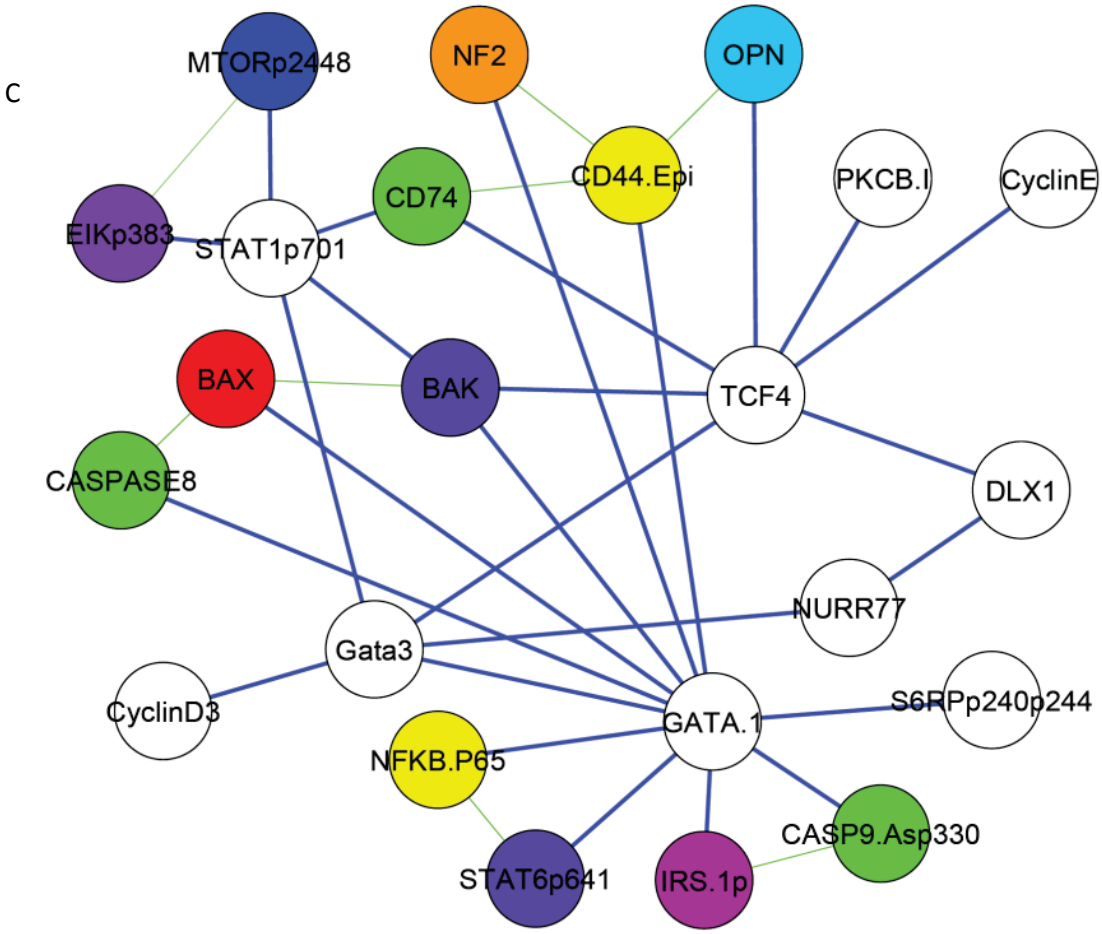
A



B


Stathmin
CASPASE8
Egln1
VEGFR2
CyclinB1
BAX
TCF4
STAT1p701
Cavelin.1
CASP9.Asp330
DLX1
CD44.Epi
IRS.1p
GATA.1
SMAD6
EGFR
PI3Kp110

C


MTORp2448
NF2
OPN
CD44.Epi
PKCB.I
CyclinE
CD74
EIKp383
STAT1p701
BAX
BAK
TCF4
CASPASE8
DLX1
NURR77
Gata3
CyclinD3
GATA.1
S6RPp240p244
NFKB.P65
CASP9.Asp330
STAT6p641
IRS.1p

Supplemental Figure 3. Calculations of the edge weights for the combined t-test and lasso regression network graphs. The current weighting factor is a 50% weight for lasso regression at $\lambda = 2.84$ and 50% weight for the t-test at $\alpha = 1 \times 10^{-10}$.

Weighting values were determined by normalizing the exponential fits of the number of significant protein pairs found vs. p-value (A) and the number of significant protein pairs found vs. λ value (B). $R^2 > 0.98$ in both fits. The final equation (Eqn. S3) allowed us to weight the data within a range of edge values that would be visible when applied to the network graphs in Cytoscape.

$$K = (17021e^{0.6257\alpha}) * (34161e^{0.6257\lambda}) \quad \text{(S1)}$$

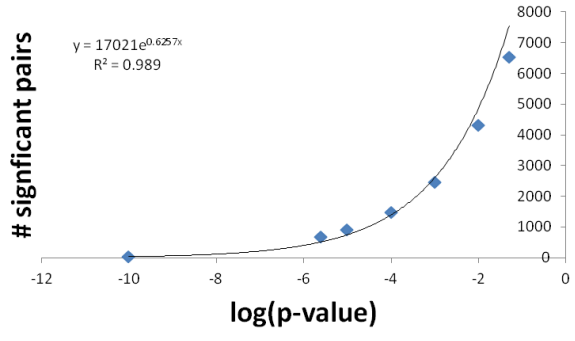
$$\frac{1}{K_{max}} * K \quad \text{(S2)}$$

$$10 * \log(100 * \frac{1}{K_{max}} * K) \quad \text{(S3)}$$

Below a calculated weight of 2.5, all patient data edge widths are set to 2.5. All protein databases edges have a width of 1.5.

α	λ	Equation Value	Calculated Weight	Actual Edge Weight
1E-10	2.84	0.0016	1.0000	20.0
2.44E-06	1.89	0.0261	0.0624	8.0
0.00001	1.42	0.0439	0.0370	5.7
0.0001	1.13	0.0908	0.0179	2.5
0.001	0.95	0.1723	0.0094	2.5
0.01	0.71	0.3500	0.0046	2.5
0.05	0.57	0.5592	0.0029	2.5
1E-10	0	1.6677	0.0010	2.5

A



B