# INTEGRATING FUNCTIONAL GENOMICS, PROTEOMICS AND COMPUTATIONAL ANALYSIS FOR THE CHARACTERIZATION OF CELLULAR NETWORKS

APPROVED BY SUPERVISORY COMMITTEE

Michael A. White, Ph.D

Jerry Shay, Ph.D

Melanie Cobb, Ph.D

Kevin Rosenblatt, M.D. Ph.D

Richard Scheuermann, Ph.D

# ACKNOWLEDGMENTS

First of all, I would like to thank my mentor Dr. Michael White, in whom I see not only an exceptional scientist, but also an extraordinary person. His wisdom and patience, more than anything, have been crucial in my education to become a mature scientist. His support of my computational projects, even though out of the scope of his immediate interests, was instrumental for me to be able to develop computational skills that constitute a major part of systems biology.

I would also like to thank Dr. Kevin Rosenblatt, both a committee member and a collaborator, in whose lab I learned and performed the reverse-phase protein arrays. I also thank all White lab members for their understanding and kindness, especially Jackie and Kiran for their help during my initial years, Tzuling for her help with RT-PCRs; and Rosenblatt lab members Johanne and Prem for helping me learn the protein arrays. My other committee members Drs. Melanie Cobb, Richard Scheuermann and Jerry Shay were also very helpful.

The true source of my motivation was, no doubt, my family. Despite the fact that I'm their only son and they would very much wish for me to stay at home (as is the normal case in my home country), my parents encouraged me to study as

much as I can even though that would mean me leaving them for a long time. Their strong belief in me has been a great source of courage and self-confidence for me since my early childhood. Feeling their support and knowing their expectations of me was the main source of my strength throughout my education.

Finally, I want to thank my wife, Lachyn, who has opened a new era in my life. She left her comfortable life in Turkmenistan, along with her family, a lot of very close friends and a promising career, to come to USA with me. At times, her strength and selfless support for my work seemed extraordinary, therefore setting new heights for me to achieve. Through her, I found a new urge to become better, the best.

# INTEGRATING FUNCTIONAL GENOMICS, PROTEOMICS AND COMPUTATIONAL ANALYSIS FOR THE CHARACTERIZATION OF CELLULAR NETWORKS

by

KAKAJAN KOMUROV

DISSERTATION

Presented to the Faculty of the Graduate School of Biomedical Sciences

The University of Texas Southwestern Medical Center at Dallas

In Partial Fulfillment of the Requirements

For the Degree of

DOCTOR OF PHILOSOPHY

The University of Texas Southwestern Medical Center at Dallas

Dallas, Texas

August, 2008

**INTEGRATING FUNCTIONAL GENOMICS, PROTEOMICS AND COMPUTATIONAL ANALYSIS FOR THE CHARACTERIZATION OF CELLULAR NETWORKS**

KAKAJAN KOMUROV, Ph.D.

The University of Texas Southwestern Medical Center at Dallas, 2008

MICHAEL A. WHITE, Ph.D.

As the study of biological systems progresses from a molecular level to a systems level, the development of new methodology for efficient data acquisition has been a key challenge of biological research in recent years. While development of novel high throughput experimental platforms is essential for an accurate large-scale data collection, novel theoretical methodology is indispensable for proper analysis and interpretation of these data. My projects aim at both, developing novel theoretic-analytical methodology for the analysis of functional patterns in biological networks, and also establishing a high throughput experimental platform for the study of signaling pathways.

I have developed a generalized method for the analysis of functional organization in complex networks. This method makes use of several novel metrics used to characterize a node's status in the network. After the nodes are clustered according to their characteristics, statistically significant organizational patterns are revealed by random simulations of the network. Using this approach, I have found important characteristics of eukaryotic protein interaction networks that have direct implications in cellular phenomena like robustness and the efficiency of information processing. I have identified an entirely new class of functional modules with unique properties that contribute to the variability in cellular phenotypes. In addition, my analyses have uncovered a distinct pattern of organization in the protein network (called "rich club connectivity") that provides mechanistic explanations for some cell biological phenomena. This work not only reveals a highly organized functional dynamic layout of the protein interaction network, but also refines and/or corrects several notions proposed by previous studies.

Functional genomic screens are a powerful tool for finding novel components of biological networks. However, in order to make these screens effective for assays that may require multiple readouts, it is necessary to channel the assay to another high throughput platform. Here, I used high throughput RNAi as a loss-of-function screen, and reverse-phase protein arrays as a high throughput readout

tool for the study of signal transduction downstream of the EGF receptor in human cells. I knocked down the expression of each human kinase in A431 cells, and measured the response of each perturbation to EGF stimulation by using reverse phase protein arrays. As readouts, I used phospho-STAT3 and phospho-ERK, two important signaling molecules downstream of the EGF receptor. In addition to identifying some of the obligate as well as novel components of the EGF signaling network, the screen also revealed some novel global characteristics of EGF signaling. By using a network-based bioinformatics approach, I was able to extract some important relationships regarding the differential regulation of STAT3 and ERK pathways in response to EGF. This work shows a high promise of integration of genomics and proteomics platforms for the study of signal transduction.

## TABLE OF CONTENTS

## PREVIOUS PUBLICATIONS

Komurov K, Pastor J, Chen T, Rosenblatt KP and White MA (2008). Kinome-wide RNAi screen for the regulators of EGF signaling using reverse-phase protein arrays. **In preparation.**

Komurov K, Gunes MH, Sarac K, White MA (2008). Fine-scale dissection of functional protein network organization by statistical network analysis. **submitted**

Komurov K, White MA (2007). Revealing static and dynamic modular architecture in the eukaryotic protein interaction network. Mol Syst Biol 3:110

# List of Figures

# List of Tables

# Abbreviations

| | |
|---|---|
| CGH: | comparative genome hybridization |
| PCC: | Pearson Correlation Coefficient |
| avPCC: | average neighborhood PCC |
| nPCC: | neighborhood PCC |
| EV: | Expression Variance |
| nEV: | neighborhood EV |
| RPPA: | Reverse phase protein arrays |
| RNAi: | RNA interference |
| TNK1: | Thirty-eight Negative Kinase 1 |
| PTPRR: | Protein Tyrosine Phosphatase, non-receptor type |
| ERK: | Extracellular-regulated kinase |
| STAT: | Signal transducer and activator of transcription |
| HUNK: | Hormonally-upregulated Neu kinase |
| DAPK: | Death-activated protein kinase |
| DGKD: | Diacyl-glycerol kinase delta |
| PFKFB: | Phosphofruktokinase, bisphosphatase |
| MERTK: | c-mer proto-oncogene tyrosine kinase |
| UCK: | Uridine-cytidine kinase |
| EGF: | Epidermal growth factor |
| EGFR: | EGF receptor |
| CSF1R: | colony-stimulating factor 1 receptor |
| MAPK: | Mitogen-activated protein kinase |
| PLCγ: | phospholipase C gamma |
| PLD: | phospholipase D |
| GRB: | growth factor receptor bound protein |
| SOS: | son of sevenless |
| NF-κB: | nuclear factor kappa B |
| SNCA: | synuclein alpha |
| MAP3K: | mitogen-activated protein kinase kinase kinase |
| PIK: | phospho-inositide kinase |
| CSNK: | casein kinase |
| Sur: | Suppressor of Ras |

# CHAPTER 1

# Introduction

## 1.1 FROM SYSTEMS TO MOLECULES AND BACK

The second half of the past century has been a stage for major developments in biology. From the discovery of the DNA structure to the sequencing of the human genome, we have witnessed the evolution of biology from macro-scale studies, where only system-level observations could be made, to molecular biology, where we began unraveling the molecular bases of macroscopic biological phenomena. We came to appreciate the incredibly intricate mechanisms driving organismal development, very fine-tuned interplay of cell cycle and cell death pathways to govern the cell fate, and genetic mechanisms of many diseases like cancer.

The advent of high throughput technologies in the past decade, and subsequent accumulation of biological data, has helped the next transition of biological research back to systems-level analyses. However unlike the initial years of biological research, systems-level analyses in so-called “Systems Biology” is a study of systems properties arising from a specific pattern of underlying molecular interactions. Therefore, Systems biology is a natural extension of molecular biology in a way where we first uncover the bits and pieces making up

the system (molecular biology), and then we put them back together to get a greater understanding of the system from a global perspective (systems biology). While molecular biology seeks to find the specific molecular mechanisms of the biological phenomena, systems biology is interested in the systems mechanisms that arise out of the collection of the molecular phenomena.

### 1.1.1 The era of Genomics: rise of Systems Biology

Perhaps the first systems-level analyses of the cell were made possible by the invention of cDNA microarrays, whereby the expression levels of every mRNA in the cell could be, in principle, measured. This led to the definition of so-called gene expression "patterns", or "signatures" that are specific to a given cell type (DeRisi, et al., 1996). In their seminal study, Golub et al for the first time described the classification of acute leukemias into distinct classes based on their systems-level gene expression patterns (Golub, et al., 1999). Development of another technology, high throughput Yeast 2 Hybrid (Y2H), allowed for massive screening of proteins for their physical interactions with each other (Ito, et al., 2001; Ito, et al., 2001; Uetz, et al., 2000), which opened the door for network analysts in other disciplines to flood into biology. Barabasi, a statistical physicist, first described some of the global properties of protein interaction networks that particularly make them robust to random mutations (Barabasi and Albert, 1999;

Jeong, et al., 2001; Jeong, et al., 2000). Shortly afterwards, analyses combining gene expression patterns with the protein interaction data emerged as an attempt to derive global properties of biological systems dynamics with implications in cellular robustness, evolvability and information processing (Bader, et al., 2004; Batada, et al., 2006; de Lichtenberg, et al., 2005; Han, et al., 2004; Harbison, et al., 2004; Ihmels, et al., 2002; Ihmels, et al., 2004; Komurov and White, 2007; Luscombe, et al., 2004). These pioneering studies were followed by many others utilizing various high throughput approaches to study cellular processes at a systems-level, and those utilizing novel theoretical tools for the interpretation of the increasingly accumulating data. This transition of molecular-level approach to a systems-level approach in biology has created a large necessity for the development of novel tools for systems biology analysis.

## 1.2 METHODOLOGY OF SYSTEMS BIOLOGY

A major shift in paradigm from a molecular-level research to a systems-level research (Hartwell, et al., 1999) has ushered a need for novel experimental tools for efficient and reliable large-scale data acquisition as well as theoretical tools for the analysis of the sea of biological data accumulated through these efforts.

### 1.2.1 High throughput technology for efficient data acquisition

Large-scale data collection in an unbiased manner lies at the heart of systems biology. Recent advances in genomics and proteomics and their integration have radically transformed the process of data acquisition for systems-level analyses. Most experimental approaches in systems biology are aimed at identifying components of biological networks and the interactions between them. The process of identification of novel components of protein networks and their post-translational modifications has been drastically accelerated thanks to the development of high throughput genetic screens, mass spectrometry-based methodology and protein arrays (Cho, et al., 2006). Each has been applied within various settings for successful elucidation of biological networks.

*RNAi screens*

Genetic screens are the most widely used tool for the identification of novel components of biological networks. These screens have been essential in the identification of key biological processes like cell cycle and apoptosis and of their components in model organisms. However approaches for routine large-scale genetic screens in mammalian cells were lacking. Perhaps the most popular genetic screen in mammalian cells in the recent years has been made possible by the development of high throughput RNA interference screens. The discovery that short double-strand RNA can silence the expression of specific genes through

RNA interference (RNAi) led to the extensive harnessing of this mechanism for functional analysis of genes in cultured cells (Hannon and Rossi, 2004). RNAi has become a primary genetic tool for loss-of-function approaches in mammalian cells, and even has the potential to be exploited therapeutically (Hannon and Rossi, 2004).

High throughput RNAi libraries have generated a unique platform for functional genomics where each gene in the genome can be individually interrogated for its role in a given process (Silva, et al., 2004). RNAi molecules each targeting a specific gene are introduced into cells in mass or one by one, and the cells can be screened for a phenotype of interest to identify genes involved in the given process. Many studies have utilized this platform for genome-wide screens for successful identification of novel components of various cellular networks (Iorns, et al., 2008; MacKeigan, et al., 2005; Moffat and Sabatini, 2006; Whitehurst, et al., 2007). Most of these studies, however, utilized a simple phenotypic readout: cell death versus cell proliferation. The ability to assess more complex phenotypes in a high throughput platform will allow for the design of genome-wide screens with more potential information about the underlying cellular networks. Several emerging technologies in high throughput proteomics may offer a unique advantage towards achieving such a goal.

*Protein arrays*

The progress in proteomic research giving insight into the organization of cellular networks has been enabled not only by the rapidly maturing mass spectrometry-based methods (Cox and Mann, 2007), but also by protein arrays, which have progressed considerably over the past years (Liotta, et al., 2003). There are typically two types of protein arrays depending on whether the antigen is fixed or soluble, so-called forward phase protein arrays and reverse-phase protein arrays. In forward-phase arrays, also called antibody arrays, antibodies are arrayed on a slide and immobilized, and a cellular lysate is applied on top in order to detect the abundance of antigens in the lysate as reported by the antibodies on the array. In reverse-phase arrays, however, a large number of samples are arrayed and immobilized on a slide, and a specific antibody of interest is applied on top in order to detect the abundance of the antigen of interest in each of the samples. As opposed to forward-phase protein arrays, which are used to measure a multitude of antigens within a single sample, reverse-phase protein arrays measure a smaller number of antigens within a large number of samples (Liotta, et al., 2003). Reverse-phase protein arrays can uniquely quantify protein levels, post-translational modifications and cleaved products from a limited amount of sample. As opposed to gene expression microarrays, which can only be used for cell samples, reverse-phase protein arrays have also been successfully applied for serum samples or body fluids for the identification of various biomarkers

(VanMeter, et al., 2007). The amenability of these arrays to high throughput formats makes them especially attractive for integrative approaches.

### 1.2.2 Numerical methodology for multi-dimensional data analysis

Accumulating body of biological data warrants automated computational techniques for their analysis, interpretation and model predictions. The prime goal of biological research is the understanding of biological systems. Although molecular biology has uncovered the multitude of biological facts, such as gene functions, protein properties and cellular processes, experimentation alone is not sufficient for understanding the complexity of biological systems. Due to the inherent complexity of biological systems, a combination of quantitative experimentation with proper computational tools is necessary to illuminate the principles underlying biology at genetic, molecular and cellular level.

Computation is used to build mathematical models using the vast experimental data in order to make testable predictions as well as to gain more insight into the underlying biological system behavior. The most widely used application of computational biology is data mining, which aims at extracting hidden patterns with potential functional consequences from huge quantities of experimental data. Biological data mining usually involves the already existing analytical

methodology in other theoretical disciplines like computational chemistry, theoretical physics or computer science and most often contains a heavy statistical component. This sort of analyses is most useful for generating testable hypotheses about the nature of biological systems that are being analyzed.

Some well-known examples to this class of techniques would include sequence analysis, protein structure prediction or gene finding. Perhaps the best known computational methodology in this class, however, would be microarray gene expression analysis techniques that are aimed at inferring gene regulatory networks out of high throughput microarray data (Allison, et al., 2006). Starting from techniques as simple as clustering (Eisen, et al., 1998), principal component analysis (Brown, et al., 2000) or linear modeling to more sophisticated methodology like probabilistic graphical modeling (Friedman, 2004), molecular concept maps (Tomlins, et al., 2007) or information-theoretic inference (Basso, et al., 2005), microarray gene expression analysis has evolved considerably during the past decade. Comprehensive microarray gene expression data have been used to reconstruct complex genetic regulatory networks and to extract systems-level properties of gene expression regulation under various conditions in yeast (Luscombe, et al., 2004; Segal, et al., 2003; Stuart, et al., 2003). Extensive analyses of microarray data from human cancer cell lines has also led to the identification of gene regulatory modules as well as gene expression networks

playing roles in cancer progression (Basso, et al., 2005; Segal, et al., 2005; Tomlins, et al., 2007). Recent extension of microarray technology to detect DNA copy number changes (CGH arrays) and single-nucleotide polymorphisms (SNP arrays) has enabled the analyses of human diseases yet at another dimension involving genetic alterations (Kallioniemi, 2008).

*Network biology*

It is becoming increasingly clear that biological functions are rarely a result of the action of a single molecule. Instead, biological behavior of a cell arises from the many interactions between its constituents like proteins, DNA, RNA, lipids and small molecules. Therefore, the study of the structure and dynamics of complex intracellular web of interactions between biological molecules is a key challenge for systems biology (Barabasi and Oltvai, 2004).

Systems of interactions are represented by graphs in mathematics, where nodes (vertices) represent species, and edges (arcs) represent specific interactions between species. Although the term "network" traditionally only refers to a specific type of graphs, it has become a standard term for representing systems of complex interactions in various fields like physics, biology, computer science, social sciences and engineering (Wasserman and Faust, 1994). Bulk of work on the properties of complex networks and on data mining through networks has

been done under social network analysis. Many of the standard network characteristics like centrality, clustering, importance and network size and their functional implications have been laid out by social network analysts over the past few decades (Wasserman and Faust, 1994). With the recent advances in the study of complex networks towards uncovering the organizing principles that govern the formation and evolution of social and technological networks, we have gained a considerable understanding of general characteristics of complex networks. As the construction of biological networks became possible by the accumulation of high throughput biological interaction data, the already existing network analysis methodology from other disciplines was applied for the initial analyses of biological networks. Remarkably, it was found that many of the architectural features of biological networks are shared to a large degree by other complex systems, such as the Internet, food web and social networks (Barabasi and Oltvai, 2004), thereby allowing for the network analysis expertise in other fields to be used in biology.

Networks can be directed or undirected. In directed networks, the edges between nodes have directions thereby showing the direction of impact, whether it is a personal influence, signal flow or enzymatic catalysis. Undirected networks usually depict interactions that lack such a sense of directionality, such as a molecular interaction or a scientific collaboration. Networks can also be single

partite or bi-partite, depending on the nature of their nodes. Networks consisting of a single type of nodes are single-partite, and those with two types of nodes are bi-partite. Most of the current network theory is built on the study of simple networks with a single type of nodes and interactions. Although such networks, especially in case of biological networks, are far from faithfully encompassing the characteristics of most complex networks, much has been learnt from their studies in the past (Albert and Barabasi, 2002). In the case of biological networks, their integration with other data types has especially yielded useful insights into the organization and dynamics of biological networks.

Perhaps one of the pioneering studies to integrate heterogeneous biological data into network analyses was that by Han *et al*, where they integrated gene expression information with the architectural characteristics in networks to uncover a highly specific dynamic organization pattern in the protein interaction network of yeast (Han, et al., 2004). Later, similar approaches with combining different datasets were taken to further analyze the complex dynamic properties of protein-protein interaction networks (Ihmels, et al., 2004; Kharchenko, et al., 2005; Komurov and White, 2007) and to deduce sub-networks responsible for cancer progression (Boehm, et al., 2007; Chuang, et al., 2007; Rhodes, et al., 2005; Tomlins, et al., 2007). Recently, more sophisticated approaches to biological network analysis have been taken by integrating a number of

heterogeneous biological data to reconstruct the behavior of simple bacterial organisms (Bonneau, et al., 2007; Ishii, et al., 2007). Further advancement of biological network theory will greatly aid in the quantitative modeling of biological systems.

The work that will be presented in the following chapters deals with the two fundamental aspects of systems biology described above, namely development of computational analytical methodology for the analysis of cellular networks, and development of highly efficient integrated high throughput platform for the analysis of cellular processes. The computational methodology described in chapter 3 is aimed at detecting statistically significant patterns in the protein interaction networks, although it can be extended to other types of networks given the right input metrics. In chapter 4, I describe a pioneering study of integration of two high throughput platforms, a high throughput RNAi screen of the human kinome, and reverse-phase protein arrays for a large-scale detection of phospho-signals within a signaling network. Future work to further integrate the computational framework with a multidimensional high throughput platform will no doubt be invaluable in terms of establishing standardized methodology for improved understanding of cellular systems.

# CHAPTER 2

# Methods

## 2.1 COMPUTATIONAL TOOLS FOR DECIPHERING PROTEIN NETWORK ORGANIZATION

### 2.1.1. General methodology

First, biological network metrics are constructed to measure a specific property of each node in the network. Biological network metrics (see below) are then used to assign a "dynamic profile" to each protein. Dynamic profile $D_i$ of a node $x_i$ based on $n$ biological network metrics is defined as

$$D_i = \{f_1(x_i), f_2(x_i), ..., f_n(x_i)\}$$

where $f_k$ corresponds to a $k^{th}$ metric (one of below). So a dynamic profile of a protein in the network is the set of its values as given by the biological network metrics. Next, proteins are clustered according to the similarity of their dynamic profiles into several distinct groups. The statistical significance of each group is tested by comparing the dynamic profiles of each group with that of randomly re-wired networks. In detail:

1. A model to account for the dynamic profiles of the group *i* is created by calculating the means and standard deviations of each metric within the group, so that $\boldsymbol{\mu_i}$ and $\boldsymbol{\sigma_i}$ correspond to the means and standard deviations of the metrics in the group *i.*
2. The original network is permuted so that its node positions are randomly shuffled
3. For each position in the randomized network, a dynamic profile is created
4. The number of dynamic profiles in the randomized network with metrics corresponding to the interval $(\boldsymbol{\mu_i - \sigma_i}, \boldsymbol{\mu_i + \sigma_i})$ is calculated.

After sufficiently large iterations (~100) of the above procedure, a null model for each group is created, which represents the chance of occurrence of the dynamic profile represented by the group in a randomly re-wired network. The statistical significance (*P* value) of the dynamic profile of the group is then given by the fraction of the observations in the null model that are greater or equal to the number of proteins corresponding to the interval $(\boldsymbol{\mu_i - \sigma_i}, \boldsymbol{\mu_i + \sigma_i})$ in the original network. The lower the *P* value the less likely that the dynamic profile is likely to occur by chance in the given network.

### 2.1.2 Biological network metrics

First, we consider an |N| x |N| adjacency matrix $A$ of the network with a node set $N$ and an |N| x |N| expression correlation matrix $C$, constructed by calculating all pair-wise Pearson Correlation coefficients (PCC) of expression profiles of genes using our microarray compendium. $A$ is such that $A_{i,j}$ is 1 if proteins $i$ and $j$ interact, and 0 otherwise. $C$ is such that $C_{i,j}$ is the variance of the expression profile of gene $i$ if $i = j$, otherwise it is the Pearson correlation coefficient of expression profiles of genes $i$ and $j$. Variances of expression profiles of genes in the diagonal of $C$ are normalized so that their values reflect their quantile in the whole distribution of variances (i.e. these values range from 0 to 1).

The biological network metrics mentioned above can be divided into two categories; those that are node-based (i.e. inherent properties of a gene/protein that is not dependent on its connectivity properties in the network), and those that are network-based (i.e. properties that arise from the gene's/protein's properties that arise from its specific positioning in the network).

*Node-based metrics*

These can be any function that quantitatively describes a protein's property, for example its domain composition or its size in kilo-Daltons. Since we are dealing with the dynamic behavior of the protein's mRNA concentration in the cell, we defined two metrics that describe a protein's inherent properties.

Expression Variance: Expression variance is a gene-specific property that describes overall tendency of a gene's mRNA level to change in response to an external stimulus. This property can be useful in determining whether a gene's activity is likely to be regulated at the mRNA level or not. We defined Expression Variance, or simply *EV*, to be the statistical variance of a gene's expression levels across all the experimental conditions in our microarray compendium (see later), so that a low EV indicates that the gene has a static expression pattern and therefore its mRNA levels may not be regulated, while a high EV indicates a highly regulated expression pattern.

$$EV_i = \sigma_i = C_{i,i}$$

Dynamic degree: Dynamic degree (*yK*) is a dynamic equivalent of node degree in social networks. It is defined as the sum of its absolute PCC values with all proteins in the network,

$$yK_i = \sum_{j \in N} \left| C_{i,j} \right|$$

and reflects the number of proteins that it is co-regulated with. Formally, *yK* measures the size of the co-expression neighborhood of a protein, so that a protein with a high *yK* is probably a member of a gene expression program with many genes and therefore its expression may be tightly regulated, while a low *yK* would indicate that the protein's expression is not coupled to the expressions of other proteins in the network.

*Network-based metrics*

These metrics are the ones that are dependent on the local or global structural properties of the network. For the present moment, our network-based metrics take into account the local neighborhood structure of the protein.

Neighborhood EV: Neighborhood EV (*nEV*) is the average EV in the immediate neighborhood of a protein and is defined as

$$nEV_i = \frac{\sum_{j \in N} C_{i,j} A_{i,j}}{\sum_{j \in N} A_{i,j}}$$

Neighborhood EV reflects the expression variances of a protein's neighbors in the network. We will show that nEV can be particularly informative about a protein's location in the network; low nEV of proteins being a strong indicator that the protein is located within densely connected modules in the network (i.e. set of proteins dedicated to a specific cellular process) (see next chapter).

Variance in neighborhood EV ($v^{EV}$): This is variance of EV values in the neighborhood and is defined as:

$$v_i^{EV} = \frac{\sum_{j \in N} A_{i,j} (C_{j,j} - nEV_i)^2}{\sum_{j \in N} A_{i,j}}$$

where $nEV_i$ is the neighborhood EV of gene *i*. This function describes how variable the neighborhood of a protein is in terms of their EV, so that a

neighborhood with high $v^{EV}$ would suggest that the neighborhood of the protein is composed of proteins with non-similar expression variances and may indicate that the protein is not located within a module.

Neighborhood Pearson correlation coefficient (*nPCC*): This is the average expression correlation between neighbors of a protein and is defined as:

$$nPCC_i = \frac{\sum_{j,k \in n} C_{j,k}}{n^2}$$

where *n* is the set of neighbors of protein *i*. nPCC is a dynamic equivalent of the clustering coefficient in social networks, and as opposed to *connectivity* coherence in social networks, it shows the extent of *expression* coherence in a protein's neighborhood. High nPCC indicates that the neighbors of the protein are highly co-regulated and we will show that these proteins are located within a dynamically regulated module (dynamic module) whose protein constituents are highly co-expressed.

2nd neighborhood Pearson Correlation Coefficient (*nPCC2*): This is the average nPCC among neighbors of a protein. nPCC2 reflects the extent of co-regulation in the second neighborhood of a protein. A protein with high nPCC2 but low nPCC is most likely to be located "just outside" of a dynamic module and interacting with one or more proteins inside the module.

Variance in the neighborhood Pearson Correlation Coefficient ($v^{PCC}$): This is variance in correlation between neighbors of a protein and is defined as:

$$v_i^{PCC} = \frac{\sum_{j,k \in n} (C_{j,k} - nPCC_i)^2}{\sum_{j,k \in n} A_{j,k}}$$

PCC variance ($v^{PCC}$), reflects the variation in the co-regulation of proteins in the neighborhood. Like $v^{EV}$, $v^{PCC}$ shows how variable the neighborhood is, but unlike $v^{EV}$, $v^{PCC}$ also reports how similar or dissimilar the expression profiles of the neighbors are.

### 2.1.3 Ancillary materials and methods

*Datasets*

Microarray datasets: The microarray gene expression datasets from various conditions (Cell Cycle, Sporulation, Stress Response, Unfolded Protein Response and Diauxic Shift) were obtained from the Saccharomyces Genome Database (ftp://ftp.yeastgenome.org/yeast/). In order for the data to account for true fold differences in the expression of genes relative to the control (i.e. 0' time point), 0' time points were removed from the datasets, and the corresponding later time points were zero-transformed by subtracting the expression values at these time points from those at the 0 time point.

Protein-protein interaction dataset: Protein interaction network was compiled from studies of Krogan *et al* (2006) (high quality binary interaction data) and Bader *et al* (2004) (high quality interactions with a confidence cut-off of 0.65).

Protein expression noise values: For protein expression noise values, we used the values derived by a large-scale single cell proteomic analysis of Newman *et al* (2006). They defined protein expression noise as coefficients of variation (standard deviation divided by mean expression) of protein expression between cells in a population.

*Neighborhood function homology*

Let $G_i$ be the set of Gene Ontology (GO) terms assigned to protein $i$ that has a node degree of $k$. Neighborhood function homology $F_i$ of the protein $i$ is defined as

$$F_i = \frac{\sum_{j=1}^{k} \left| G_i \cap G_j \right|}{k \cdot \left| G_i \right|}$$

where $G_j$ is the set of GO terms assigned to the $j$th neighbor of protein $i$. $F_i$ ranges from 0, where there are no shared GO terms between protein $i$ and its neighbors, to 1, where all GO terms assigned to the protein $i$ are also present in all of its neighbors.

*In silico loss of function method for network connectivity analysis*

$D$ is an |N| x |N| matrix of shortest path distances (see Appendix) between all node pairs in the network, where N is the set of nodes. Let $D^{\mathbf{x}}$ be the distance matrix of a network formed by the deletion of a set $\boldsymbol{x} \subseteq N$ of nodes from the original network. A difference matrix $\Delta^k$ is such that $\Delta^k_{i,j} = D^{\mathbf{x}}_{i,j} - D_{i,j}$ if and only if $D_{i,j} \leq k$, and 0 otherwise. Value $k$ denotes the distance of interest. For example if $k = 2$ (our case), differences in distances between node pairs that are 1 node apart in the original network are considered, so that if $\Delta^k_{i,j} > 0$, we conclude that some node(s) in $\boldsymbol{x}$ are directly linking nodes $i$ and $j$ in the original network. If all distances are to be considered, $k = \infty$ should be chosen.

We consider a null model for matrix $\Delta^k$, by performing 20 random deletions of $|\boldsymbol{x}|$ number of proteins with node degrees similar to $\boldsymbol{x}$. $\Delta^k_{null}$ is such that

$$\Delta^k_{null\_i,j} = \left\langle D^{null}_{i,j} - D_{i,j} \right\rangle,$$

where $D^{null}$ is the distance matrix of network formed by a random deletion of $|\boldsymbol{x}|$ number of nodes, of which there are 20. Normalized form of the difference matrix therefore becomes

$$\Delta^k_{norm\_i,j} = \log\left( \frac{\Delta^k_{i,j}}{\Delta^k_{null\_i,j}} \right)$$

where each $i, j$ position gives the amount of impact on the path length between nodes $i$ and $j$ relative to what would be expected by chance.

*Rich club coefficients*

Rich club coefficient ($\varphi$) is defined as the density of interactions between nodes having node degrees larger than a specific value,

$$\phi_{>k} = \frac{2E_{>k}}{n_{>k}(n_{>k} - 1)}$$

where $E_{>k}$ is the number of edges between, and $n_{>k}$ is the number of, nodes that have node degrees higher than $k$. We define rich club coefficient within the group as

$$\phi_S = \frac{2E_S}{n_S(n_S - 1)}$$

where $E_S$ is the number of edges between, and $n_S$ is the number of, nodes in group $S$. A null model is considered by randomly shuffling the positions of nodes at one side of the adjacency matrix 100 times (equivalent to random rewiring of each node's connections), and calculating the corresponding rich club coefficients at each time. Normalization of the within-group rich club coefficients against null model is performed by

$$\phi_S^{'} = \frac{\phi_S - \mu_{null}}{\sigma_{null}}$$

where $\mu_{null}$ is the mean and $\sigma_{null}$ is the standard deviation of the distribution of the null model.

*Network modularity*

First, a function similarity matrix was constructed by measuring all pair-wise function similarities between proteins in the network. The pair-wise function similarity between proteins *i* and *j* was defined as

$$S_{i,j} = \frac{\left| G_i \cap G_j \right|}{\left| G_i \cup G_j \right|}$$

where $G_i$ and $G_j$ are the sets of GO terms assigned to proteins *i* and *j*, respectively. For a network of *n* proteins, the function similarity matrix *S* would be a matrix of dimensions *n* x *n*. The network modularity *M* is calculated by summing the pair-wise function similarities between every interacting pair of proteins in this network and dividing by the total number of interactions in the same network.

$$M = \frac{\sum_{i<j} A_{i,j} \cdot S_{i,j}}{\sum_{i<j} A_{i,j}}$$

where *A* is the adjacency matrix of the network and has the same dimensions as *S*. *A* is boolean, $A_{i,j}$ being 1 only if proteins *i* and *j* interact, and 0 otherwise.

*Partial correlation analysis*

Linear correlation between two variables *a* and *b* is given by

$$r_{a\sim b} = \frac{\mathrm{cov}(a,b)}{\sqrt{\mathrm{var}(a)\,\mathrm{var}(b)}}$$

where cov(*a*,*b*) is covariance between *a* and *b*, and var(*a*) is variance of *a*. Partial correlation between *a* and *b* while controlling for a variable *c* is given by

$$r_{a\sim b,c} = \frac{r_{a\sim b} - r_{a\sim c}r_{b\sim c}}{\sqrt{(1-r_{a\sim c}^2)(1-r_{b\sim c}^2)}}$$

## 2.2. HIGH THROUGHPUT SCREEN FOR STUDYING SIGNAL TRANSDUCTION

### 2.2.1 General scheme

The general format of our screen is given in Figure 2.1. Briefly, cells are transfected in 96-well plates and after 3 days of incubation they are stimulated and lysed. Lysates are printed onto nitrocellulose-coated slides and blotted with pERK, pSTAT3 and Actin antibodies, imaged and analyzed.

### 2.2.2 High throughput reverse transfection

*Reverse Transfection:* For transfections, we used the kinome siRNA library from Dharmacon, which consists of 10 96-well plates. Each well in this library contains a pool of 4 oligos targeting one specific gene and the kinome library targets a total of 672 distinct genes in the human kinome. In each well of our 96-well assay plates, 5 microliters of 2 micromolar siRNA (10 picomoles) from the library plates was mixed with 30 microliters of the Dharmacon Cell Culture Reagent (DCCR) and incubated for 5 minutes to make our siRNA solution. As a transfection reagent, we used DharmaFECT 3 transfection reagent (DF3) from Dharmacon. For each well, 0.3 microlitters of DF3 was diluted in 10 microliters

of DCCR and mixed with the siRNA solution prepared earlier. After letting the mixture sit for 30 minutes, 160 microliters of the cell solution is added. The cell solution is prepared by resuspending cells in DMEM supplemented with 10%

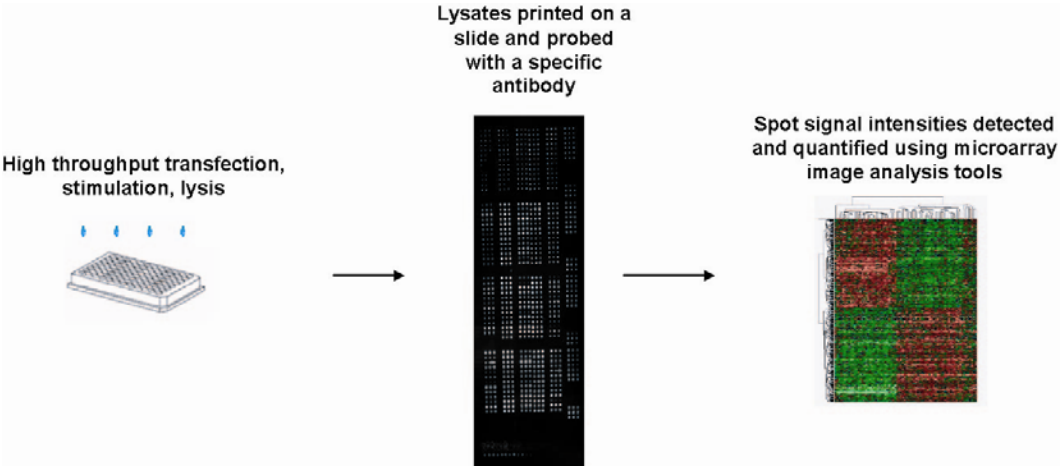


**Figure 2.1.** General scheme of our screen.

Fetal Bovine Serum (no antibiotics) to make the final concentration of 50,000 cells per milliliter. Therefore, each well was seeded with 8,000 cells during the transfection. Each plate was transfected in triplicate for biological reproducibility.

For our primary screen, preparation of siRNA solutions was performed with a high throughput liquid handling platform Biomek FX robotic liquid handler

(Beckman Coulter), while all the other high throughput dispensions were performed with TiterTek Multidrop.

*Stimulation and lysis:* After 3 days of incubation, 50 microliters of 250 ng/ml EGF (Sigma) in DMEM (final concentration of 50 ng/ml EGF) is added into each well and incubated for 5 minutes. Then, the cell media is dumped and 100 microliters of the 2% SDS lysis buffer (2% SDS, 5% Glycerol, 1% beta-mercapto ethanol, 68 mM Tris pH 6.8) is added into each well. Plates are immediately taken into -80C until the spotting procedure.

### 2.2.3 Reverse-phase protein microarrays

*Plate pre-processing:* To facilitate lysis and eliminate viscosity, plates are incubated at 80C for 45 minutes and briefly spinned to collect the drops. Lysates are then filtered using 96-well 0.7 micron filter plates.

*Slide printing:* For the primary screen, we used nitrocellulose-coated FAST slides from Whattman. Printing was performed in SpotArray 24 (PerkinElmer) using 4 pins. Slides were spotted 9 plates per slide, each plate in single column, each well was spotted in triplicate. In total, per each slide, we had 9 x 96 x 3 = 2592 spots. Slides are dried for 2 hours at room temperature or at 4C overnight before processing.

*Reverse-phase protein arrays:* After slides are dried, they are incubated in 1x Re-blot Plus Mild (Chemicon) for 7 minutes, and washed 3 times in 1x TBST (Dako Cytomation). Then, slides are blocked in SEA BLOCK (Pierce) for 2 hours at room temperature or at 4C overnight. The next incubation steps are carried in the following order: 5 minutes in 0.03% Hydrogen Peroxide (Dako Cytomation), 10 minutes in Avidin block (Dako Cytomation), 10 minutes in Biotin block (Dako Cytomation), 15 minutes in Protein block (Dako Cytomation), 45 minutes in primary antibody (pERK (Cell signaling), pSTAT3 (Cell signaling), Actin (Sigma)) diluted 1:1000 in Antibody diluent (Dako Cytomation), 20 minutes in respective secondary antibody at 1:5000 in Antibody diluent (Dako Cytomation), 15 minutes Streptavidin complex (Dako Cytomation, prepared exactly as described in the kit manual), 15 minutes Amplification Reagent (Dako Cytomation), 15 minutes Streptavidin-conjugated 655 nm Quantum-dots (Invitrogen) after which slides are washed in PBS. Between each step above, slides were washed 3 times in TBS supplemented with 0.1% Tween for 5 minutes each. Finally, slides are briefly rinsed in milli-Q water and dried by centrifugation.

*Imaging:* Slides are scanned with a 480nm laser in ProScanArray microarray scanner (PerkinElmer) and quantified using ProScanArray Express software (PerkinElmer).

*Data processing:* First, the triplicate spots are averaged for each well. Normalization of the data points are carried out by dividing each data point (i.e. averaged triplicate spot) by the average of the data points 3 above and 3 below on the slide (see Figure 4.3). Then, the biological triplicates are averaged for each gene. Those genes with discordant values across the 3 replicates were discarded from further analyses.

### 2.2.4 Ancillary materials and methods

*Cell culture:* A431 cells are grown in DMEM supplemented with 10% Fetal Bovine Serum and 1x Antibiotic/antimycotic, at 37C, 5% CO2.

*Tools for computational network analysis:* The protein-protein interaction network of human proteins was obtained from HPRD, MINT, BIND, Entrez-Gene and INTACT. Cancer mutation data was obtained from COSMIC and from Vogelstein et al. Enrichment analyses were performed using the hypergeometric distribution equation.

*Network connectivity analysis of hits:* Each protein in the network is assigned a p-value for their enrichment of the hits using hypergeometric distribution. A cut-off p-value is determined by Benjamini-Hochberg transformation of the p-values at q = 0.01. Initially, the target network only contains hits, then other proteins are added into this network one by one starting with the lowest p-value. At each step,

the new network is evaluated based on whether any previously disconnected components in the network from the previous step have been connected to each other by the addition of the protein. If addition of the protein connects any new components to each other the protein is retained in the network, otherwise not. This is determined by calculating the all-pair shortest distance matrix of the network at each run, and subtracting it from the distance matrix of the network from the previous run. If the sum of the difference is greater than zero, it is an indication that some previously disconnected nodes have been connected to each other by the addition of the new node. This continues till the specified p-value at q=0.01 is reached. The resultant network contains hits and proteins with q-values smaller than 0.01.

# CHAPTER 3

# Fine-scale dissection of organizational principles in the protein interaction networks

## 3.1 INTRODUCTION

### 3.1.1 Problems with the conventional methodology

Despite major advancements in the development of analytical techniques for the analysis of networks in graph theory and statistical physics, there is no established standard methodology for the analysis of protein networks. This is largely due to the incredible complexity of protein interaction networks that makes their study using conventional methodology less useful when it comes to extracting biologically meaningful information from these networks. Networks studied in physics are usually relatively simple both in terms of the types of nodes and the types of connections between nodes (typical networks are composed of nodes and edges between these nodes), whereas in biological networks, each protein and interaction between proteins possess specific properties that distinguish them from other proteins/interactions in the network. This highly heterogeneous nature of the composition of protein networks prevents their representations and analyses

within a conventional framework where proteins are denoted by nodes and their interactions are depicted by edges. For example, if a kinase and a scaffolding protein are represented by the same type of node in the network, an ambiguity will arise regarding their roles in their network localities as the biological significance of interactions of a kinase can be entirely different from that of a scaffolding protein. Connections between nodes are not of trivial nature either, and can be highly heterogeneous in their biological functions. A connection between two proteins can mean a catalytic reaction, formation of an active complex, formation of an inactive complex, scaffolding, tethering…; each having a distinct biological function. As if this is not complicated enough, both the protein properties and the connection properties are highly dynamic and can depend on the specific biological context. For example, a given protein can assume several functional states depending on, just to name a few, its subcellular localization, post-translational modifications or its expression level. Similarly, a connection between two proteins can highly depend on factors like local ion concentrations or presence of scaffolding proteins. Conventional network analysis methodology developed for the analysis of social networks is far from encompassing the above-mentioned aspects of protein interaction networks.

Nevertheless, a significant effort has been spent in the past on characterizing complex properties of biological networks through analyses of connectivity

patterns of gene and protein networks, some even resulting in major observations regarding the structure-function relationships in these networks (Albert, et al., 2000; Ihmels, 2002; Jeong, et al., 2001; Jeong, et al., 2000; Klemm and Bornholdt, 2005; Maslov and Sneppen, 2002; Milo, 2002; Yu, et al., 2007). Perhaps the most prominent finding is the observation that protein interaction networks display a type of node degree (i.e. number of interactions) distribution where the network is dominated by few proteins with a high number of connections whereas the majority of proteins have low node degrees (Jeong, et al., 2001; Jeong, et al., 2000). This so-called "scale-free" distribution of node degrees has been implemented in various cell biological phenomena like robustness and lethality in single-gene knock-outs (Jeong, et al., 2001; Jeong, et al., 2000). In an attempt to address more complex characteristics of protein networks, Han et al have incorporated gene expression data into their analyses of protein interaction networks and have found intriguing relationships regarding specific positioning of proteins in the network and their co-regulation with proteins in their network neighborhood (Han, et al., 2004). However, despite providing elegant hypotheses regarding some of the important cell biological phenomena, most of these studies have been met with criticisms in the field (Batada, et al., 2006; Batada, et al., 2007; Colizza, et al., 2006; Komurov and White, 2007; Nunes Amaral and Guimera, 2006; Przulj, et al., 2004). It has been suggested that the distribution of node degrees in a cellular protein interaction network may not be scale-free

(Przulj, et al., 2004), and moreover, centrality in biological networks may not at all correlate with essentiality in terms of cellular survival (Komurov and White, 2007). Similarly, the type of network organization suggested by Han et al in their seminal work has been suggested to be an artifact of their particular methodology and/or datasets (Batada, et al., 2006; Batada, et al., 2007). This controversy regarding global properties of protein interaction networks underlines the complexity of biological networks and the need for more rigorous and more comprehensive approaches towards analyzing biological networks. In our opinion, sensitivity of the observations to data handling and also on slight variations in the methodology is a consequence of an inappropriately focused approach, and that a more comprehensive rigorous approach would be robust to these variations in terms of the results. Our method defining central network characteristics stems from this need of a multidimensional analysis and, importantly, it can detect biologically significant patterns using multiple available datasets where other approaches fail.

### 3.1.2 Our Methodology

Our methodology is flexible enough to incorporate any desired number of data dimensions in the form of *biological metrics* (see Methods). Biological metrics are functions describing a specific property of each protein in the protein

interaction network, like expression variance, expression similarity with its neighbors, etc… It can also be a function describing the protein's molecular properties or its molecular relationship with its interactors. As a proof of principle analysis, we incorporated mRNA expression information together with the protein interaction data for an initial test of our method. In our method, proteins are assigned a *profile* based on the metrics, and the significance of each profile is evaluated using a large number of randomized instances of the underlying protein interaction network. Significance profiles are then identified and analyzed in more detail.

*Derivation of network metrics:* The structure of the protein interaction network can be defined by its adjacency matrix, while the dynamic properties of individual proteins can be captured by an mRNA expression correlation matrix created using a compendium of microarray data (see previous chapter). In order to account for the dynamic properties of proteins as well as their dynamic relationship with their neighbors in the network, we used the network architectural and expression information of proteins from these two matrices to derive 9 biological network metrics that describe the dynamic behavior of a protein and of its neighborhood in the network (see previous chapter). Briefly, we defined expression variance (EV) in order to capture the variability of a protein's expression across multiple conditions, neighborhood EV in order to describe the neighborhood of a protein in

terms of their EV, neighborhood EV variance ($v^{EV}$) to account for variability of EVs of neighbors of a protein, average interactor Pearson correlation coefficient (avPCC) to describe how a protein is co-regulated with its neighbors, neighborhood PCC (nPCC) to ask if the neighbors of a protein are co-expressed with each other, nPCC2 to describe co-expression of proteins in the second neighborhood of a protein, neighborhood PCC variance ($v^{PCC}$) to account for variability of expression profiles of proteins in the neighborhood, dynamic degree (*yK*) to ask if a protein is co-regulated with other proteins in the network, and neighborhood *yK* (*nyK*) to account for average *yK* in the neighborhood. These metrics are explained in detail in the Methods section of the previous chapter.

## 3.2 REVEALING STATIC AND DYNAMIC MODULES

### 3.2.1 Initial clustering of dynamic profiles

First, a dynamic profile (see Methods) was assigned to each protein in the network based on these metrics. Then, we performed a hierarchical clustering of proteins in order to identify distinct classes of dynamic profiles in the network and to test if they represent specific functions of proteins in the network. We only

evaluated highly connected proteins (i.e. those that have >6 interaction partners, which is the upper 30$^{th}$ percentile of the node degree distribution), as they

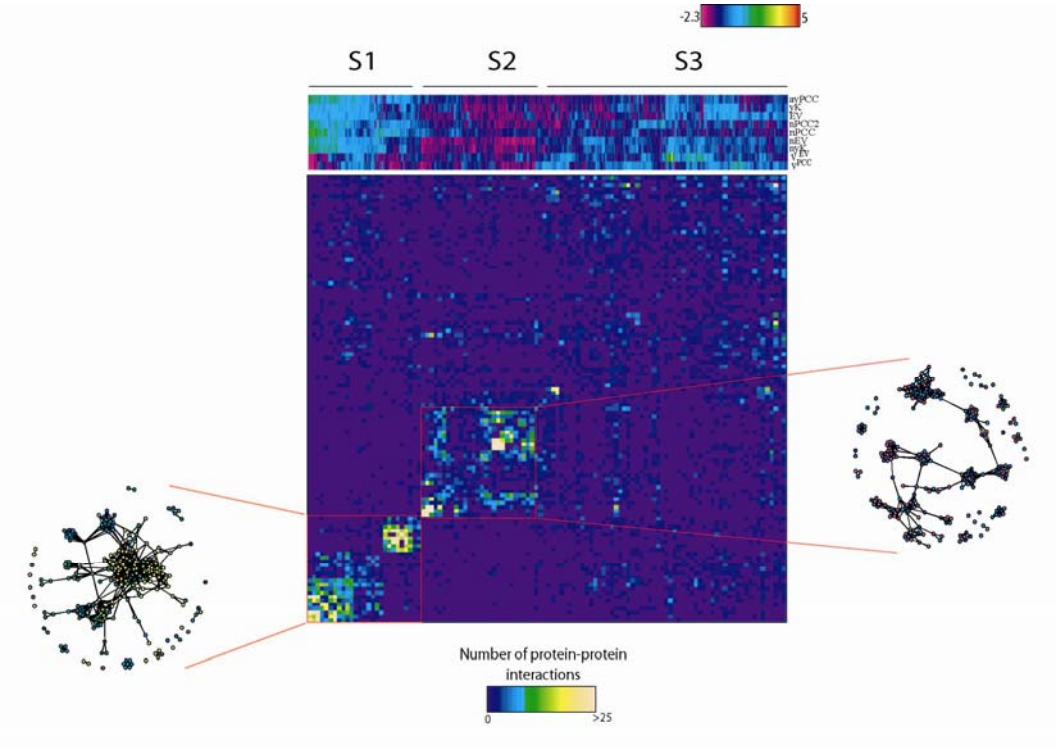


**Figure 3.1**. Dissection of proteins into dynamical classes. Hierarchical clustering of proteins by their dynamic profiles (upper panel) and the interaction matrix showing the protein-protein interaction patterns between different dynamic profiles (lower panel). In order to make clustering possible, we normalized each row to have a mean of 0 and a variance of 1. For the lower panel, proteins were binned into 114 bins with the exact ordering as in the clustering in the upper panel. Each square in the matrix represents the number of interactions between respective bins.

produced best clustering with these values when compared to the clustering performed by proteins having lower node degrees (not shown). From the graphical representation of the clustering, it is possible to dissect three main groups of proteins (Fig. 3.1). Group S1 is characterized by the highest nPCC, avPCC, nEV and EV values, while S2 has the lowest values in these categories.

An obvious distinguishing feature of S1 and S2 from the group S3 is their lower $v^{EV}$ values, indicating that S1 and S2 proteins are located in neighborhoods with homogeneous expression profiles. Despite having higher variation in terms of most values, S3 proteins consistently have higher $v^{EV}$ values, suggesting that these proteins are located within highly variable neighborhoods. Although the dynamic profiles of S3 as a whole are likely to be found in a randomized network, the profiles of S1 and S2 are significantly overrepresented in our original network as compared with randomly rewired networks, indicating that these groups represent biologically significant populations of proteins.

In order to get a first impression about the connectivity profiles of these groups in the network, we examined the protein-protein interactions of these groups within and between each other. For this purpose, we binned the proteins into 114 bins of 10 proteins, respecting the order of proteins in the clustering in Figure 1a, and calculated the number of interactions between every bin pair. Strikingly, we see a significant interaction density within the S1 and S2 groups, but not in S3 or between S1 and S2, immediately suggesting the existence of densely connected clusters in these groups (Figure 3.1, lower panel). Indeed, a network plot of these groups reveals densely connected clusters of high and low EVs respectively (Figure 3.1), indicating that the groups S1 and S2 are mainly composed of highly

co-regulated dynamic and non-variant static densely connected clusters of proteins.

Densely connected clusters are likely to represent specialized modules in the cell (Spirin and Mirny, 2003). Indeed, S1 proteins have significantly higher nPCC and nPCC2 values, which indicate that S1 proteins represent dynamically expressed modules. In addition, these proteins have higher nEV and EV values, pointing to their highly dynamic expression pattern. S2 proteins, on the other hand, have the lowest EV, nEV, $v^{EV}$, $v^{PCC}$, nPCC and nPCC2 values, which strongly suggests that these proteins are located in neighborhoods with non-variant expression patterns. We hypothesized that these two groups of densely connected proteins may represent two distinct types of biologically significant functional modules and therefore performed further analysis on them.

### 3.2.2 Functional specialization of static and dynamic neighborhoods

The current notion of functional modules predicts that a set of interacting proteins that are highly co-expressed is likely to be specialized to a specific process, and some studies suggest that the protein interaction network is enriched for interactions between co-regulated proteins (Ge, et al., 2001; Ihmels, et al., 2004). We show that sets of interacting static proteins, which are supposedly

constitutively present in the cell but do not have high statistical correlation in their expressions, also may represent specialized functional modules, and that they are at least as abundant in the cell as the sets of interacting dynamic proteins that are highly co-expressed. In order to test this hypothesis, a simple function that compares a protein's Gene Ontology (Ashburner, et al., 2000) (GO) annotations with that of its neighbors, was derived to quantitate the functional specialization of a protein's neighborhood. This function, "neighborhood function homology" (see Methods), generates values in the range from 0 (no shared GO terms between a protein and its neighbors) to 1 (all of GO terms assigned to a protein are shared with its neighbors).

Neighborhood function homology of static proteins (EV < 0.25, i.e. lower quartile of genomic distribution) in the network negatively correlates with their neighborhood EV with a high significance (Spearman's $\rho = -0.41$, $p = 1.8 \times 10^{-18}$), suggesting that static proteins interacting with other static proteins are found in functionally specialized neighborhoods. On the other hand, neighborhood function homology of dynamic proteins (EV > 0.75, higher quartile of EV distribution) positively correlates with their neighborhood EV (Spearman's $\rho = 0.40$, $p = 1.5 \times 10^{-12}$). This indicates that dynamic proteins, in contrast to static proteins, are more functionally homologous to their neighbors when they are in dynamic neighborhoods. Neighborhood function homology of dynamic proteins

correlates even more significantly with their average interactor Pearson Correlation Coefficients (avPCC) (Spearman's $\rho = 0.57$, $p = 8x10^{-27}$), a measure of how well a protein is co-expressed with its neighbors (Han, et al., 2004), (see Methods). Together, these observations suggest that network neighborhoods composed of constitutively expressed proteins (static neighborhoods) are highly specialized modules, much like the neighborhoods of highly co-expressed proteins (dynamic neighborhoods).

### 3.2.3 Identification of static and dynamic modules and their functions

Past studies have measured statistical correlation of gene expression in order to assign proteins to specific modules and also to assign new functions to previously uncharacterized proteins (Ihmels, et al., 2002; Segal, et al., 2003; Segal, et al., 2003). Since static neighborhoods also seem to be functionally coherent, it should be possible to assign proteins to specific modules by the virtue of their associations with static neighborhoods. To this end, all static neighborhoods in our network were identified by compiling all the interactions between static proteins in the network (static network, 491 proteins connected by 897 interactions). The static network consists of 82 distinct disconnected sub-networks

ranging in size from 2 to 86 proteins (see Table 3.1[1]). The functional annotations associated with these static sub-networks appear to be functionally coherent, representing various functions including mRNA transcription and splicing, vesicle transport and cell cycle regulation (see Table 3.1). This apparent functional coherence suggests that the static network is enriched for functional modules. In order to test the significance of modular composition of the static network and to see if it is possible to achieve a similar level of functional coherence in a network generated by random draws of interactions, a network modularity metric was defined to measure functional specialization of the interactions in a network (see Methods). The static network shows significantly higher network modularity than what would be expected by random draws of interactions from the large network (Figure 3.2), suggesting that the association of functionally coherent sets of proteins with each other within static neighborhoods reflects a biological phenomenon. We compared the static network modularity with that of the network formed by highly co-expressed proteins, which is expected to be enriched for functional modules, in agreement with the previous studies showing modularity of co-expressed proteins (Han, et al., 2004; Segal, et al., 2003). We identified the dynamic network by taking all interacting pairs of proteins that also have pair-wise Pearson correlation coefficients of at least 0.65 (383 proteins connected by 777 interactions). This dynamic network, therefore, contains

[1] For Tables 3.1, 3.2 and 3.3 see http://www.nature.com/msb/journal/v3/n1/suppinfo/msb4100149_S1.html

interactions only between proteins that are also highly transcriptionally co-regulated. The dynamic network consists of 77 sub-networks mainly composed of dynamic proteins (not shown) and, as expected from previous publications, the dynamic sub-networks are highly functionally coherent (Table 3.2). The dynamic network also shows a significantly high network modularity that is comparable to

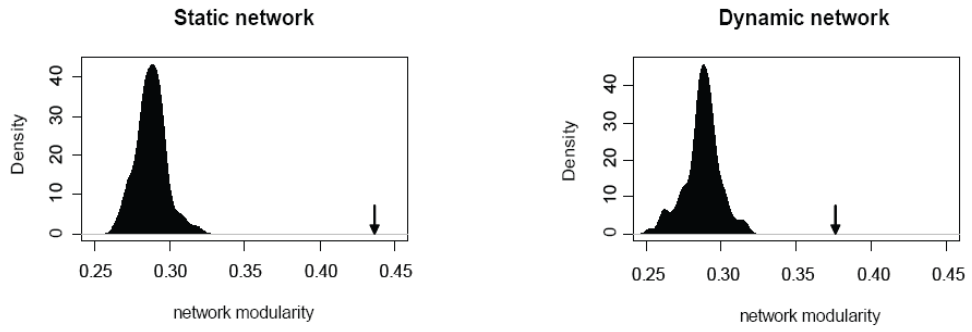


**Figure 3.2.** Functional specialization in the static and dynamic networks. Comparison of network modularity in the static and dynamic networks with that of 100 networks formed by random draws of interactions from the original network. The plot shows the distributions of network modularity values for random draws of 897 (left, for comparison with the static network) and 777 (right, for comparison with the dynamic network) protein-protein interactions out of the original network. Arrows show the actual network modularity values of the static and dynamic networks ($P < 0.01$ in both cases).

that of the static network (Figure 3.2). This indicates that both networks are enriched for functional modules. The fact that only 15 proteins and 6 interactions are common to both networks indicates that the modules in the two networks are distinct, and that the high modularity of the static network is not a consequence of a significant overlap with the dynamic network. These observations argue that the static protein neighborhoods represent functional modules, and it should be

possible to assign proteins to functional modules by virtue of their association with static proteins.

In order to see if either network is specifically enriched for certain cellular functions, we performed enrichment analyses of the two networks for over-representations of MIPS functional categories (Mewes, et al., 2004). Interestingly, the most significant relative enrichment is seen in the functional categories related to mRNA transcription and processing (static network) and rRNA transcription and processing as well as translation (dynamic network) (Table 3.3). The static network is enriched for general mRNA transcription (RNA Polymerase II holoenzyme complexes), splicing (the pre-mRNA splicing complex) and processing (CPF and CCR4-NOT) as well as co-regulator complexes like the chromatin remodeling complexes (SWI/SNF and INO80), histone acetyl-transferase complexes (SAGA and NuA4), histone methylase (COMPASS) as well as mRNA nuclear export (TREX) (see Table 3.1). The dynamic network, in addition to RNA Polymerase I and III components, contains modules like the SSU processome, involved in rRNA processing, and translation initiation factor complexes (see Table 3.2), which is consistent with studies reporting extensive regulation of these modules under various stress conditions (Gasch, et al., 2000; Warner, 1999). In addition, the dynamic network contains most of the

proteasomal proteins, while the static network also contains many of the mitochondrial ribosomal proteins.

There are many modules in the two networks that also seem to perform similar functions. For example, components of the mitotic cohesin complex, which holds sister chromatids together, and the septin ring complex, which is required for cytokinesis, are in the dynamic network (sub-networks 63 and 55, Table 3.2) while the DASH complex, which plays a role in chromosome segregation, and the COMA complex, which is involved in the kinetochore assembly, are in the static network (sub-networks 51 and 58, Table 3.1). Components of the Anaphase Promoting Complex (APC) are also static (sub-network 4), as also reported previously (de Lichtenberg, et al., 2005). These complexes are all involved in the final stages of cell division, yet their regulation is markedly different. Another potentially interesting correlation relates to vesicle trafficking, where proteins associated with clathrin-coated vesicles (AP-1 and AP-3 complex proteins) seem to be static (sub-networks 9 and 69), while those associated with coatomer protein-coated vesicles that are involved in vesicle transport between Golgi and ER (COPI and COPII complex proteins) are dynamic (sub-networks 10, 45 and 76). The dynamic expression pattern of the latter may stem from the involvement of the early secretory pathway in various stress responses like unfolded protein response or osmotic stress (Higashio and Kohno, 2002; Lee and Linstedt, 1999;

Sato, et al., 2002), whereas clathrin-coated vesicles may play role in constitutive transport. These examples suggest that although some functions in the cell can be classified as static or dynamic (like mRNA and rRNA synthesis, respectively), many others are carried out through dynamic interplay between distinct static and dynamic modules. A closer analysis of expression dynamics of functional modules under various conditions may provide an in-depth insight into the regulation of cellular behavior by transcriptional programs.

In their study, Han *et al* (2004) defined hubs that are highly co-expressed with their neighbors as "party" hubs, which are modular, and those that are not co-expressed with their neighbors as "date" hubs, which they reported as central. However, the set of date hubs also contains hubs that are found within static modules (where there is also no statistical correlation of expression among neighbors). Therefore based on our findings, we propose that static hubs interacting with static proteins within static modules be excluded from date hubs and, in analogy to the party-date hub terminology, be named "family" hubs, as they are always present in the network and interact with their neighbors constitutively. Therefore, family hubs and party hubs form static and dynamic modules, respectively, while date hubs (family hubs excluded) organize the network.

### 3.2.4 Protein expression noise and evolutionary rate in the static and dynamic networks

Based on the classification of hubs by Han *et al* (2004), it was suggested that centrally positioned hubs in the network evolve faster than hubs in modules, and that modularity imposes a constraint on the evolvability of proteins, suggesting an evolutionary scenario where protein networks evolve mainly by modifying their central coordinators (Fraser, 2005). We examined this hypothesis in the context of our modified hub classification, and also find that party hubs evolve at a significantly slower rate than other hubs (Figure 3.3a). However, surprisingly, family hubs do not evolve slower than date hubs (Figure 3.3a). By extrapolation, this suggests that hubs present in dynamic modules are evolutionarily constrained, while those present in static modules are not. Accordingly, proteins in the dynamic network have significantly lower evolutionary rates than proteins in the static network ($p < 1x10^{-16}$, Wilcoxon test), and there is a significant negative correlation between EV values of proteins and their rates of evolution (Spearman's $\rho = -0.21$, $p = 4.5x10^{-15}$). These results suggest that proteins in static modules have more freedom of variation than proteins within dynamic modules.

Less evolutionary constraint of static modules may indicate that proteins in these modules are largely dispensable for the module function due to compensation in

the network. A prediction of this hypothesis is that proteins in static modules are less likely to be essential for cell survival, perhaps due to their functional redundancy. Indeed, proteins in dynamic modules are almost twice as likely to be essential as proteins in static modules (Figure 3.3b), which is also true for party hubs when compared to family hubs (not shown), indicating that the cell is highly tolerant of the loss of proteins in static modules, a property which may allow them to evolve at a faster rate than proteins in dynamic modules.

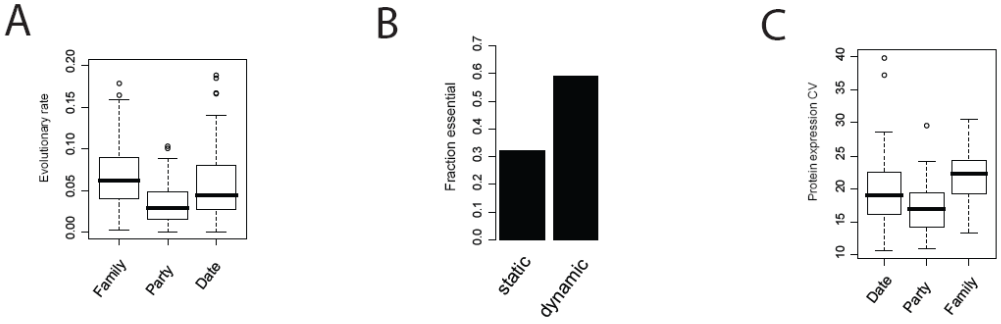


**Figure 3.3.** Evolutionary rate and expression noise of the static and dynamic modules. Evolutionary rates of yeast proteins derived by Hirsh *et al* (2005) (Hirsh, et al., 2005) were used. A) Boxplot of evolutionary rates of family, party and date hubs. Family hubs are static hubs with neighborhood EVs of <0.3, party hubs are hubs with avPCC > 0.45, and date hubs are those with neighborhood EV>0.3 and avPCC < 0.45. B) Fractions of proteins in the static and dynamic networks whose gene deletion is lethal to yeast. C) Boxplot of protein expression noise in the different hub classes.

The significant correlation of protein EVs with their evolutionary rates suggest that the static network may be a buffer of evolutionary variations in the protein interaction network, granting static proteins a role as evolutionary modifiers of

cell behavior. We reasoned that if the cell is more tolerant of genetic variations in the components of static modules, then the cell may also be more tolerant of variations in the expression of these proteins within a cell population. Expression variation of proteins between cells within a population, or protein expression noise, is a major factor contributing to the variations of cell behavior among cells within a cell population (Blake, et al., 2003; Raser and O'Shea, 2005). Therefore, we compared the expression noise of proteins in static modules with that of proteins in dynamic modules. Using the coefficients of variation of protein expression levels (CV values) within a clonal cell population derived by a recent study (Newman, et al., 2006), we find that proteins in dynamic modules are significantly less noisy in their expression when compared to proteins in static modules ($p = 3x10^{-15}$, Wilcoxon test), indicating that the expression levels of static proteins are the ones that show most cell-to-cell variations within a population. Accordingly, family hubs have significantly higher CV values than other hubs (Figure 3.3c). It is surprising to find that proteins with least variable mRNA expression patterns are most variable between cells and during evolution. These observations argue that static components of the eukaryotic protein interaction network are a source of robustness in cell regulatory networks that allows for evolutionary as well as populational variations in cell behavior (see Discussion).

### 3.2.5 Expression levels of static and dynamic modules

Expression variance of genes positively correlates with their mRNA abundance (Spearman's $\rho = 0.21$, $p < 1x10^{-16}$), and accordingly, proteins in static modules are expressed at a significantly lower level than those in dynamic modules (Wilcoxon test, $p < 1x10^{-16}$). This may suggest that the correlation of EV with the organizational layout in the protein interaction network may be a reflection of the effect of expression levels of proteins rather than their EV. The expression levels of proteins does seem to contribute to the protein network layout, as there is a high positive correlation between mRNA abundance values of hub proteins and that of their neighbors in the network (i.e. average neighborhood mRNA abundance, Spearman's $\rho = 0.43$), although the correlation is significantly less than that between EV and neighborhood EV (Spearman's $\rho= 0.61$). This correlation is not surprising given that the expression levels of proteins participating in the same protein complex are generally similar (Papp, et al., 2003). The relatively low correlation between EV and mRNA abundance and the fact that the correlation of mRNA abundance between neighboring proteins is less than that of EV suggests that our observations with EV above are not an artifact of the underlying mRNA abundance values. In order to rule out the possibility that our observations with EV values of proteins presented above are an artifact of their expression levels, we performed partial correlation analyses (see Methods)

between EV, neighborhood EV, neighborhood function homology and mRNA abundance values of proteins. Partial correlation between EV and neighborhood EV while controlling for mRNA abundance is almost as high ($r_{EV\sim neigh.EV,\ mRNA}$ = 0.66) as their normal correlation ($r_{EV\sim neigh.EV}$ = 0.67). Similarly, partial correlation between neighborhood function homologies of static proteins with their neighborhood EV while controlling for mRNA abundance or average neighborhood mRNA abundance is almost as high as their normal correlations (not shown). These observations argue that the observed effects of EV on the organizational layout of the protein interaction network are not an artifact of expression levels of proteins, and that proteins segregate into different modules according to their expression variances.

## 3.3 FUNCTIONAL ANALYSIS OF PROTEIN NETWORK ORGANIZATION

Of S1, S2 and S3, dynamic profiles of S3 proteins are the most disparate. The only common characteristic of proteins in this group seems to be the almost invariant high $v^{EV}$ or $v^{PCC}$ values, which excludes these proteins from modules, where expression properties of proteins are similar. The disparity of the dynamic

profiles of these proteins may stem from the versatility of their functions, as they are located more centrally in the network (as judged from their betweenness centrality scores, not shown) and therefore may have functions in multiple processes. However, a close analysis of the clusters generated by hierarchical clustering of S3 reveals subgroups of proteins with distinct dynamic profiles (see below). We hypothesized that these different dynamic profiles may correspond to different functional classes of S3 proteins and therefore analyzed them in more depth.

### 3.3.1 Dynamic classes have distinct roles in network connectivity

In order to analyze if the dynamic profiles have distinct roles in network connectivity, we separated S1 and S3 groups into more subgroups based on their dynamic profiles. These classes are distinguished from each other by one or more characteristics that give insights about the dynamic nature of their neighborhood and suggest specific functions that these proteins may be performing in their respective localities in the network.

*Clustering of S1 into 3 subgroups*

In order to get a higher resolution clustering of proteins according to their dynamic profiles, we performed further clustering of each of the groups. Group

S2 did not show any further significant clustering (not shown), but group S1 has three subgroups distinguished by their EV, $v^{EV}$ and nEV values (Figure 3.4a). Based on their network plots and interaction profiles in Figure 3.1, subgroups S1.1 and S1.3 seem to be distinct densely connected highly co-expressed modules; one subgroup with lower EV and the other with higher EV values (see Figure 3.4a). Interestingly, the higher EV subgroup is almost exclusively composed of modules involved in rRNA synthesis/processing and protein translation, while those with moderate EV values are mainly proteasomal proteins as well as some other small dynamic modules (not shown). This indicates that genes involved in ribosome biogenesis and protein synthesis constitute a distinct subclass of modules that are more robustly regulated when compared to other dynamically expressed modules. These genes have been shown to be extensively regulated as a cellular energy-saving mechanism during stress, which probably contributes to their high EV when compared to genes in other dynamically expressed modules. The S1.2 proteins have more variable neighborhoods as evidenced by their high $v^{EV}$ and $v^{PCC}$ and low nPCC, indicating they are not found within modules like proteins in S1.1 and S1.3. An interaction pattern of the S1 proteins with each other shows a significant interaction density among proteins of S1.1 and S1.2 (Figure 3.4b), indicating that S1.2 proteins are preferentially interacting with S1.1 proteins, and therefore may be located outside the modules

formed by S1.1 proteins. This may also suggest a function for S1.2 proteins as coordinators of S1.1 proteins in the network (see below).

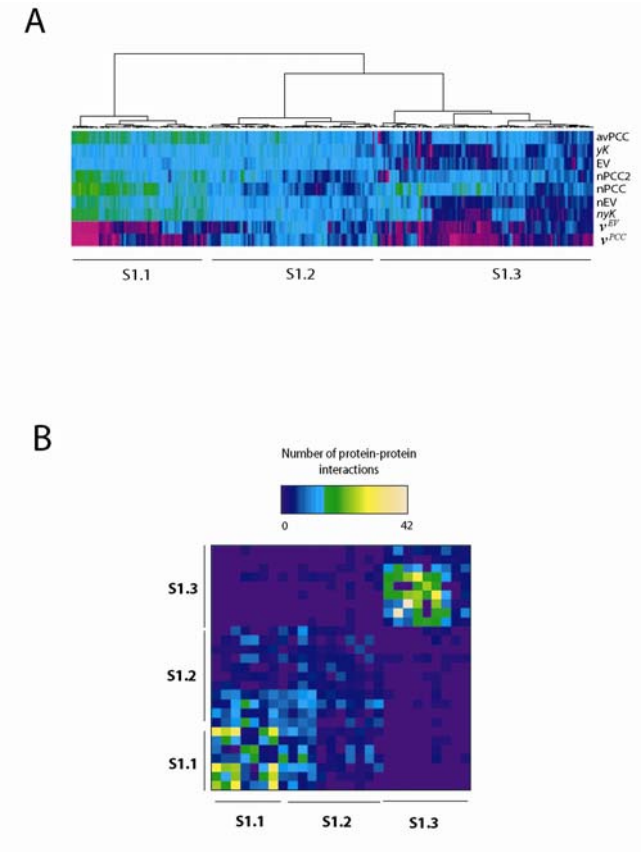


**Figure 3.4.** Characterization of S1. **A)** Hierarchical clustering of S1 into 3 subgroups. **B)** Interaction matrix of subgroups in S1.

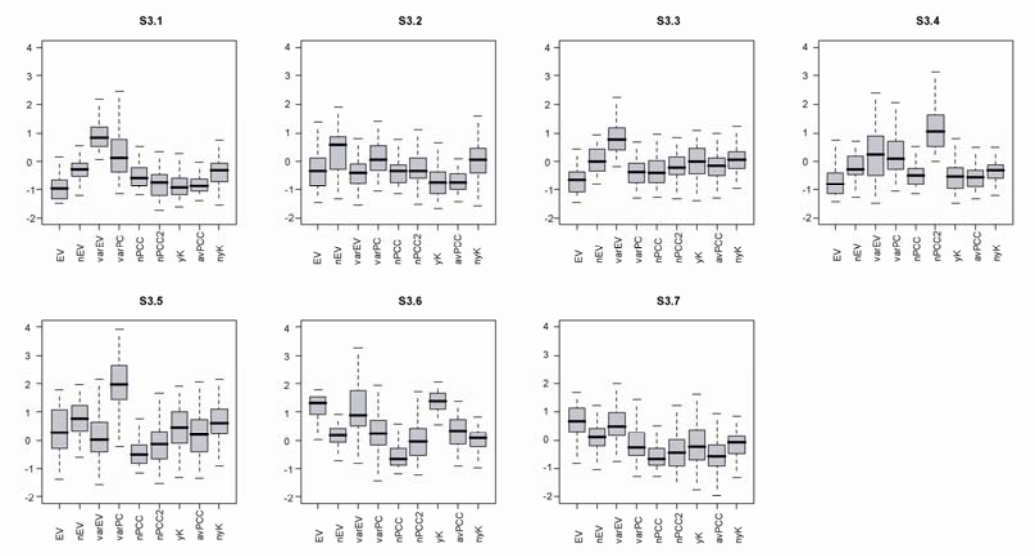


**Figure 3.5.** Dynamic profiles of each S3 subgroup. Y axes represent normalized values of each metric (see Methods).

*Clustering of S3 into 7 subgroups*

Detailed dynamic profiles of each subgroup are seen in Figure 3.5.

S3.1: This subgroup has most of its values low just like S2, with the exception that it has relatively high $v^{EV}$ and $v^{PCC}$ (shown as varEV and varPC in the figure). Low nEV of this group indicates that proteins in this group are surrounded mostly by static (low EV) proteins, however their high $v^{EV}$ and $v^{PCC}$ suggest that their neighbors in the network have variable EV values and dissimilar expression profiles. This dynamic profile suggests that these proteins may be located at the boundaries of static modules. A network plot of a representative S3.1 protein

(IES1, a subunit of the INO80 chromatin remodeling complex) shows that this protein links the module constituents to other proteins in the network (Figure 3.6a).

S3.2: This subgroup has relatively higher nEV values, but low $v^{EV}$, suggesting that these proteins are mostly interacting with high-EV (dynamic) proteins. However their relatively higher $v^{PCC}$ also shows that neighbors of these proteins have dissimilar expression profiles, therefore indicating that S3.2 proteins are not found within modules. It follows that S3.2 proteins are coordinating functions of various dynamic proteins. An example to this subgroup is RHO1, a small GTPase involved in cytoskeleton signaling. This protein is mainly surrounded by dynamic proteins of various functions (Figure 3.6B).

S3.3: Dynamic profile of this subgroup is very similar to that of S3.1, with a subtle difference that this subgroup has higher $yK$ and lower $v^{PCC}$ values (Fig. S1). Higher $yK$ values indicate that, unlike S3.1, proteins in this group may be co-expressed with other proteins in the network and thus may be a part of cellular gene expression programs. An example to this subgroup of proteins is KIN2, a protein kinase involved in the regulation of exocytosis (Figure 3.6C).

S3.4: The most striking feature of proteins in this subgroup is their high nPCC2 but low nPCC (Fig. S1), indicating that neighbors of these proteins have high nPCC values but low PCC with each other. This profile suggests that S3.4 proteins are found outside dynamic modules, perhaps bridging them. An example to this subgroup is SBA1, a co-chaperone that binds and regulates Hsp90 family chaperones. Its network plot shows that this protein is interacting with UFD1/NPL4/CDC48 complex involved in protein transport from ER to the cytosol, and also with a nuclear pre-ribosomal complex containing NUG1, NOG1, RIX1, etc… (Figure 3.6D).

S3.5: The most obvious feature of this subgroup is its highest $v^{PCC}$ values. This indicates that neighbors of these proteins have highly dissimilar expression profiles, with probably negative co-expressions. Therefore, these proteins interact with some of their neighbors under one condition, and with the others under another condition, but never with both sets at the same time. A representative of this subgroup is BCY1, a regulatory subunit of cAMP-dependent protein kinase (PKA) (Figure 3.6E). It associates with either of the three catalytic subunits of PKA: TPK1, TPK2 or TPK3. TPK3 is negatively co-expressed with the other two subunits, indicating that TPK3 may have an opposing function to the other kinases (Figure 3.6E). Indeed, TPK3 has been shown to inhibit pseudohyphal growth in yeast whereas TPK2 promotes it(Robertson and Fink, 1998).

S3.6: Proteins in this subgroup have high EVs, high $v^{EV}$ and high $yK$ and avPCC (Fig. S1). Therefore, these proteins are dynamic, are co-regulated with other proteins in the network, they may also be co-regulated with their neighbors. However unlike dynamic modules, these proteins have high $v^{EV}$ values, indicating that these proteins interact with proteins of various expression patterns. Proteins

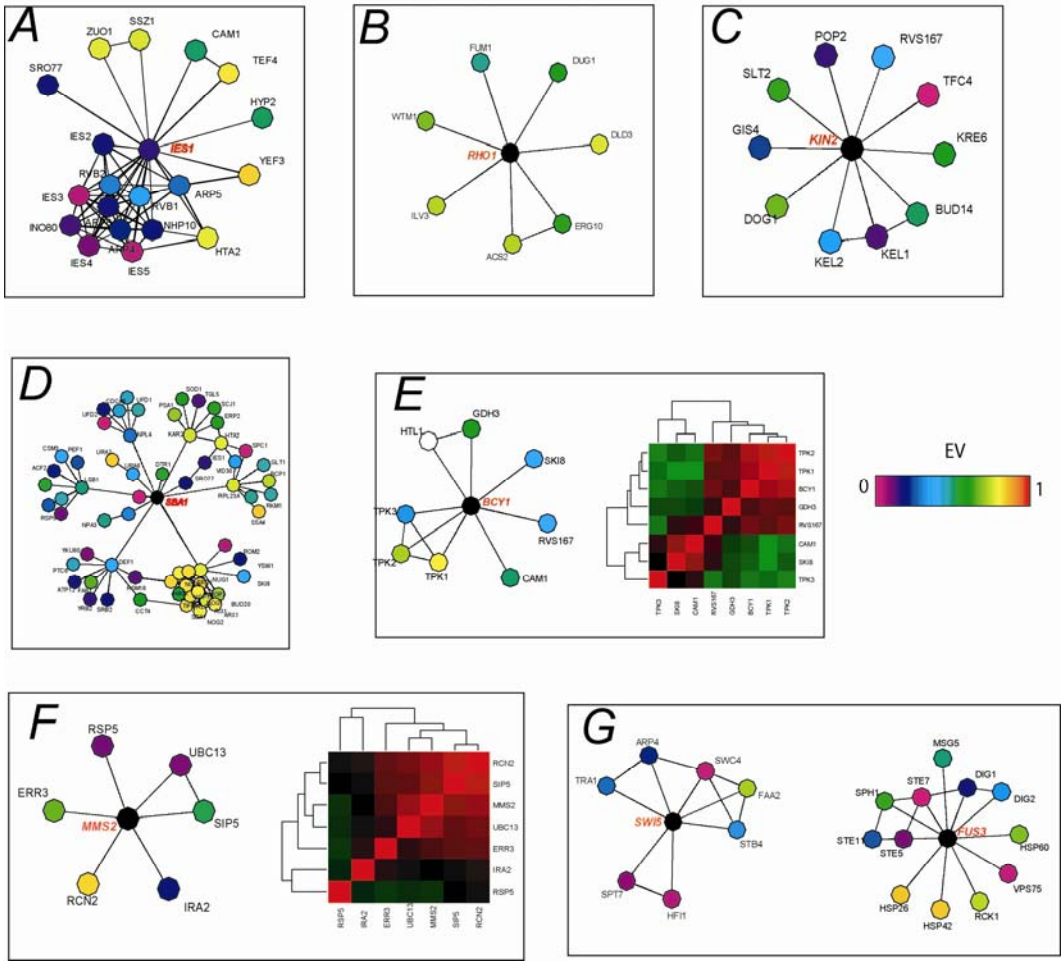


**Figure 3.6.** Representatives of each dynamic class. **A-G** show representative protein neighborhoods from subgroups S3.1 through S3.7, respectively. Heatmaps in **E** and **F** show expression correlation matrices of the neighbors of the respective proteins, where shades of green indicate strength of negative correlation while shades of red indicate strength of positive

correlation. Nodes in the network plots are colored according to their EV (color key shown); nodes of interest (center nodes) are colored in black for convenience.

in this subgroup may be located at the boundaries of small dynamic modules. An example to this subgroup is MMS2, a protein involved in post-replication DNA repair (Figure 3.6F). The co-expression profile of its neighbors suggests that MMS2 is co-expressed with UBC13, RCN2 and SIP5. MMS2 is known to form a complex with UBC13, however the functional significance of its interactions with the RCN2 and SIP5 are not yet known.

S3.7: Proteins in this subgroup have high EV and $v^{EV}$ values and moderate nEV values; the rest of its values are relatively low. This is the only subgroup where EV and *yK* values do not correlate (see Figure 3.5). EV highly correlates with *yK* (Spearman's $\rho = 0.68$, $P < 1x10^{-16}$), which is not surprising as highly regulated proteins are more likely to be co-regulated with other proteins in the network. This high correlation is not a consequence of high variance *per se*, as Pearson correlation coefficient is defined as the ratio of covariance of two variables to their individual variances. The fact that S3.7 proteins are highly regulated but still are not co-regulated with many proteins in the network suggests that the expression profiles of these proteins are highly specific, which may be the case for master regulators of cellular processes in the cell. Accordingly, S3.7 contains

proteins like FUS3 and SWI5, master regulators of pheromone response and cell cycle, respectively (Figure 3.6G).

*Deletion analyses*

In order to analyze the specific roles of these dynamic classes in the organization of the protein network, we undertook an *in silico* loss-of-function approach where we remove the desired set of proteins from the network and observe where the connectivity has been perturbed in the network (see Methods). We removed each dynamic class of proteins from our network, and measured where the network path lengths of proteins has increased. An increase in the path length between two nodes *a* and *b* upon removal of a node *c* indicates that the node *c* lies on the path between nodes *a* and *b*. Here, we only measured changes in path lengths between proteins that are separated by one node in the original network (path length = 2). Thus, when we remove a group *c* of proteins from the network and see that the network paths from a group *a* of proteins to another group *b* of proteins has been increased, we conclude that the group *c* proteins are directly linking proteins of groups *a* and *b*.

Figure 3.7A shows the results for the removal of each dynamical class from the network as compared to the removal of the same number of randomly selected proteins of similar node degrees. The removal of subgroup S1.1, which mainly

contains ribosome biosynthesis dynamic modules, impairs the connection between S1.2 proteins as well as the connection of other proteins to S1.2. Removal of S1.2 has an even stronger impact on the connectivity of S1.1 proteins to each other as well as to most of the rest of the network. These results indicate that S1.2 and S1.1 proteins are inter-linked to each other. However, since S1.2 proteins are not densely connected to each other and are not modular in terms of their neighborhoods (not shown), it can be concluded that S1.2 proteins are mainly found outside S1.1 modules and they have a major role in connecting these modules to the rest of the network and to each other.

Removal of subgroups S1.3 and S2 does not seem to impact the connectivity of the network, corroborating with the idea that these proteins are isolated modules with highly specialized functions. Connections of S2 to S3.6 and S3.7 are however impaired by the removal of S3.1, which is in accordance with the dynamic profile of this subgroup, which shows that although these proteins are mainly surrounded by static proteins, they are also interacting with dynamic proteins (see Appendix). An important role of S3.1 proteins in the network may be in connecting the static modules to the rest of the network.

The dynamic profile of S3.2 suggests that these proteins interact with dynamic proteins that are not modular (see above). Figure 3.7A shows that their removal

has the most significant impact on the connection of S1.2 to the proteins of S3.6 and S3.7, indicating that S3.2 proteins are coordinating the connections between proteins in S1.2 and proteins in S3.6 and S3.7.

The most significant feature of proteins in S3.4 is their high nPCC2 but low nPCC values, which suggest that these proteins are found "just outside" of dynamic modules (see above). Accordingly, their removal from the network results in an impaired connectivity between dynamic modules of S1.3 and the proteins in S3.6 and S3.7. Therefore, it can be concluded that S3.4 proteins are playing a role as coordinators of dynamic modules in S1.3.

Although the overall betweenness centrality values of groups S3.5 through S3.7 are not significantly different from each other (not shown), removal of each has markedly different effects on the network connectivity. While removal of S3.5 does not significantly affect the network connectivity when compared to randomly selected proteins, removal of S3.6 proteins seems to affect the connectivity of most of S3 proteins to each other as well as to S1.1 and S1.2 (Figure 3.7A). However the most potent effect on the network connectivity is seen

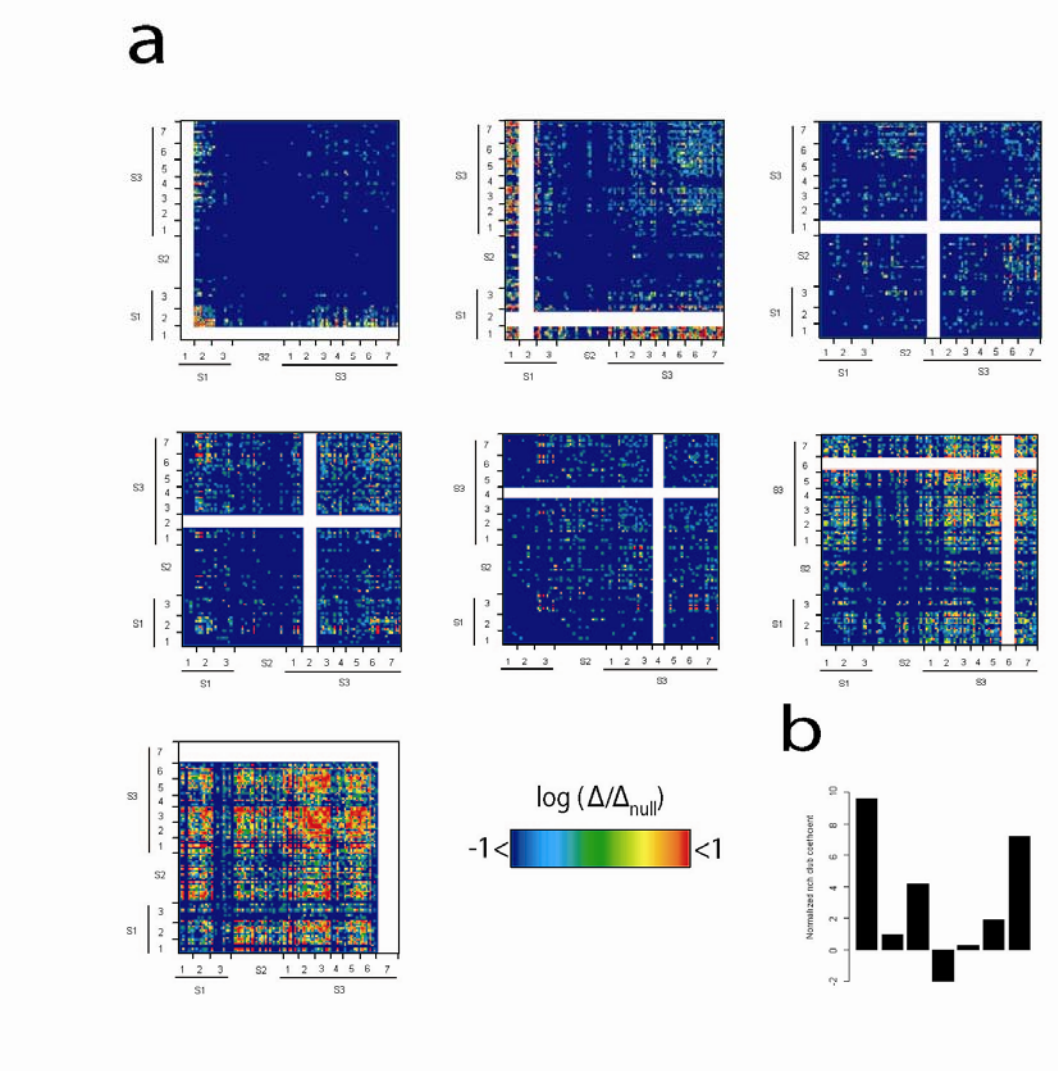


**Figure 3.7.** Characterization of roles of subgroups in the network connectivity. **a)** Deletion profiles of select subgroups. White stripes in the heatmaps indicate the deleted group. Please see Methods for a detailed description of the deletion profiles. **b)** Normalized rich club coefficients (see Methods) of each group.

with the removal of S3.7, where connections within and between almost every group of proteins becomes impaired (Figure 3.7A). This observation argues that S3.7 proteins may be the most centrally located proteins in the network as their

deletion results in a severe network disorganization. Since our *in silico* loss-of-function approach only takes into account node pairs that are only 1 node apart in the original network (see above and Methods), the profile of S3.7 deletion may indicate that these proteins are highly dispersed throughout the network as opposed to more localized positioning of other groups. Given its significantly higher impact on the network connectivity as compared to other groups, we hypothesized that S3.7 may contain proteins that play roles as the central coordinators of cellular events, and therefore analyzed this group in more depth.

### 3.3.2 S3.7: a "rich club" of central organizers in the network

Although deletion of S3.7 from the network results in a significantly greater disintegration of connectivity among other groups, S3.7 proteins are not significantly more centrally located in the network as judged from their betweenness, degree or closeness centralities (three metrics commonly used to measure a node's centrality in the network(Wasserman and Faust, 1994)) (not shown). This is surprising at first sight, because betweenness centrality of a node measures the frequency of paths between all node pairs that pass through that node, and Figure 3.7A shows that the paths between most node pairs get impaired upon removal of S3.7 proteins. It is conceivable, therefore, that S3.7 proteins may not be as central individually as they are as a group. In order to test this, we

calculated group betweenness values (measures the centrality of a group of proteins) of all groups, and find that S3.7 proteins have a more significant group betweenness than other S3 groups (not shown). However surprisingly, S2 has as high group betweenness as S3.7, although their individual betweenness values are the lowest among all (not shown), suggesting that static modules may also play central roles in the regulation of cellular processes (see below).

An observation that a group of nodes are significantly central as a group but not as individuals suggests that there is some redundancy among group members regarding connectivity of the network. This notion requires that the group members are tightly connected to each other so that the absence of one node would be compensated by another in the network. Indeed, a network plot of the dynamic classes shows that S3.7 has a considerable within-group interaction density as compared to others (Figure 3.8), which corroborates with a possibility of a within-group redundancy in terms of connectivity. Interestingly, among the S3 groups, only S3.7 seemed to be displaying significant within-group connectivity (Figure 3.8). In order to test if the observed density of interactions among S3.7 proteins is expected by chance, we compared within-group interaction densities of the dynamical groups with those in 100 randomized instances of the network, and find that S3.7 proteins are significantly more inter-connected than what would be expected by chance (Figure 3.7B). Only S3.1 and

S3.3 groups have within-group interaction densities close to that of S3.7. However unlike S3.7, where proteins are inter-linked to each other predominantly in a single connected web, S3.1 and S3.3 groups contain some proteins that form small dense clusters with each other, thus contributing to their high densities of within-group interactions (Figure 3.8). Therefore, it follows that S3.7 proteins

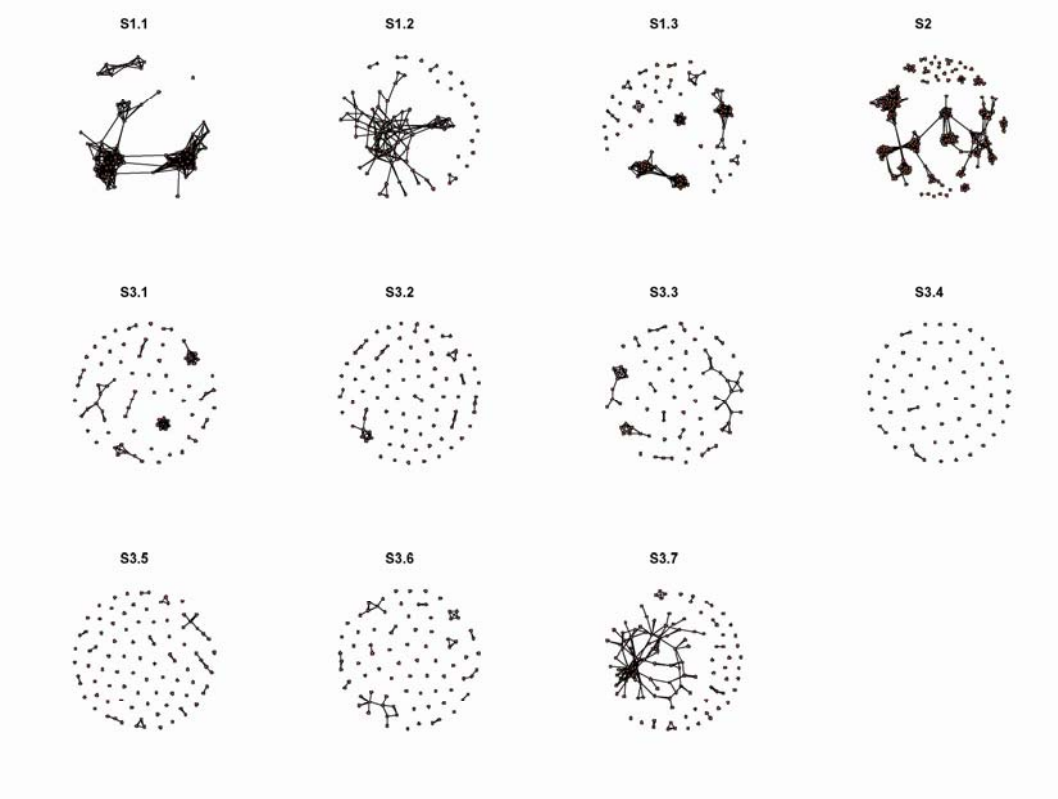


**Figure 3.8. Network plots of S3 groups.**

form an inter-connected web at the core of the cellular network thereby regulating the connectivity among different classes of proteins. This specific connectivity pattern, where instead of being dispersed in the network, central proteins are tightly inter-connected in a web, resembles so-called "rich-club" connectivity pattern in social networks and may have important implications about the cellular mechanisms of regulating information flow within the protein network (see Discussion).

Another striking feature of S3.7 is that proteins in this group are highly regulated as evidenced from their high EV, but nevertheless are not subject to a significant co-regulation with other genes in the network as evidenced from their low *yK* (see above). This indicates that S3.7 proteins are not likely to be a part of cellular gene expression programs and therefore have less constraint in their expression when high EV proteins (compare to S1, S3.5 and S3.6). This property may corroborate with the notion that these proteins are the central regulators of cellular processes (see above).

### 3.3.3 Static modules function as the central phenotypic enhancers

Our observations so far have shown that proteins with different dynamical properties have distinct roles in the network connectivity. Next, we wanted to see

if different dynamic classes have different roles in the regulation of cell behavior. First, we wanted to check which of the dynamic classes are involved in the cell-to-cell variability of cell behavior, which has been mainly attributed to gene expression noise (Blake, et al., 2003; Raser and O'Shea, 2005). We reasoned that proteins with most "noise" or variability in their expressions from cell-to-cell would also be most responsible for the cell-to-cell variations in cell behavior. For this purpose, we used the protein expression variation values from the extensive study of Newman et al (2006)(Newman, et al., 2006) and compared these values in the dynamic classes. We see that the most variable group in terms of their protein expression is S2 or static modules (Figure 3.3). This indicates that most of the cell-to-cell variations in cell behavior are due to the variations in the expression levels of static modules. This is particularly interesting given that static modules are the ones that are subject to the least amount of transcriptional regulation in response to an external stimulus.

Static modules are primarily those involved in the regulation of mRNA synthesis, splicing and/or transport (see above). A recent study in *C. elegans* has suggested a role for co-regulators of transcription as common modifiers of cell behavior as these proteins were involved in genetic interactions within diverse signaling pathways (Lehner, et al., 2006). We wanted to check the genetic interactions in

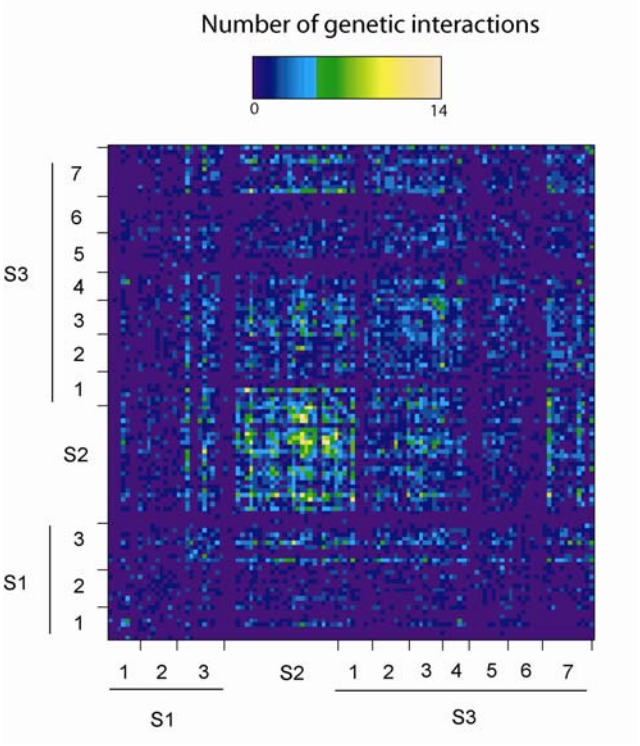


**Figure 3.9.** Genetic interaction matrix of proteins with different dynamic profiles. Each square *i, j* in the matrix was normalized to the total number of genetic interactions of proteins in bins *i* and *j*.

yeast and see if the same pattern as in worm can be reproduced. For this purpose, we compiled all the genetic interaction data from the BioGRID database (Stark, et al., 2006) and constructed a genetic interaction matrix conserving the ordering of proteins in the clustering in Figure 3.1. Most of the genetic interactions observed are among the proteins of S2, even after normalizing for the total number of genetic interactions (Figure 3.9), indicating that S2 proteins are the hubs of the genetic interaction network. This observation corroborates with the study in *C. elegans* (Lehner, et al., 2006), and together with the observation that S2 is one of

two most central groups in the protein interaction network (see above), shows that static modules are the common modifiers of cell behavior (see Discussion).

### 3.3.4 The dynamic organization pattern is reproducible across different datasets

An important factor to be considered in protein network studies is the high rate of false positives in high throughput protein-protein interaction data. Even though our dataset contains only high quality data(Bader, et al., 2004; Krogan, et al., 2006), we wanted to check if the dynamic profiles in this study and their interaction profiles can be reproduced using other high quality datasets. For this purpose, we used high quality datasets from two recent studies that reported contradictory findings with respect to each other about network modularity (Batada, et al., 2007; Bertin, et al., 2007). A clear separation of S1 and of its subgroups, S2 and S3 groups as well as their interaction patterns very similar to the one in Figure 3.1 can be seen in both datasets (Figure 3.10A). Out of each dataset, we extracted a cluster that most resembled S3.7 according to their dynamic profiles. Our criterion for S3.7 was that the cluster must have a high EV, low *yK*, low nPCC and avPCC, moderate nEV and high $v^{EV}$ or $v^{PCC}$, in accordance with Figure 3.1. In both datasets, the cluster we extracted had a significantly

higher within-group density of interactions than what would be expected by chance (not shown), supporting our observations above.

In addition, the genetic interaction pattern in Figure 3.7 can also be seen in both datasets with some differences (Figure 3.10B). These observations show that the dynamic organization pattern identified in this study is likely to reflect true biology rather than some artifact in the interaction data. Although S2 and subgroups of S1 can be clearly distinguished from the profiles in these datasets, not all of the subgroups of S3 may be readily distinguishable. We suggest that, in addition to the obvious differences between datasets, this is because dynamic profiles of S3 are most subtle and may be more difficult to capture than in the case with S1 and S2.

Using their high quality dataset, Batada *et al* (2007) argued against the model of organized modularity in the protein interaction network(Batada, et al., 2006; Batada, et al., 2007) that was proposed earlier(Han, et al., 2004). Interestingly, using our approach, we show that their dataset in fact supports the model of organized dynamic modularity. We suggest that our multi-dimensional approach can resolve the discrepancy in literature by providing a more comprehensive view of the protein network characteristics.

## 3.4 DISCUSSION AND FUTURE PERSPECTIVES

In this work, we first derived several novel graph theoretical metrics to explain the dynamic behavior of a protein and of its neighborhoods in the network, then out of a global distribution of dynamic profiles of proteins we identified dynamical classes with distinct dynamic profiles, then we characterized some of the important properties of these classes, and finally proposed a functional dynamic layout model of the protein network.

### 3.4.1 Static and dynamic modules

A proper stoichiometry in the expression levels of components of a module is essential as an imbalance in the levels of the module constituents can be deleterious (balance hypothesis) (Papp, et al., 2003). A priori, there are two simple ways to control stoichiometry of module components at the level of transcription: by maintaining constant expression, or by co-regulated expression of all components. Both mechanisms are apparently employed for the design of cellular modules, leading to an organizational model of the network resembling a circuit board with integrated “built-in” as well as removable “plug-and-play” components. For example, the functionally ubiquitous process of mRNA

synthesis and splicing is carried out by proteins organized in modules with apparent invariant expression. The highly dynamic nature of ribosome biogenesis modules, on the other hand, has been suggested to be a mechanism of energy preservation for the cell under stress, as transcription of ribosomal genes accounts for around 80% of all RNA synthesis in the cell (Warner, 1999).

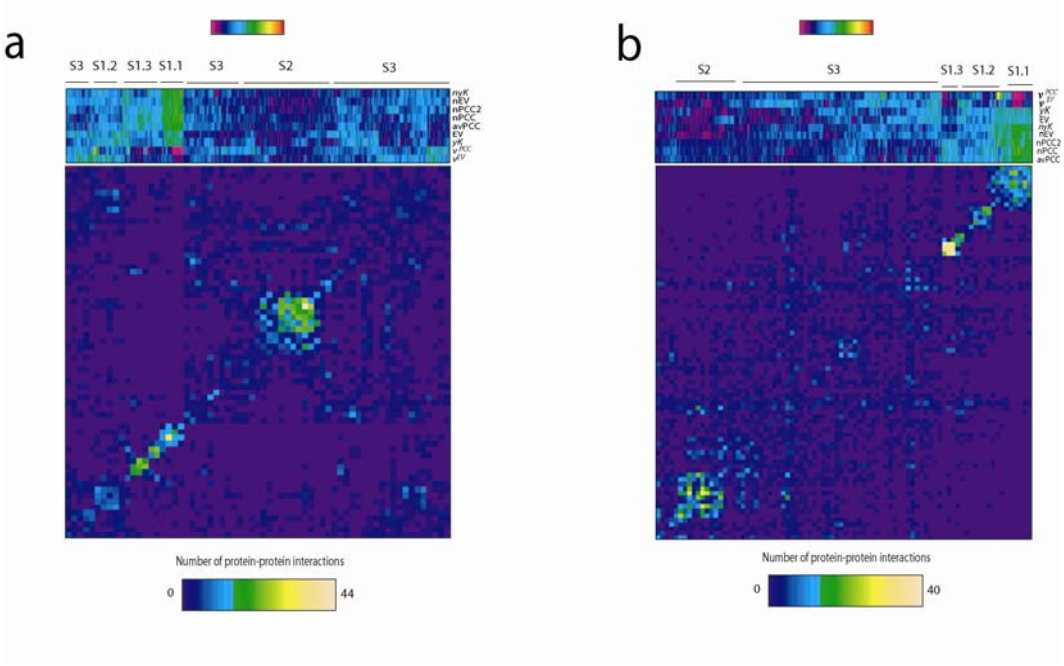


**Figure 3.10.** Heatmaps for the dynamic profiles of proteins in two independent datasets and their protein-protein interaction profiles. **a)** High confidence dataset from Bertins *et al* (2006). **b)** High confidence dataset from Batada *et al* (2006). Ordering of proteins in bins of the interaction matrices are exactly like in the heatmaps above each matrix. Clustering was done the same way as for our dataset (see text).

While the expression variations of modular proteins are constrained by those of their neighbors, central proteins, which are versatile in their functions, are also more versatile in their expression patterns. The existence of both static and

dynamic central hubs, which are presumably the coordinators of cellular processes, suggests that some connections between processes in the cell are "hard-wired", while some are adjustable depending on the cellular requirements. For example, the sub-network 2 in our static network (Table 3.1) indicates that the TFIID/SAGA complex is hard-wired to the nuclear proteasomal complex, suggesting an integral function of the proteasome in sequence-specific transcription, consistent with previous reports (Auld, et al., 2006; Lee, et al., 2005). This sub-network also indicates an integral connection of vesicle trafficking with general mRNA synthesis, a relationship that to our knowledge has not yet been explored. Therefore, in addition to revealing some novel architectural characteristics of the protein network, the analysis employed in this study also helps reveal how local dynamics of the network architecture may shape cell behavior.

The faster evolutionary rate and higher expression noise in static modules suggests that robustness to variations in these modules may be a selected trait during evolution. Since expression noise may contribute to population fitness of unicellular organisms (Kaern, et al., 2005; Raser and O'Shea, 2005), localization of noise to static modules may reflect a specific fitness advantage to the population. An interesting observation consistent with this hypothesis is that proteins functioning in the regulation of mRNA synthesis, which are mostly static

in yeast, have been found to be phenotypic enhancers of genetic mutations in worm as well as of oncogenic mutations in human cancers (Lehner, et al., 2006). High variability in the levels of common modifiers of cell behavior is of particular importance. Cell-to-cell variations in these modules would enhance or dampen the phenotypes of other specific expression variations in the cell. This would invariably lead to a wider and a more robust distribution of cellular phenotypes within a clonal cell population. Similarly, genetic variations in static modules during evolution may result in the phenotypic enhancement of other mutations in the cell, which may facilitate adaptation. Since mRNA abundance is a major factor contributing to protein expression noise (Newman, et al., 2006) and evolutionary rate (Pal, et al., 2001), it is conceivable that relatively lower expression levels of static modules is an evolutionarily selected trait to maximize variations in these modules.

### 3.4.2 The Rich Club phenomenon in protein interaction networks

We have found that the most central organizers of the protein interaction network have a high preference of interactions for each other and therefore form a highly connected web at the core of the network. This connectivity pattern is reminiscent of the “rich club” phenomenon in complex networks, which is characterized by a significant connection density among “important” hubs (i.e. “rich” nodes) in the

network (hence “rich club”), and has implications in the network routing efficiency, redundancy and/or robustness (Colizza, et al., 2006; Zhou and Mondragon, 2004). The rich-club in the internet network has been suggested to serve as a super traffic hub and provide a large selection of shortcuts for a greater efficiency and flexibility of the traffic routing (Zhou and Mondragon, 2004). In the case of protein interaction networks, dense connectivity between central proteins indicates fast communication between different parts of the network, which may be necessary for an efficient coordination of cellular processes carried out by modules that can be far apart in the network.

Another dimension to this intriguing scenario is added by the consideration of highly regulated expression pattern of S3.7 proteins, as evidenced from their high EV (see Appendix). While allowing for fast information flow within the network, the pattern of signal transduction between different parts of the network may be regulated by modulating the expression levels of central proteins, thereby fine-tuning network behavior according to the conditions at hand. Therefore, it is tempting to speculate that the presence of rich clubs among highly dynamic proteins in the protein interaction networks of eukaryotes may be an evolutionarily selected mechanism of highly efficient yet regulated signal propagation across the network.

### 3.4.6 Emerging and disappearing paradigms

Since the initial observation of differential positioning of proteins in the network according to their expression profile based on a single metric (avPCC)(Han, et al., 2004), there has been some debate regarding whether the original observations by Han *et al* (2004) reflected an artifact of the specific network they used for their study (Batada, et al., 2006; Batada, et al., 2007). By utilizing a more comprehensive survey of expression characteristics of proteins as well as of their immediate network localities in several datasets, our study confirms the notion of dynamic modularity in the eukaryotic protein interaction network. We propose that the discrepancy in the literature is due to a relatively narrow view of the network characteristics that was offered by the analyses of respective groups in their studies. We show that a more comprehensive analysis can resolve the discrepancy between these studies by offering a higher resolution view of the dynamic network organization. For example, the initial proposition of so-called "date" hubs to be central proteins by Han *et al* is refined in this study by showing that date hubs also contain highly modular static proteins as well as non-central organizer proteins. Moreover, most of the characteristics attributed to date hubs (like higher evolutionary rate, higher synthetic lethality rate, higher density of genetic interactions) turn out to be the characteristics of proteins in static modules, which, importantly, logically dissociates the notion of centrality from

the variability in protein networks suggested earlier(Fraser, 2005). In addition, suggestion that the protein network lacks an organized pattern(Batada, et al., 2006; Batada, et al., 2007) (and hence displays a disorganized highly inter-connected "stratus" pattern) is shown to be incorrect in this study by using a more comprehensive approach, even using the same dataset as in the original study of Batada *et al*(Batada, et al., 2007).

Extrapolating from notions in social networks (Wasserman and Faust, 1994), centrality in biological networks has been linked with "importance", or essentiality, of genes in yeast (Albert, et al., 2000; Jeong, et al., 2001; Yu, et al., 2007). Most of these studies used node degrees of proteins (i.e. number of interactions) to infer centrality in protein networks, which is not informative about the true centrality of proteins, as highly connected proteins can also be in modules ((Han, et al., 2004) and this study). Here, by using a relatively unbiased approach, we identify true central hubs of protein networks and show that these proteins, which appear to link most classes of proteins to each other, are not more important (if not less important) for the survival of yeast than other highly connected proteins. In fact, the group that is least central both individually and as a group is most enriched for essential genes. We propose that this observation dissociates the previously proposed connection between topological centrality and biological importance.

Most importantly, our study emphasizes the complexity of biological networks and suggests that the complex network characteristics ought to be addressed by analyses that take this fact into account. Although several previous studies have addressed the issue of developing proper methodology for better computational analysis of biological data(Breitkreutz, et al., 2003; Demir, 2004; Endy and Brent, 2001; Hu, et al., 2007; Shannon, 2003), most of these efforts have been towards better graphical representation of biological networks rather than for better computational data mining. To our knowledge, this study is the first to develop novel theoretical formalisms for studying protein network dynamics. We believe that further development of novel methodology for the analysis of biological networks is crucial for systems biology to be successful in discovering the complex fabric of life.

# CHAPTER 4

# High throughput siRNA screen to identify components of EGF signaling

## 4.1 COMBINING LOSS-OF-FUNCTION GENOMICS WITH PROTEOMICS

Our understanding of signal transduction has evolved based on thousands of studies over the last couple of decades. The picture of the resultant network of signaling proteins is significantly large as many of the mechanisms of signaling that give rise to diverse cellular phenotypes have been elucidated at the molecular level, creating opportunities for developing strategies for targeted therapeutic interventions. However, despite this progress, the signaling network generated through these efforts is far from including the whole repertoire of signaling molecules involved in signal transduction, and is also far from encompassing the myriad of intricate relationships giving rise to the network's complex behavior, which underline the need for more comprehensive studies. Moreover, our notions of mechanisms of signaling have been mainly derived from low-throughput biochemistry experiments with a limited focus, which despite their experimental

accuracy, are inherently biased towards the specific approach taken. As a result, much of our current knowledge in the field reflects a large historical bias.

With the goal of taking an unbiased global approach to the identification of signaling components in the signal transduction network, we sought to undertake a large-scale genetic screen where we could interrogate a large number of genes individually for their roles in signaling processes. Genetic loss-of-function assays have been widely employed in the signal transduction field to successfully study the mechanisms of action of genes in signaling processes, but a large-scale assay in this direction has never been done, probably owing to the difficulty of measuring multidimensional readouts in a large-scale platform. Here, we overcome this difficulty by coupling our high throughput genetic screen to reverse-phase protein arrays (RPPAs) where a large-scale readout becomes practical while preserving accuracy. While high throughput siRNA screen allows for a massive parallel knock-down of genes each one at a time, RPPAs can offer a quantitative measurement of the signaling changes in response to each of the genetic perturbations in a highly reproducible and sensitive manner.

### 4.1.1 Use of RNAi screens in functional genomics

Genome-wide RNAi offers a unique advantage of an unbiased interrogation of each gene in the human genome for their role in a given biological process. Several studies in the recent past have made an extensive use of this powerful platform for tackling biologically as well as clinically relevant problems. Using large-scale RNAi, Whitehurst et al and Iorns et al have identified several genes that are important for the resistance of lung and breast cancers to chemotherapeutic agents, respectively (Iorns, et al., 2008; Whitehurst, et al., 2007). Several others have used genome-wide RNAi to study cell biological processes like cell division, cell death and endocytosis (Kittler, et al., 2007; MacKeigan, et al., 2005; Pelkmans, et al., 2005). The specificity and high throughput offered by RNAi screens has made them a primary tool for large-scale loss-of-function screens in culture cells.

For our study, we employed the kinome siRNA library from Dharmacon, which is a collection of all human kinases (human kinome set) and some of the associated canonical proteins, like some phosphatases and small GTPases. Each well of 96-well plates in this library contains pools of 4 oligos targeting a specific gene. In order to control for plate-to-plate differences in transfection efficiency, each plate contains positive and negative controls for transfection. Overall, the library targets 772 genes.

### 4.1.2 Reverse-phase protein arrays for quantitative large-scale profiling

As opposed to forward-phase protein arrays, where antibodies are arrayed on a glass-supported nitrocellulose slide, and cell lysates are applied on top, in reverse-phase protein arrays (RPPA), cell lysates are arrayed on a slide and assayed with a specific antibody of interest. While the former is mainly useful for a low throughput assaying of a large number of antigens, the latter is ideal for a large-scale assaying of relatively few number of antigens. In RPPA, the basic principle is very much like that in western blotting, where immobilized proteins from a sample are assayed with a primary antibody followed by secondary and then a fluorophore for signal detection, the main difference being a large number of samples in RPPA instead of a few in western (Liotta, et al., 2003).

Traditionally, reverse-phase protein arrays have been used in proteomic profiling of large number of samples obtained from clinical specimens. RPPA offers a high sensitivity through signal amplification, and a large dynamic range. Only a few nanoliters of the sample is required for array printing, which is especially important when the samples are limited. RPPA is a novel rising technology with an increasing popularity in functional proteomics (Liotta, et al., 2003).

### 4.1.3 Our strategy

Our protocol combines kinome-wide RNAi with RPPA in a high throughput platform. As our target network for the development and validation of our platform, we chose the EGF signaling network. EGF (Epidermal Growth Factor) is a growth factor that activates several signaling pathways through its receptor, EGFR (EGF Receptor). EGF signaling has become popular in the cancer community because of the frequent genetic alterations found in pathway

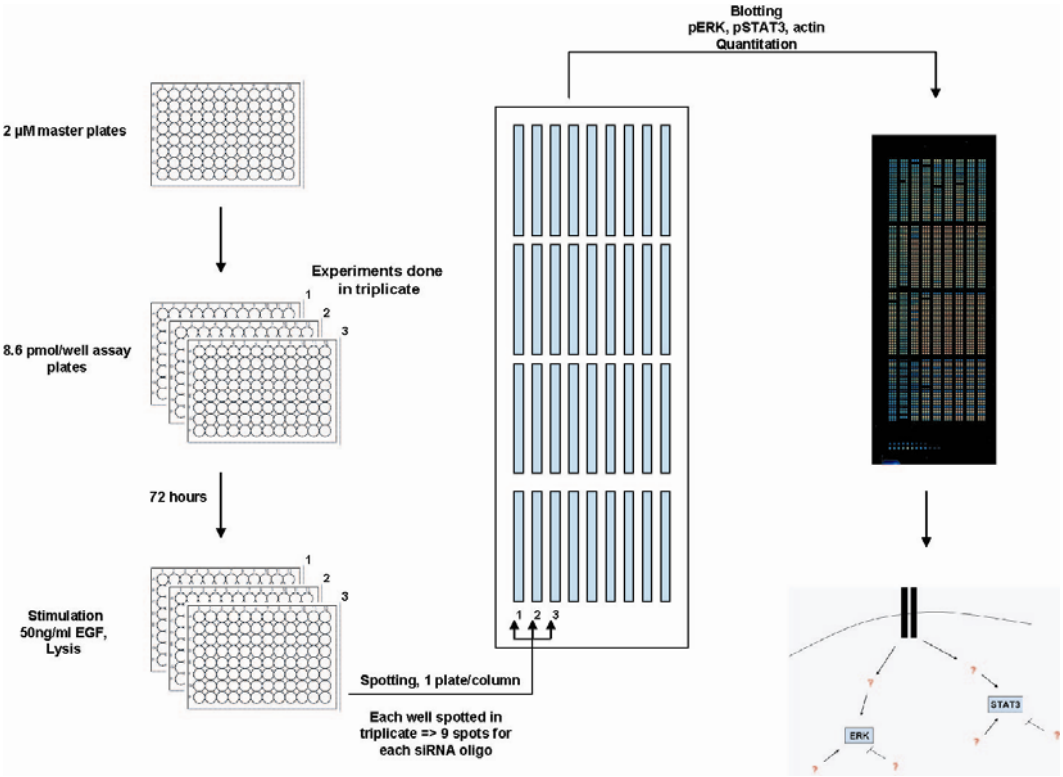


**Figure 4.1. Work scheme for the RNAi screen for regulators of EGF signaling.**

components of this network in many cancers that lead to their hyper-activations. Consequently, the EGF signaling network is arguably the best-studied signaling

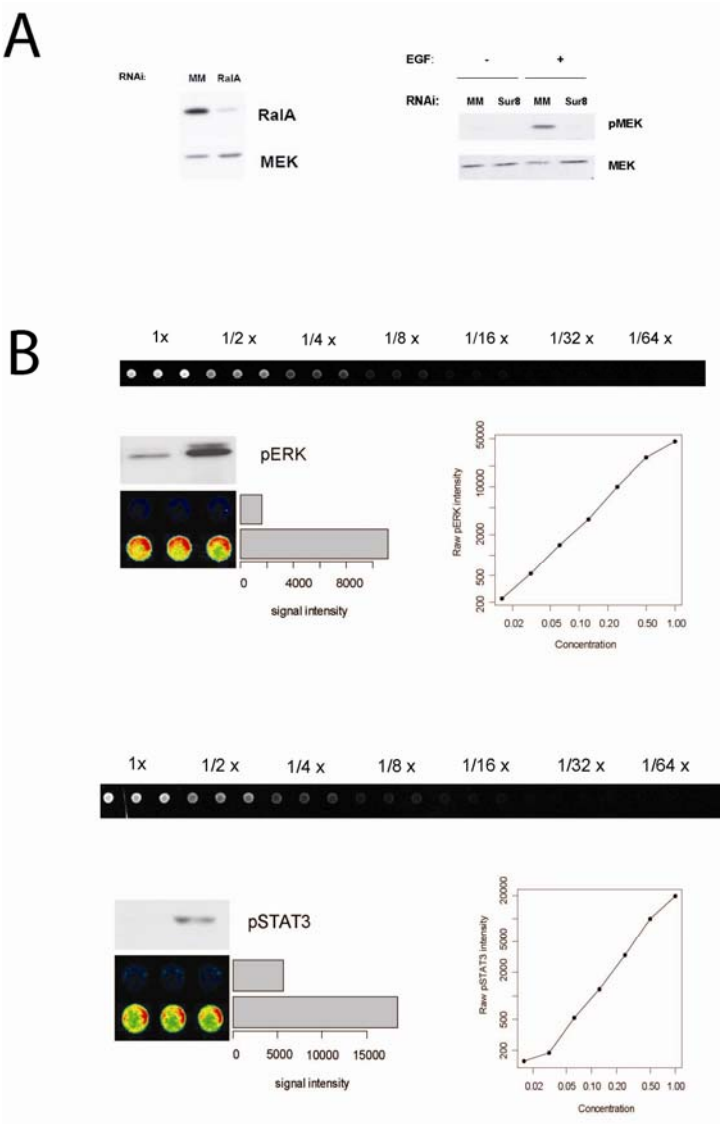


**Figure 4.2.** Platform optimizations. **A)** Transfection optimization using RalA and Sur8 knock-down. Sur8 knock-down blocks signal propagation from EGF to MEK. **B)** Optimization of RPPA protocols for a reproducible spotting and imaging. Upper panel, pERK dilutions showing the dynamic range of the pERK antibody in our platform. Upper band in the pERK plot is ERK1 (MAPK3), and the lower band is ERK2 (MAPK1). Bottom panel, dilution curve and dynamic range for pSTAT3.

network, which makes it an especially attractive target for a pilot study using a novel experimental platform. Therefore, we chose to study signaling downstream of EGF with our combined platform of high throughput RNAi and RPPAs. The work-scheme for the screen is depicted in Figure 4.1 and experimental details are explained in the Methods in Chapter 2. Briefly, we make transfections in triplicate for biological reproducibility, and after 72 hours of transfection, cells are stimulated with 50 ng/ml EGF for 5 minutes. After cells are lysed and processed, lysates are printed onto nitrocellulose-coated slides for RPPA. Slides are processed according to the RPPA procedure and blotted either with pERK, pSTAT3 or Actin antibody and imaged. Spot intensities in the resultant images are quantified, and the data are analyzed as described below.

In this screen, we are expecting to find genes whose deletion affects EGF-dependent signaling to STAT3 or ERK. However, we are also aware that these genes may either be direct modulators of signaling (signal transduction kinases), or those that affect signaling indirectly, may be by the virtue of their role in the transcriptional regulation of some important signaling proteins. Keeping in mind that the potential hits from the screen may be of either type (i.e. direct or indirect modulators of signaling) will help in the interpretation of results.

## 4.2 RESULTS

### 4.2.1 Platform validations

We optimized our transfection procedure for a reproducible knock-down that was efficient in eliciting a signaling phenotype using a previously identified obligate component of the EGF signaling pathway (Figure 4.2a). Lysis, pre-spotting and

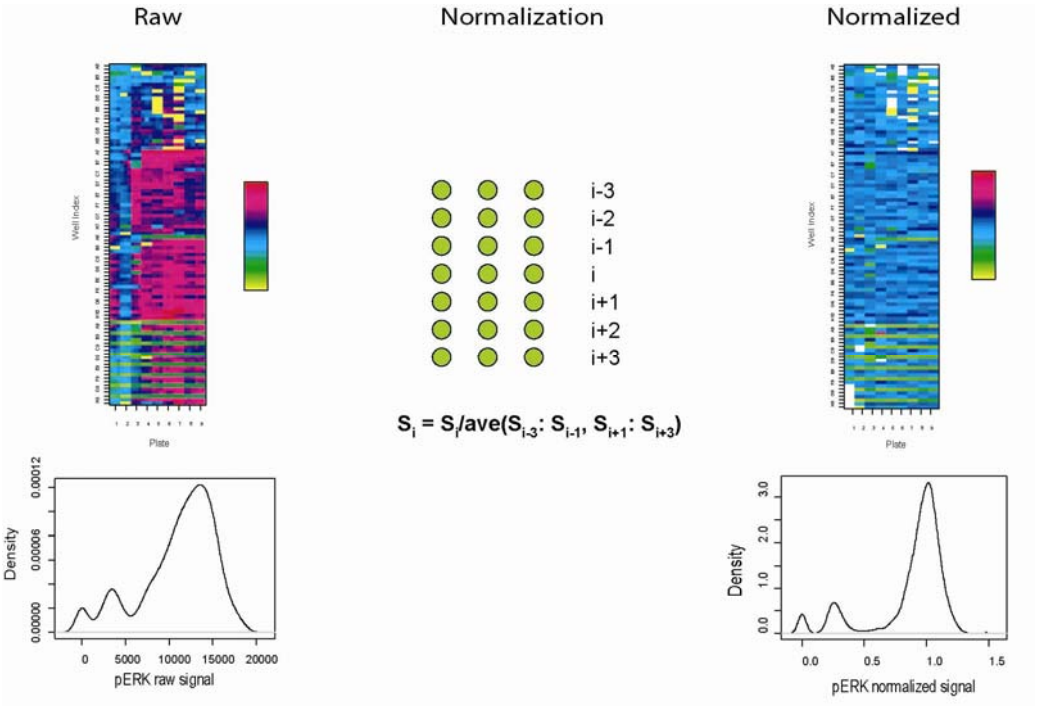


**Figure 4.3.** Normalization of slides for position-specific effects on the slide. The heatmap on the left shows raw signal distribution across one of the slides for pERK. Yellow and white colors indicate missed spots (no signal). The plot below the heatmap shows distribution density of raw pERK signal. Normalization is carried out as described in Methods in Chapter 2. Briefly, each signal ($S_i$) is divided by the average of 6 neighboring spots (3 signals below, and 3 signals above $S_i$). The heatmap on the right shows the distribution of normalized signal intensities across the same slide. The "bumps" on the plots of density distributions are missing values (those around zero) or samples that were stimulated with EGF.

spotting protocols were also optimized to achieve a reproducible wide dynamic range close to 2 logs with both pERK and pSTAT3 antibodies (Figure 4.2b). Importantly, the coefficient of variation between replicate spots was on average 5% for both antibodies, showing high technical accuracy.

*Transfections:* Transfections of A431 cells were performed as described in Methods (Chapter 2). After 72 hours of incubation, cells were stimulated with 50 ng/ml EGF for 5 minutes, and lysed. In order to minimize handling errors when dealing with a large number of plates, we did not starve cells in a serum-free media before stimulation with EGF as we found no difference in the dynamic range of stimulation with or without starvation, possibly due to substantial serum-depletion of the media by cells during 3 days of transfection (not shown).

*Data processing:* In order to eliminate the position-specific differences in the signal intensities on the slide, we normalized each signal to the neighboring spots on the slide (see Methods in Chapter 2 and Figure 4.3), where an assumption was made that the majority of the neighboring samples can be regarded as functionally unrelated. This assumption is not far-fetched in our case as two neighboring samples on a slide are those that are 4 wells away from each other on the assay plate due to the array printing protocol. Accordingly, normalized value distributions for pERK, pSTAT3 and Actin are centered around 1 with a relatively

narrow overall variance with position-specific effects on the slide minimized (Figure 4.3). Samples that did not reproduce in at least 2 of 3 biological replicates, or those with dubious spots were excluded from further analyses. Finally, the biological replicates were averaged to get a single value for each sample for each antibody.

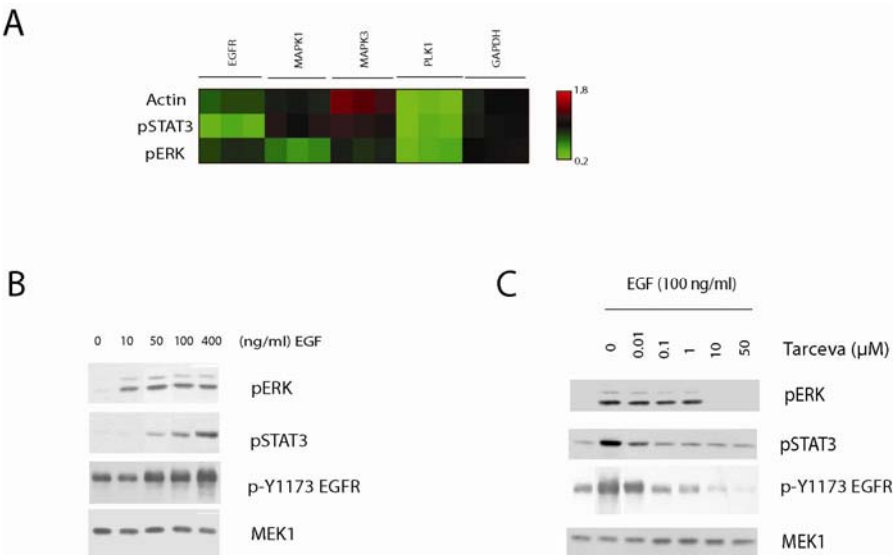


**Figure 4.4.** Observations for some of the canonical components of the EGF signaling network. **A)** Heatmap for the normalized signals for the genes from the screen. Results for all three biological replicates are shown for each gene. **B)** Differential response of ERK and STAT3 phosphorylation in A431 cells to increasing concentrations of EGF at 5 minutes after stimulation. **C)** Differential sensitivity of ERK and STAT3 phosphorylation to specific EGFR inhibitor Tarceva® at 5 minutes after stimulation with 100ng/ml EGF.

*Transfection validation:* Each of our assay plates contains internal positive and negative controls for transfection. PLK1 is included in each plate, and has

significant toxicity to A431 cells upon successful transfection (not shown). In all of our assay plates, we observed a significant reduction in the Actin levels of PLK1 wells (Figure 4.4A), indicating a successful transfection. In addition, some of the canonical members of the EGF signaling network like the EGF receptor (EGFR), ERK2 (MAPK1) and ERK1 (MAPK3) had significant effects on pSTAT3 or pERK levels (Figure 4.4A). MAPK3 did not have a significant effect on pERK levels. We suggest that this is due to the large bias of the pERK antibody towards ERK2 (see Figure 4.2). Surprisingly, EGFR knock-down had no effect on pERK levels, despite its substantial effect on pSTAT3 levels, an effect that can also be seen on western blot (not shown). We hypothesized that this reflects the differential sensitivity of ERK and STAT3 pathways to EGFR depletion. In order to test this hypothesis, we checked how ERK and STAT3 respond to increasing concentrations of EGF, which reflects increasing concentrations of active EGFR. We found that ERK displays a hypersensitivity to EGF, showing near maximal activation at 10 ng/ml EGF at 5 minutes, whereas STAT3 phosphorylation above baseline is undetectable at 10 ng/ml EGF at 5 minutes (Figure 4.4B). Moreover, there is at least 100-fold difference in the sensitivity of ERK and STAT3 pathways to the inhibition of EGFR activity, as shown by their responses to specific EGFR inhibitory drug Tarceva® (Figure 4.4C). STAT3 tyrosine phosphorylation is completely inhibited at 0.1μM concentration of the drug, while ERK phosphorylation is only inhibited at 10μM

concentration. This differential sensitivity of the two pathways to EGFR activation/inhibition not only provides an explanation to why EGFR knock-down (which presumably does not completely eliminate EGFR protein levels despite a considerable reduction) only affects STAT3, but also may give some insight into the modes of regulation of the two pathways by EGFR (see below). Interestingly, although STAT3 phosphorylation correlates with EGFR auto-phosphorylation (Y1173 site), ERK phosphorylation does not seem to (see Figure 4.4B-C).

### 4.2.2 Revealing potential regulators of EGF signaling

In order analyze genes that potentially play roles in the signal propagation downstream of EGF to STAT3 or ERK, we identified genes whose knock-downs reproducibly had at least 20% effect on pSTAT3 or pERK levels after having controlled for Actin levels (see Methods in Chapter 2, Figure 4.5). The resultant 51-gene list contains several known components of the EGF network other than those mentioned above, like phospho-inositol pathway components (PIK3C2A, PIK3C2B, PIP5K1A, PKN3), MAP3 Kinases (MAP3K2, MAP3K3, MAP3K7, MAP3K13), MAP kinases (MAPK4, MAPK12), MAP kinase phosphatases (PTPRR, DUSP1), a small GTPase and a GEF (RAC1, RAPGEF3), and a scaffolding protein (CNKSR1). In addition, other than EGFR, there are other receptor tyrosine kinases, FGFR1, MERTK, CSF1R, EPHA3 and EPHA7. Also,

there are kinases with little or no known function in human cells, like MGC16169, LRRK1, TNK1, HUNK and NYD-SP25.

Interestingly, none of the central Ras/ERK pathway components, such as Raf and MEK family members, made it into the list. This is not surprising as we have shown that no single Raf family member RNAi knock-down affects activation of ERK by EGF in HeLa cells (unpublished data). In addition, none of the Janus kinases (JAKs), canonical STAT tyrosine kinases in cytokine signaling, appeared to have a significant effect on STAT3 tyrosine phosphorylation in response to EGF. This is not surprising either, as activation

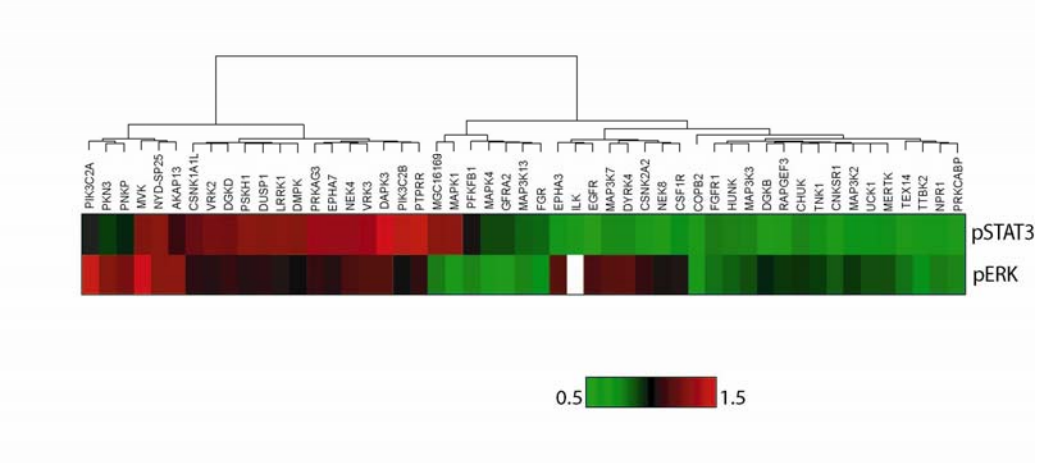


**Figure 4.5.** Heatmap of the final list of 51 candidate genes selected from the screen. Data are normalized to actin. White color for pERK in ILK column reflects a missing value (i.e. experimental error).

of STAT3 by EGF has been shown to be independent of JAK activity, although it is highly dependent on the EGFR tyrosine kinase activity (Leaman, et al., 1996), in accordance with our primary screen data.

Surprisingly, the list of 51 genes seems to contain more than twice as many genes that affect only STAT3 as there are genes that affect only ERK (see Figure 4.5), suggesting that STAT3 pathway is more sensitive to perturbations. Also, pSTAT3 signal has significantly more variance between samples than pERK signal (not

| Genes | Number of cases | pERK value | pSTAT3 value |
|---|---|---|---|
| MAP3K13 | 1 | 0.73 | 0.80 |
| NPR1 | 2 | 0.68 | 0.61 |
| PNKP | 1 | 0.84 | 0.64 |
| TTBK2 | 2 | 0.90 | 0.79 |
| MGC16169 | 3 | 0.84 | 1.36 |
| FGR | 1 | 0.78 | 0.90 |
| AKAP13 | 3 | 1.22 | 1.03 |
| CHUK | 5 | 0.70 | 0.59 |
| FGFR1 | 5 | 0.75 | 0.74 |
| MAP3K2 | 3 | 0.97 | 0.75 |
| MAP3K3 | 1 | 0.80 | 0.70 |
| NEK8 | 3 | 1.11 | 0.73 |
| TEX14 | 8 | 0.91 | 0.55 |
| TNK1 | 1 | 0.92 | 0.75 |
| CSF1R | 41 | 0.99 | 0.70 |
| DGKB | 8 | 0.86 | 0.60 |
| DYRK4 | 1 | 1.01 | 0.66 |
| EGFR | 2783 | 0.82 | 0.39 |
| EPHA3 | 8 | 0.89 | 0.48 |
| LRRK1 | 4 | 1.04 | 1.29 |
| CSNK1A1L | 1 | 1.10 | 1.24 |
| NEK4 | 1 | 1.04 | 1.22 |
| PIK3C2B | 1 | 0.92 | 1.33 |
| DAPK3 | 3 | 1.16 | 1.55 |
| DGKD | 2 | 1.18 | 1.34 |
| DMPK | 1 | 1.11 | 1.38 |
| EPHA7 | 4 | 1.00 | 1.26 |

| | | | |
|---|---|---|---|
| MERTK | 2 | 0.91 | 0.76 |

**Table 4.1.** Genes found mutated in cancers as reported in the COSMIC database and in the recent large-scale studies (see text).

shown). Enrichment of STAT3 hits over ERK hits is not likely to be an artifact of antibody sensitivity, as pERK antibody has more dynamic range than pSTAT3 (see Figure 4.2b). This observation may be consistent with the above-proposed notion that the ERK pathway may be less sensitive to upstream perturbations than the STAT3 pathway.

*Candidate hits mutated in cancers*

EGF-activated pathways play an important role in cancer progression. We asked whether our candidate hit list contains genes that are mutated in cancers and whether their ERK or STAT3 signaling phenotype is informative of their role in cancers. We derived a list of cancer-mutated genes from COSMIC database (Forbes, et al., 2008) (see Methods in Chapter 2) and from a recent large-scale sequencing analysis (Sjoblom, et al., 2006). We find that 28 of 51 candidate hits are also found mutated in various cancers (Table 4.1). Of these, EGFR and CSF1R are known proto-oncogenes frequently mutated in many cancers, and accordingly, they seem to be required for STAT3 signaling by EGF. EPHA3 has also been reported to be frequently activated in breast and colorectal cancers, and the expression of EPHA7 has been found to be frequently suppressed by promoter hyper-methylation. Intriguingly, EPHA3 knock-down reduces STAT3 activation

by EGFR, while EPHA7 knock-down causes an increase in STAT3 tyrosine phosphorylation in response to EGF, which is in accordance with their proposed roles as an oncogene and a tumor suppressor, respectively. Although DAPK3 has not been explicitly reported to be mutated in cancers, its ortholog DAPK1 is a known tumor suppressor in some tumor types (Simpson, et al., 2002; Tada, et al., 2002), and DAPK3 knock-down causes a significant increase in STAT3 activity in response to EGF. Some of the kinases with no known function in human cells, TNK1, MGC16169 and TEX14, are also in the list.

### 4.2.3 Network analysis of hits

Next, we wanted to analyze our potential hits in the context of their network neighborhoods in the protein interaction network. This would enable us to deduce functional relationships between hits as well as to derive hypotheses about the possible mechanism of their action in the regulation of ERK and/or STAT3 pathways. For this purpose, we compiled a large protein-protein interaction network from different public databases. Our network mostly contains previously reported interactions in the literature for human, inferred interactions from other species as well as some primary Yeast 2 Hybrid data (see Methods in Chapter 2). Overall, our protein-protein interaction network consists of 12396 proteins connected by ~ 60,000 interactions.

Although 47 out of 51 hits are represented in our network, only 13 are involved in interactions with each other (Figure 4.6A). MAP3K2 (MEKK2), MAP3K3 (MEKK3), MAP3K7 (TAK1) and CHUK (IKKα) are known to cooperate in the activation of NF-κB in response to some stimuli, and here, we find that they have some role in EGF signaling, particularly in the activation of STAT3. A more obvious subnetwork is the one formed by ILK (integrin-linked kinase), COPB2 (coatomer protein B2) and CSNK2A2 (casein kinase 2), all three of which have a profound effect on STAT3 activation. Interactions of COPB2 with ILK and CSNK2A2 are derived from high throughput studies and therefore no information exists about the functional significance of these interactions. Therefore, a direct implication of these interactions in the regulation of STAT3 activation constitutes a novel relationship. The third subnetwork is composed of EGFR, MAPK1, two phospho-inositide 3-kinase II catalytic subunits PIK3C2A and PIK3C2B, and two MAPK phosphatases PTPRR and DUSP1. Interestingly, DUSP1, dual-specificity phosphatase that has been implicated in ERK dephosphorylation, only seems to affect STAT3. PTPRR is a tyrosine phosphatase that also has been implicated in the regulation of MAP kinase signaling. Here, as in the case of DUSP1, we see that it only affects STAT3 and not ERK, suggesting some role for these phosphatases in the regulation of STAT3 signaling.

The subnetworks we derived here only consist of direct interactions between hits. However, the hits may be involved in common pathways through intermediary proteins that are not represented in the screen. In order to examine what common processes our hits may be involved in, we first identified most common proteins that the hits for ERK or STAT3 interact with. We did this by calculating the enrichment of interactions of each protein in the network with hits from the screen

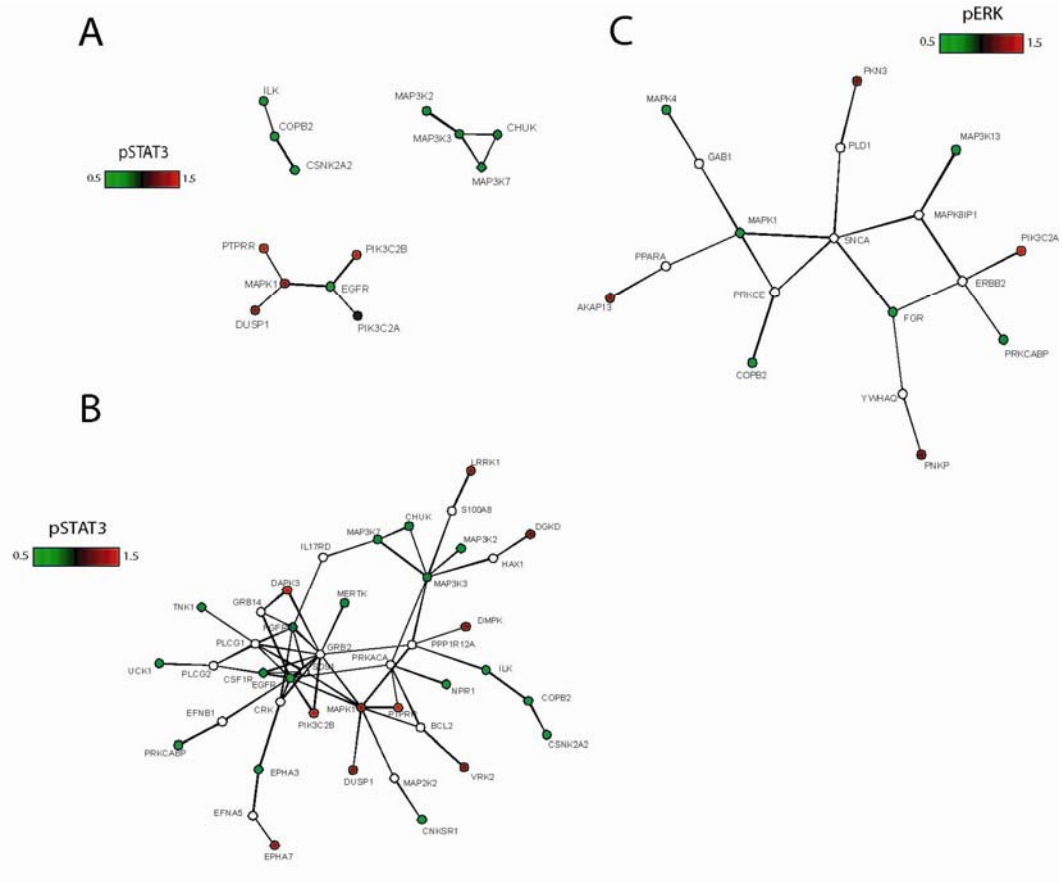


**Figure 4.6.** Network analysis of pathways regulating EGF signaling. **A)** Network connecting hits directly to each other. **B)** STAT3 network formed by including most significant interactors of hits. **C)** ERK network formed in the same as for STAT3 network.

by using hypergeometric formula (see Methods in Chapter 2). Then, we selected the most significant interactors ($q < 0.05$, see Methods in Chapter 2) and constructed a common network that connects most of these hits to each other through these significant intermediary proteins (see Methods in Chapter 2). We constructed a network that connects hits that affected STAT3 phosphorylation or those that had effects on pERK, so that we could do a comparative analysis of two pathways (Figure 4.6B-C).

*STAT3 network*

The STAT3 network seems to have several hubs that connect other hits to themselves through intermediary proteins and a dense cluster of interactions around EGFR (see Figure 4.6B). These hubs and their network vicinities seem to reflect distinct branches of signaling pathways; NF-κB pathway (MAP3K3/CHUK/MAP3K2/MAP3K7), Cytoskeleton/Vesicle trafficking (COPB2/ILK/DMPK/PPP1R12A), MAPK1/2 signaling (CNKSR1/MAP2K2/DUSP1/MAPK1/PTPRR) and EGFR signaling. NF-κB pathway may be involved in the regulation of STAT3 signaling by the virtue of transcriptional regulation of some signaling components, as has been shown previously for NF-κB-dependent activation of STAT3 through transcriptional activation of IL-6 (Squarize, et al., 2006). Regulation of cytoskeleton and endocytosis are intimately coupled and are crucial in modulation of EGFR

signaling (Wiley, 2003). Also, ERK1/2 signaling has long been known to be a critical negative feedback regulator of EGF signaling in general (Buday, et al., 1995; Rozakis-Adcock, et al., 1995) or of STAT3 activation directly (Jain, et al., 1998). The dense cluster of interactions around EGFR may indicate that most of the hits in our screen may be affecting STAT3 phosphorylation by modulation of EGFR activity, which would be consistent with the EGFR knock-down phenotype above. Some of the interactions within these pathways revealed by our network analysis are consistent with their STAT3 phenotypes within this hypothesis. GRB14, a GRB7 family member, has been shown to be a negative regulator of several Receptor Tyrosine Kinases (Cariou, et al., 2004) and DAPK3 (ZIP) has been shown to facilitate its inhibitory activity (Cariou, et al., 2002). We show that DAPK3 knock-down causes an elevation of STAT3 phosphorylation, which is concordant with its possible inhibitory role on EGFR activity. MAPK1 can downregulate EGFR signaling in a negative-feedback loop involving phosphorylation-dependent dissociation of SOS1 from EGFR (Buday, et al., 1995; Rozakis-Adcock, et al., 1995), and accordingly, MAPK1 knock-down causes a slight elevation in STAT3 activity in response to EGF.

Another intriguing observation is the existence of 4 receptor tyrosine kinases other than EGFR in this network, EPHA3, MERTK, CSF1R, FGFR1, all of which appear to negatively affect STAT3 phosphorylation upon their knock-down.

Concomitant co-activation of multiple RTKs in tumors and the need for combined inhibition of multiple RTKs for effective inhibition of oncogenic signaling has been reported (Stommel, et al., 2007). However in our case, inhibition of any one of the given RTKs seems to effectively reduce STAT3 activation in response to EGF, although none seems to significantly affect ERK activation.

A rather unexpected observation is that knock-down of a MAP kinase phosphatase DUSP1 (dual-specificity phosphatase 1) does not affect pERK levels, but causes an increase in pSTAT3 levels. PTPRR (a protein tyrosine phosphatase) also can act as a MAPK phosphatase; however its knock-down gives a similar phenotype to that of DUSP1. Reduction of MAPK1 leads to an increased STAT3 phosphorylation similar to that of DUSP1 and PTPRR, which suggests that the phenotype of DUSP1 and PTPRR knock-down is not a consequence of increased MAPK activity that may result from the depletion of these phosphatases. Therefore, it is possible that the knock-down of DUSP1 and PTPRR leads to an increase in STAT3 phosphorylation through a different effector pathway than the ERK pathway. This would suggest additional roles for DUSP1 and PTPRR as possible modulators of STAT3 signaling independent of their roles in ERK signaling. Indeed, a close homolog of PTPRR, PTPRT, a frequently mutated oncogene in human cancers, has been shown to be a tyrosine phosphatase for STAT3 (Zhang, et al., 2007).

Some of the interactions shown in the STAT3 network are either novel or have not been studied for their functional significance. TNK1-PLCG1 interaction has been reported in mouse cells, but functional implications of this interaction have not been investigated (Felschow, et al., 2000). Similarly, UCK1-PLCG2 interaction is derived from a high throughput Yeast 2 hybrid screen and has not been validated. TNK1 is a non-receptor tyrosine kinase of ACK family, and UCK1 is a uridine/cytidine kinase with no known function in signal transduction. Intriguingly, knock-down of TNK1 or UCK1 causes a reduction of STAT3 phosphorylation in response to EGF, implying a positive role for these kinases in STAT3 signaling. TNK2 (ACK1), a close homolog of TNK1, is known for its role in EGF signaling through its direct association with the EGFR (Shen, et al., 2007). However, TNK1 has not been implicated in EGF signaling. Our network suggests that TNK1 and UCK1 may be employing a common mechanism in the regulation of STAT3 signaling.

Not represented in the network is MGC16169, a novel protein kinase with an unknown function. MGC16169 RNAi causes an increase in STAT3 phosphorylation and a reduction in ERK phosphorylation, suggesting a dual role for MGC16169 as a negative regulator of STAT3 but a positive regulator of ERK activation. In addition to its dual specificity kinase domain, this protein contains a

TBC domain, a domain with Rab GTPase activator function, which may implicate MGC16169 in vesicle trafficking. Rab5 is an essential mediator of EGFR endocytosis after EGF stimulation (Barbieri, et al., 2000), and it has been shown to be required for EGFR signaling (Barbieri, et al., 2004). Therefore, the role of MGC16169 in STAT3 and ERK signaling revealed in our screen may reflect its role in EGFR internalization through its TBC domain.

Another protein not represented in the network is HUNK (hormonally upregulated Neu-associated kinase), a serine-threonine kinase with little known function. Interestingly, HUNK has also been implicated in endocytosis through its interaction with Rabaptin-5 (Korobko, et al., 2000), a Rab GTPase effector protein involved in endocytosis. HUNK knock-down causes a reduction in both pSTAT3 and pERK levels. Therefore, HUNK and MGC16169 may be kinases involved in the regulation of EGF signaling at the level of regulation of endocytosis.

*ERK network*

Next, we constructed a network based on hits that had effects on pERK levels using the same method as above (Figure 4.6C). This network is considerably smaller compared to that of STAT3, reflecting less number of hits associated with pERK (see above). The most central proteins in this network are MAPK1 and

SNCA (synuclein α). SNCA is a gene that is involved in synaptic vesicle trafficking and its accumulation plays an important role in the pathogenesis of Parkinson's disease and Lewy body disease. SNCA has been reported to regulate ERK signaling by modulation of caveolin levels (Hashimoto, et al., 2003). SNCA can also inhibit phospholipase D activity (Ahn, et al., 2002), which is crucial for EGF-dependent recruitment of Sos1 and subsequent activation of Ras (Zhao, et al., 2007). FGR is a non-receptor tyrosine kinase of SRC family, which can phosphorylate SNCA at Y125 (Ellis, et al., 2001). Tyrosine phosphorylation of SNCA has been shown to prevent its self-oligomerization, a key event in its pathogenicity (Negro, et al., 2002). Our finding of reduced ERK activity in response to FGR knock-down may be in agreement with a negative regulatory role of SNCA on ERK signaling. Therefore, SNCA may be a modulator of ERK activity.

Among the ERK hits that are not represented in the network is PFKFB1 (phospho-fructokinase bisphosphatase), an apical component of the glycolytic pathway. PFKFB family of enzymes (PFKFB1-4) control the rate of glycolysis by modulating the concentrations of Fructose-2,6-bisphosphate, a potent allosteric activator of the key glycolytic enzyme, PFK (phosphofructokinase). Increased rate of glycolysis is considered a hallmark of cancer (Warburg effect), and is important for cancer cell survival during energy deprivation due to hypoxia.

Accordingly, some of PFKFB genes are frequently upregulated in tumors (Bartrons and Caro, 2007). As another explanation of selective advantage of the Warburg effect, where tumors maintain a high glycolysis rate regardless of oxygen levels, has been provided by the demonstration of activation of the Akt pathway by NADH, a product of glycolytic pathway (Pelicano, et al., 2006), suggesting that the glycolytic pathway can directly interact with the signal transduction machinery and modulate its activity. Therefore, our finding that PFKFB1 seems to be important for the activation of ERK by EGF may reflect an important mechanism of PFKFB action in tumors.

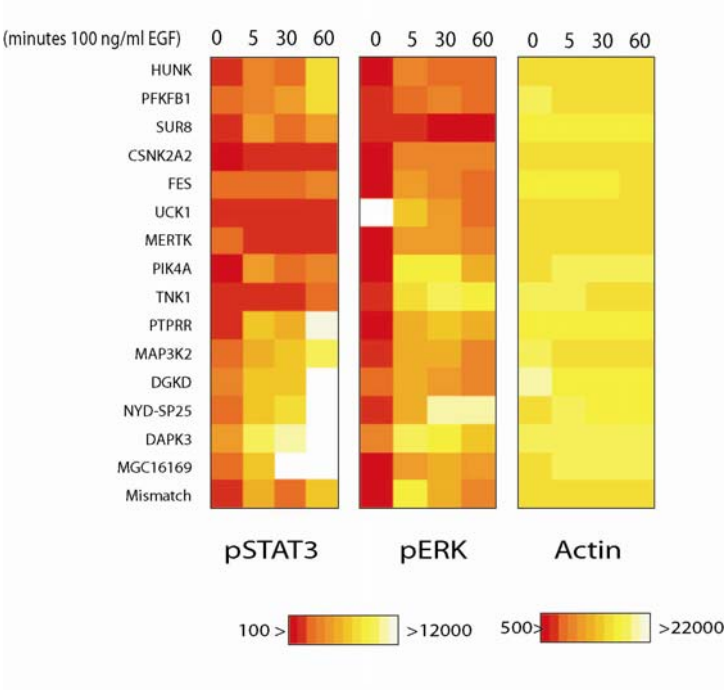


**Figure 4.7. Time course of 100 ng/ml EGF stimulation with potential hits from the screen.** White square for pERK/UCK1 is a sample with missing spots.

### 4.2.4 Analysis of time-course of ERK and STAT3 signaling

Our hits from the primary screen reveal genes that have potential impacts on signal transduction downstream of EGF at an early timepoint. However, we do not know how these genes might be regulating the kinetics of signaling by EGF. For example, a gene may be affecting the amplitude, duration or the shape of the signaling curve in an EGF-dependent of independent manner. Knowing how our hits are impacting signaling by EGF will have important mechanistic implications. In order to get an insight into how our hits may be modulating the kinetics of signaling, we performed another screen using only 16 genes from our 51 list that had most profound effects on STAT3 or ERK. For our secondary analysis, we chose to check ERK and STAT3 signaling activity at 0, 5, 30 and 60 minutes in response to 100ng/ml EGF. At this concentration, ERK phosphorylation levels peak at 5 minutes of stimulation, and gradually go down at 30 and 60 minutes. STAT3 phosphorylation however, increases at 5 minutes, goes down at 30 and then rises again at 60 minutes of EGF stimulation (see Figure 4.7). Interestingly, while the largest dynamic range of ERK phosphorylation between different knock-downs is seen at 5 minutes of stimulation, the largest dynamic range between pSTAT3 signals among the hits is seen at 60 minutes,

possibly reflecting the different kinetics of regulation of signaling in the two pathways.

As expected, knock-down of Sur8, an adaptor protein essential for ERK activation by EGF, blocks activation of ERK by EGF, indicative of successful transfection and assay execution. PFKFB1 and HUNK knock-down had the most significant negative effects on ERK activation in response to EGF, while NYD-SP25 knock-down caused an increase in ERK phosphorylation. The most potent negative effect on STAT3 phosphorylation was observed with TNK1, UCK1, MERTK and CSNK2A2, while the knock-down of DAPK3, MGC16169, PTPRR and DGKD caused an increase in STAT3 phosphorylation.

Although TNK1, UCK1 and MERTK do not seem to affect background levels of STAT3 phosphorylation, DGKD and DAPK3 seem to significantly increase STAT3 phosphorylation in the absence of EGF. MGC16169, HUNK and CSNK2A2 also seem to affect background phospho-STAT3 levels to a lesser degree. These observations suggest that TNK1, UCK1 and MERTK are specifically blocking the ability of EGF to activate STAT3, rather than reducing the overall phospho-STAT3 levels. In contrast, DGKD and DAPK3 seem to cause a hyper-activation of STAT3 even in the absence of EGF, implying that their effect on STAT3 may be EGF-independent. A notable observation here is that

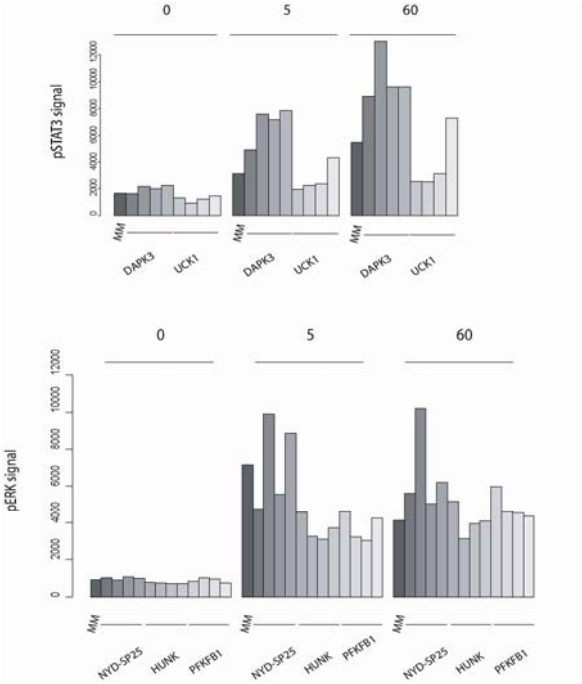


**Figure 4.8** Off – target effect validation of hits using 4 independent oligos for each. Only results for DAPK3, UCK1, NYD-SP25, HUNK and PFKFB1 are shown.

most of variation in STAT3 signaling between STAT3 hits is seen not early during EGF response, but late as shown by the 60 minutes time point. Although STAT3 hits that were causing increased STAT3 activity upon their knock-down do not significantly affect STAT3 phosphorylation as compared to the control at 5 minutes of stimulation, there is a significant difference at 60 minutes. This may indicate that STAT3 activation may be near peak at its early time points making it more difficult to hyperactivate it, while the later time points are significantly more

sensitive to activating perturbations. A similar scenario is seen with the effect of NYD-SP25 on ERK phosphorylation. Although the effect of NYD-SP25 cannot be clearly seen at 5 minutes, it can clearly be seen at 30 and 60 minutes of stimulation, again implying a relative saturation of ERK phosphorylation at early time points. Effects of inactivating perturbations on both pathways, as seen with TNK1, UCK1, MERTK, PFKFB1, HUNK and CSNK2A2 can be seen at all time points.

### 4.2.5 Controlling for off-target effects

The main potential problem of large-scale RNAi screens is their susceptibility to off-target effects of knock-downs. Although oligos in our RNAi library have been optimized to minimize off-target effects, it is necessary to check if our observations with the screen hits may be constituting off-target effects. In order to check for off-target effects of our hits, we tested 4 independent oligos for each of 8 genes from our primary screen that gave most potent phenotypes both in the primary screen as well as with the different time points discussed above. Using RPPAs, we measured pERK and pSTAT3 levels in response to EGF at 5 minutes in each case. Importantly, in 7 out of 8 genes, we got expected results with at least 2 independent oligos, and in one case (TNK1), only one oligo reproduced the primary screen phenotype (Figure 4.8).

### 4.2.6 Involvement of STAT3 hits in the regulation of EGFR

Differential sensitivity of STAT3 and ERK pathways to the perturbations in the EGFR signaling as we have shown above with different doses of EGF as well as EGFR inhibitors (Figure 4.4B), suggests that STAT3 signaling can be specifically modulated by regulating EGFR activity. Moreover, our network plot of STAT3 hits (Figure 4.6) shows that many of STAT3 hits are clustered around EGFR through either direct interactions or indirect interactions. Given these observations, we hypothesized that some of our hits may be regulating EGFR activity thereby affecting STAT3 or ERK signaling. In order to check this hypothesis, we checked if any of our hits may have a role in regulating EGFR activity by testing EGFR tyrosine phosphorylation at some functionally important sites in response to EGF. We tested EGFR phosphorylation at Y1173: an EGFR autophosphorylation site, Y1045: the binding site for the c-Cbl ubiquitin kinase, and Y845: a c-Src target site. Interestingly, the most obvious observation is that the knock-down of UCK1 and MERTK leads to a significant reduction in EGFR protein levels, which results in a reduced STAT3 activation in response to EGF while having no effect on ERK (Figure 4.9). TNK1 seems to be required for Y845 and Y1045 phosphorylation, while HUNK also seems to be required for Y845 phosphorylation. Interestingly, knock-down of DGKD, DAPK3 and MGC16169,

three kinases whose knock-down causes increased STAT3 activation, leads to elevated EGFR Y1045 phosphorylation at 5 minutes of EGF stimulation, suggesting that these kinases may be suppressing the phosphorylation of this site under normal conditions. Based on these observations, we suggest that tyrosine phosphorylation of EGFR at 845 and 1045 may be required for STAT3 activation by EGFR.

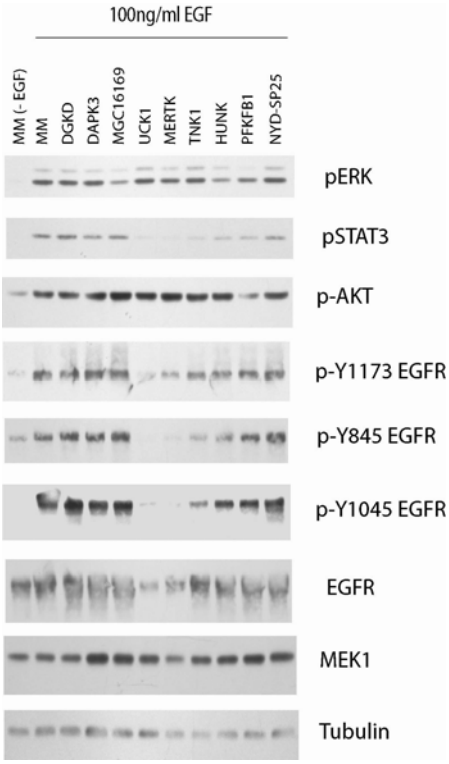


**Figure 4.9.** Effect of knock-down of some hits on EGFR tyrosine phosphorylation.

## 4.3 DISCUSSION

In this work, we combined a functional genomics screen with functional proteomics for the identification of novel regulatory components of the EGF signaling network. In addition, by using a computational systems biology tool that we developed, we were able to extract novel relationships regarding the regulation of STAT3 and ERK pathways by EGF. This work shows high promise of integrating high throughput genomics and proteomics for studying signaling networks.

### 4.3.1 Revealing essential regulators of STAT3 and ERK signaling

Out of 772 genes that we screened in this work, we found that 51 genes have an effect on STAT3 or ERK signaling upon their knock-down. Although there are genes whose knock-down affects both ERK and STAT3 phosphorylation, majority of hits affect only one of them. This may be an indication that the two signaling pathways diverge at a very early step in the EGF signal transduction process, so that there are relatively few genes involved in the signaling processes that are common to both pathways. Divergence of some signaling pathways downstream of EGFR as early as at the level of EGF receptor phosphorylation or endocytosis have been reported. Interestingly, phosphorylation of Y845 on EGFR

has been shown to be essential for EGF-dependent STAT3 activation in A431 cells, but not for activation of ERK (Sato, et al., 2003), suggesting that STAT3 and ERK pathways may diverge at the level of EGFR phosphorylation. In addition, GRB2, an adaptor protein that binds to the EGFR and initiates Ras/ERK signaling, has been shown to compete with STAT3 for the same site on activated EGFR, thereby potentiating ERK pathway activation while suppressing STAT3 tyrosine phosphorylation (Zhang, et al., 2003). The latter observation suggests that ERK and STAT3 signaling can be mutually exclusive on an EGFR molecule, thereby implying that ERK and STAT3 signaling may initiate from different EGF receptor pools on the cell membrane. This hypothesis may not be far-fetched, as EGFR has been reported to be found in both clathrin-coated pits and caveolae, two distinct endocytosis-competent membrane microdomains (Xiao, et al., 2008). Moreover, differential signaling from the two membrane microdomains has been reported, with PLCγ activity being primarily localized to caveolae (Jang, et al., 2001), and ERK signaling to clathrin-dependent endocytosis (Carpenter, 2000). Overall, these observations suggest that different EGF receptor pools may be responsible for the activation of STAT3 and ERK pathways, and that some of the hits from our screen affecting ERK, STAT3 or both signaling pathways may be functioning at the level of EGFR phosphorylation, localization and/or endocytosis.

One interesting observation to support this hypothesis is the differential sensitivity of ERK and STAT3 pathways to EGF doses. ERK can be activated at very low EGF doses, while STAT3 activation requires high EGF concentrations (see above). The same conclusion can be obtained retrospectively with specific EGFR inhibitors, where very low doses of EGFR inhibitors can inhibit STAT3 signaling, while high doses of inhibitors are required to inhibit ERK signaling (see above). These observations suggest that high concentrations of activated EGFR are required to activate STAT3 in response to EGF. Remarkably, high doses of EGF have also been shown to specifically cause relocation of activated EGFR molecules into caveolae, while low EGF doses cause relocation of EGFR molecules into clathrin pits (Sigismund, et al., 2005; Xiao, et al., 2008), which may strongly implicate caveolae in STAT3 signaling, and clathrin-coated pits in ERK signaling.

*Potential regulators of endocytosis*

Interestingly, several of our hits either have been implicated in endocytosis or are likely to be involved in it. Genes that seemed to affect STAT3 or ERK signaling in response to EGF like HUNK and DGKD have been shown to be required for endocytosis (Kawasaki, et al., 2008; Korobko, et al., 2000). In addition, some of the intermediary proteins that preferentially interact with the hits from our screen (see above), such as PLCγ1-2, PLD1-2, GRB2 or SOS1, have also been

implicated in endocytosis of the EGFR in response to EGF stimulation (Carpenter, 2000). Furthermore, still some other hits from the screen have been found to be involved in clathrin- or caveolae/raft-mediated endocytosis in a recent large-scale screen of human kinases for regulators of endocytosis (Pelkmans, et al., 2005) (see Table 4.2). Importantly, there seems to be some correlation between the direction of impact on ERK/STAT3 signaling and the direction of impact on the efficiency of endocytosis with some of the hits (see Table 4.2). For example, DGKD (diacyl-glycerol kinase delta) is required for caveolae/raft-mediated endocytosis, while DGKB (diacyl-glycerol kinase beta) suppresses it. Accordingly, DGKD and DGKB have opposing roles in STAT3 phosphorylation in response to EGF. Since endocytosis is an important component of EGF signaling (Carpenter, 2000), it is reasonable to suggest that many genes affect EGF signaling by the virtue of their involvement in an endocytic process.

This hypothesis is further supported by our finding that several of our hits may be affecting EGFR Y1045 and Y845 phosphorylation, both of which have been implicated in lipid raft-dependent endocytosis (Jang, et al., 2001; Sigismund, et al., 2005). Phosphorylation of EGFR at Y1045, the binding cite for Cbl ubiquiting kinase, has been shown to be required for lipid raft-mediated EGFR internalization (Sigismund, et al., 2005), and we have shown above that TNK1 is required for Y1045 phosphorylation, while DGKD, DAPK3 and MGC16169 are

inhibiting the phosphorylation of Y1045. These observations implicate Cbl binding, EGFR ubiquitination and endocytosis in STAT3 signaling (see Figure 4.10).

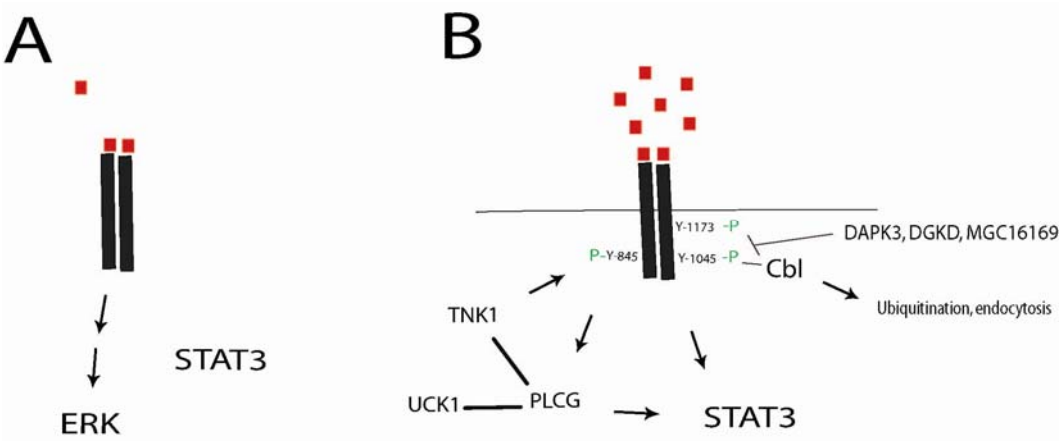


**Figure 4.10** A model of EGFR signaling at low and high EGF concentrations. A) At low EGF concentrations, only ERK pathway is activated. B) At high concentrations, STAT3 is also activated, which strongly correlates with the tyrosine phosphorylation of EGFR at 1173, 1045 and 845 residues. Y1045 is a binding cite for Cbl, and it also seems to be the site of action of at least a few of our hits. Y845 is required for PLCγ activation by EGFR, which only occurs in caveolae (Jang, et al., 2001), and TNK1 seems to be required for Y845 phosphorylation, thereby implicating this site in STAT3 activation by EGFR.

### 4.3.2 Experimental platform

We have presented a pioneering study of combination of functional genomics screen with a functional proteomics platform for an efficient high throughput interrogation of human genes for their roles in signal transduction. As our tools, we chose human kinome siRNA library and reverse-phase protein arrays with

three readouts, phosphorylated ERK, phosphorylated STAT3 and Actin. We also developed a computational data mining strategy based on the protein interaction data from literature to infer functional relationships between our candidate hits from the screen.

Although kinases are *bona fide* signal transducers, our experience shows that kinases are less likely to produce significant signal transduction phenotypes in RNAi screens. This may be due to a high redundancy between kinases, or due to the difficulty in achieving a knock-down level by RNAi that is sufficient to obtain a loss-of-function phenotype with many kinases. This can be observed in the RNAi-mediated knock-down of ERK pathway components in response to EGF, very few of which seem to affect ERK activation in response to EGF. In this study, we show that even the knock-down of the signal source, EGFR, does not affect ERK activation, although a high concentration of a small molecule inhibitor of EGFR does. In contrast, even a partial knock-down of SUR8, an essential scaffolding protein in Raf/ERK pathway, can cause a complete block of ERK activation in response to EGF. Therefore, it is possible that many of the essential kinases playing roles in EGF signal transduction have not been revealed in this study, and that the repertoire of kinases involved in the process of STAT3 and ERK signaling may be much wider than what has been shown here. An extension of our screen to include more gene families (like GTPases, scaffolding proteins,

etc…) may provide a significantly larger collection of potential regulators of EGF-mediated signal transduction.

| | CME | C/R-E | pERK | pSTAT3 |
|---|---|---|---|---|
| MAP3K13 | 2.36 | 0.59 | 0.77 | 0.85 |
| MAP3K7 | 2.34 | 1.68 | 1.16 | 0.77 |
| MVK | 0.86 | 2.97 | 1.50 | 1.27 |
| NPR1 | 0.08 | 0.92 | 0.79 | 0.71 |
| PNKP | 0.85 | 0.46 | 1.25 | 0.96 |
| TTBK2 | 0.78 | 5.19 | 0.73 | 0.64 |
| MGC16169 | 0.93 | 1.17 | 0.80 | 1.29 |
| NYD-SP25 | 1.14 | 0.55 | 1.29 | 1.29 |
| FGR | 0.65 | 0.58 | 0.72 | 0.83 |
| MAPK1 | 1.84 | 0.55 | 0.65 | 1.29 |
| MAPK4 | 0.62 | 2.25 | 0.59 | 0.89 |
| PFKFB1 | 2.33 | 1.01 | 0.77 | 1.03 |
| CHUK | 0.33 | 0.62 | 0.93 | 0.78 |
| FGFR1 | 2.8 | 1.68 | 0.82 | 0.80 |
| HUNK | 0.66 | 1.61 | 0.85 | 0.78 |
| ILK | 1.66 | 2.11 | NaN | 0.58 |
| MAP3K2 | 0.98 | 0.1 | 0.92 | 0.71 |
| MAP3K3 | 1.83 | 0.49 | 0.88 | 0.77 |
| NEK8 | 0.51 | 3.28 | 1.03 | 0.68 |
| TEX14 | 0.56 | 2.5 | 0.81 | 0.50 |
| TNK1 | 0.51 | 0.53 | 0.92 | 0.75 |
| UCK1 | 1.02 | 1.7 | 0.89 | 0.71 |
| CSF1R | 0.08 | 0.18 | 1.04 | 0.74 |
| CSNK2A2 | 0.49 | 0.6 | 1.09 | 0.65 |
| DGKB | 1.1 | 0.11 | 0.96 | 0.67 |
| DYRK4 | 1.03 | 0.71 | 1.18 | 0.77 |
| EGFR | 1.86 | 1 | 1.14 | 0.54 |
| EPHA3 | 0.77 | 0.94 | 1.17 | 0.63 |
| NEK4 | 0.98 | 1.24 | 1.13 | 1.33 |
| PSKH1 | 0.55 | 0.43 | 1.06 | 1.28 |
| VRK2 | 0.84 | 2.28 | 1.06 | 1.24 |
| VRK3 | 0.48 | 0.66 | 1.15 | 1.36 |
| DAPK3 | 1.11 | 1.1 | 1.16 | 1.50 |
| DGKD | 0.13 | 3.55 | 1.09 | 1.23 |
| DMPK | 0.45 | 1.27 | 1.02 | 1.28 |
| EPHA7 | 3.39 | 1.46 | 1.06 | 1.34 |
| MERTK | 1.94 | 0.62 | 0.88 | 0.74 |

**Table 4.2.** Endocytosis data for our hits from Pelkmans et al (2005). CME indicates clathrin-mediated endocytosis, C/R-E indicates caveolae-raft mediated endocytosis. Values smaller than 0.33 here indicates required for endocytosis, while larger than 3 indicate suppresses endocytosis. pERK and pSTAT3 are final neighbor and actin-normalized values from our screen.

Our high density reverse phase protein arrays have made a high throughput profiling of large numbers of screen samples possible in a highly reproducible and quantitative manner. We addressed a problem of position-specific effects on the slide by applying a "smoothing" algorithm where we normalize each spot to its neighboring spots on the array, thereby essentially comparing each sample to its surrounding samples on the array. While this can open doors to other potential problems, such as false positives or negatives due to the nature of neighboring samples, we have obtained data that are biologically meaningful and reproducible within different experimental setups. Probably an ideal control, however, would be a multiplex measurement of slides with at least two antibodies, one of which will be a control for protein load like actin. In this way, each phospho-signal can be normalized to its actin signal without the need for complicated procedures of controlling for position-specific effects on the slide. A limiting factor to this approach is the tyramide-mediated signal amplification step in the staining procedure for reverse arrays (see Chapter 2). The use of two different signal amplification agents (like tyramide conjugated to biotin or fluorescein) can make multiplexing possible.

In order to get insights into potential mechanisms of action of our hits in the screen, we employed a network-based approach where we examined our hits within their context in the protein-protein interaction network. By extracting the most significant intermediary proteins that link hits to each other in the network, we were able to propose some hypotheses regarding the mechanistic and functional distinction of STAT3 and ERK pathway activations in response to EGF. To construct our network, we compiled protein-protein interaction data from many databases, most of which comes from literature-based curation. Perhaps the most obvious limitation in these data is the absence of information about the direction of impact in protein-protein interactions. For example, an interaction between a kinase and another protein can mean a phosphorylation, scaffolding or inhibition, each having an entirely different biological significance. Incorporating these data into our platform would greatly aid in the hypothesis generation about the possible mechanisms of action of hits. Another important addition to this platform would be incorporation of additional heterogeneous datasets for an improved data mining. Integration of gene expression data, other functional genomics data, sequence data, protein domain composition data and alike will no doubt expand the field of view offered by this type of network approach.

Our network-based method for extracting functional relationships between potential hits heavily relies on the existing information on protein interactions in the literature as well as the databases that compile these data. The existing information in the literature is likely to be biased towards the history of research interests in biology. Moreover, it is likely that many interactions reported in the literature are cell type or condition-specific, and therefore may not be applicable to the conditions in a given screen. A more ideal situation would be to be able to compile biological information about the relevant processes in an unbiased manner and that also will be compatible with the screen conditions at hand. A particularly exciting approach in this direction is offered by the recently developed mass spectrometry-based proteomics tools that have been extensively used for extracting condition-specific protein interactions (Blagoev, et al., 2003; Blagoev, et al., 2004; Cox and Mann, 2007; Dengjel, et al., 2007). A preliminary "screen" using these technologies about the biological processes of interest to the RNAi screen at hand would greatly aid in the interpretation of the subsequent data as well as in designing more effective experimental strategies for the screen.

### 4.3.3 Future work

In this study, we were able to identify several potential novel regulators of the EGF signal transduction pathway like HUNK, MGC16169, TEX14, NYD-SP25,

TNK1 and DAPK3. In addition, we proposed previously unknown mechanisms of differential pathway activations by the EGF receptor, like clathrin or caveolae-mediated endocytosis, PLCγ activity and/or EGFR phosphorylation. These potential mechanisms of EGF-mediated activation of ERK and STAT3 have to be investigated further, and the relevant roles of our potential hits have to be identified in these processes.

Questions posed above will be primarily addressed by applying a correlative strategy, where correlation of STAT3 and/or ERK activation phenomena will be correlated with differential EGFR phosphorylation, endocytosis and/or PLCγ activity. Although there are many clues as to differential signaling of EGFR by each of these mechanisms in the literature, no study has addressed this issue with a sufficient rigor. We have shown that STAT3 activation requires high concentrations of active EGFR as evidenced from the STAT3 activation curves with different EGF and EGFR inhibitor concentrations (see above). First of all, the same experiment will be performed in a non-transformed human cell line in order to check if this phenomenon is specific for EGFR-overexpressing cells or if it reflects normal cell biology. Then, this experiment will be repeated by extending the number of readouts to include efficiency of endocytosis (both clathrin-dependent and caveolae-dependent), a panel of EGFR phosphorylation sites and a panel of other downstream pathways like AKT, p38 and JNK. In

addition, inhibition of clathrin-mediated endocytosis by clathrin RNAi, of caveolae-mediated endocytosis by specific small-molecule inhibitors of lipid raft formation like beta-cyclodextrin and of PLCγ activity by specific small-molecule inhibitor will performed with the same readouts as above. Finally, RNAi knock-down of several of our hits will be performed using these same readouts. Primary goal of these experiments will be to extract correlations of each of these readouts with STAT3 and ERK signaling phenotype so as to get insights into the mechanisms of regulation of these two pathways by EGFR.

# Bibliography

Ahn, B.H., Rhim, H., Kim, S.Y., Sung, Y.M., Lee, M.Y., Choi, J.Y., Wolozin, B., Chang, J.S., Lee, Y.H., Kwon, T.K., Chung, K.C., Yoon, S.H., Hahn, S.J., Kim, M.S., Jo, Y.H. and Min, D.S. (2002) alpha-Synuclein interacts with phospholipase D isozymes and inhibits pervanadate-induced phospholipase D activation in human embryonic kidney-293 cells, *J Biol Chem*, **277**, 12334-12342.

Albert, R. and Barabasi, A.L. (2002) Statistical mechanics of complex networks, *Rev. Mod. Phys.*, **74**, 47-97.

Albert, R., Jeong, H. and Barabasi, A.L. (2000) Error and attack tolerance of complex networks, *Nature*, **406**, 378-382.

Allison, D.B., Cui, X., Page, G.P. and Sabripour, M. (2006) Microarray data analysis: from disarray to consolidation and consensus, *Nat Rev Genet*, **7**, 55-65.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M. and Sherlock, G. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium, *Nat Genet*, **25**, 25-29.

Auld, K.L., Brown, C.R., Casolari, J.M., Komili, S. and Silver, P.A. (2006) Genomic association of the proteasome demonstrates overlapping gene regulatory activity with transcription factor substrates, *Mol Cell*, **21**, 861-871.

Bader, J.S., Chaudhuri, A., Rothberg, J.M. and Chant, J. (2004) Gaining confidence in high-throughput protein interaction networks, *Nat Biotechnol*, **22**, 78-85.

Barabasi, A.L. and Albert, R. (1999) Emergence of scaling in complex networks, *Science*, **286**, 509-512.

Barabasi, A.L. and Oltvai, Z.N. (2004) Network biology: understanding the cell's functional organization, *Nat Rev Genet*, **5**, 101-113.

Barbieri, M.A., Fernandez-Pol, S., Hunker, C., Horazdovsky, B.H. and Stahl, P.D. (2004) Role of rab5 in EGF receptor-mediated signal transduction, *Eur J Cell Biol*, **83**, 305-314.

Barbieri, M.A., Roberts, R.L., Gumusboga, A., Highfield, H., Alvarez-Dominguez, C., Wells, A. and Stahl, P.D. (2000) Epidermal growth factor and membrane trafficking. EGF receptor activation of endocytosis requires Rab5a, *J Cell Biol*, **151**, 539-550.

Bartrons, R. and Caro, J. (2007) Hypoxia, glucose metabolism and the Warburg's effect, *J Bioenerg Biomembr*, **39**, 223-229.

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

Batada, N.N., Reguly, T., Breitkreutz, A., Boucher, L., Breitkreutz, B.J., Hurst, L.D. and Tyers, M. (2006) Stratus not altocumulus: a new view of the yeast protein interaction network, *PLoS Biol*, **4**, e317.

Batada, N.N., Reguly, T., Breitkreutz, A., Boucher, L., Breitkreutz, B.J., Hurst, L.D. and Tyers, M. (2007) Still stratus not altocumulus: further evidence against the date/party hub distinction, *PLoS Biol*, **5**, e154.

Bertin, N., Simonis, N., Dupuy, D., Cusick, M.E., Han, J.D., Fraser, H.B., Roth, F.P. and Vidal, M. (2007) Confirmation of organized modularity in the yeast interactome, *PLoS Biol*, **5**, e153.

Blagoev, B., Kratchmarova, I., Ong, S.E., Nielsen, M., Foster, L.J. and Mann, M. (2003) A proteomics strategy to elucidate functional protein-protein interactions applied to EGF signaling, *Nat Biotechnol*, **21**, 315-318.

Blagoev, B., Ong, S.E., Kratchmarova, I. and Mann, M. (2004) Temporal analysis of phosphotyrosine-dependent signaling networks by quantitative proteomics, *Nat Biotechnol*, **22**, 1139-1145.

Blake, W.J., M, K.A., Cantor, C.R. and Collins, J.J. (2003) Noise in eukaryotic gene expression, *Nature*, **422**, 633-637.

Boehm, J.S., Zhao, J.J., Yao, J., Kim, S.Y., Firestein, R., Dunn, I.F., Sjostrom, S.K., Garraway, L.A., Weremowicz, S., Richardson, A.L., Greulich, H., Stewart, C.J., Mulvey, L.A., Shen, R.R., Ambrogio, L., Hirozane-Kishikawa, T., Hill, D.E., Vidal, M., Meyerson, M., Grenier, J.K., Hinkle, G., Root, D.E., Roberts, T.M., Lander, E.S., Polyak, K. and Hahn, W.C. (2007) Integrative genomic approaches identify IKBKE as a breast cancer oncogene, *Cell*, **129**, 1065-1079.

Bonneau, R., Facciotti, M.T., Reiss, D.J., Schmid, A.K., Pan, M., Kaur, A., Thorsson, V., Shannon, P., Johnson, M.H., Bare, J.C., Longabaugh, W., Vuthoori, M., Whitehead, K., Madar, A., Suzuki, L., Mori, T., Chang, D.E., Diruggiero, J., Johnson, C.H., Hood, L. and Baliga, N.S. (2007) A predictive model for transcriptional control of physiology in a free living cell, *Cell*, **131**, 1354-1365.

Breitkreutz, B.J., Stark, C. and Tyers, M. (2003) Osprey: a network visualization system, *Genome Biol.*, **4**, R22.

Brown, M.P., Grundy, W.N., Lin, D., Cristianini, N., Sugnet, C.W., Furey, T.S., Ares, M., Jr. and Haussler, D. (2000) Knowledge-based analysis of microarray gene expression data by using support vector machines, *Proc Natl Acad Sci U S A*, **97**, 262-267.

Buday, L., Warne, P.H. and Downward, J. (1995) Downregulation of the Ras activation pathway by MAP kinase phosphorylation of Sos, *Oncogene*, **11**, 1327-1331.

Cariou, B., Bereziat, V., Moncoq, K., Kasus-Jacobi, A., Perdereau, D., Le Marcis, V. and Burnol, A.F. (2004) Regulation and functional roles of Grb14, *Front Biosci*, **9**, 1626-1636.

Cariou, B., Perdereau, D., Cailliau, K., Browaeys-Poly, E., Bereziat, V., Vasseur-Cognet, M., Girard, J. and Burnol, A.F. (2002) The adapter protein ZIP binds Grb14 and regulates its inhibitory action on insulin signaling by recruiting protein kinase Czeta, *Mol Cell Biol*, **22**, 6959-6970.

Carpenter, G. (2000) The EGF receptor: a nexus for trafficking and signaling, *Bioessays*, **22**, 697-707.

Cho, C.R., Labow, M., Reinhardt, M., van Oostrum, J. and Peitsch, M.C. (2006) The application of systems biology to drug discovery, *Curr Opin Chem Biol*, **10**, 294-302.

Chuang, H.Y., Lee, E., Liu, Y.T., Lee, D. and Ideker, T. (2007) Network-based classification of breast cancer metastasis, *Mol Syst Biol*, **3**, 140.

Colizza, V., Flammini, A., Serrano, M.A. and Vespignani, A. (2006) Detecting rich-club ordering in complex networks, *Nat Phys*, **2**, 110-115.

Cox, J. and Mann, M. (2007) Is proteomics the new genomics?, *Cell*, **130**, 395-398.

de Lichtenberg, U., Jensen, L.J., Brunak, S. and Bork, P. (2005) Dynamic complex formation during the yeast cell cycle, *Science*, **307**, 724-727.

Demir, E. (2004) An ontology for collaborative construction and analysis of cellular pathways, *Bioinformatics*, **20**, 349-356.

Dengjel, J., Akimov, V., Olsen, J.V., Bunkenborg, J., Mann, M., Blagoev, B. and Andersen, J.S. (2007) Quantitative proteomic assessment of very early cellular signaling events, *Nat Biotechnol*, **25**, 566-568.

DeRisi, J., Penland, L., Brown, P.O., Bittner, M.L., Meltzer, P.S., Ray, M., Chen, Y., Su, Y.A. and Trent, J.M. (1996) Use of a cDNA microarray to analyse gene expression patterns in human cancer, *Nat Genet*, **14**, 457-460.

Eisen, M.B., Spellman, P.T., Brown, P.O. and Botstein, D. (1998) Cluster analysis and display of genome-wide expression patterns, *Proc Natl Acad Sci U S A*, **95**, 14863-14868.

Ellis, C.E., Schwartzberg, P.L., Grider, T.L., Fink, D.W. and Nussbaum, R.L. (2001) alpha-synuclein is phosphorylated by members of the Src family of protein-tyrosine kinases, *J Biol Chem*, **276**, 3879-3884.

Endy, D. and Brent, R. (2001) Modelling cellular behavior, *Nature*, **409**, 391-395.

Felschow, D.M., Civin, C.I. and Hoehn, G.T. (2000) Characterization of the tyrosine kinase Tnk1 and its binding with phospholipase C-gamma1, *Biochem Biophys Res Commun*, **273**, 294-301.

Forbes, S.A., Bhamra, G., Bamford, S., Dawson, E., Kok, C., Clements, J., Menzies, A., Teague, J.W., Futreal, P.A. and Stratton, M.R. (2008) The Catalogue of Somatic Mutations in Cancer (COSMIC), *Curr Protoc Hum Genet*, **Chapter 10**, Unit10 11.

Fraser, H.B. (2005) Modularity and evolutionary constraint on proteins, *Nat Genet*, **37**, 351-352.

Friedman, N. (2004) Inferring cellular networks using probabilistic graphical models, *Science*, **303**, 799-805.

Gasch, A.P., Spellman, P.T., Kao, C.M., Carmel-Harel, O., Eisen, M.B., Storz, G., Botstein, D. and Brown, P.O. (2000) Genomic expression programs in the response of yeast cells to environmental changes, *Mol Biol Cell*, **11**, 4241-4257.

Ge, H., Liu, Z., Church, G.M. and Vidal, M. (2001) Correlation between transcriptome and interactome mapping data from Saccharomyces cerevisiae, *Nat Genet*, **29**, 482-486.

Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A., Bloomfield, C.D. and

Lander, E.S. (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring, *Science*, **286**, 531-537.

Han, J.D., Bertin, N., Hao, T., Goldberg, D.S., Berriz, G.F., Zhang, L.V., Dupuy, D., Walhout, A.J., Cusick, M.E., Roth, F.P. and Vidal, M. (2004) Evidence for dynamically organized modularity in the yeast protein-protein interaction network, *Nature*, **430**, 88-93.

Hannon, G.J. and Rossi, J.J. (2004) Unlocking the potential of the human genome with RNA interference, *Nature*, **431**, 371-378.

Harbison, C.T., Gordon, D.B., Lee, T.I., Rinaldi, N.J., Macisaac, K.D., Danford, T.W., Hannett, N.M., Tagne, J.B., Reynolds, D.B., Yoo, J., Jennings, E.G., Zeitlinger, J., Pokholok, D.K., Kellis, M., Rolfe, P.A., Takusagawa, K.T., Lander, E.S., Gifford, D.K., Fraenkel, E. and Young, R.A. (2004) Transcriptional regulatory code of a eukaryotic genome, *Nature*, **431**, 99-104.

Hartwell, L.H., Hopfield, J.J., Leibler, S. and Murray, A.W. (1999) From molecular to modular cell biology, *Nature*, **402**, C47-52.

Hashimoto, M., Takenouchi, T., Rockenstein, E. and Masliah, E. (2003) Alpha-synuclein up-regulates expression of caveolin-1 and down-regulates extracellular signal-regulated kinase activity in B103 neuroblastoma cells: role in the pathogenesis of Parkinson's disease, *J Neurochem*, **85**, 1468-1479.

Higashio, H. and Kohno, K. (2002) A genetic link between the unfolded protein response and vesicle formation from the endoplasmic reticulum, *Biochem Biophys Res Commun*, **296**, 568-574.

Hirsh, A.E., Fraser, H.B. and Wall, D.P. (2005) Adjusting for selection on synonymous sites in estimates of evolutionary distance, *Mol Biol Evol*, **22**, 174-177.

Hu, Z., Mellor, J., Wu, J., Kanehisa, M., Stuart, J.M. and DeLisi, C. (2007) Towards zoomable multidimensional maps of the cell, *Nat Biotech*, **25**, 547-554.

Ihmels, J. (2002) Revealing modular organization in the yeast transcriptional network, *Nat. Genet.*, **31**, 370-377.

Ihmels, J., Friedlander, G., Bergmann, S., Sarig, O., Ziv, Y. and Barkai, N. (2002) Revealing modular organization in the yeast transcriptional network, *Nat Genet*, **31**, 370-377.

Ihmels, J., Levy, R. and Barkai, N. (2004) Principles of transcriptional control in the metabolic network of Saccharomyces cerevisiae, *Nat Biotechnol*, **22**, 86-92.

Iorns, E., Turner, N.C., Elliott, R., Syed, N., Garrone, O., Gasco, M., Tutt, A.N., Crook, T., Lord, C.J. and Ashworth, A. (2008) Identification of CDK10 as an important determinant of resistance to endocrine therapy for breast cancer, *Cancer Cell*, **13**, 91-104.

Ishii, N., Nakahigashi, K., Baba, T., Robert, M., Soga, T., Kanai, A., Hirasawa, T., Naba, M., Hirai, K., Hoque, A., Ho, P.Y., Kakazu, Y., Sugawara, K., Igarashi, S., Harada, S., Masuda, T., Sugiyama, N., Togashi, T., Hasegawa, M., Takai, Y., Yugi, K., Arakawa, K., Iwata, N., Toya, Y., Nakayama, Y., Nishioka, T., Shimizu, K., Mori, H. and Tomita, M. (2007) Multiple high-throughput analyses monitor the response of E. coli to perturbations, *Science*, **316**, 593-597.

Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M. and Sakaki, Y. (2001) A comprehensive two-hybrid analysis to explore the yeast protein interactome, *Proc Natl Acad Sci U S A*, **98**, 4569-4574.

Ito, T., Chiba, T. and Yoshida, M. (2001) Exploring the protein interactome using comprehensive two-hybrid projects, *Trends Biotechnol*, **19**, S23-27.

Jain, N., Zhang, T., Fong, S.L., Lim, C.P. and Cao, X. (1998) Repression of Stat3 activity by activation of mitogen-activated protein kinase (MAPK), *Oncogene*, **17**, 3157-3167.

Jang, I.H., Kim, J.H., Lee, B.D., Bae, S.S., Park, M.H., Suh, P.G. and Ryu, S.H. (2001) Localization of phospholipase C-gamma1 signaling in caveolae:

importance in EGF-induced phosphoinositide hydrolysis but not in tyrosine phosphorylation, *FEBS Lett*, **491**, 4-8.

Jeong, H., Mason, S.P., Barabasi, A.L. and Oltvai, Z.N. (2001) Lethality and centrality in protein networks, *Nature*, **411**, 41-42.

Jeong, H., Tombor, B., Albert, R., Oltvai, Z.N. and Barabasi, A.L. (2000) The large-scale organization of metabolic networks, *Nature*, **407**, 651-654.

Kaern, M., Elston, T.C., Blake, W.J. and Collins, J.J. (2005) Stochasticity in gene expression: from theories to phenotypes, *Nat Rev Genet*, **6**, 451-464.

Kallioniemi, A. (2008) CGH microarrays and cancer, *Curr Opin Biotechnol*, **19**, 36-40.

Kawasaki, T., Kobayashi, T., Ueyama, T., Shirai, Y. and Saito, N. (2008) Regulation of clathrin-dependent endocytosis by diacylglycerol kinase delta: importance of kinase activity and binding to AP2alpha, *Biochem J*, **409**, 471-479.

Kharchenko, P., Church, G.M. and Vitkup, D. (2005) Expression dynamics of a cellular metabolic network, *Mol Syst Biol*, **1**, 2005 0016.

Kittler, R., Pelletier, L., Heninger, A.K., Slabicki, M., Theis, M., Miroslaw, L., Poser, I., Lawo, S., Grabner, H., Kozak, K., Wagner, J., Surendranath, V., Richter, C., Bowen, W., Jackson, A.L., Habermann, B., Hyman, A.A. and Buchholz, F. (2007) Genome-scale RNAi profiling of cell division in human tissue culture cells, *Nat Cell Biol*, **9**, 1401-1412.

Klemm, K. and Bornholdt, S. (2005) Topology of biological networks and reliability of information processing, *Proc Natl Acad Sci U S A*, **102**, 18414-18419.

Komurov, K. and White, M. (2007) Revealing static and dynamic modular architecture of the eukaryotic protein interaction network, *Mol Syst Biol*, **3**, 110.

Korobko, I.V., Korobko, E.V. and Kiselev, S.L. (2000) The MAK-V protein kinase regulates endocytosis in mouse, *Mol Gen Genet*, **264**, 411-418.

Krogan, N.J., Cagney, G., Yu, H., Zhong, G., Guo, X., Ignatchenko, A., Li, J., Pu, S., Datta, N., Tikuisis, A.P., Punna, T., Peregrin-Alvarez, J.M., Shales, M., Zhang, X., Davey, M., Robinson, M.D., Paccanaro, A., Bray, J.E., Sheung, A., Beattie, B., Richards, D.P., Canadien, V., Lalev, A., Mena, F., Wong, P., Starostine, A., Canete, M.M., Vlasblom, J., Wu, S., Orsi, C., Collins, S.R., Chandran, S., Haw, R., Rilstone, J.J., Gandi, K., Thompson, N.J., Musso, G., St Onge, P., Ghanny, S., Lam, M.H., Butland, G., Altaf-Ul, A.M., Kanaya, S., Shilatifard, A., O'Shea, E., Weissman, J.S., Ingles, C.J., Hughes, T.R., Parkinson, J., Gerstein, M., Wodak, S.J., Emili, A. and Greenblatt, J.F. (2006) Global landscape of protein complexes in the yeast Saccharomyces cerevisiae, *Nature*, **440**, 637-643.

Leaman, D.W., Pisharody, S., Flickinger, T.W., Commane, M.A., Schlessinger, J., Kerr, I.M., Levy, D.E. and Stark, G.R. (1996) Roles of JAKs in activation of STATs and stimulation of c-fos gene expression by epidermal growth factor, *Mol Cell Biol*, **16**, 369-375.

Lee, D., Ezhkova, E., Li, B., Pattenden, S.G., Tansey, W.P. and Workman, J.L. (2005) The proteasome regulatory particle alters the SAGA coactivator to enhance its interactions with transcriptional activators, *Cell*, **123**, 423-436.

Lee, T.H. and Linstedt, A.D. (1999) Osmotically induced cell volume changes alter anterograde and retrograde transport, Golgi structure, and COPI dissociation, *Mol Biol Cell*, **10**, 1445-1462.

Lehner, B., Crombie, C., Tischler, J., Fortunato, A. and Fraser, A.G. (2006) Systematic mapping of genetic interactions in Caenorhabditis elegans identifies common modifiers of diverse signaling pathways, *Nat Genet*, **38**, 896-903.

Liotta, L.A., Espina, V., Mehta, A.I., Calvert, V., Rosenblatt, K., Geho, D., Munson, P.J., Young, L., Wulfkuhle, J. and Petricoin, E.F., 3rd (2003) Protein microarrays: meeting analytical challenges for clinical applications, *Cancer Cell*, **3**, 317-325.

Luscombe, N.M., Babu, M.M., Yu, H., Snyder, M., Teichmann, S.A. and Gerstein, M. (2004) Genomic analysis of regulatory network dynamics reveals large topological changes, *Nature*, **431**, 308-312.

MacKeigan, J.P., Murphy, L.O. and Blenis, J. (2005) Sensitized RNAi screen of human kinases and phosphatases identifies new regulators of apoptosis and chemoresistance, *Nat Cell Biol*, **7**, 591-600.

Maslov, S. and Sneppen, K. (2002) Specificity and stability in topology of protein networks, *Science*, **296**, 910-913.

Mewes, H.W., Amid, C., Arnold, R., Frishman, D., Guldener, U., Mannhaupt, G., Munsterkotter, M., Pagel, P., Strack, N., Stumpflen, V., Warfsmann, J. and Ruepp, A. (2004) MIPS: analysis and annotation of proteins from whole genomes, *Nucleic Acids Res*, **32**, D41-44.

Milo, R. (2002) Network motifs: simple building blocks of complex networks, *Science*, **298**, 824-827.

Moffat, J. and Sabatini, D.M. (2006) Building mammalian signalling pathways with RNAi screens, *Nat Rev Mol Cell Biol*, **7**, 177-187.

Negro, A., Brunati, A.M., Donella-Deana, A., Massimino, M.L. and Pinna, L.A. (2002) Multiple phosphorylation of alpha-synuclein by protein tyrosine kinase Syk prevents eosin-induced aggregation, *FASEB J*, **16**, 210-212.

Newman, J.R., Ghaemmaghami, S., Ihmels, J., Breslow, D.K., Noble, M., DeRisi, J.L. and Weissman, J.S. (2006) Single-cell proteomic analysis of S. cerevisiae reveals the architecture of biological noise, *Nature*, **441**, 840-846.

Nunes Amaral, L.A. and Guimera, R. (2006) Complex networks: Lies, damned lies and statistics, *Nat Phys*, **2**, 75-76.

Pal, C., Papp, B. and Hurst, L.D. (2001) Highly expressed genes in yeast evolve slowly, *Genetics*, **158**, 927-931.

Papp, B., Pal, C. and Hurst, L.D. (2003) Dosage sensitivity and the evolution of gene families in yeast, *Nature*, **424**, 194-197.

Pelicano, H., Xu, R.H., Du, M., Feng, L., Sasaki, R., Carew, J.S., Hu, Y., Ramdas, L., Hu, L., Keating, M.J., Zhang, W., Plunkett, W. and Huang, P. (2006) Mitochondrial respiration defects in cancer cells cause activation of Akt survival pathway through a redox-mediated mechanism, *J Cell Biol*, **175**, 913-923.

Pelkmans, L., Fava, E., Grabner, H., Hannus, M., Habermann, B., Krausz, E. and Zerial, M. (2005) Genome-wide analysis of human kinases in clathrin- and caveolae/raft-mediated endocytosis, *Nature*, **436**, 78-86.

Przulj, N., Corneil, D.G. and Jurisica, I. (2004) Modeling interactome: scale-free or geometric?, *Bioinformatics*, **20**, 3508-3515.

Raser, J.M. and O'Shea, E.K. (2005) Noise in gene expression: origins, consequences, and control, *Science*, **309**, 2010-2013.

Rhodes, D.R., Kalyana-Sundaram, S., Mahavisno, V., Barrette, T.R., Ghosh, D. and Chinnaiyan, A.M. (2005) Mining for regulatory programs in the cancer transcriptome, *Nat Genet*, **37**, 579-583.

Robertson, L.S. and Fink, G.R. (1998) The three yeast A kinases have specific signaling functions in pseudohyphal growth, *Proc Natl Acad Sci U S A*, **95**, 13783-13787.

Rozakis-Adcock, M., van der Geer, P., Mbamalu, G. and Pawson, T. (1995) MAP kinase phosphorylation of mSos1 promotes dissociation of mSos1-Shc and mSos1-EGF receptor complexes, *Oncogene*, **11**, 1417-1426.

Sato, K., Nagao, T., Iwasaki, T., Nishihira, Y. and Fukami, Y. (2003) Src-dependent phosphorylation of the EGF receptor Tyr-845 mediates Stat-p21waf1 pathway in A431 cells, *Genes Cells*, **8**, 995-1003.

Sato, M., Sato, K. and Nakano, A. (2002) Evidence for the intimate relationship between vesicle budding from the ER and the unfolded protein response, *Biochem Biophys Res Commun*, **296**, 560-567.

Segal, E., Friedman, N., Kaminski, N., Regev, A. and Koller, D. (2005) From signatures to models: understanding cancer using microarrays, *Nat Genet*, **37 Suppl**, S38-45.

Segal, E., Shapira, M., Regev, A., Pe'er, D., Botstein, D., Koller, D. and Friedman, N. (2003) Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data, *Nat Genet*, **34**, 166-176.

Segal, E., Wang, H. and Koller, D. (2003) Discovering molecular pathways from protein interaction and gene expression data, *Bioinformatics*, **19 Suppl 1**, i264-271.

Segal, E., Yelensky, R. and Koller, D. (2003) Genome-wide discovery of transcriptional modules from DNA sequence and gene expression, *Bioinformatics*, **19 Suppl 1**, i273-282.

Shannon, P. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks, *Genome Res.*, **13**, 2498-2504.

Shen, F., Lin, Q., Gu, Y., Childress, C. and Yang, W. (2007) Activated Cdc42-associated kinase 1 is a component of EGF receptor signaling complex and regulates EGF receptor degradation, *Mol Biol Cell*, **18**, 732-742.

Sigismund, S., Woelk, T., Puri, C., Maspero, E., Tacchetti, C., Transidico, P., Di Fiore, P.P. and Polo, S. (2005) Clathrin-independent endocytosis of ubiquitinated cargos, *Proc Natl Acad Sci U S A*, **102**, 2760-2765.

Silva, J., Chang, K., Hannon, G.J. and Rivas, F.V. (2004) RNA-interference-based functional genomics in mammalian cells: reverse genetics coming of age, *Oncogene*, **23**, 8401-8409.

Simpson, D.J., Clayton, R.N. and Farrell, W.E. (2002) Preferential loss of Death Associated Protein kinase expression in invasive pituitary tumours is associated with either CpG island methylation or homozygous deletion, *Oncogene*, **21**, 1217-1224.

Sjoblom, T., Jones, S., Wood, L.D., Parsons, D.W., Lin, J., Barber, T.D., Mandelker, D., Leary, R.J., Ptak, J., Silliman, N., Szabo, S., Buckhaults, P., Farrell, C., Meeh, P., Markowitz, S.D., Willis, J., Dawson, D., Willson, J.K., Gazdar, A.F., Hartigan, J., Wu, L., Liu, C., Parmigiani, G., Park, B.H., Bachman, K.E., Papadopoulos, N., Vogelstein, B., Kinzler, K.W. and Velculescu, V.E. (2006) The consensus coding sequences of human breast and colorectal cancers, *Science*, **314**, 268-274.

Spirin, V. and Mirny, L.A. (2003) Protein complexes and functional modules in molecular networks, *Proc Natl Acad Sci U S A*, **100**, 12123-12128.

Squarize, C.H., Castilho, R.M., Sriuranpong, V., Pinto, D.S., Jr. and Gutkind, J.S. (2006) Molecular cross-talk between the NFkappaB and STAT3 signaling pathways in head and neck squamous cell carcinoma, *Neoplasia*, **8**, 733-746.

Stark, C., Breitkreutz, B.J., Reguly, T., Boucher, L., Breitkreutz, A. and Tyers, M. (2006) BioGRID: a general repository for interaction datasets, *Nucleic Acids Res*, **34**, D535-539.

Stommel, J.M., Kimmelman, A.C., Ying, H., Nabioullin, R., Ponugoti, A.H., Wiedemeyer, R., Stegh, A.H., Bradner, J.E., Ligon, K.L., Brennan, C., Chin, L. and DePinho, R.A. (2007) Coactivation of receptor tyrosine kinases affects the response of tumor cells to targeted therapies, *Science*, **318**, 287-290.

Stuart, J.M., Segal, E., Koller, D. and Kim, S.K. (2003) A gene-coexpression network for global discovery of conserved genetic modules, *Science*, **302**, 249-255.

Tada, Y., Wada, M., Taguchi, K., Mochida, Y., Kinugawa, N., Tsuneyoshi, M., Naito, S. and Kuwano, M. (2002) The association of death-associated protein

kinase hypermethylation with early recurrence in superficial bladder cancers, *Cancer Res*, **62**, 4048-4053.

Tomlins, S.A., Mehra, R., Rhodes, D.R., Cao, X., Wang, L., Dhanasekaran, S.M., Kalyana-Sundaram, S., Wei, J.T., Rubin, M.A., Pienta, K.J., Shah, R.B. and Chinnaiyan, A.M. (2007) Integrative molecular concept modeling of prostate cancer progression, *Nat Genet*, **39**, 41-51.

Uetz, P., Giot, L., Cagney, G., Mansfield, T.A., Judson, R.S., Knight, J.R., Lockshon, D., Narayan, V., Srinivasan, M., Pochart, P., Qureshi-Emili, A., Li, Y., Godwin, B., Conover, D., Kalbfleisch, T., Vijayadamodar, G., Yang, M., Johnston, M., Fields, S. and Rothberg, J.M. (2000) A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae, *Nature*, **403**, 623-627.

VanMeter, A., Signore, M., Pierobon, M., Espina, V., Liotta, L.A. and Petricoin, E.F., 3rd (2007) Reverse-phase protein microarrays: application to biomarker discovery and translational medicine, *Expert Rev Mol Diagn*, **7**, 625-633.

Warner, J.R. (1999) The economics of ribosome biosynthesis in yeast, *Trends Biochem Sci*, **24**, 437-440.

Wasserman, S. and Faust, K. (1994) *Social Network Analysis*. Cambridge University Press, Cambridge.

Whitehurst, A.W., Bodemann, B.O., Cardenas, J., Ferguson, D., Girard, L., Peyton, M., Minna, J.D., Michnoff, C., Hao, W., Roth, M.G., Xie, X.J. and White, M.A. (2007) Synthetic lethal screen identification of chemosensitizer loci in cancer cells, *Nature*, **446**, 815-819.

Wiley, H.S. (2003) Trafficking of the ErbB receptors and its influence on signaling, *Exp Cell Res*, **284**, 78-88.

Xiao, Z., Zhang, W., Yang, Y., Xu, L. and Fang, X. (2008) Single-molecule diffusion study of activated EGFR implicates its endocytic pathway, *Biochem Biophys Res Commun*, **369**, 730-734.

Yu, H., Kim, P.M., Sprecher, E., Trifonov, V. and Gerstein, M. (2007) The importance of bottlenecks in protein networks: correlation with gene essentiality and expression dynamics, *PLoS Comput Biol*, **3**, e59.

Zhang, T., Ma, J. and Cao, X. (2003) Grb2 regulates Stat3 activation negatively in epidermal growth factor signalling, *Biochem J*, **376**, 457-464.

Zhang, X., Guo, A., Yu, J., Possemato, A., Chen, Y., Zheng, W., Polakiewicz, R.D., Kinzler, K.W., Vogelstein, B., Velculescu, V.E. and Wang, Z.J. (2007) Identification of STAT3 as a substrate of receptor protein tyrosine phosphatase T, *Proc Natl Acad Sci U S A*, **104**, 4060-4064.

Zhao, C., Du, G., Skowronek, K., Frohman, M.A. and Bar-Sagi, D. (2007) Phospholipase D2-generated phosphatidic acid couples EGFR stimulation to Ras activation by Sos, *Nat Cell Biol*, **9**, 706-712.

Zhou, S. and Mondragon, R.J. (2004) The rich-club phenomenon in the Internet topology, *IEEE Commun. Lett.*, **8**, 180-182.