GATA6 Crosstalk between Insulin Signaling and Tumor Necrosis Factor alpha-induced Gene Expression

A Dissertation

Presented to

the faculty of the School of Art and Science

University of Virginia

in partial fulfillment

of the requirements for the degree

Doctor of Philosophy

Department of Pharmacology

by

Zeinab Chitforoushzadeh

August

2016

Approval Sheet

This dissertation is submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Department of Pharmacology

Zeinab Chitforoushzadeh

This dissertation has been read and approved by the Examining Committee:

Dr. Kevin A. Janes (Advisor)

Dr. Thurl E. Harris (Chair)

Dr. David L. Brautigan

Dr. Ira G. Schulman

Dr. Christopher Deppmann

August 2016

Abstract

Many diseases involve malfunction of several signaling pathways. After close to a century of signaling research, we have access to a wealth of information about many individual signaling pathways. However, in order to take our understanding to the next level, we need to study signaling pathways in the context of each other. These types of systematic studies require simultaneous measurement of multiple entities across several time points. The resulting big and complex data sets can be simplified and decoded by means of computational techniques such as mathematical modeling. The work in my thesis builds upon a prediction from a data-driven statistical model using bioinformatics tools to extract hypotheses. These hypotheses are further tested in vitro by means of molecular biology techniques and pharmacological perturbations. Specifically, the model predicted that an early-phase, Akt-associated signal downstream of insulin repressed a set of transcripts induced by TNF. Through bioinformatics and cell-based experiments, we identified the Akt-repressed signal as glycogen synthase kinase-3 (GSK3)-catalyzed phosphorylation of Ser^{37} on the long form of the transcription factor GATA6. Phosphorylation of GATA6 on Ser^{37} promoted its degradation, thereby inhibiting the ability of GATA6 to act as a repressor of transcripts that are induced by TNF and attenuated by insulin. Our analysis showed that insulin-induced signaling activity and TNF-induced transcriptional regulation is integrated through phosphorylation of $\mathrm{GATA6}_L$.

Acknowledgments

This work would not have been possible without the continued support from my mentors, colleagues, friends, and family.

First, I thank God for everytning that I have including being healthy and able to pursue my academic goals. I hope that he helps me to use what I have learned to increase the quality of life for mankind.

I thank my graduate advisor, Professor Kevin Janes for his invaluable and consummate mentorship teaching me science and life lessons. Special thanks to my PhD thesis committee, Professors David Brautigan, Thurl Harris, Ira Schulman and Christopher Deppmann for their support, guidance and helpful suggestions. Past and current members of the Janes lab provided tremendous support by maintaining such a positive, nurturing environment for learning and success that was crucial in completing this dissertation. I appreciate contributions of undergraduate researchers Zi Ye and Sonya Sheng to my project. I also thank my fellow BIMS and biomedical engineering graduate students at UVA that supported me during my studies. I also need to specially thank my fellow iranian friends in Charlottesville who helped me to withstand the turbulences of graduate school.

Finally, I thank my family and friends for their support and motivation. I would like to express my immense gratitude to my parents, for their continuous support and prayers. I wish I could spend more time with them in the recent years and I appreciate that they have put up with the long distance. Throughout my life, they have always ensured that every opportunity is available to me. I also thank my lovely sisters, Fatemeh and Maryam for their constant encouragement and support. I would like to thank my great uncles, Professor Ahmad Jarrahbashi-Razavi and Mohammad-Ali Jarrahbashi-Razavi for introducing me to critical thinking and always embracing my questions since I was a little kid. They helped me to identify my interest in pursuing the science path.

Last but certainly not least, special thanks go out to my wonderful husband, Hamid, for his constant encouragement, sacrifice and love along the way. Also I thank my newborn son, Alireza, for his patience allowing me to finish my studies.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Cells are exposed to a sea of mixed, diverse and sometimes conflicting messages. These signals shape the information network inside the cells including signaling pathways and gene expression programs which drive cellular decisions. A large body of research has been conducted trying to understand individual signaling pathways in isolation which has substantially increased our knowledge about the way a cell processes each individual message. This information has translated into the discovery and development of many drugs. Nevertheless ample scientific evidence suggests that these pathways never work in isolation. One example is the mechanism of acquired drug resistance in cancer therapy when only one pathway is targeted. Such observations trigger scientific curiosity to pursue several intriguing question: How do cells process one stimulus in the context of others in these mixed sea of cues and yet make the right decision? How do signaling pathways cross-communicate with each other? What are those specific nodes of crosstalk between signaling pathways and how can these be investigated in a systematic way? The emergence and accessibility of new high throughput technologies provides opportunities for scientists to dive in and study these cellular pathways in the context of each other at a systems level. These systems analysis methods enable researchers to quantify the information at different levels, as it flows through the cell from the extracellular environment through the cytoplasm into different compartments. Moreover, high throughput systems methods enable scientists to take a global look at what is happening inside the cells by generating big datasets. One challenge though would be to parse out the interesting information within the data and discover new biology. The work in this dissertation demonstrates an example where applying a combination of data-driven statistical modeling and bioinformatics methods enabled us to extract testable hypotheses from signaling and transcriptomics datasets. In addition, testing these hypotheses using molecular biology techniques helped us to make a new discovery regarding an originally hidden crosstalk node.

1.1 Network biology

Common diseases such as cancer, diabetes and asthma develop due to a complex interplay between many genes, proteins and environmental factors. Reductionist molecular biology approaches that study and target one factor at a time have failed to increase the number of efficacious drugs as expected in the postgenomics era [1]. The emergence and increased accessibility of high-throughput technologies has facilitated the adoption of network-based approaches. These approaches allow for a more holistic understanding of biological systems which is the goal of systems biology. Systems biology seeks answer to many unsolved problems in biology. For example, what happens if the effects that biologists are seeking by performing experiments are not 10 fold or higher? Should one just leave that particular scientific question or nuances and choose another question? It seems that scientific community has paid the price for not focusing on these nuances and employing only reductionist traditional approaches.

1.2 Pathway crosstalk

A topic of great interest to both systems biologists and signaling biologists is pathway crosstalk [2]. Receptor-mediated signaling pathways are densely connected and exhibit nonadditive behaviors when stimulated with multiple input ligands [3–5]. Complexity increases even further when considering the consequences of signaling on gene regulation [6–8]. Fortunately, signaling synergy-antagonism is not typically observed with more than two inputs [9–12], suggesting that stimulus pairs are sufficient to assess potential crosstalk.

1.2.1 Crosstalk types

Robin Donaldson and Mufty Calder [13] characterized five types of pathway crosstalk. The authors further validated their proposed crosstalk categories using published experimental data [14]. These crosstalk categories are enumerated below and illustrated in Figure 1.1, adopted from Harvey [15].

- Signal flow crosstalk: a molecule in a pathway changes the activity of signaling molecule (s) in another pathway.

- Substrate availability crosstalk: this type of crosstalk happens when two pathways share common components which they compete for.

- Receptor function crosstalk: a receptor's function is altered in a way that signaling can hppen in the absence of ligand.

- Gene expression crosstalk: two distinct signaling pathways have mutual effects on a transcription factor and the subsequent gene expression.

- Intracellular communication crosstalk: this type of crosstalk happens when one signaling pathway changes the amount of substrate availability for another pathway.

It is important to note that these suggested mechanisms are not the only mechanisms of pathway cross-communication. Among others, signal from two inputs can be integrated at the level of gene expression. In this type of crosstalk, different signals can activate pathways that each perturb a different transcription factor that both mutually modulate the same set of genes. Alternatively, the two pathways can modify the same common transcription factor post-translationally, thus modulating the transcription of a set of genes (Figure 1.1D).

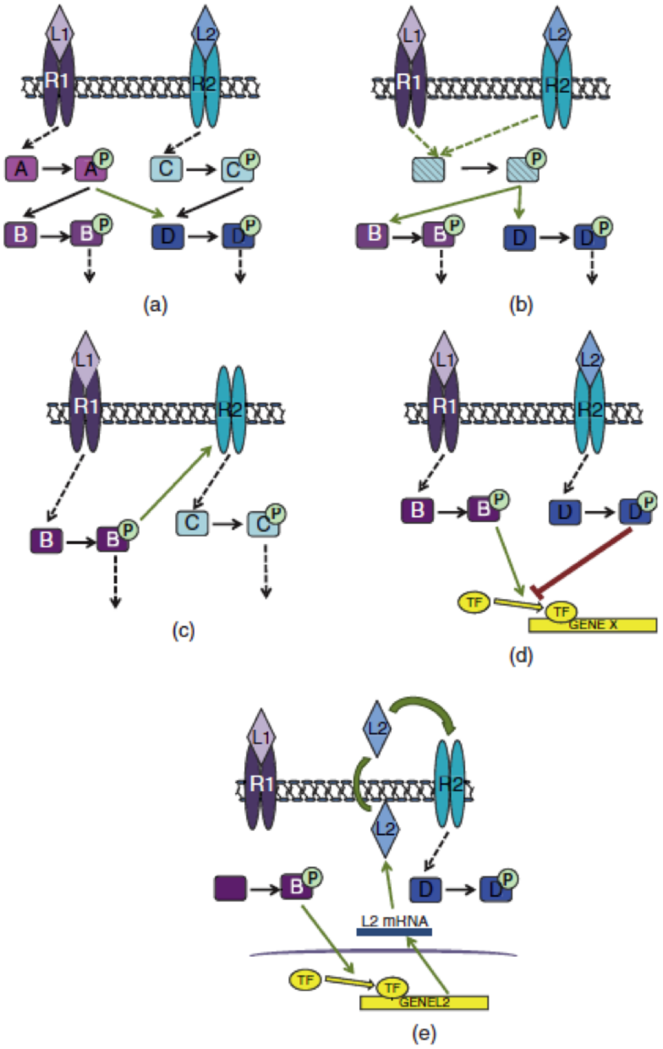


Figure 1.1: Schematic representation on the different types of signaling crosstalk, adopted from Harvey [15]. In signal flow crosstalk (a) a molecule in one pathway affects the rate of activation of signaling molecules in a second pathway. Substrate availability crosstalk (b) occurs when two pathways 'compete' for common components. In receptor function crosstalk (c) an individual receptor's ability to detect a ligand is altered and signaling can occur in ligand absence. When gene expression crosstalk (d) occurs, two pathways have reciprocal effects on transcription factor activation and subsequent gene expression. Lastly, intracellular communication crosstalk (e) occurs when one pathway affects the amount of available ligand for a second pathway.

1.2.2 TNF-Insulin pathway crosstalk

Cytokine-cytokine crosstalk has an important role in both normal and disease states. For example in Inflammatory Bowl Disease (IBD), TNF is a key player in mediating the inflammation of the colonic epithelium [16]. TNF when combined with other inflammatory cytokines can cause cultured colonic epithelial cells to apoptose [17]. On the other hand, Insulin-like growth factor and Insulin signaling stimulate growth in these same cells [18,19]. The opposing communication between the proinflammatory cytokine TNF and insulin has been observed in other cell types as well. For example in adipocytes it has been established that TNF antagonizes the effect of insulin [20,21]. Thus TNF and insulin are good candidate cytokines for studying pathway crosstalk due to their documented potent and antagonistic effect in several cell and disease contexts. A better understanding of TNF and insulin crosstalk can yield novel therapeutics for conditions such as obesity, diabetes and inflammatory bowl disease.

1.3 GATA family of transcription factors

GATA is an evolutionary conserved family of zinc-finger transcription factors that bind to the nucleotide sequence (A/T)GATA(A/G) in the regulatory region of genes and activate or repress their expression. GATA transcription factors have been studied in the context of development, differentiation and oncogenesis. There are currently six members in vertebrates. Based on initial studies of their tissue specific expression patterns they have been divided into two subfamilies. GATA1, 2, and 3 belong to the hematopoietic subfamily known to be expressed in hematopoitic stem cells. GATA4, 5, and 6 are expressed mainly in mesoderm and endoderm-derived tissues such as heart, gut, liver, lung and gonad [22]. However recent studies have shown a much broader expression pattern for GATA factors beyond this initial categorization. One example is the expression and important role of GATA3 in non-hematopoitic cells such as mammary epithelial cells [23]. Another example is the important role of GATA6 in the regulation of tissue macrophage proliferative renewal [24]. Genetic perturbation studies have demonstrated that loss of all GATA factors other than GATA5 is embryonically lethal [25–31].

At the amino acid level, GATA transcription factors are well conserved in their two zinc finger domains (Figure 1.2A). However the sequence outside the zinc finger, vary a lot among the different family members (Figure 1.2B). On the other hand each GATA factor's sequence seems to be well conserved across the different vertebrate species. The variation in the sequence outside the zinc finger domain in GATA proteins points to an existing regulatory potential in these regions.

1.4 GATA6

GATA6 is the sixth member of the GATA family of zinc finger transcription factors. A number of perturbation studies point out the important role of this transcription factor in early embryogenesis and tissue specification. Among other GATAs, GATA6 is required earliest in embryogenesis. According to global gene knockout studies

```
A
hGATA1 ECVNCGATATPLWRRDRTGHYLCNACGLYHKMNGQNRPLIRPKKRLIVSKRAGTQCTNCQTTTTTLWRRNA
hGATA2 ECVNCGATATPLWRRDGTGHYLCNACGLYHKMNGQNRPLIKPKRRLSAARRAGTCCANCQTTTTTLWRRNA
hGATA3 ECVNCGATSTPLWRRDGTGHYLCNACGLYHKMNGQNRPLIKPKRRLSAARRAGTSCANCQTTTTTLWRRNA
hGATA4 ECVNCGAMSTPLWRRDGTGHYLCNACGLYHKMNGINRPLIKPQRRLSASRRVGLSCANCQTTTTTLWRRNA
hGATA5 ECVNCGALSTPLWRRDGTGHYLCNACGLYHKMNGVNRPLVRPQKRLSSSRRAGLCCTNCHTTNTTLWRRNS
hGATA6 ECVNCGSIQTPLWRRDGTGHYLCNACGLYSKMNGLSRPLIKPQKRVPSSRRLGLSCANCHTTTTTLWRRNA
       ******:  ******* ************ **** .***::*::*:  ::* *  *:**:**.*******:

hGATA1 SGDPVCNACGLYYKLHQVNRPLTMRKDGIQTRNRK
hGATA2 NGDPVCNACGLYYKLHNVNRPLTMKKEGIQTRNR-
hGATA3 NGDPVCNACGLYYKLHNINRPLTMKKEGIQTRNR-
hGATA4 EGEPVCNACGLYMKLHGVPRPLAMRKEGIQTRKRK
hGATA5 EGEPVCNACGLYMKLHGVPRPLAMKKESIQTRKRK
hGATA6 EGEPVCNACGLYMKLHGVPRPLAMKKEGIQTRKRK
       .*:********* *** : ***:*:*:.****:*

B
hGATA1 ASGKGKKKRGSSLGGTGAAEGPAGGFMVVAGGSGSGNCGEVASGLTLGPPGTAHLYQGLGPVVL---SGPV
hGATA2 -----------------------------KMSNKSKKSKKGAECFEELSKCMQEKSSPFSAAALAGHMAPV
hGATA3 -----------------------------KMSSKSKKCKKVHDSLEDF-----PKNSSFNPAALSRHMSSL
hGATA4 --PKNLNKSKTPAAPSGSESLP---PA-SGASSNSSNATTSS------SEEMRPIKTEPGLSSHYGHSSSV
hGATA5 --PKTIAKARGSSGSTRNASAS---PSAVASTDSSAATSKAK------PSLASPVCPGPSMAP--------
hGATA6 --PKNINKSKTCSGNSNN-SIP---MTPTSTSSNSDDCSKNT------SPTTQPTASG-------------
                                        . *

hGATA1 SHLMPF---------PGPLLGSPTGSFPTGP----------MPPTTSTTVVAPLSS---------------
hGATA2 GHLPPFSHSGHILPTPTPIHP--SSSLSFGH----------PHPSSMVTAMG-------------------
hGATA3 SHISPFSHSSHMLTTPTPMHP--PSSLSFGP----------HHPSSMVTAMG-------------------
hGATA4 SQTFSVSAMSGHG---PSIHP-VLSALKLSPQG-----YASPVSQSPQTSSKQDSWNSLVLADSHGDIITA
hGATA5 ------QASGQED---DSLAP-GHLEFKFEPEDFAFPSTAPSPQAGLRGALRQEAWCALALA---------
hGATA6 ---AGAPVMTGAG---ESTNP-ENSELKYSGQDGLYIGVSLASPAEVTSSVRPDSWCALALA---------
                                  :
```

Figure 1.2: Domain conservation in the sequence of human GATA transcription factors. Multiple sequence alignment of the zinc finger region (A) and the C-terminal region (B) of human GATA factors using 'Clustal Omega'.
* (asterisk) indicates positions which have a single, fully conserved residue.
: (colon) indicates conservation between groups of strongly similar properties.
. (period) indicates conservation between groups of weakly similar properties.

in mice, GATA6$^{-/-}$ mouse embryos die at E4.5-7.5 due to inability to develop to gastrulation [30, 31].

GATA6 was first PCR-cloned from pig and rat stomach extracts (initially named GATA-GT1) along with GATA4 and GATA5 by Tamura S. et al. in Masamitsu Futai lab in Osaka University in 1993. Later in the 90s, human GATA6 cDNA encoding a 449-amino acid protein was cloned by several research groups [32–34]. Right around the same time, GATA6 cDNAs encoding similar-length proteins were cloned from other vertebrates including xenopus, chicken, and mouse [30,35,36]. The sequence of this form of GATA6, alternatively called S-type or short form GATA6 (GATA6$_S$) has a high homology to other closely related GATAs, GATA4 and GATA5, especially in the two zinc finger domains (Figure 1.2A). In 1999, Brewer et al. [37] cloned full length human and mouse GATA6 cDNA which encode a longer version of GATA6 protein. GATA6$_L$ or L-type GATA6 as first mentioned by Takeda et al. in 2004, contains a 146-amino acid long N-terminal extension which is absent in GATA6$_S$. Both forms of GATA6 protein are translated from the same mRNA. The second in frame methionine codon, Met^{147} is selected through a phenomenon called ribosomal leaky scanning which results in the production of the short form of GATA6 protein [38]. Due to early misannotations of nonhuman genomes and a lack of proper reagents, the majority of GATA6 literature has focused on GATA6$_S$ while the role of the highly conserved N-terminal extension remains unclear.

1.5 Post-translational regulation of transcription factors

Site-specific DNA-binding transcription factors function at the interface between signaling pathways and gene expression programs. These important regulators integrate external signal information relayed from signaling pathways into gene expression programs towards a particular cell fate. The activity of transcription factors (TF) can be modulated by signaling pathways through post-translational modifications (PTM) [39].

Among other modifications, phosphorylation of serine, threonine and tyrosine residues in proteins is an evolutionary conserved mechanism used by cells to positively or negatively regulate the activity of transcription factors. This reversible modification is a common way that extracellular signals are integrated into changes in gene programs that leads to appropriate cellular behaviours. Phosphorylation of a transcription factor can affect its stability, localization, protein-protein or protein-DNA interaction thus changing the function of these important regulators [40]. Identifying functional phosphorylation sites on transcription factors provides a good opportunity for perturbing these hardly druggable molecules.

1.6 GATA6 post-translational regulation

A number of unbiased proteomics studies have reported phosphorylation sites within GATA6 protein [41–46]. The majority of published studies regarding GATA6 phosphorylation has been focused on the role of MAP kinases [47–49]. Adachi et al.

reported a Ras-MEK-ERK mediated phosphorylation of GATA6 on Serine^{120} which is Serine^{266} of the full length human GATA6 protein. Phosphorylation of Serine^{266} has also been reported by a number of proteomics studies [41, 45, 46]. Ushijima et al. [48] suggested a JNK-mediated phosphorylation on GATA6 however they failed to provide any site-specific evidence. It is noteworthy to mention that the GATA6 plasmid constructs used in these studies encode the short form of GATA6 protein but not the full length protein that contains an extended N-terminal region.

1.7 Data driven modeling approaches

An ongoing challenge for systems and network biology is the integration of different levels of biological information. Here we used a novel data driven modeling approach to overcome this challenge and integrate signaling network data into gene expression data. An important feature of these two datasets is that they are structurally similar which enabled us to integrate them by data-driven statistical modeling. The interrelation of these two databases by a type of regression called “partial least square regression” modeling approach helped us to first predict the gene fluctuations from signaling information. More importantly, we were able to better grasp the complex information network in response to complex inputs and examine how the statistical model makes prediction(s). Thus, the modeling approach combined with bioinformatics tools, further enabled us to generate hypotheses that were tested experimentally by employing molecular biology techniques. The combinatorial approach led to novel discoveries regarding hidden nodes in the information network with no direct biological measurements in the input databases.

The availability of high-throughput technologies for quantifying signaling, gene expression and cellular responses has made it possible to collect large datasets on different levels of the information flow inside the cells. These big and complex datasets can be advantageous in that they provide a more holistic view of the information flow inside the cells; however they can also bring confusion. Thus, there is a need to employ approaches that simplify these datasets allowing us to grasp the data and extract biological insight.

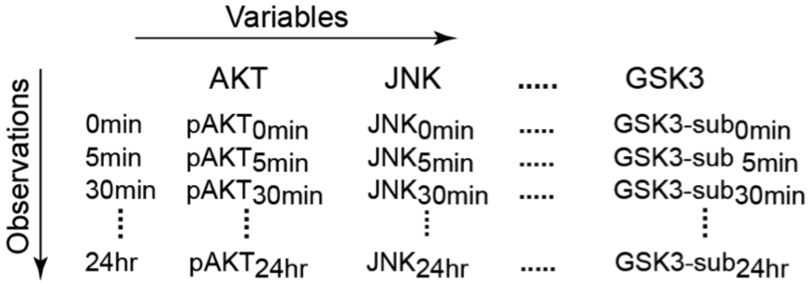
One important tool that can reduce the complexity of these big datasets is computation and modeling approaches that systems biologists are equipped with. There are diverse computational modeling approaches that can provide biological insights if carefully used depending on the biological question being asked and the structure of the data. One group of modeling approaches is “theory-driven” meaning that it is rooted in prior biological knowledge of the pathways under study. In this approach, the mathematical model specifications such as input-output relationship and model parameters are defined based on published literature and previous experimental results. The adoption of these knowledge-based approaches have provided mechanistic biological insight which are hard to achieve by pure experimental work [50].

However, even for well-studied pathways such as receptor tyrosine kinases [51], these hypothesis driven models, including ordinary differential equation (ODE) , partial differential equation (PDE) and stochastic models, quickly uncover gaps in our understanding [52]. Often, the phenomenon of interest is so poorly characterized that we only really have a sense of the pathways that are important and a rudimentary rule set for how they could interact [53, 54].

In these circumstances, it can be advantageous to pursue statistical models that do not prescribe mechanisms but allow the data to define the system of interest [55, 56]. In statistical modeling, one must first collect a systematic dataset that has been designed to capture as many relevant variations and covariations as possible among genes, proteins, and cellular phenotypes [57, 58]. Although not absolutely required, it is strongly recommended that the statistical approach be chosen conceptually before the data acquisition. Each class of models has its own set of strengths and weaknesses [55], and ideally the dataset should be tailored to exploit a model's strengths and avoid its weaknesses.

Among the different statistical modeling approaches, techniques like partial least squares are ideal for predicting new behaviors [55]. Statistical models may be "mechanism free", but it is possible to guide models toward identifying new mechanisms by selecting the right biomolecular measurements and designing the experiments appropriately [56, 59].

With current technologies in molecular biology, any laboratory can now generate datasets that are highly multivariate. Statistical modeling serves as a powerful way to extract as much information as possible from these often expensive and difficult-to-



	AKT	JNK	.....	GSK3
0min	pAKT$_{0\text{min}}$	JNK$_{0\text{min}}$	.....	GSK3-sub$_{0\text{min}}$
5min	pAKT$_{5\text{min}}$	JNK$_{5\text{min}}$	.....	GSK3-sub $_{5\text{min}}$
30min	pAKT$_{30\text{min}}$	JNK$_{30\text{min}}$	.....	GSK3-sub$_{30\text{min}}$
⋮	⋮	⋮		⋮
24hr	pAKT$_{24\text{hr}}$	JNK$_{24\text{hr}}$	.....	GSK3-sub$_{24\text{hr}}$

Figure 1.3: A data matrix of time points and three intracellular signals: v-akt murine thymoma viral oncogene homologue (AKT) activity, C-jun N-terminal kinase (JNK) activity, Glycogen synthase kinase-substrate (GSK3-sub).

conceptualize datasets. The resulting patterns and relationships identified by statistical models are not always apparent when analyzing the full spectrum of the dataset, as it often contains measurements not significant to the system. Thus, the class of statistical models that we will discuss in this chapter centers around those that build simplified representations of data to give a clearer picture of possible mechanisms underlying the system.

Usually, in modern biological datasets, we have many more variables per observation than observations of each variable. These “short and fat” data tables (or matrices) (Figure 1.7) are inherently underconstrained; in frequentist statistics, it is equivalent to having fewer than zero degrees of freedom. Consequently, many of the dimensions are redundant with one another, in that they can be expressed as linear combinations of other variables. This redundancy allows the data matrix to be “reduced” in interesting and useful ways, depending on the type of statistical model and the overall goals of the study.

Here, we will review three main categories of statistical models that reduce the dimensions of multivariate datasets. We begin with singular value decomposition (SVD), which draws on the concept of eigenvalues and eigenvectors in linear algebra to decompose a matrix according to its eigenvalue spectrum. Then, we will discuss principal components analysis (PCA), which is conceptually akin to SVD but yields a factorized model that is more directly interpretable with respect to the starting dataset. Finally, we will link reduced dimensions to the concept of predictive statistical modeling through partial least squares regression (PLSR). The statistical model that will be discussed in chapter 2 is a more modern implementation of PCA and PLSR that involves tensor decomposition of data cubes or hypercubes of structured datasets.

1.8 Singular value decomposition

Before going into detail about singular value decomposition (SVD) computation, it is important to introduce some basic concepts from vector and matrix algebra. Most datasets can be organized as matrices with the rows indicating experimental observation, such as treatments and time points and the columns indicating variables, such as enzymatic activity and phosphoprotein levels. One way to simplify multidimensional data is to focus on parts of the data that show the most variation. Linear algebra serves this purpose by finding orthogonal or linearly independent vectors in the data matrix. Since orthogonal vectors have zero projections into one another, they can act as latent variables onto which the data can be mapped [60]. Orthogonal vectors of a data matrix can be identified by calculating eigenvectors. The nonzero eigenvector (x) of matrix A satisfies equation 1.1:

$$Ax = \lambda x \tag{1.1}$$

Where A is a square matrix and λ is a scalar called “eigenvalue”. An eigenvector can serve as a new dimension along which the data can be projected. By definition, matrix A is an n × n square matrix. However, typical biological datasets have fewer observations than variables and thus are rarely square matrices with full rank. One way to solve this problem is by factorizing the data matrix using singular value decomposition.

1.8.1 SVD: mathematical framework

Suppose that we define an m × n data matrix A that can be broken down into the product of three other matrices U, S, and V. This factorization results in the following equation:

$$A_{m \times n} = U_{m \times l} S_{l \times l} V^T_{l \times n} \tag{1.2}$$

Where U is an m × l left-singular matrix, S is a square l × l diagonal matrix, V is an l × n right matrix, and U and VT are orthogonal matrices. The diagonal entries in S are the singular values of A (square roots of non-zero eigenvalues of U

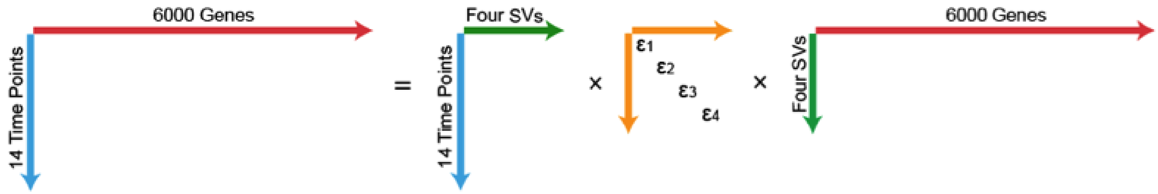


Figure 1.4: SVD Decomposition Schematic. Decomposition of the yeast elutriation data from Spellman et al. [62] into a left singular-value matrix, a square matrix of eigenvalues (four eigenvalues shown), and a right singular-value matrix.

and VT) descending in magnitude from top left to bottom right, the columns in VT are right-singular vectors and the columns in U are left-singular vectors [61]. Once singular vectors are extracted, the significant ones can be determined and used for visualizing the data.

1.8.2 Application of SVD to gene expression data analysis

Gene expression data is a good candidate for singular value decomposition based analysis due to the inherent noise in the measurements that makes the detection of small signals rather difficult. Alter et al. performed SVD analysis on the budding yeast elutriation gene microarray data [62]. The elutriation dataset used by Alter et al. contained 5,981 genes (n=5,981 genes) captured over the course of one yeast cell cycle (fourteen time points; m=14). The dataset can be tabulated to an n × m matrix with each row reflecting the expression of a single gene in 14 different time points (14-arrays) and each column showing the expression of n-genes in a single array (timepoint). SVD transforms this dataset from an n × m space to a reduced l-eigengenes × l-eigenarrays subspace where l= [min m, n] (Figure 1.4). The diagonals in the l × l matrix ε are eigenvalues here called "eigenexpression levels" $[\varepsilon_l]$ which can be used to calculate "fractions of eigenexpression" for the lth eigenvalue from the equation below:

$$p_l = \frac{\varepsilon_l^2}{\sum_{k=1}^{l} \varepsilon_k^2} \tag{1.3}$$

Alter et al. used fractions of eigenexpression as a mean to infer the significance of eigengenes and their corresponding eigenarrays (singular vectors). Once the significance of singular values (SVs) was determined, the relationship between these mathematical concepts and biological processes or cellular states, in this case cell cycle, were investigated. To this end, the authors visualized individual singular values by plotting the expression level of each eigengene over time. Since the authors were interested in gene programs involved in a specific cellular state, they filtered out the first singular vector because it followed a steady state expression pattern. The next three SVs showed biologically meaningful oscillations during cell cycle. The oscillations of the second and fourth SVs at early time points corresponded to a transient response to elutriation. Thus SVD naturally decomposed the dynamical patterns of gene expression in the yeast cell cycle.

1.9 Principal components analysis

Following SVD, principal components analysis (PCA) can be used to compress a dataset to relevant measurements that approximate the data. Both computational and visual analysis is often hard to do in higher order datasets as each measurement (observation) constitutes its own dimension in space and the value of each sample (variable) constitutes a point in each of these dimensions. By transforming the data using PCA, we can identify important relationships in the data.

First, eigenvalues and eigenvectors are derived from the data covariance matrix [63, 64]. These eigenvectors make up an orthogonal basis set, or set of linearly independent vectors that, when combined, can describe the data. The eigenvectors paired with the smallest eigenvalues are eliminated to yield a compressed basis set. This basis set of eigenvectors is then used to generate a transformed data matrix, the dimensions of which are called latent dimensions or principle components (LVs or PCs) [63–65]. A principal component is by analogy a singular vector in SVD.

A latent dimension is a new dimension created to capture the majority of information in multiple of the original dimensions [66] . Mathematically, a principle component is a linear combination of the original data dimensions, weights for which are determined by the magnitude of the eigenvector corresponding to that principle component [64, 67]. The eigenvector paired with the largest eigenvalue defines the first principal component and captures the greatest amount of variance in the data [64, 65]. In this component, the original dimensions with the most variance in variable data will have the largest weighting. Because the PCs are orthogonal, the second principal component will point in a direction perpendicular to the first component and capture the majority of the leftover variance. This iteration continues for all subsequent PCs. Thus, the transformed dataset is usually only made up of a handful of latent dimensions because they can capture the majority of the data variance eliminating any statistical noise from subsequent PCs. This filtering makes relevant relationships between samples more readily apparent.

Further, one can create predictive models with latent dimensions by searching for relationships between PCs using principal components regression (PCR). This method utilizes established regression techniques to find the relationship between several variables (predictor variables) and dependent variables not included in the predictors [67]. PCR uses the first few principal components to simplify the analysis of many variables to linear or multilinear regression between the components (predictors) and the desired measurements [64, 67]. Resulting coefficients of the PCs, fitted using least-squares approaches, can be decomposed to regression coefficients of each of the original variables in the component. The variable with the largest magnitude coefficient is the most correlated to the desired dependent variable while the sign of the coefficient indicates positive or negative correlation [67]. In this way, decomposition by PCR can be used to extract relationships between different variables in the dataset. Thus, PCA and PCR can be used not only to generate hypotheses about sample relationships but also to generate data-driven predictions. PCA is often used to analyze DNA (or cDNA) microarrays by clustering observational data such that relevant coregulations of genes or relevant similarities or disparities between cellular samples such as different cancer tumors are exposed [68–70].

Principal component analysis: mathematical framework

First, the dataset should be mean-centered so that the mean of each variable across all observations is zero. This adjustment greatly simplifies the covariance matrix calculation as well as eigenvector determination. For centering, the means of each variable (column) should be subtracted from each observation of that variable (row) in an element-wise manner as shown in Eqn (1.4).

$$\begin{bmatrix} M_{1,1} & M_{1,2} \\ \\ M_{2,1} & M_{2,2} \end{bmatrix} - \begin{bmatrix} \overline{M_1} & \overline{M_2} \\ \\ \overline{M_2} & \overline{M_2} \end{bmatrix} = \begin{bmatrix} A_{1,1} & A_{1,2} \\ \\ A_{2,1} & A_{2,2} \end{bmatrix} \tag{1.4}$$

Here M is a representative 2 × 2 data matrix and another matrix of the same size, containing the means of each sample (columns), is subtracted to generate the adjusted data matrix A. Notice that the mean of all columns should now be zero.

Now the covariance matrix of the dataset can be found from A. While in SVD the original dataset was used for decomposition, in PCA the chief interest is in the covariance of the data not the absolute magnitude [66]. Thus, the sample covariance matrix is used for decomposition as shown in Eqn (1.5).

$$C = \begin{bmatrix} cov(1,1) & K & cov(1,N) \\ \\ M & O & M \\ \\ cov(M,1) & K & cov(M,N) \end{bmatrix} = \frac{1}{N-1}\sum_{i=1}^{N}(A_i - \overline{A})(A_i - \overline{A})^T \tag{1.5}$$

which simplifies to $C = AA^T/(N-1)$

Here C is a symmetric sample covariance matrix, where the elements of the matrix are the covariances of each observation (row) with every other variable dimension, M denotes the number of observations, and N is the number of variables (columns) in A. Because A is mean-centered ($\overline{A} = 0$), this equation simplifies to $AA^T/(N-1)$. Using this notation, we can find the eigenvectors of C by decomposing it into a diagonal matrix D. [71].

We can rewrite Eqn (1.2) as,

$$C = VDV^T \tag{1.6}$$

such that the columns of V are the eigenvectors of A which correspond to the eigenvalues in the diagonal matrix D. Here, eigenvalues correspond to the contribution of that eigenvector to the reconstruction of C from the decomposition. For the covariance matrix, small eigenvalues correspond to eigenvectors that contain a small amount of the variance in the data. Thus, columns corresponding to low-magnitude eigenvalues can be eliminated from V to yield a compressed eigenvector matrix (B) that will make up the basis set of the data A [71].

Multiplying the compressed eigenvector matrix B with A transforms the adjusted data into principle component space as given by Eqn (1.7).

$$P = B^T A \tag{1.7}$$

Here P is the approximated data matrix where the rows correspond to latent dimensions or principle components and the columns correspond to samples. The elements of the matrix are the values of samples in each component. As previously mentioned, the eigenvectors are ordered from greatest corresponding eigenvalue to smallest. Therefore, the first principle component (first eigenvector) accounts for the most variance in the data. If P is composed of three or fewer principal components, the sample values can be plotted in a 2D or 3D fashion to group covarying samples.

1.10 Principal component regression (PCR) using total least squares

After the principal components have been defined, there may be instances in which knowing the relationship between principle components or principle components and an independent observation dimension are useful. Linear or planar orthogonal regression techniques can be used to determine these relationships [64, 67, 72]. In this section we focus on total least squares regression (TLSR).

As opposed to ordinary least squares regression, TLSR aims to minimize the perpendicular residual error from the regression fit [72]. This is an important distinction as it implies variance or measurement error in all the dimensions. Measurement inaccuracies create uncertainty or associated variance in the position of each data point in principal component space. Therefore, regression models should take this into account when minimizing residual error to create an unbiased fit. For PCA, all the observation dimensions used to create latent dimensions are subject to measurement error or variance [65].

First, appropriate PCs must be chosen as predictor variables. In most cases choosing the first one or two principle components is the most relevant [64, 67]. However this is not always the case and a more in-depth discussion of choosing appropriate PCs can be found in [64]. Once predictor variables have been chosen, iterative computational optimization algorithms, in environments like Matlab, can be used to identify the best-fit line or plane. In general these computational methods attempt to minimize Eqn (1.8).

$$E = \sum_{i=1}^{N} |r_i|^2 \tag{1.8}$$

where E is the residual error and r_i is the orthogonal distance of a data point (o) from the regression. The schematic Figure 2 illustrates the orthogonal distance (r_i) of a representative data point (o) from the linear regression line.

While PCA is an unsupervised decomposition method that does not take into account the inherent variance between variables, it can extract valuable information from multidimensional datasets. This information can be used to generate simplified regression models by using the principal components themselves as predictors rather than the original observations.

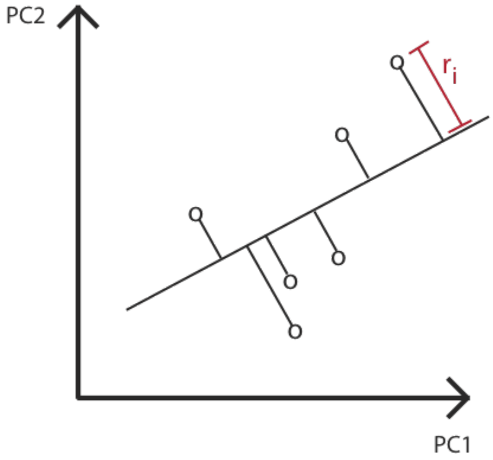


Figure 1.5: TSLR Orthogonal Residual Schematic. TLSR uses orthogonal residuals (red–r_i) to fit a regression line to data (o) displayed in PC space.

1.11 Partial least squares regression (PLSR)

As mentioned in the previous section with data matrices, PCA defines principal components that are optimized to capture the overall variance in the data matrix A. However, this does not mean that the resulting principal components are optimally interpretable, nor that they are the best regressors for predicting another data matrix. In such circumstances, it is preferred to rotate the leading principal components [60], which is easily achieved in two dimensions with the following linear operator:

$$\begin{bmatrix} cos\theta & -sin\theta \\ sin\theta & cos\theta \end{bmatrix} \tag{1.9}$$

Similar operators can be defined for rotations in three dimensions. A key point is that this "subspace rotation" does not affect the overall variance captured by the PCA model, because the solution is rotationally degenerate. Rather, it rebalances the variance among the retained principal components. Subspace rotation is commonly employed when building statistical models of biological processes [63, 73].

For statistical modeling of signal transduction, PLSR has proved widely useful and informative. Successful models have been built to link signaling to cell death [63, 74–77], cell-cycle progression [76, 78], proliferation [79–81], and cytokine secretion [74, 77, 80, 82]. More-recent theoretical work has suggested that, because of the fundamental chemical-reaction kinetics of biochemical networks, PLSR is virtually guaranteed to reduce a signaling circuit down to a handful of principal components for follow-on analysis [83]. Of course, there are caveats about framing a proper $X \rightarrow Y$ hypothesis [56], but it is reassuring to know that the approach is fundamentally sound and highly versatile. Consequently, PLSR has entered into the standard curricula for many systems-biology courses [84].

PLSR : mathematical framework

For regression modeling within high-dimensional datasets, there is a more effective way of identifying correlated principal components than PCR followed by subspace rotation. In partial least squares regression (PLSR), principal components are identified numerically that maximize the covariance between an independent data matrix (X) and a dependent data matrix (Y). (Note the distinction from PCA, which simply maximizes capture of the overall variance of a single data matrix.) Computationally, PLSR arrives at a covariance model by jointly factorizing X and Y as follows:

$$X = TP^T \tag{1.10}$$

$$Y = UQ^T \tag{1.11}$$

Where T and U are scores vectors and P and Q are loading matrices

The regression between X and Y is linear between the “scores vectors” of the independent and dependent matrices:

$$U = TB \tag{1.12}$$

Thus,

$$Y = TBQ^T \tag{1.13}$$

The simplest protocol for building a PLSR model is by using the nonlinear iterative partial least squares (NIPALS) algorithm. In this algorithm, a row from Y is randomly chosen as the first guess for a scores vector (u), and then X is projected onto u to define the first guess at a “loadings vector”, *p*. Here, the exchange of scores vectors (using u with X and t with Y) is critical for linking the two matrices together and building a PLSR model that maximizes the covariance between X and Y [85].

The first iteration of the loadings vector is then normalized and projected onto X to define a provisional t, which is subsequently projected onto Y to calculate the first iteration of its loadings vector, *q*. This loadings vector is normalized to unit length as done previously for p, and then the normalized q is projected onto Y to define the second iteration of u. This process continues until u converges to a fixed value within a specified tolerance. Software for building PLSR models is readily available in MATLAB, R, as well as independent commercial platforms [56].

1.12 Summary

Abnormal cross-communication between conflicting stimuli is the underlying cause for many common diseases. In order to study the crosstalk between pathways, one needs to get a more holistic view of cellular pathways. This view requires thousands of measurements to be made which is readily possible with the increasing availability of high-throughput technologies. The resulting big and complex datasets then need to be viewed in a simple way in order to extract biological insight. Systems biology is equipped with computational tools including the statistical methods mentioned in this chapter thus well suited to handle these emerging complex datasets. Age-old statistical techniques such as SVD, PCA and PLSR are useful tools that can reduce the dimentionality of these big dataset and make predictions from the data.

Specifically, the statistical models introduced in this chapter are among the simplest linear methods for reducing complex datasets. Simple models are more easily interpretable–by generating principal components that can be immediately mapped back onto the primary data, the models stay grounded in what they were derived from. Statistical models therefore avoid the pitfalls of machine-learning approaches, such as support-vector machines and neural networks, which can make remarkable predictions but leave the user confused about how the predictions were made [86, 87]. Moreover, the iterative methods for PCA and PLSR are very scalable to large datasets, because they do not require calculating the covariance matrix as with older implementations of SVD. Biologists are already accustomed to looking at their results–statistical models provide a more-formal way of inspecting complex data and illustrating the power of computation in real terms [88]. Just as it is difficult to imagine life now without a computer or a smartphone, biological research will soon become unfathomable without the aid of statistical models.

Chapter 2

Linking Signaling and gene expression datasets through data driven modeling and bioinformatics

Receptor-mediated signaling pathways are densely connected and exhibit non-additive behaviors when two ligands activating different receptors are applied simultaneously [3–5]. Downstream of receptor activation and signal transduction, complexity increases even further when considering the consequences of signaling on gene regulation [6–8]. Fortunately, new signaling synergy or antagonism does not typically emerge with more than two inputs [9–12], suggesting that stimulus pairs are sufficient to assess potential crosstalk. Towards systems-level discovery of crosstalk within the signal-transduction and transcription networks, a quantitative signaling data compendium as well as a condition-matched transcriptomics dataset were used to build a data-driven model. The model was created by multilinear partial least square regression (PLSR) and used to extract crosstalk hypothesis. The hypothesis was further developed by bioinformatics. In this chapter we will expand upon the crosstalk hypothesis generation process using a combination of PLSR and bioinformatics techniques.

2.1 Cytokine combinations elicit complex changes in signaling and transcript abundance

For the signaling pathway measurements, we used data from our own previous studies [58, 89], which enabled us to combine these data with the transcriptomic data generated here under the same experimental conditions. In the previous studies, HT-29 cells were exposed to IFNγ, which primes HT-29 cells to apoptose upon stimulation with TNF [90], and subsequently stimulated with saturating or subsaturating doses of TNF, EGF, or insulin alone or in combination. Lysates were profiled at 13 time points over 24 hours for 19 intracellular signaling events measured by kinase assay, immunoblot, or antibody array (Figure 2.1A and B) [91–93]. These data provided quantitative, systematically collected information on phosphorylation-mediated regulatory events, changes in protein abundance, and cleavage-dependent protein activation.

To determine how the signaling events altered gene expression, we complemented the signaling compendium with a matched set of transcriptomic profiles (Figure 2.1A). IFNγ-pretreated HT-29 cells were exposed to the same combinations of TNFα-EGF-insulin and analyzed at a subset of the time points from the previous signaling studies (Figure 2.1B). We collected transcriptomic profiles by microarray at intermediate-to-late times after cytokine simulation, thereby avoiding the bursts of immediate-early transcripts that are already well characterized [94–96]. We selected the time points for transcriptomic analysis according to earlier modeling of the signaling compendium, which showed that signaling from 4-16 hours did not predict apoptosis accurately [58]. We reasoned that the loss of predictive ability was because prolonged cytokine stimulation had transmitted the relevant information to the downstream transcriptional network.

The microarray data revealed extensive transcriptional alterations with time and stimulus condition (Figure 2.1C). Among 14,541 probe sets identified as present in at least one sample, we identified significant changes in 10,319 with time, 4948 upon TNF stimulation, 75 upon EGF stimulation, and 15 upon insulin stimulation after correction for multiple hypothesis testing [four-way analysis of variance (ANOVA), false-discovery rate = 5%]. One unanticipated complication was that many transcript abundances changed with time in the mock stimulation condition lacking TNF, EGF, and insulin (Figure 2.1C, leftmost column). We attributed these background transcriptional dynamics to the ongoing IFNγ exposure. Extensive background drifts in transcript abundance can confound interpretations from standard analyses of differentially expressed genes that focus on time-dependent changes [97, 98]. Therefore, alternative methods were required to identify meaningful changes in transcriptional regulation and link them to the upstream signaling network.

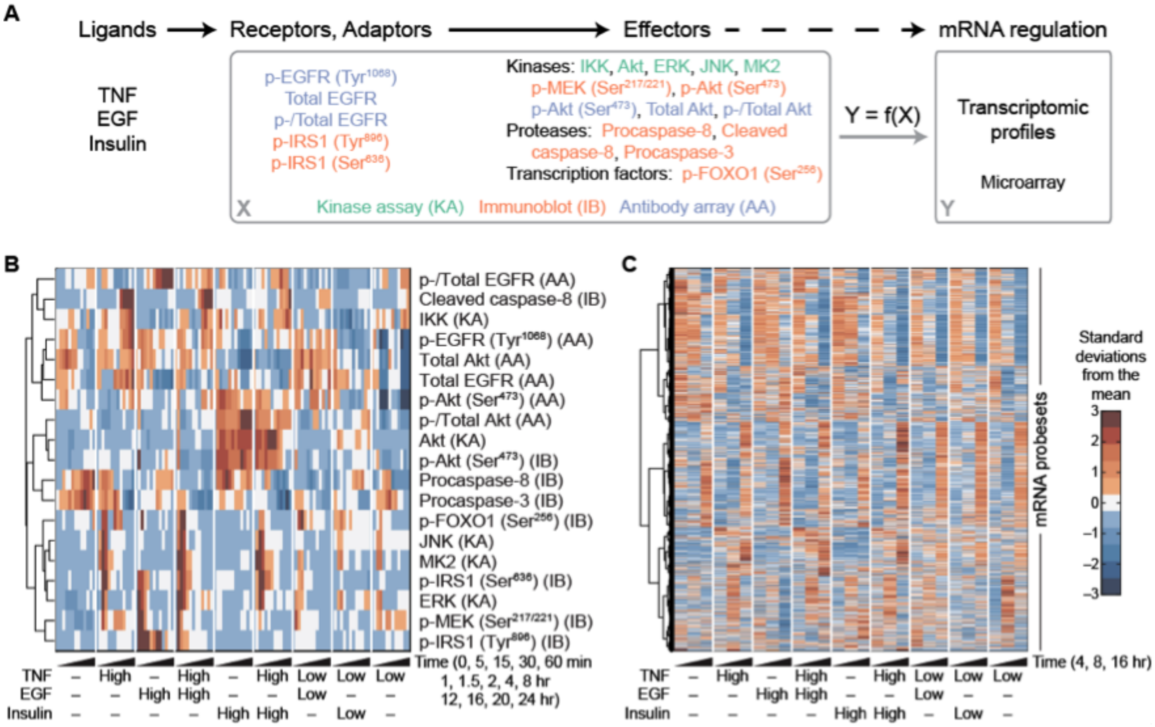


Figure 2.1: A compendium of ligand-induced signals and transcriptional responses. (A) Overview of the experimental design. HT-29 cells were pretreated with IFNγ, stimulated with various combinations and concentrations of TNF, EGF, and insulin, and profiled for the indicated signaling receptors, adaptors, and effectors by kinase assay (KA), immunoblot (IB), or antibody array (AA) and for the associated transcriptomic signatures by microarray. The goal is to determine whether global ligand-induced mRNA regulatory states (Y) can be predicted from the upstream signaling network activation ($\underline{X}$). (B) Hierarchical clustering of the signaling compendium for saturating (High) and subsaturating (Low) concentrations of TNF, EGF, and insulin [58, 89]. Data are shown as the mean of n = 3-6 independent biological replicates. (C) Hierarchical clustering of the dynamic transcriptomic responses resulting from the ligand combinations in (B). Data are shown as the mean of n = 2 independent biological replicates.

2.2 Models of dynamic, multivariate datasets are properly structured as data tensors

Systematic biology experiments monitor the same signaling events or transcripts over multiple time points and across multiple stimulus conditions [57,99]. Each data point thus contains information about the various "modes" of its acquisition; for example, which stimulus was added (Mode 1), the time after stimulus (Mode 2), and the signal or transcript measured (Mode 3). Additional modes are possible if multiple pharmacologic perturbations or cell types [76] are profiled systematically along with the modes listed above.

For systematic measurements, the acquisition modes create a data structure that is very powerful mathematically, because it conveys how different data points are related to one another. This structure vanishes when, for example, a data cube is sliced along one of its modes and "unfolded" end-to-end as a series of matrices (Figure 2.2A). When matrix-based algorithms are applied to unfolded data, each unfolded measurement variable is treated independently and Modes 2 and greater are lost. Using Figure 2-2A as an example, AKT measurements at two and four hours post-stimulation (same signal, two time points) are not handled any differently than AKT and epidermal growth factor receptor (EGFR) measurements at two hours post-stimulation (two signals, same time point). The result of unfolding is a model that is less interpretable because of too many fitted regression coefficients [100].

The alternative to unfolding is to retain datasets as cubes (three modes) or hy-

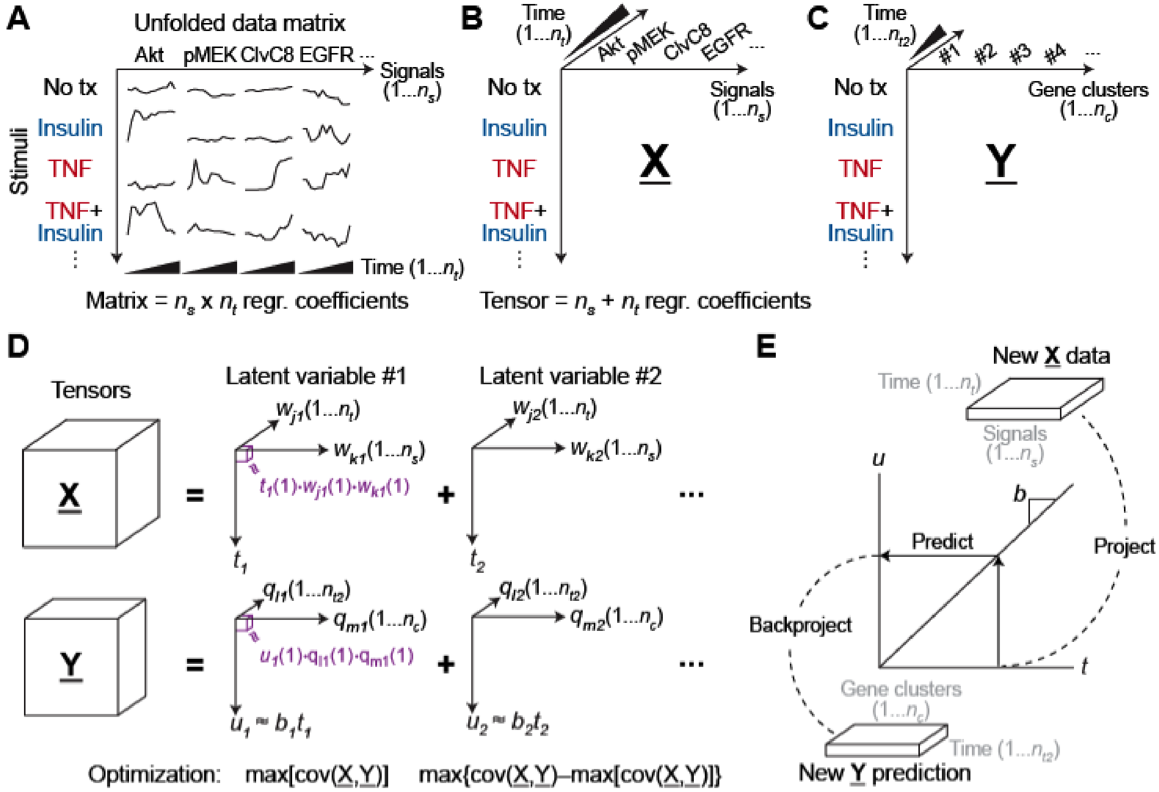


Figure 2.2: Structuring and modeling biological datasets as tensors. (A) Structured datasets are conventionally unfolded with time to create a concatenated data matrix of n_s signals and n_t time points. Using the unfolded matrix, data-driven modeling approaches [56] treat each time point of each signal as a separate predictor variable, yielding $n_s \times n_t$ regression (regr) coefficients that must be inferred. (B) Recasting stimulus-signal-time datasets as a third-order tensor. The tensor structure ($\underline{X}$) considers each time point as a predictor variable for all signals and each signal as a predictor variable for all time points, resulting in $n_s + n_t$ regression coefficients and thus a more parsimonious model. (C) A dependent third-order transcriptomic tensor ($\underline{Y}$) structured by stimulus, n_c gene clusters, and n_{t2} time points. (D) Decomposing third-order data tensors as sums of latent variables comprised of triple products. The decomposed tensor for each latent variable is reconstructed as the triple product (purple) of a scores vector (t or u) and two weight vectors (w_j and w_k or q_l and q_m). Latent variables are iteratively calculated to capture the maximum covariance between $\underline{X}$ and $\underline{Y}$ that remains from the preceding latent variable. $\underline{X}$ and $\underline{Y}$ are connected by a linear inner relationship between t and u with slope = b. (E) Prediction with tensor models involves projecting a new stimulus onto the latent variables of $\underline{X}$, predicting the dependent scores vector u from the linear inner relationship (u = bt), and then backprojecting onto the latent variables of $\underline{Y}$.

percubes (4+ modes) in the form of data “tensors”, which are the higher-dimensional generalization of vectors (one mode) and matrices (two modes). For example, the TNF-EGF-insulin signaling compendium naturally organizes as a third-order tensor defined by stimulus, time point, and measured signaling event (Figure 2.2B). The transcriptomic profiles likewise arrange as a third-order tensor according to stimulus, time point, and transcript or cluster of transcripts (Figure 2.2C). Tensors reduce the parameterization of a data-driven model, because free regression coefficients remain fixed across the other acquisition modes of each tensor (Figure 2.2B) [101, 102]. In this instance, the stimulus-time point-signaling tensor (the “regressor” tensor) is linked to the stimulus-time point-transcript tensor (the “regressand” tensor) by the regression coefficients.

Biological data tensors have been used successfully for unsupervised purposes, such as singular value decomposition [103], to analyze transcriptional kinetics during DNA replication origin firing [104] and to identify consistent copy-number changes across different array-based comparative genomic hybridization platforms [105]. Here, we sought a supervised method that could connect the signaling tensor to the transcriptomic tensor and predict gene-expression patterns from signaling-network dynamics. This application is ideal for the tensor generalization of partial least squares regression (PLSR), a matrix implementation that has been used widely to model signaling networks [56, 60, 63, 74–82, 84, 106–110].

Tensor PLSR (equivalently, “multilinear PLS” [111]) is an established method that creates a data-driven model by jointly factorizing an independent “predictor” tensor ($\underline{\mathrm{X}}$; here, the signaling tensor) and a dependent “predicted” tensor ($\underline{\mathrm{Y}}$; the transcriptomic tensor). $\underline{\mathrm{X}}$ and $\underline{\mathrm{Y}}$ are factorized as an element-by-element product of vectors, where the number of vector elements multiplied is equal to the number of dimensions in the data tensor. Thus, if $\underline{\mathrm{X}}$ is a third-order tensor, then X(1,1,1) [the tensor element in $\underline{\mathrm{X}}$ occupying the first position in Mode 1 (stimulus), the first position in Mode 2 (time point), and the first position in Mode 3 (signal)] is factorized as: $\mathrm{X}(1,1,1) = \mathrm{t}(1) \bullet w_j(1) \bullet w_k(1)$ (Figure 2.2 D, purple). In the factorization, t(1) is the first element of a “scores“ vector (t) that relates to the stimulus conditions that are shared with the $\underline{\mathrm{Y}}$ tensor. $w_j(1)$ and $w_k(1)$ are the first elements of two “weight” vectors (w_j and w_k) that relate to Modes 2 and 3 of the tensor (here, time and signal). A similar calculation is performed for $\underline{\mathrm{Y}}$ by factorizing it into its own scores (u) and weight (q_1 and q_m) vectors. $\underline{\mathrm{X}}$ and $\underline{\mathrm{Y}}$ are linked by an “inner relationship” between their respective scores vectors: $\mathrm{u} = \mathrm{bt}$, where b is a linear regression coefficient determined by the model. The inner relationship implies that how a stimulus projects on t [through the signaling (w_k) that occurs over time (w_j)] is directly proportional to its projection on u and thus how that stimulus changes gene expression (q_m) with time (q_1).

The factorization of the two tensors is posed as a numerical optimization that seeks to capture as much of the inner relationship between $\underline{\mathrm{X}}$ and $\underline{\mathrm{Y}}$ as possible. The best first set of scores and weight vectors defines the first “latent variable” of the tensor PLSR model. Residual information (covariation) in $\underline{\mathrm{X}}$ and $\underline{\mathrm{Y}}$ not captured by the first latent variable is then subjected to a second factorization, which is optimized to capture as much covariance in the residual as possible (Figure 2.2D). By repeating the algorithm, latent variables are iteratively calculated until there are no predictive inner relationships remaining between the $\underline{\mathrm{X}}$ and $\underline{\mathrm{Y}}$ data tensors [56, 63].

Predictions with a tensor PLSR model use w_j and w_k from each latent variable to project an X-like observation onto t (Figure 2.2E). Then, the $\mathrm{u} = \mathrm{bt}$ inner relationship is used to predict u, which is backprojected with q_l and q_m to yield a predicted set of values in the form of $\underline{\mathrm{Y}}$ (time-dependent gene expression). The project-predict-backproject sequence is important for making independent predictions with new data and for crossvalidation of the model to identify the optimum number of latent variables [56, 63, 85, 88].

2.3 Tensor PLSR modeling identifies predictive links between signaling and transcriptional dynamics

We first constructed a tensor PLSR model of three latent variables that predicted the 14,541 probe set fluorescence intensities of the transcriptomic dataset. Although crossvalidated predictions of the model were 99% accurate (Figure 2.3A), the model was strongly biased toward the differences in fluorescence intensities across probe sets.

Consequently, changes in probe set intensities across treatment conditions were

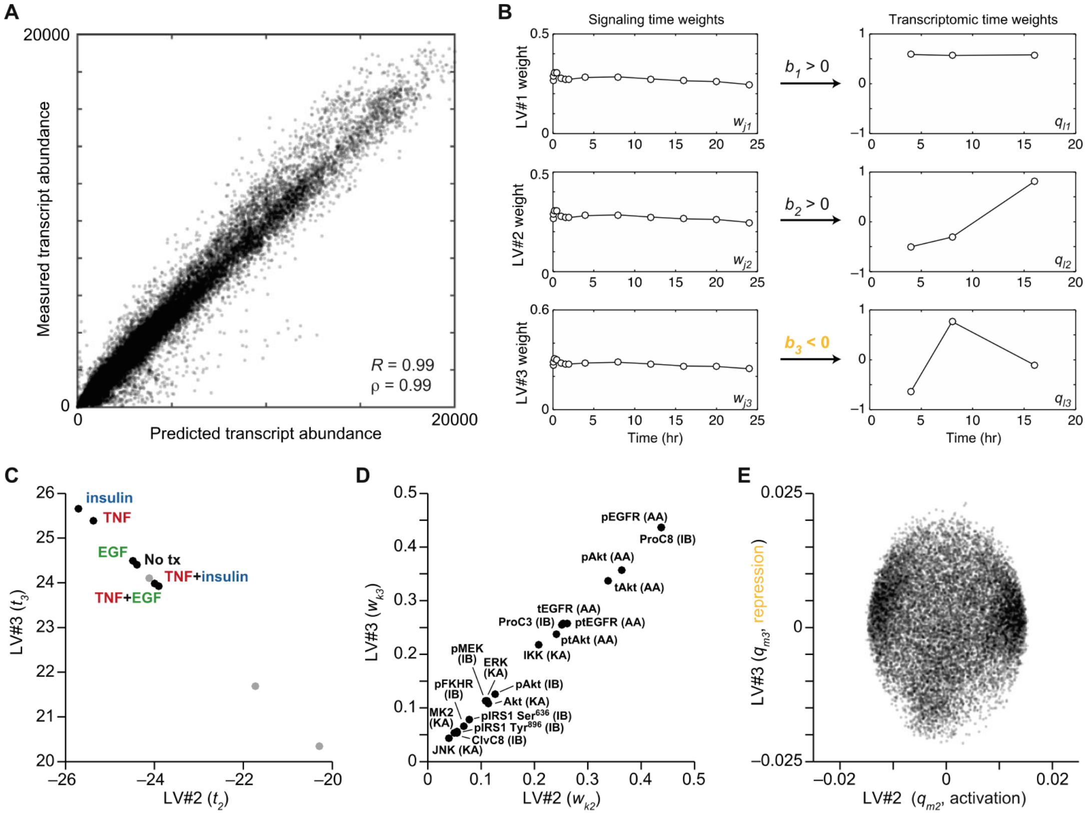


Figure 2.3: Tensor PLSR modeling predicts overall transcript abundance but cannot link changes in transcript abundance to cytokine-induced signaling. (A) Measured probeset intensities compared to crossvalidated predictions of the tensor PLSR model. Pearson (R) and Spearman (ρ) correlations are shown. (B) Latent variable (LV) time weights for the signaling and transcriptomic tensors. The third LV has a negative inner relationship (yellow) indicating that LV #3 signaling is anticorrelated with LV #3 transcription. (C to E) Projections of the indicated stimulus conditions (C), signals (D), and transcriptional clusters (E) onto the second and third LVs. Anticorrelated scores in (C) indicate a poorly posed model.

overlooked, and the resulting components of the model were uninterpretable (Figure 2.3B to E). To focus on recurrent stimulus-dependent changes in transcript abundance shared by multiple genes, we condensed the transcriptomic dataset by using the unbiased CLuster Identification via Connectivity Kernels (CLICK) algorithm [112]. Among the transcripts profiled, CLICK identified nine separable clusters comprised of dozens to hundreds of genes, the mean trajectories of which were organized as the $\underline{\mathrm{Y}}$ data tensor (Figure 2.2C). Using the entire signaling compendium as $\underline{\mathrm{X}}$, we constructed a tensor PLSR model of four latent variables that predicted gene-cluster dynamics to within 72% (Figure 2.4A and Figure 2.5A). Although the model did not predict certain cytokine-induced changes for some gene clusters (Figure 2.4A, see EGF and insulin stimuli of Cluster #3), we considered the overall accuracy of predictions remarkable considering that the model involved 10-fold fewer parameters than previous PLSR models of TNF-induced apoptosis [58, 63, 107].

Our principal motivation for building the tensor PLSR model was to use the model to reveal undiscovered mechanisms of how signaling alters gene expression. To identify which latent variables captured both signaling and gene-cluster dynamics, we analyzed the time weights (w_j and q_l) and inner regression coefficients for $\underline{\mathrm{X}}$ and $\underline{\mathrm{Y}}$ (Figure 2.4B). The leading two latent variables (LV#1 and LV#2) harbored signaling time weights (w_{j1} and w_{j2}) that were nearly constant from 0-24 hours, indicating that time-dependent changes in signaling did not determine the projection of $\underline{\mathrm{X}}$ along LV#1 or LV#2. Accordingly, time weights for the gene clusters (q_{l1} and q_{l2}) were time

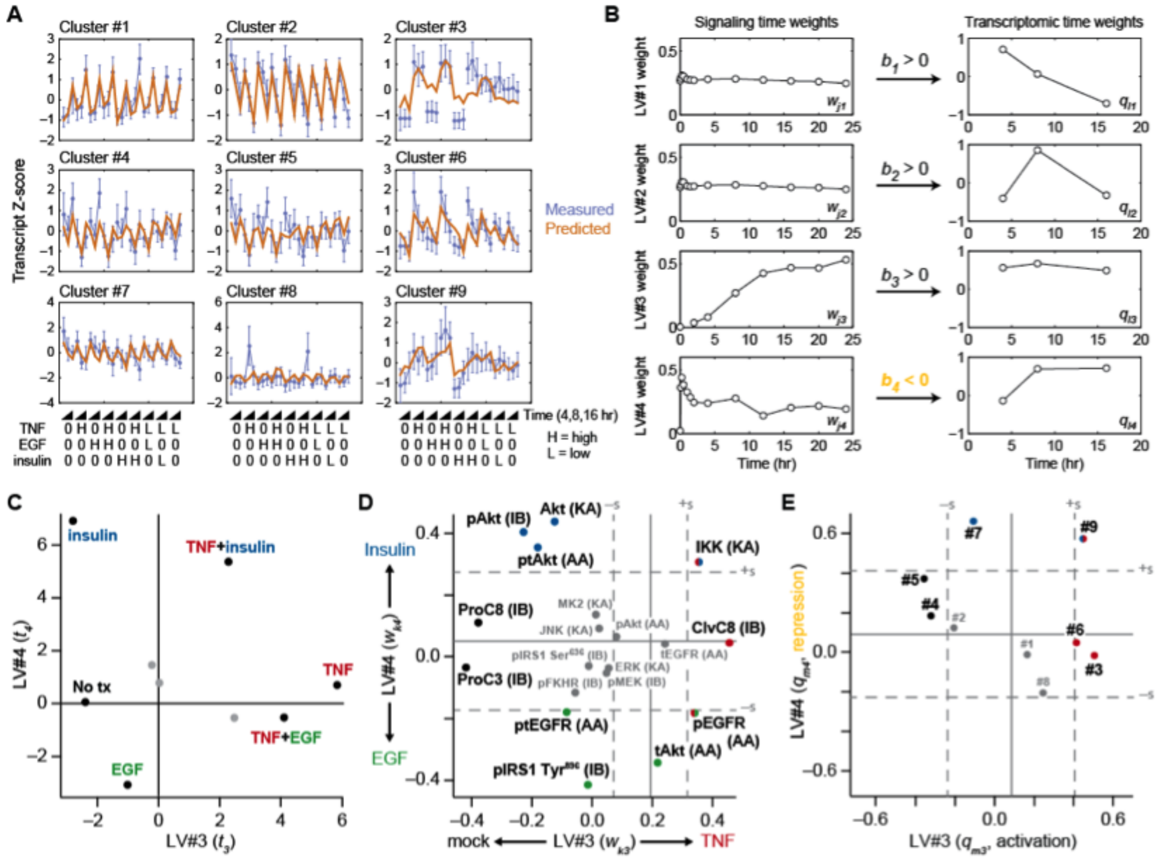


Figure 2.4: A tensor PLSR model linking ligand-induced signaling and changes in transcript abundance. (A) Time-unfolded measurements of transcriptional clusters (blue) compared to crossvalidated predictions of the tensor PLSR model (brown). Standardized Z-scores of measured transcriptional clusters are shown as the mean SD of n = 897 (#1), 841 (#2), 119 (#3), 106 (#4), 66 (#5), 49 (#6), 42 (#7), 33 (#8), and 26 (#9) probe sets. High (H) indicates saturating concentration of ligand, 0 indicates absence of ligand, and low (L) indicates subsaturating concentration of ligand. (B) Latent variable (LV) time weights for the signaling and transcriptomic tensors. The fourth LV has a negative inner relationship (orange), indicating that LV#4 signaling is anticorrelated with LV#4 transcription. (C to E) Projections of the indicated stimulus conditions (C), signals (D), and transcriptional clusters (E) onto the third and fourth LVs. For (D) and (E), the null projections of reshuffled data tensors are shown as the mean (solid gray) SD (dashed gray) of n = 500 randomizations [113]. In D, the type of assay used to measure the signaling protein is indicated in parentheses (see Figure 2.1A for details). ClvC8, cleaved caspase 8; ProC3, procaspase 3; ProC8, procaspase 8; Lower case p prefix represents phosphorylated protein; lowercase t prefix represents total protein; lowercase pt prefix represents the ratio of phosphorylated protein to total protein.

variant and derived from the stimulus-independent transcriptional changes of Clusters #1, #2, and #7 (Figure 2.4, A and B), presumably resulting from IFNγ pretreatment. Because latent variables are calculated iteratively (Figure 2.2D), LV#1 and LV#2 eliminated the TNF-, EGF-, and insulin-independent transcriptional changes, revealing paired signaling and gene-cluster dynamics in the third and fourth latent variables (LV#3 and LV#4). LV#3 harbored time weights of late-phase signaling (w_{k3}) and sustained transcriptional activation (q_{l3}). Conversely, LV#4 was weighted with early-phase signaling (w_{k4}) and late-phase transcriptional regulation (q_{l4}). The inner regression coefficient for this fourth latent variable was negative (Figure 2.4, orange), implying a link between early-phase signaling and downstream transcriptional repression.

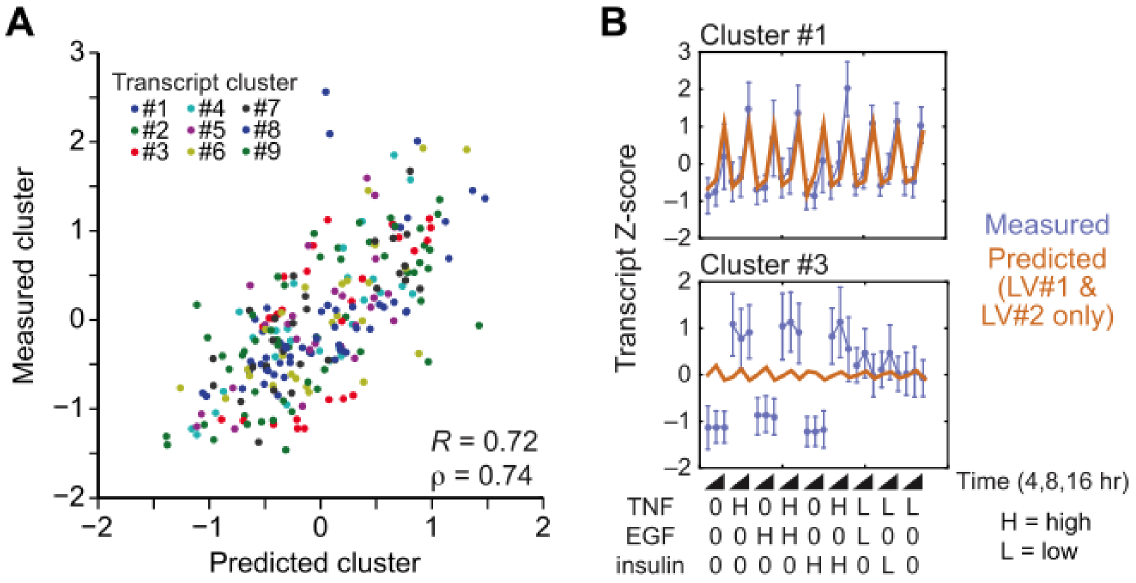
Focusing on LV#3 and LV#4, we evaluated the relationship between the treatment scores (t3 and t4; (Figure 2.4C). Relative to mock treatment, saturating TNF stimulation projected almost entirely along LV#3, suggesting that LV#3 represented a TNF “axis“. In contrast to TNF, we found that EGF and insulin projected in opposite directions along LV#4, indicating that this latent variable distinguished between the two growth-factor stimuli. Combinatorial stimulations exhibited intermediate scores that approximately averaged the scores of the individual stimuli. For example, TNF+EGF projected positively along LV#3 (like TNF) and negatively along LV#4 (like EGF). The interpolated response observed here for gene regulation contrasts with prior work on apoptosis in which EGF and insulin each nonlinearly antagonized

TNF-induced cell death [63].

To connect specific signals and gene clusters with the prevalent cytokine-induced dynamics, we evaluated the signaling and gene-cluster weights along LV#3 (TNF, late-phase signaling axis: w_{k3} and qm3) and LV#4 (EGF-insulin, early-phase signaling axis: wj4 and qm4) (Figure 2.4C, D and E). Multiple signals–such as Ser636-phosphorylated insulin receptor substrate-1 [pIRS1 (Ser636)], c-jun N-terminal kinase (JNK) activity, and mitogen-activated protein kinase-activated protein kinase-2 (MK2) activity–were negligibly weighted (Figure 2.4D), implying that these early-phase TNF-induced signals (Figure 2.1B) were statistically uninformative for predicting $\underline{\mathrm{Y}}$ (the transcriptional response). Gene cluster #1 was also unweighted along the third and fourth latent variables, because its dynamics were almost entirely captured by the first and second latent variables (Figure 2.4A, B, and E and Figure 2.5B).

To filter the weight vectors further, we randomly shuffled the signaling, gene-cluster, and time information within each cytokine stimulation (Mode 1 slice) of $\underline{\mathrm{X}}$ and $\underline{\mathrm{Y}}$ [113]. With hundreds of shuffled tensor PLSR models, we estimated a null projection for the weight vectors of LV#3 and LV#4 (Figure 2.4D and E, gray line). We considered signals and gene clusters outside one standard deviation (σ, gray dashed lines) of the null projection as weighted strongly enough to warrant further analysis (Figure 2.4D and E).

Among the strongest signaling weights, we found clear agreement with known mechanisms of signal transduction (Figure 2.4D). For example, cleavage of apoptotic



C

Cluster 3&6 promoter analysis

TF binding site	Occurrence (rank)	Importance (rank)
NFKAPPAB	25% (1)	0.44 (1)
NFKAPPAB65	23% (2)	0.30 (3)
NFKAPPAB50	20% (3)	0.24 (4)
CETS1P54	19% (4)	0.05 (23)
AP1	18% (5)	0.09 (13)
ELK1	17% (6)	0.001 (109)
GATA1	16% (7)	0.04 (35)
MAF	16% (7)	0.09 (14)
HMG1Y	16% (7)	0.13 (12)
OCT1	15% (10)	0.05 (24)

Figure 2.5: Accuracy of tensor PLSR predictions. (A) Measured CLICK cluster dynamics plotted versus the leave-one-out crossvalidated predictions of the tensor PLSR model. Pearson (R) and Spearman (ρ) correlations are shown. (B) The first and second latent variables (LV#1 and LV#2) are sufficient to predict gene cluster #1 but not cluster #3. Time-unfolded measurements of transcriptional clusters (blue) are compared to crossvalidated predictions of a tensor PLSR model comprised of LV#1 and LV#2 only (brown). (C) DiRE promoter analysis [21] of the transcripts in Clusters #3 and #6 that map strongly to TNF stimulation in the tensor PLSR model. NF-κB subunits are highlighted in red.

caspases [negative weighting for procaspase-8 (ProC8) and procaspase-3 (ProC3) and positive weighting of cleaved caspase-8 (ClvC8)] was strongly aligned along LV#3, which is consistent with late-phase caspase activation triggered by TNF [90, 114]. Along LV#4, insulin stimulation coincided with positive weights for three complementary measures of AKT activation, a recognized effector pathway [115]. Likewise, multiple measures of EGFR phosphorylation were weighted in a direction that corresponded to EGF stimulation. Also strongly associated with EGF signaling was phosphorylated insulin receptor substrate-1 [pIRS1 (Tyr^{896})], consistent with reports that this site may be directly phosphorylated by active EGFR [58, 116]. Not all signaling events were associated with individual stimuli. For instance, the weights associated with inhibitor of nuclear factor-κB kinase (IKK) activation mapped not only to TNF but also to insulin stimulation, possibly because AKT signaling can activate IKK in certain contexts [117, 118]. Similarly, phosphorylated EGFR [pEGFR (Tyr^{1068})] projected with early-phase EGF and also late-phase TNF signaling. The latter is probably due to autocrine signaling by transforming growth factor-β, an EGFR ligand that is released after TNF stimulation [89, 119]. Together, the weights of LV#3 and LV#4 provided a condensed map of the signaling compendium that was optimized for predicting the observed transcriptomic profiles.

The inner regression coefficient (b_3) connecting $\underline{\mathrm{X}}$ and $\underline{\mathrm{Y}}$ along LV#3 was a positive value, indicating gene activation; whereas the inner regression coefficient b4 was negative, indicating that signaling along LV#4 resulted in gene repression (Figure 2.4B and E). Contrary to that of the signaling compendium, the projection of gene clusters along LV#3 and LV#4 was surprising (Figure 2.4E). Amidst thousands of time-dependent transcriptional changes, few clusters were weighted toward specific stimuli. Clusters #3 and #6 were primarily weighted along LV#3, indicating an association with TNF stimulation. Accordingly, promoter analysis [120] of the transcripts in these two clusters revealed a strong overrepresentation of binding sites for nuclear factor-κB (NF-κB) (Figure 2.5C). Cluster #7 mapped along LV#4 because of the mild suppression of transcripts observed with saturating insulin alone (Figure 2.4A and E). Only Cluster #9 projected strongly along both latent variables, indicating that the transcripts in this cluster were induced by TNF and repressed by insulin (Figure 2.6). TNF antagonism of insulin function has been well documented in adipocytes [20, 21], but there are few reports of insulin antagonizing TNF [121]. Given this predicted TNF and insulin "crosstalk cluster", we used the tensor PLSR model to investigate its mechanism of regulation by the upstream signaling network.

With respect to its latent-variable projections, Cluster #9 was cartographically most similar to IKK (Figure 2.5D and E). If these shared projections were indicative of mechanism, however, it would imply that early-phase IKK activity (downstream of TNF and insulin signaling) represses transcription of the crosstalk cluster, whereas late-phase IKK (downstream of TNF signaling) promotes it. Repress-then-activate kinetics are opposite of the prevailing view of IKK signaling [122]. Accordingly, we found that TNF-induced responses of 85% of transcripts in Cluster #9 were not

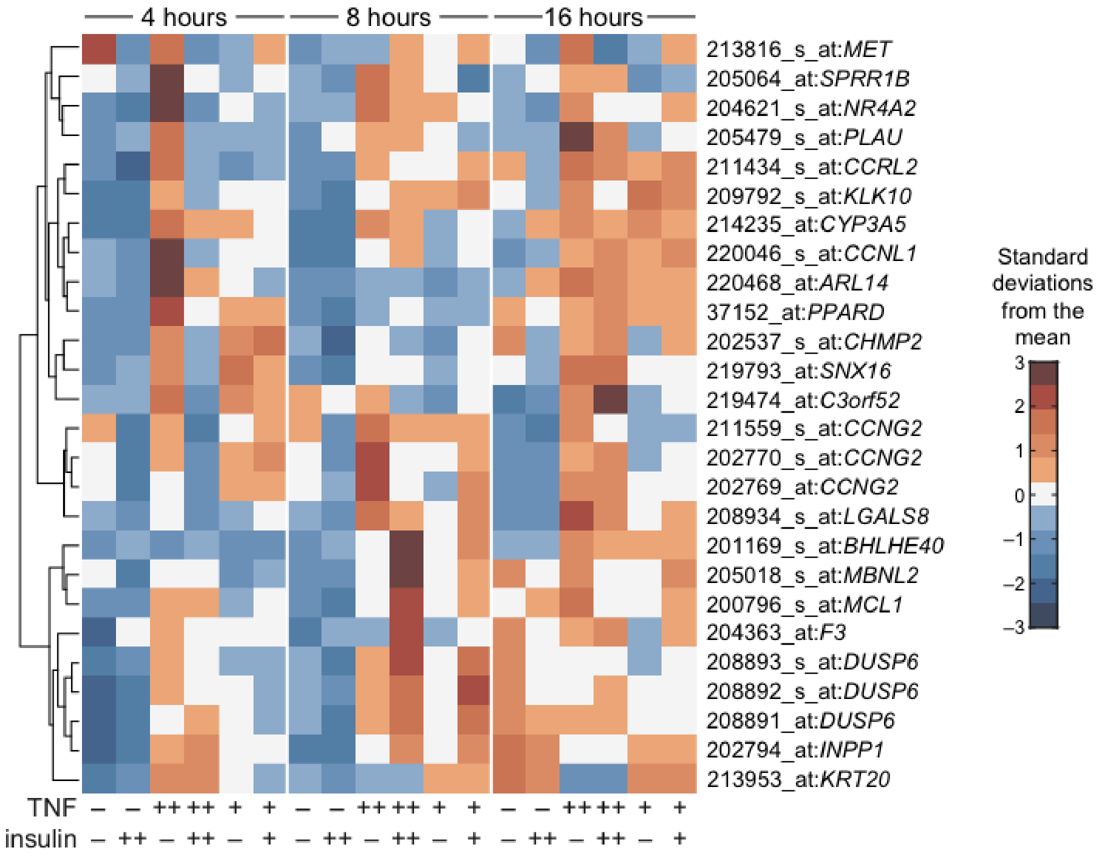


Figure 2.6: Induction of Cluster #9 probesets by TNF and repression by insulin. Hierarchical clustering of probesets from Figure 2.1C corresponding to CLICK Cluster #9 (Figure 2.4B) and their response to saturating (++) or subsaturating (+) concentrations of TNF or insulin for the indicated times. Data are shown as the standardized mean probeset fluorescence of n = 2 independent biological replicates.

significantly affected when a phosphorylation- and degradation-resistant mutant of IκBα was ectopically expressed in HT-29 cells (log-transformed Welch's t test, false-discovery rate = 15%; (Figure 2.7). We therefore considered alternatives that were consistent with the tensor PLSR model.

One possible explanation was that the crosstalk cluster integrated two distinct signaling inputs. An activating input could arise from a TNF-specific signal that was either not measured or not projected strongly on the third and fourth latent variables. In parallel, the cluster could be transcriptionally inhibited by an insulin-specific signal, such as AKT (Figure 2.4D) or a downstream effector pathway of AKT [125].

2.4 Promoter and signaling bioinformatics suggest a link between GSK3 and the crosstalk cluster through GATA6

First, we determined the reliability and generality of TNF-insulin crosstalk among transcripts in Cluster #9. We repeated the stimulation experiments with an independently obtained vial of HT-29 cells and assayed individual transcripts by quantitative reverse transcription polymerase chain reaction (qRT-PCR). This analysis confirmed the expression of over 90% of the 22 Cluster #9 transcripts (SPRR1B and PPARD from Figure 2.6 were false positives), and we detected an antagonistic interaction between TNF and insulin from 2-8 hours after stimulation (Figure 2.8 and Figure 2.11A and B). At individual time points for specific genes, we observed

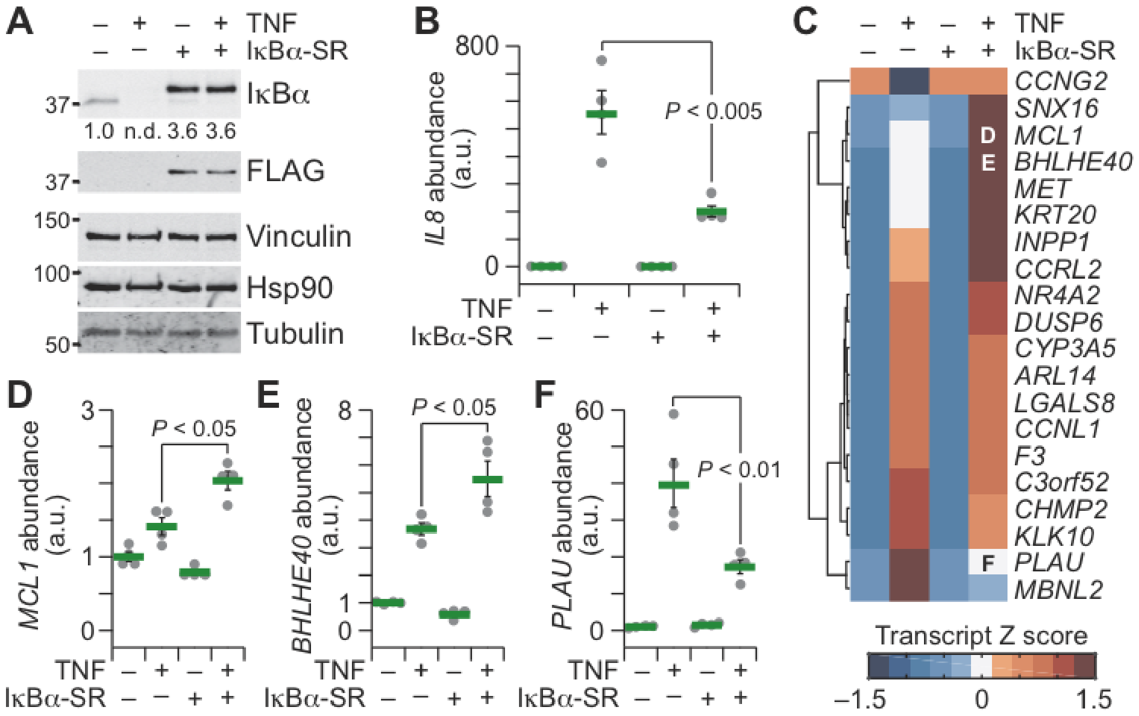


Figure 2.7: Disruption of NF-κB signaling does not widely affect the TNF-induced transcriptional response of Cluster #9. (A) HT-29 cells stably expressing FLAG-tagged IκBα super-repressor (IκBα-SR) or pBabe puro control were stimulated with 100 ng/ml TNF for 15 minutes and immunoblotted for total IκBα and FLAG with vinculin, Hsp90, and tubulin used as loading controls. Data are representative of n = 2 stable cell lines. (B) Confirmation of IκBα-SR perturbation of qRT-PCR analysis of the classic NF-κB target gene, IL8 [123]. (C) Hierarchical clustering of qRT-PCR measurements of Cluster #9 transcripts in control or IκBα-SR cells stimulated with 100 ng/ml TNF for two hours. Significant perturbations in TNF response were assessed by Welch's two-sided t test after log transformation (FDR = 15%). (D to F) Expanded view of the significant perturbations in MCL1 (D), BHLHE40 (E), and PLAU (F), a recognized NF-κB-dependent transcript [124]. qRT-PCR data are shown as the geometric mean ± log-transformed SEM of n = 4 biological replicates, with changes in geometric means assessed by Welch's two-sided t test.

instances of antagonism represented by significant interaction P value ($P_{int} < 0.05$, two-way ANOVA; (Figure 2.11A and B), an indication that TNF and insulin have nonadditive effects on the expression of those genes. We also observed nonlinear significant differences in gene expression for other transcripts even when the P_{int} was not significantly different (Figure 2.11C and D). The qRT-PCR data thus confirmed the microarray results and the tensor PLSR model, showing an early-phase suppression of TNF-induced Cluster #9 genes by insulin (Figure 2.4B to E). Furthermore, these data indicated that the TNF-insulin crosstalk cannot be predicted by adding the effect of insulin to the TNF response.

To identify candidate mediators of TNF-insulin crosstalk, we analyzed the expression-verified transcripts of Cluster #9 with three orthogonal promoter-analysis algorithms [119,120,126]. Only two transcription factors were suggested as candidate regulators by all three algorithms: T-cell factor 4 (TCF4) and GATA (Figure 2.11E). HT-29 cells harbor a truncating mutation in APC, a protein that inhibits the β-catenin pathway, and this truncation renders β-catenin and its transcriptional partner TCF4 constitutively active [127]. Moreover, no changes in β-catenin localization were observed upon TNF simulation with or without insulin (Figure 2.10). Therefore, we focused on GATA, a family of six transcription factors that are important for development and differentiation [128].

Using qRT-PCR [129,130], we quantified the relative copy numbers of the GATA family and found that GATA6 was the most abundant isoform (Figure 2.11 and

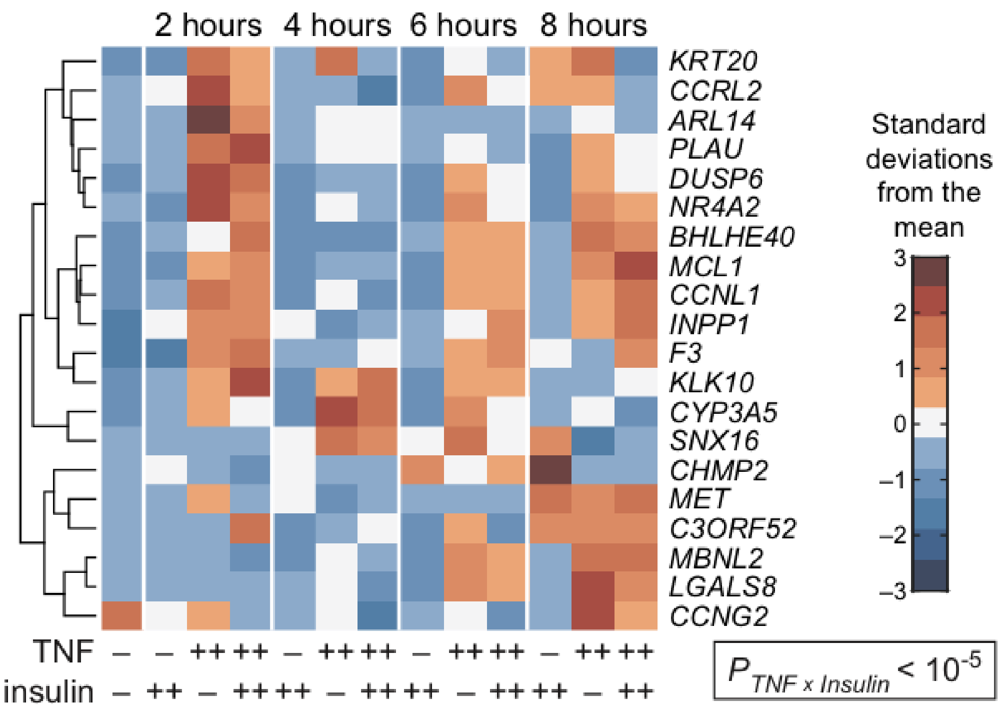


Figure 2.8: Widespread TNF-insulin crosstalk among genes in transcriptional Cluster #9. qRT-PCR validation of the Cluster #9 transcripts in HT-29 cells pretreated with 200 U/ml IFNγ for 24 hours and stimulated with 100 ng/ml TNF with or without 500 ng/ml insulin for the indicated times. Data are shown as the geometric mean n = 3-16 biological replicates, with the interaction between TNF and insulin assessed by log-transformed four-way ANOVA with the following factors: transcript, TNF, insulin, and time.

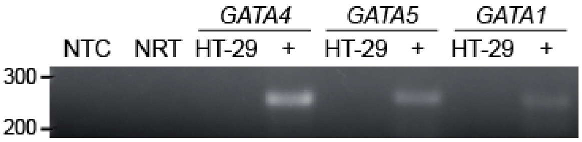


Figure 2.9: HT-29 cells lack GATA1, GATA4, and GATA5. Presence of the indicated GATA isoforms was assessed by RT-PCR followed by agarose gel electrophoresis. 293T cells (GATA4, GATA5) and K562 cells (GATA1) were used as positive controls (+) for detection. Data are representative of n = 2 agarose gels and n = 5 biological replicates. NTC, representative no template control. NRT, representative no reverse transcription control.

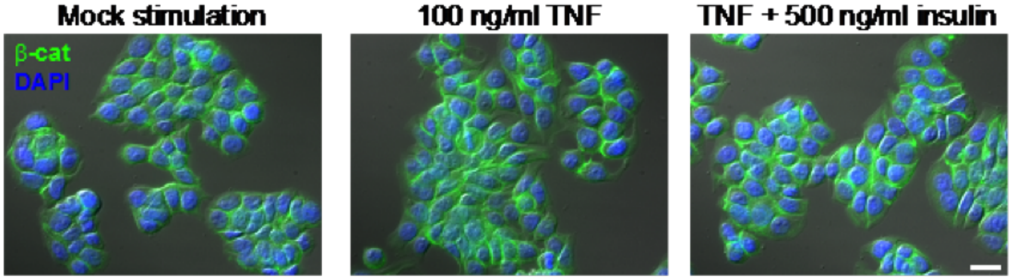


Figure 2.10: Immunolocalization of β-catenin is not altered by TNF stimulation or insulin costimulation of HT-29 cells. Cells were sensitized with 200 U/ml interferon-γ for 24 hours and then treated with 100 ng/ml TNF ± 500 ng/ml insulin for one hour, fixed, and immunostained for β-catenin (green). Nuclei were counterstained with DAPI (blue) before imaging by widefield immunofluorescence. Data are representative of three exposures from n = 2 biological replicates. Fluorescence channels are shown overlaid on the corresponding differential inference contrast image (gray). Scale bar is 20 μm.

Figure 2.9). Copies of GATA6 transcript also remained high during the early phase of TNF-only and TNF + insulin stimulation, whereas GATA2 and GATA3 were reduced two- to fourfold (Figure 2.11G to I). Notably, bioinformatic analysis [131] of the full-length GATA6 protein sequence uncovered a cluster of highly conserved serine residues that were candidate phosphorylation sites for GSK3 (Figure 2.11J). Because GSK3 is a recognized substrate of AKT [115], these results suggested that GATA6 could be an insulin-dependent regulator of the crosstalk cluster.

Full-length GATA6 (GATA6$_L$) is distinguished by a long N-terminal extension that is conserved across vertebrates but missing in other GATA family members (Figure 2.12). In addition, leaky ribosome scanning [38] onto an in-frame methionine (Met^{147}) gives rise to a short-form of GATA6 (GATA6$_S$) that is comparable in size (45 kD) to the other GATA isoforms. GATA6$_S$ lacks the candidate GSK3 phospho-

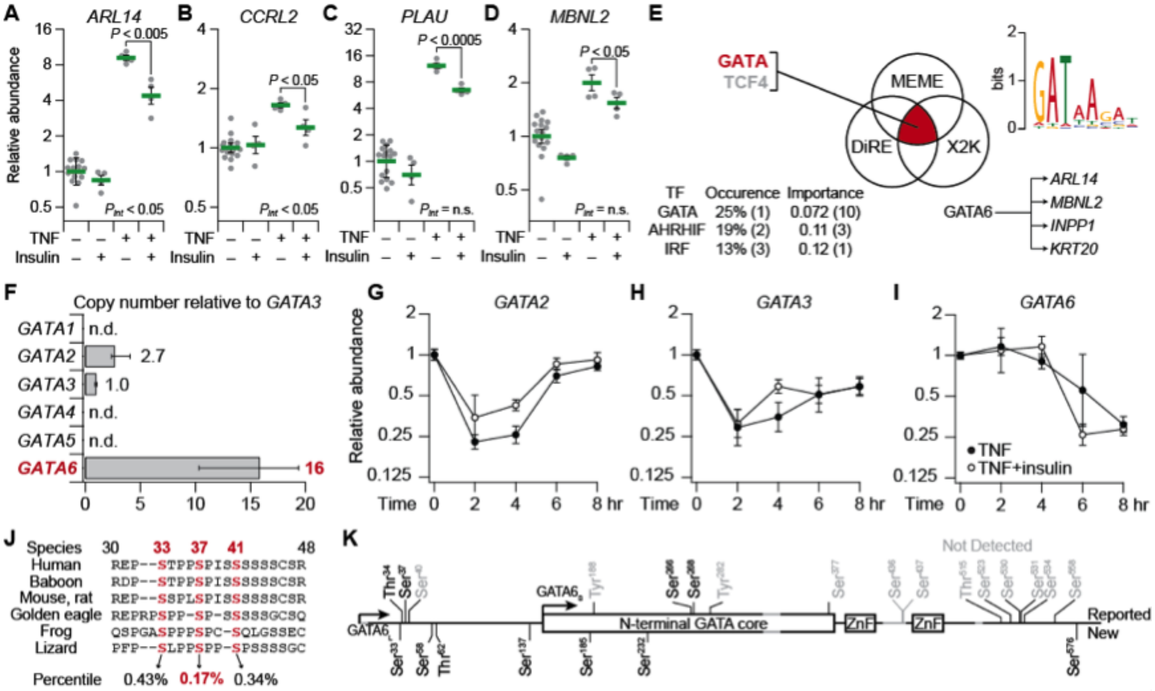


Figure 2.11: Multipronged bioinformatics of TNF-insulin crosstalk suggests post-translational regulation from GSK3 to GATA6. (A to D) qRT-PCR validation of selected Cluster #9 transcripts upon pretreatment of HT-29 cells with IFNγ and stimulation with TNF with or without insulin for two (A and B) or six (C and D) hours. Data are shown as the geometric mean ± log-transformed SEM of n = 4 or 16 biological replicates. Full Cluster #9 data are shown in Figure 2.8. (E) Promoter bioinformatics [120,126,132] suggest GATA and TCF4 as candidate regulators of TNF-insulin crosstalk. (F) Relative copy-number estimates [129, 130] for the six GATA isoforms in HT-29 cells. Data are shown as the median ± range of n = 3 biological replicates. n.d., not detected. (G to I) Transcriptional dynamics of GATA isoforms in response to TNF and insulin. Data are shown as the geometric mean ± log-transformed SEM of n = 4 or 8 biological replicates. (J) Scansite [131] identification of candidate GSK3 phosphorylation sites (red). Each site's percentile rank is averaged across the indicated sequences. (K) Phospho-mass spectrometry identifies 11 phosphorylation sites on GATA6$_L$. Previously unreported sites (New) are shown below the primary sequence, those consistent with reports in the literature [133] (Reported) are shown above, and reported sites not detected in this study are gray. Start methionines (arrows) for the long and short forms are indicated along with the conserved GATA core and zinc finger (ZnF) domains.

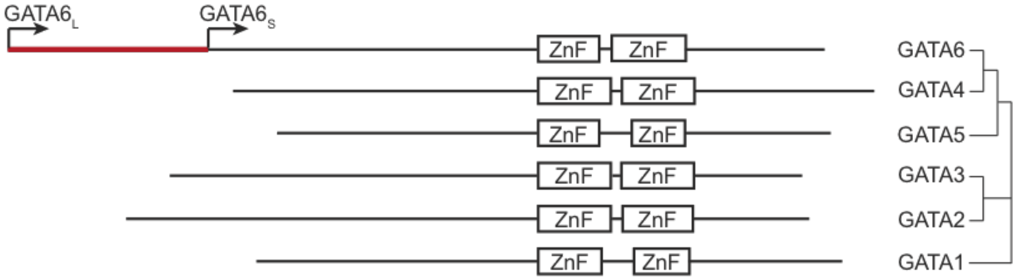


Figure 2.12: Phylogeny of the human GATA family. GATA isoforms are shown organized by their sequence similarity and aligned according to the first zinc finger (ZnF) domain. The N-terminal extension that distinguishes GATA6$_L$ from GATA6$_S$ and other family members is highlighted in red.

rylation sites that reside at the N-terminus of GATA6$_L$ (Figure 2.11J), raising the possibility of selective regulation of GATA6$_L$ by insulin.

Compared to GATA6$_S$, GATA6$_L$ has been understudied due to early misannotations of nonhuman genomes and a lack of suitable reagents. Common plasmid repositories-including Addgene [134], the human ORFeome [135], and the Mammalian Gene Collection [136]-possess only the short form or lack the gene entirely. The unusually slow electrophoretic mobility of GATA6 has created additional confusion, because commercial antibody vendors mistakenly label GATA6$_S$ as "GATA6", implying that the detected protein is the full-length form. As a result, the GATA6 literature is incredibly ambiguous, with many papers inadvertently focusing on GATA6$_S$.

To determine whether the predicted GSK3 phosphorylation sites of GATA6$_L$ could be phosphorylated in HT-29 cells, we cloned the full-length gene with N-terminal FLAG and C-terminal AU1 tags into a doxycycline (DOX)-inducible lentivector [137]. Stable HT-29 transductants were induced with DOX for 24 hours before lysis, and

FLAG immunoprecipitates were subjected to phosphorylation analysis by mass spectrometry. We achieved 92% coverage of the GATA6$_L$ sequence and identified 11 phosphorylation sites, including seven that had not been previously reported (Figure 2.11K) [133]. Among sites in the N-terminal extension specific to GATA6$_L$, two (SER^{33} and Ser^{37}) were consistent with the bioinformatic predictions of GSK3 phosphorylation (Figure 2.11J). Two N-terminal sites (Thr^{34} and Ser^{37}) were also corroborated by a proteomics study of proline-directed phosphorylation in 293T cells [42]. We concluded that SER^{33}, Thr^{34}, and Ser^{37} were the leading candidates for GATA6$_L$ phosphorylation-mediated regulation by GSK3.

2.5 Summary

With the advent of high throughput technologies, it is becoming increasingly easy for many life scientists to collect big data sets. The next big challenge is to understand these complex data and use them to make new discoveries. Systems biologists are equipped with the arsenal of computation to tackle this problem. Here we used tensor partial least square regression modeling to integrate signaling and transcriptomics data sets collected in human colonic epithelial cells. Model analysis together with bioinformatics predicted a link between Akt-glycogen synthase kinase 3 (GSK3) signaling and the inflammatory cytokine, tumor necrosis factor alpha via the endodermal transcription factor, GATA6. In addition, phosphorylation mediated regulation of GATA6 in the context of TNF and insulin can be a way that the signal from the two input ligands is integrated. Furthermore, identifying such type of crosstalk node that link a transcription factor to an upstream enzyme e.g. kinase introduces alternative drug targets for the diseases caused by these hardly druggable targets.

Chapter 3

GSK3-dependent phosphorylation of GATA6 and its effect on GATA6$_L$ turnover and downstream gene expression

Mathematical modeling approaches such as ”tensor PLSR” described in chapter 2 are quite useful in handling the ever increasing complex data in biology. A combination of modeling and bioinformatics led us to hypothesize that phosphorylation of endodermal transcription factor GATA6 within a conserved N-terminal serine stretch is a way that insulin signaling cross-communicates with TNF-induced gene expression. Since the output of a model is not more than a prediction, in this chapter we performed wet lab experiments in order to test a mechanistic crosstalk hypothesis generated via statistical modeling and bioinformatics in chapter 2.

3.1 Perturbation of basophilic kinases differentially affects GATA6$_L$ and GATA6$_S$

Studies of GATA6 phosphorylation have largely focused on its posttranslational regulation by mitogen-activated protein kinases (MAPKs) [47–49]. However, GATA6 is reportedly phosphorylated on Ser^{436} (Ser^{290} in GATA6$_S$) and stabilized upon prolonged mechanistic target of rapamycin complex 1 (mTORC1) inhibition with rapamycin and subsequent feedback activation of AKT2 in vascular smooth muscle cells (VSMCs) [138]. Our mass spectrometry data on GATA6$_L$ did not include peptides containing Ser^{436}; thus, we could not rule out a role for Akt as a GATA6 kinase activated by insulin stimulation.

To determine whether AKT-catalyzed stabilization of GATA6 was relevant to our study, we treated HT-29 cells with the mTORC1 inhibitor rapamycin for three hours and monitored targets by quantitative immunoblotting [93]. As expected, rapamycin eliminated phosphorylation of ribosomal protein S6 and increased phosphorylation of AKT by $\backsim$twofold; however, the abundances of GATA6$_S$ and GATA6$_L$ were essentially unchanged (Figure 3.1).

In the context of pretreatment and TNF-insulin stimulation, we found that rapamycin treatment for three hours altered the distribution of GATA6 forms by decreasing the abundance of GATA6$_L$ relative to the abundance of GATA6$_S$ (Figure 3.2). The mechanism reported in VSMCs [138] might be restricted to mesodermal tissues, so we repeated the experiment in AC16 ventricular cardiomyocytes [135]. In

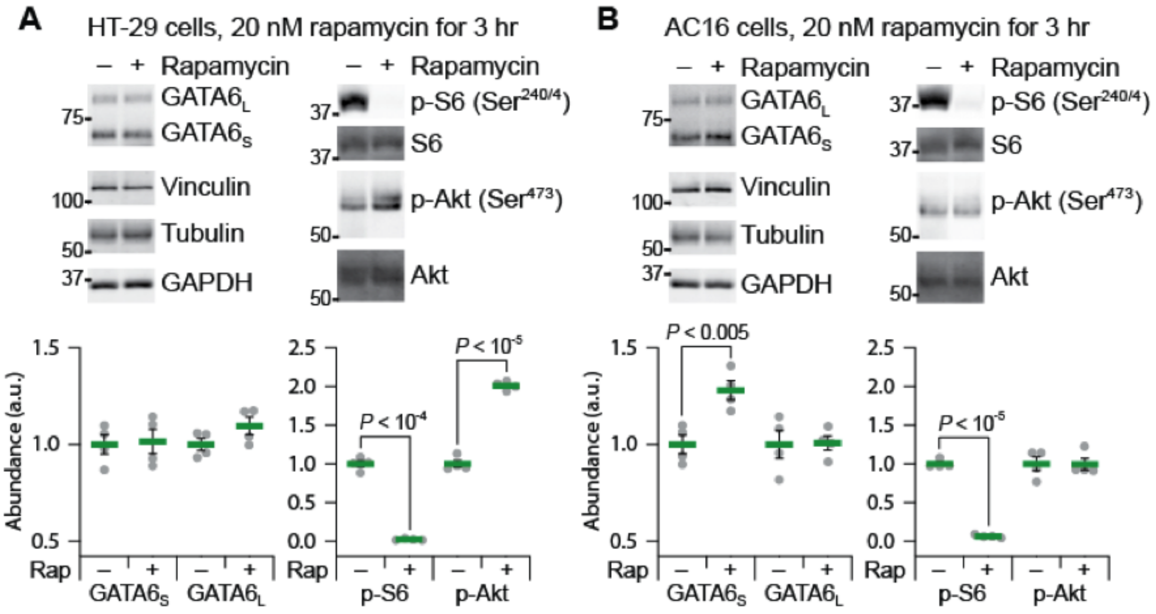


Figure 3.1: GATA6$_L$ abundance is not altered by prolonged rapamycin treatment or feedback phosphorylation of AKT. (A) HT-29 cells exhibit rapamycin-induced feedback phosphorylation of AKT, but do not stabilize GATA6. (B) AC16 cells exhibit stabilization of GATA6$_S$, but not GATA6$_L$, without rapamycin-induced feedback phosphorylation of AKT. Vinculin, tubulin, and GAPDH used as loading controls [93]. Quantitative immunoblot data are shown as the mean ± SEM of n = 4 biological replicates across two separate experiments.

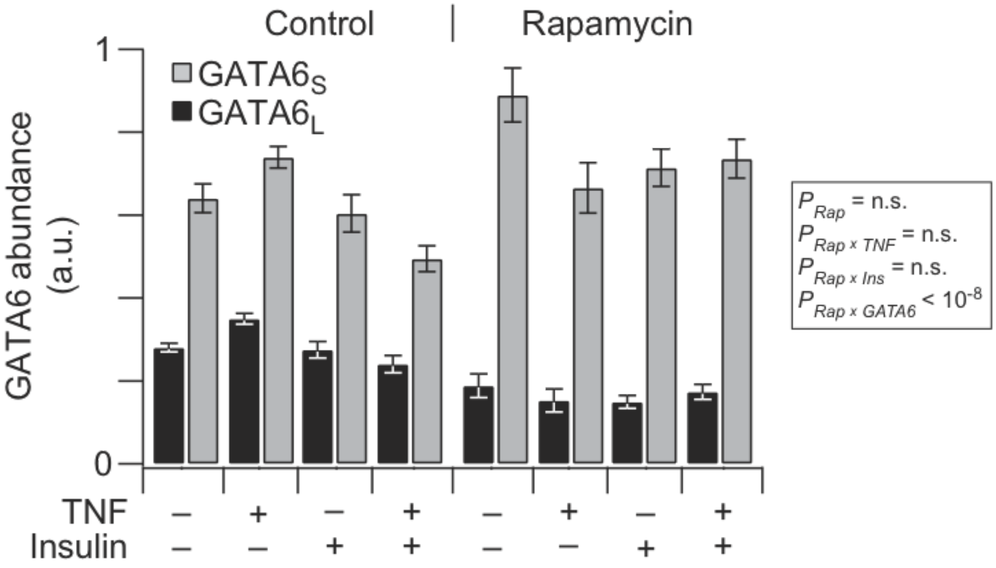


Figure 3.2: Prolonged rapamycin treatment alters the proportion of GATA6$_S$ to GATA6$_L$ independently of TNF or insulin treatment. HT-29 cells were pretreated with 200 U/ml IFNγ for 24 hours before inhibition with 20 nM rapamycin for two hours and stimulation with 100 ng/ml TNF or 500 ng/ml insulin for one hour. Quantitative immunoblot data for the relative proportions of GATA6$_S$ (gray) and GATA6$_L$ (black) are shown as the mean ± SEM of n = 6 biological replicates, with the effect of rapamycin and its interaction with TNF, insulin, and the proportion of GATA6 forms by four-way ANOVA with the following factors: rapamycin presence or absence, TNF, insulin, and GATA6 form.

these cells, S6 phosphorylation disappeared without subsequent feedback activation of AKT, and yet GATA6$_S$ abundance increased modestly after three hours as reported in VSMCs [138] (Figure 3.1B). Critically, under the same conditions in AC16 cells, we did not observe any alterations in the abundance of GATA6$_L$. These experiments illustrated that AKT feedback activation could be uncoupled from GATA6 stabilization, as could the posttranslational regulation of its long and short forms.

3.2 GSK3-dependent phosphorylation of Ser^{37} accelerates GATA6$_L$ turnover

To assess the importance of the GATA6$_L$ phosphorylation sites, we transfected single alanine mutants of Ser^{33}, Thr^{34}, and Ser^{37} or the triple mutant (3SA) into 293T cells, cells in which GATA6$_L$ phosphorylation has been detected previously (105). Compared to the wild-type allele, we noted a pronounced electrophoretic downshift in the major FLAG-immunoreactive band of the Ser^{37} and 3SA mutants (Figure 3.3A). The mobility shift was larger than that expected for a single phosphorylation site, suggesting that Ser^{37} phosphorylation was required for other phosphorylation events within GATA6$_L$. Iterative phosphorylation-dependent phosphorylation is characteristic of many GSK3 substrates, such as glycogen synthase [139].

Careful inspection of endogenous GATA6$_L$ immunoreactivity in HT-29 extracts revealed a slower migrating species that was similar to the electrophoretic shifts observed in 293T cells transfected with the phosphorylation-deficient mutants. We isolated this species from the faster migrating GATA6$_L$ through Phos-tag electrophoresis [140] followed by Gaussian mixture modeling of the densitometric traces. Acute TNF treatment reduced the upper form of GATA6$_L$ but with a concomitant increase in the lower form, such that total GATA6$_L$ abundance was not altered (Figure 3.3B). Costimulation with insulin or pretreatment with the GSK3 inhibitor CT99021 [141] did not alter the GATA6$_L$ downshift, despite insulin increasing GSK3 phosphorylation and CT99021 decreasing GSK3 activity (Figure 3.3C and D and Figure 3.4).

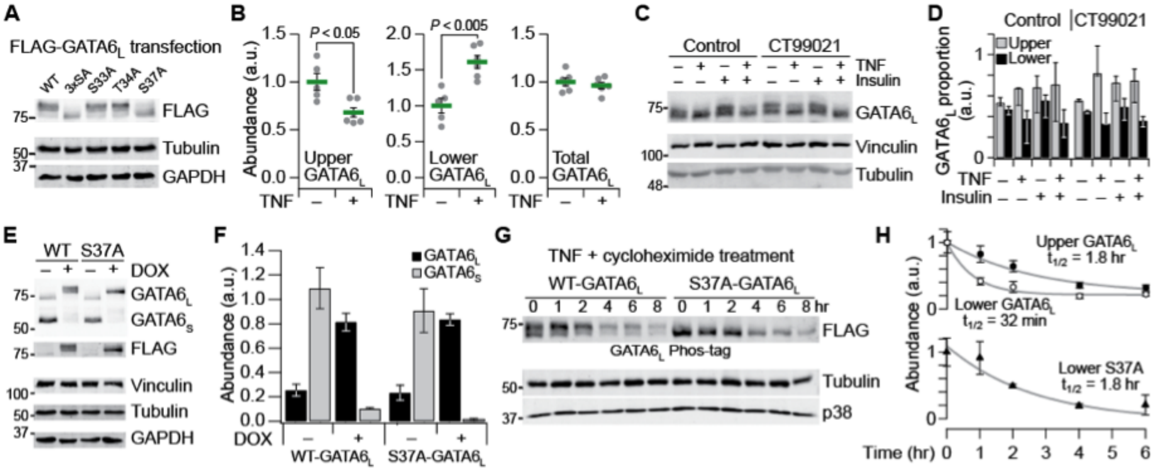


Figure 3.3: Extensive phosphorylation of GATA6$_L$ is blocked by S37A mutation, reversed by TNF stimulation, and stabilized in HT-29 cells. (A) Electrophoretic mobility of FLAG-tagged GATA6$_L$ is downshifted upon S37A mutation in lipofected 293T cells. (B) Phos-tag electrophoresis [140] reveals that TNF stimulation for one hour causes the dephosphorylation of GATA6$_L$. (C and D) Phos-tag electrophoresis (C) and quantification (D) of the upper and lower forms of GATA6$_L$ in response to IFNγ sensitization for 24 hours, pretreatment with 1 M CT99021 for one hour, and stimulation with TNF or insulin for one hour. Data are shown as the median proportion ± range of n = 3 biological replicates. (E and F) Doxycycline (DOX)-inducible addback in HT-29 cells replaces endogenous GATA6$_S$ with epitope-tagged GATA6$_L$. Cells were treated with 1 μg/ml DOX for 48 hours. (G and H) The less phosphorylated form of wild-type (WT) GATA6$_L$ is unstable. Cells were treated with 100 ng/ml TNF + 50 μM cycloheximide for the indicated times, and half-lives were estimated by nonlinear least-squares curve fitting. Quantitative immunoblot data are shown as the mean ± SEM of n = 3 (F and H) or 5–6 (B) biological replicates.

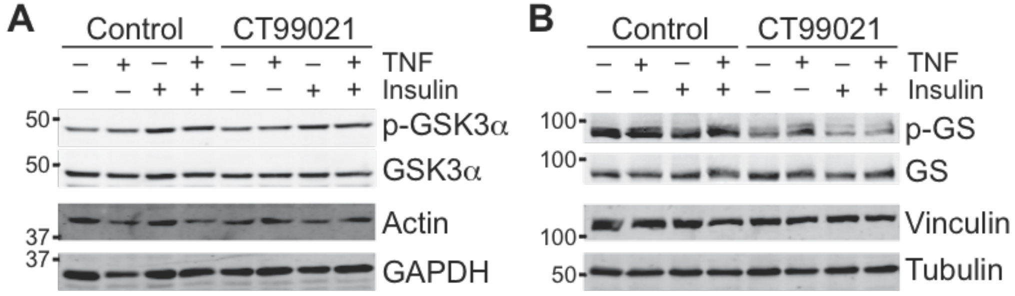


Figure 3.4: Insulin and CT99021 perturb GSK3 phosphorylation and activity. (A and B) Insulin induces GSK3α phosphorylation (A) and CT99021 inhibits GSK3-catalyzed phosphorylation of GS (B). HT-29 cells were sensitized with 200 U/ml IFNγ for 24 hours before pretreatment with 1 μM CT99021 for one hour and stimulation with 100 ng/ml TNF or 500 ng/ml insulin for one hour. Samples were immunoblotted for p-GSK3 (Ser21), total GSK3α, p-GS (Ser^{641}), and total GS with actin, GAPDH, vinculin, and tubulin used as loading controls. Immunoblots are representative of n = 23 independent biological replicates.

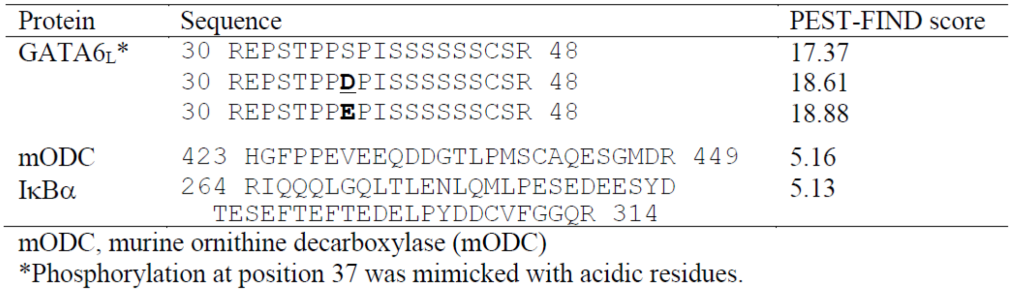
Because GATA6 mRNA was not induced by TNF (Figure 2.11I), these results indicated that GATA6$_L$ is dephosphorylated on some residues in response to TNF stimulation.

Our next goal was to evaluate the specific impact of Ser^{37} phosphorylation on GATA6$_L$ in HT-29 cells. One challenge was that the endogenous abundance of GATA6$_S$ was high compared to GATA6$_L$ (Figure 3.1), which could confound interpretations of ectopically expressed proteins. Therefore, we inducibly knocked down endogenous GATA6 with shRNA and added back epitope-tagged versions of wild-type or S37A GATA6$_L$ so that the abundance was comparable to total endogenous GATA6 (Figure 3.3E and F). DOX-induced addback in HT-29 cells recapitulated the electrophoretic mobilities of GATA6$_L$ that were observed in transfected 293T cells. The data suggested that the GATA6$_L$ modifications are not artifacts of overexpression, enabling use of the addback cells to examine the consequences of Ser^{37} phosphorylation.

Because GSK3 phosphorylation often accelerates substrate turnover [142], we combined the Phos-tag analysis with the HT-29 addback lines to estimate half-lives of wild-type and S37A GATA6_L. We combined inhibition of protein synthesis with TNF stimulation to enrich for the lower migrating form of wild-type GATA6_L in the addback cells. Under these conditions, we found that the half-life of more phosphorylated GATA6_L was more than twice that of the wild-type GATA6_L form that was less phosphorylated (Figure 3.3G and H). Surprisingly, the half-life of the S37A mutant was comparable to that of the more phosphorylated form of GATA6_L, suggesting that phosphorylation of Ser^{37} without subsequent additional phosphorylation renders GATA6_L unstable. Ser^{37} resides in the middle of a proline-glutamate-serine-threonine (PEST) degradation motif of GATA6_L, and this motif scores more strongly as a PEST motif than those in other well-known unstable proteins (Figure 2.11J and Table 3.1) [143, 144]. Ser^{37} phosphorylation might activate or expose the PEST sequence for rapid proteolytic degradation of GATA6_L, whereas additional phosphorylation at other sites could inhibit substrate recognition [143].

To monitor Ser^{37} phosphorylation specifically, we raised and affinity purified a phospho-specific antibody against a monophosphorylated peptide fragment of the PEST sequence in GATA6_L (Figure 3.5). If Ser^{37} modification were a prerequisite for subsequent phosphorylation, then the antibody would capture this initial phos-

Table 3.1: Top-scoring PEST sequences in the indicated proteins according to PEST-FIND.



Protein	Sequence	PEST-FIND score
GATA6$_L$*	30 REPSTPPSPISSSSSSSCSR 48	17.37
	30 REPSTPP**D**PISSSSSSSCSR 48	18.61
	30 REPSTPP**E**PISSSSSSSCSR 48	18.88
mODC	423 HGFPPEVEEQDDGTLPMSCAQESGMDR 449	5.16
IκBα	264 RIQQQLGQLTLENLQMLPESEDEESYD TESEFTEFTEDELPYDDCVFGGQR 314	5.13

mODC, murine ornithine decarboxylase (mODC)
*Phosphorylation at position 37 was mimicked with acidic residues.

phorylation event of GATA6$_L$, with the caveat that phosphorylation on Ser^{33} and Thr^{34} would ultimately disrupt the antibody epitope.

To evaluate the phospho-Ser^{37} antibody, we reassessed the immunoreactivity of total GATA6. The predicted molecular weights of GATA6$_S$ and GATA6$_L$ are 45.4 kD and 60 kD, respectively. However, extensive posttranslational modifications (Figure 2.11K) cause most GATA6$_S$ and GATA6$_L$ to run at an apparent molecular weight of ∽54 kD and ∽69-75 kD depending on electrophoresis conditions (Figure 3.6A). Multiply phosphorylated GATA6$_S$ (∽54 kD) can be misinterpreted as unmodified GATA6$_L$ (60 kD). Upon long exposure with a total GATA6 antibody, we revealed an additional immunoreactive band at ∽60 kD that was eliminated by GATA6 knockdown and reconstituted with addback of wild-type GATA6$_L$ and the S37A mutant (Figure 3.6A). In contrast to the 75 kD form (Figure 3.3E), the ∽60 kD form of wild-type GATA6$_L$ was significantly less abundant than the S37A mutant (Figure 3.6B), consistent with decreased stability. We interpreted the ∽60 kD band as the unmodified form of GATA6$_L$.

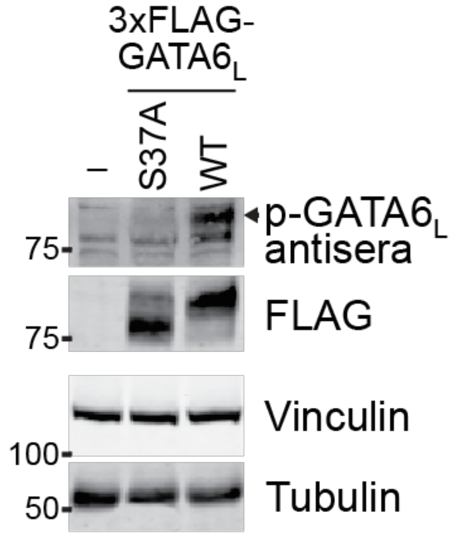


Figure 3.5: p-GATA6$_L$ (Ser^{37}) antiserum is specific for mobility-shifted wildtype GATA6$_L$ but not the S37A GATA6$_L$ mutant. 293T cells were transfected with pBabe puro control (–), 3×FLAG-tagged wildtype (WT) GATA6$_L$, or 3×FLAG-tagged S37A mutant and immunoblotted for p-GATA6$_L$ (Ser^{37}) and FLAG with vinculin and tubulin used as loading controls. Immunoblots are representative of n = 2 separate antiserum bleeds and n = 2 independent immunizations.

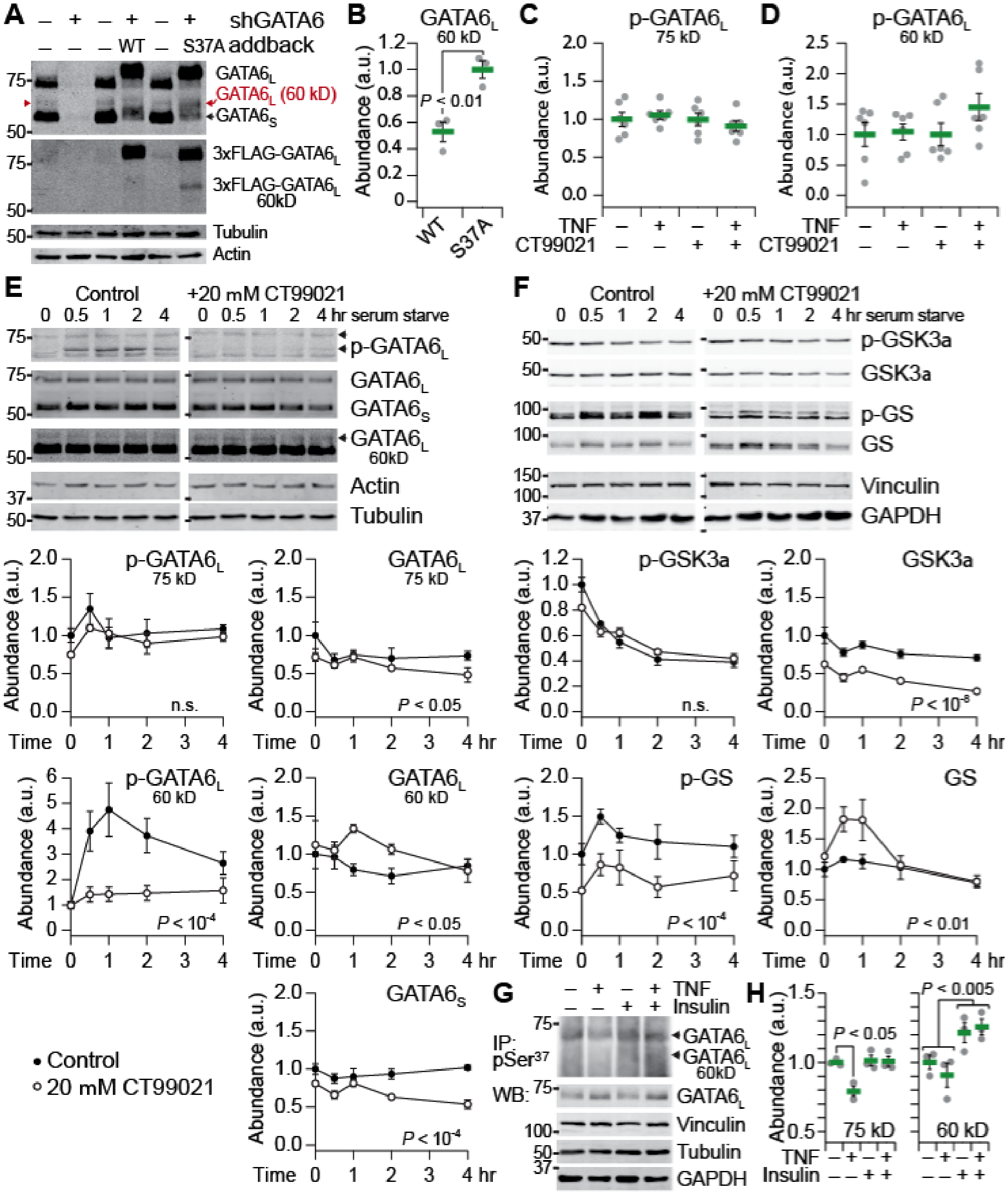

A
shGATA6
addback
WT
S37A
$GATA6_L$
$GATA6_L$ (60 kD)
$GATA6_S$
3xFLAG-$GATA6_L$
3xFLAG-$GATA6_L$ 60kD
Tubulin
Actin
75
50
B
$GATA6_L$ 60 kD
Abundance (a.u.)
$P < 0.01$
WT
S37A
C
p-$GATA6_L$ 75 kD
D
p-$GATA6_L$ 60 kD
TNF
CT99021
E
Control
+20 mM CT99021
0 0.5 1 2 4 hr serum starve
p-$GATA6_L$
$GATA6_L$
$GATA6_S$
$GATA6_L$ 60kD
Actin
Tubulin
F
p-GSK3a
GSK3a
p-GS
GS
Vinculin
GAPDH
Time
n.s.
$P < 0.05$
$P < 10^{-8}$
$P < 10^{-4}$
$P < 0.01$
p-GSK3a
GSK3a
p-$GATA6_L$ 60 kD
$GATA6_L$ 60 kD
p-GS
GS
$GATA6_S$
● Control
○ 20 mM CT99021
G
TNF
Insulin
IP: $pSer^{37}$
WB:
$GATA6_L$
$GATA6_L$ 60kD
Vinculin
Tubulin
GAPDH
H
$P < 0.05$
$P < 0.005$
75 kD
60 kD
TNF
Insulin

Figure 3.6: Phosphorylation and destabilization of endogenous GATA6$_L$ at 60 kD. (A) Knockdown (shGATA6) and FLAG-tagged addback of GATA6$_L$ at ∽60 kD (red). Samples were immunoblotted for total (modified and unmodified) GATA6 (upper), FLAG (lower), and the indicated loading controls. (B) Destabilization of the 60 kD form of wild-type GATA6$_L$ compared to the S37A-mutant addback cells. (C and D) Endogenous p-GATA6$_L$ (Ser^{37}) immunoreactivity is not detectably affected by stimulation with TNF for one hour, inhibition with 20 μM CT99021 for six hours, or both. (E and F) Phosphorylation and destabilization of the 60 kD form of GATA6$_L$ upon serum starvation. Specificity was confirmed by preincubation of cells with 20 μM CT99021 for one hour before serum starvation. (G and H) p-GATA6$_L$ (Ser^{37}) immunoprecipitation and total GATA6 immunoblot of HT-29 cells pretreated with IFNγ and stimulated with TNF ± insulin for one hour. The gamma of the immunoprecipitation image is set to 1.5 to minimize background from the immunoprecipitating antibody heavy chain. 0.5% input of each immunoprecipitate was immunoblotted for total GATA6 and the indicated loading controls. Data are shown as the mean ± SEM of n = 3 (B, E, F, H) or 6 (C and D) biological replicates.

Endogenous Ser^{37} phosphorylation of the 75 kD and 60 kD GATA6$_L$ forms was not detectably altered in response to TNF treatment for one hour or CT99021 treatment for six hours (Figure 3.6C and D). However, the endogenous 60 kD phospho-GATA6$_L$ signal was barely above the detection limit and thus highly variable [coefficient of variation (CV) ∽ 40%], yielding only ∽50% statistical power for detecting a 1.5-fold change. To evaluate phosphorylation of endogenous GATA6$_L$ on Ser^{37}, we serum starved the HT-29 cells to produce a stronger activation of GSK3 and confirmed specificity of the phosphorylation events with CT99021 pretreatment for one hour (Figure 3.6E and F). As expected, serum starvation reduced GSK3 phosphorylation (increasing GSK3 activity) and increased phosphorylation of glycogen synthase (GS) (Figure 3.6F). CT99021 reduced GS phosphorylation and total GSK3 abundance but also transiently increased total GS. Notably, within one hour of serum with-

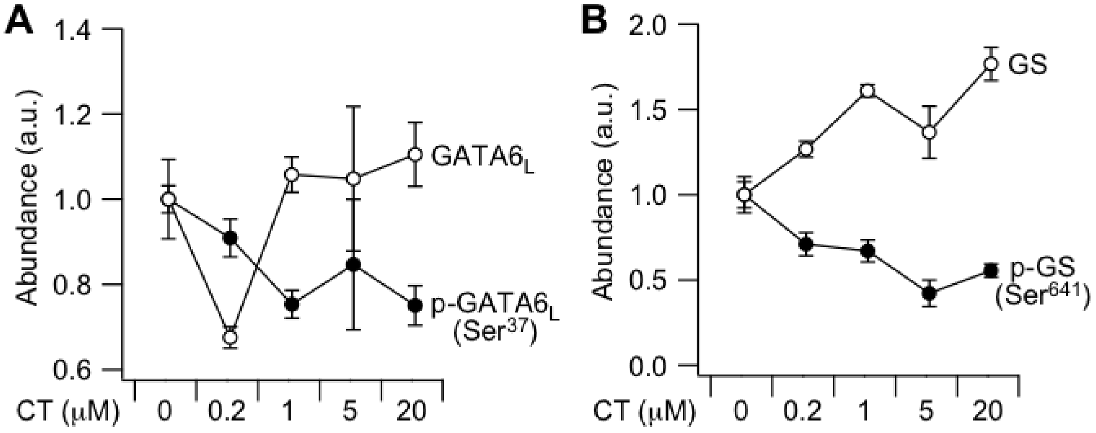


Figure 3.7: GATA6$_L$ phosphorylation on Ser^{37} and GS phosphorylation on Ser^{641} are lost in a dose-dependent manner upon treatment with the GSK3 inhibitor, CT99021. (A and B) HT-29 cells were preincubated with the indicated concentration of CT99021 (CT) for one hour, serum starved for one hour, and then analyzed for phosphorylated and total GATA6$_L$ (A) or phosphorylated and total GS (B) by quantitative immunoblotting [93]. Data are shown as the mean ± SEM of n = 3 biological replicates.

drawal, we observed a robust increase in the 60 kD form of phosphorylated GATA6$_L$, which coincided with the timing of GSK3 dephosphorylation and was blocked by CT99021 in a dose-dependent manner (Figure 3.6E and Figure 3.7). By contrast, phospho-Ser^{37} immunoreactivity of the 75 kD form of GATA6$_L$ was not altered by serum starvation or CT99021 treatment, suggesting that Ser^{37} was already stably phosphorylated in this form of GATA6$_L$.

Total abundances of the different forms of GATA6$_L$ also showed dynamic changes. GATA6$_L$ at 75 kD and GATA6$_S$ decreased significantly with CT99021 treatment, whereas GATA6$_L$ at 60 kD increased compared to uninhibited control cells that were serum starved ($P < 0.05$, two-way ANOVA). The time-dependent changes in 60 kD

$GATA6_L$ abundance mirrored the changes in total GS and inversely correlated with changes in 60 kD $GATA6_L$ phosphorylation at Ser^{37} (Figure 3.6E and F). These experiments provide further evidence that Ser^{37} is a site phosphorylated by GSK3 and that phosphorylation at this site promotes turnover of $GATA6_L$ in the absence of phosphorylation at additional sites.

With greater confidence in the p-$GATA6_L$ (Ser^{37}) antibody, we revisited the original biological context of -pretreated HT-29 cells stimulated with TNF and insulin. To enable detection, we immunoprecipitated cell extracts with p-$GATA6_L$ (Ser^{37}) antisera and immunoblotted for total GATA6 (Figure 3.6G). An extended electrophoresis was required to separate the 60 kD form from the heavy chain of the immunoprecipitating antibody, causing a smear of immunoreactivity rather than a discrete band. In response to TNF alone, we repeatedly observed a drop in 75 kD, but not 60 kD, $GATA6_L$ phosphorylation (Figure 3.6H), corroborating the dephosphorylation previously noted by Phos-tag electrophoresis (Figure 3.3B). Moreover, the reduction in 75 kD p-$GATA6_L$ (Ser^{37}) was blocked by insulin costimulation, indicating a specific point of crosstalk between TNF and insulin. Insulin, by contrast, independently elevated the abundance of 60 kD $GATA6_L$ phosphorylation, suggesting that insulin delays the turnover of this form beyond its effect on Ser^{37} phosphorylation. We conclude that the phosphorylation of endogenous $GATA6_L$ at Ser^{37} is consistent with the antagonism and linear superposition in abundance of Cluster #9 transcripts observed upon TNF + insulin stimulation (Figure 2.11A to D and Figure 2.8).

3.3 GSK3-dependent phosphorylation of Ser^{37} alleviates GATA6_L repression of transcripts in the crosstalk cluster

We investigated the role of GATA6_L phosphorylation in the regulation of transcripts subject to TNF-insulin crosstalk. We inducibly overexpressed wild-type GATA6_L in HT-29 cells (Figure 3.8A) before stimulation with TNF for two hours and transcriptomic profiling by microarray.

GATA6 can act as either a transcriptional activator or repressor [145], but we found in the GATA6_L-overexpressing HT-29 cells that the effects of GATA6_L were predominantly repressive: Without stimulation, GATA6_L overexpression induced 51 transcripts and repressed 136 transcripts at a 5% false-discovery rate ($P < 10^{-10}$, binomial test).TNF stimulation increased the number of genes affected by GATA6_L overexpression, but the bias toward repression persisted (317 induced transcripts versus 438 repressed transcripts; $P < 10^{-5}$, binomial test).

Notably, the same repressive bias was observed for transcripts in the crosstalk cluster (Figure 3.8B), and those with the strongest GATA6_L-associated repression were among the clearest examples of TNF-insulin crosstalk (Figure 2.11A to C). Using chromatin immunoprecipitation, we confirmed binding of GATA6_L to consensus sites within the promoters of many TNF-insulin crosstalk genes (Figure 3.9), suggesting that repression is direct.

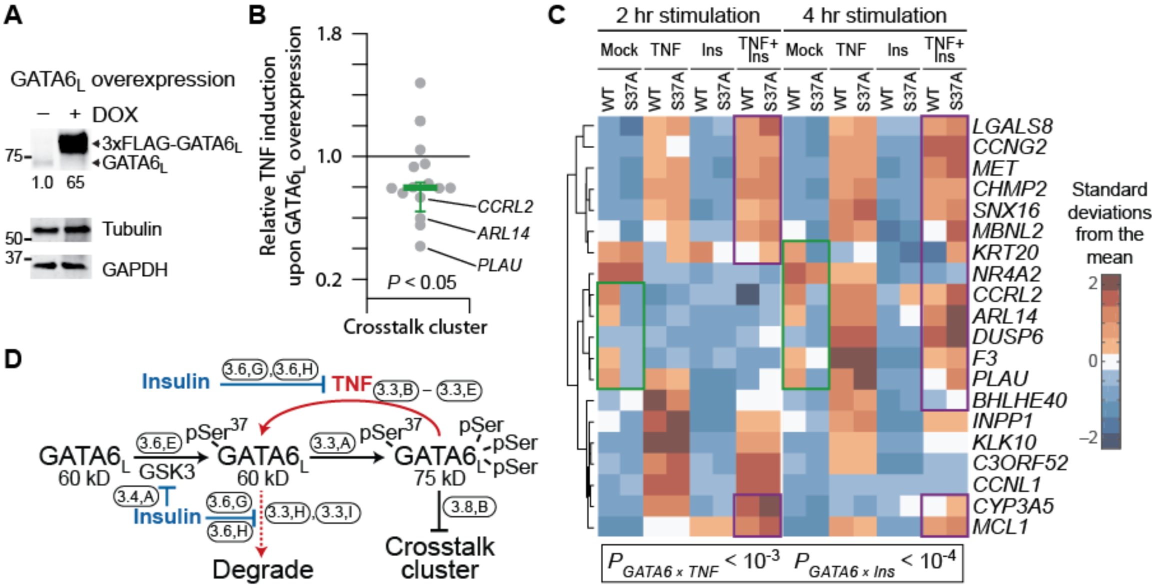


Figure 3.8: S37A mutation of GATA6$_L$ mimics and competes with the repression of transcript abundance in the TNF-insulin crosstalk cluster. (A) Doxycycline (DOX)-inducible overexpression of wild-type GATA6$_L$ in HT-29 cells. Cells were treated with 1 g/ml DOX for 24 hours. (B) Ratio of TNF-induced transcript abundance for crosstalk cluster genes in the presence or absence of GATA6$_L$ overexpression. Data are shown as the mean ratio of n = 3 independent biological samples assessed by microarray profiling, with bias in the ratio assessed by two-sided binomial test. (C) S37A mutation of GATA6$_L$ mimics insulin stimulation (green) and antagonizes TNF-insulin crosstalk (purple). qRT-PCR data for Cluster #9 genes in wild-type (WT) and S37A mutant (S37A) GATA6$_L$ addback cells pretreated with IFNγ and stimulated with TNF, insulin, or both for two or four hours. Data are shown as row-standardized geometric means of n = 6 biological replicates across two separate experiments, with interactions between GATA6 status and TNF or insulin assessed by log-transformed five-way ANOVA with the following factors: GATA6, transcript, TNF, insulin, and time. (D) Three-state conceptual model for GATA6$_L$ regulation by TNF and insulin and its relation to the crosstalk cluster of transcripts. Ovals annotate the figure subpanels supporting the links depicted.

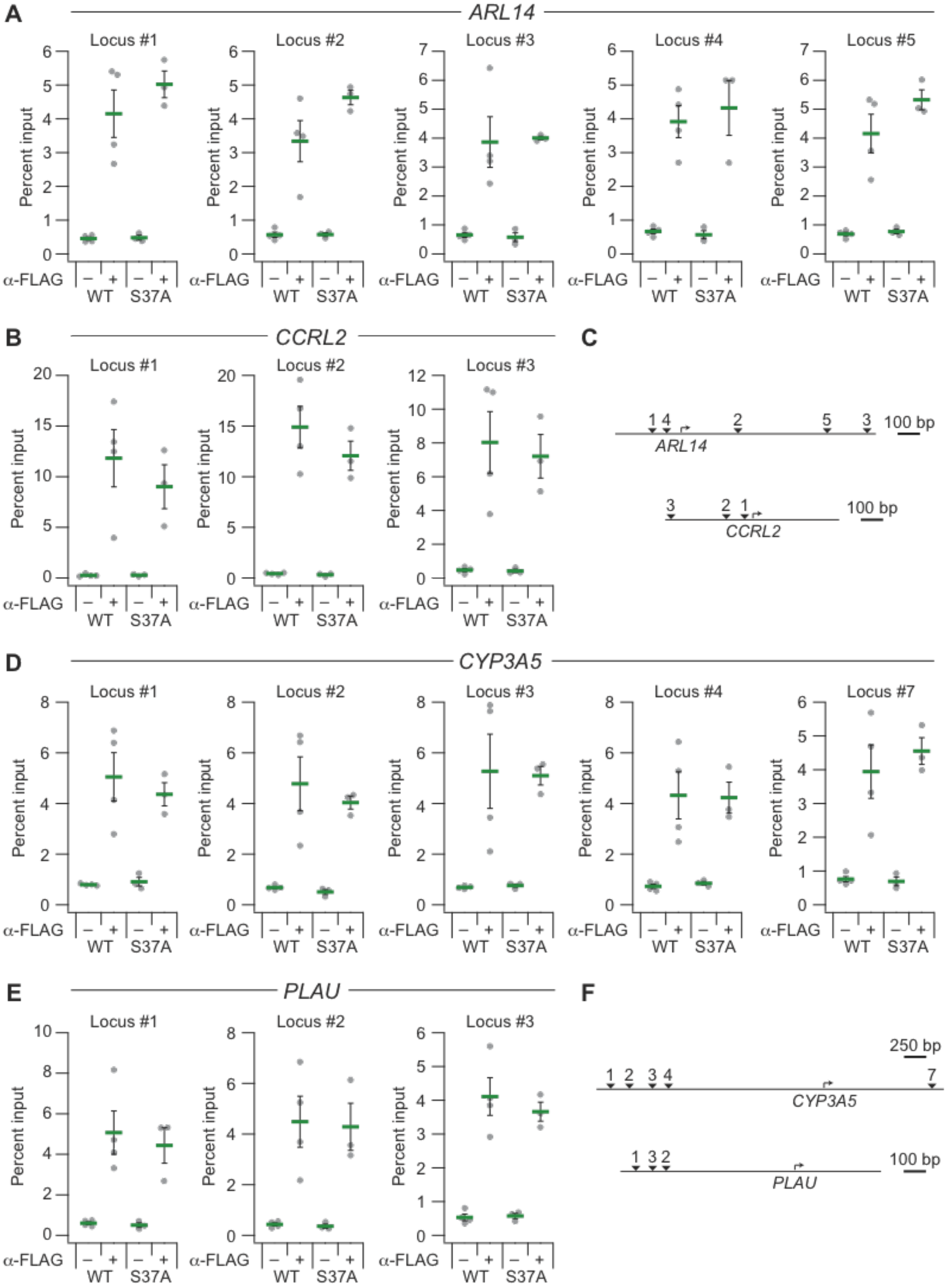

A
ARL14
Locus #1
Locus #2
Locus #3
Locus #4
Locus #5
Percent input
0
1
2
3
4
5
6
7
α-FLAG
–
+
WT
S37A
B
CCRL2
Locus #1
Locus #2
Locus #3
Percent input
0
5
10
15
20
2
4
6
8
10
12
C
1 4
2
5
3
100 bp
ARL14
3
2 1
100 bp
CCRL2
D
CYP3A5
Locus #1
Locus #2
Locus #3
Locus #4
Locus #7
Percent input
0
2
4
6
8
E
PLAU
Locus #1
Locus #2
Locus #3
Percent input
0
2
4
6
8
10
F
250 bp
1 2 3 4
7
CYP3A5
1 3 2
100 bp
PLAU

Figure 3.9: GATA6$_L$ occupies GATA binding sites in the promoters of genes within the crosstalk cluster. (A and B) Chromatin immunoprecipitation quantitative PCR (ChIP-qPCR) measurements of GATA6$_L$ DNA binding to the indicated loci of ARL14 (A) and CCRL2 (B). (C) Map indicating the positions of the loci containing GATA binding sites in ARL14 and CCRL2. (D and E) ChIP-qPCR measurements of GATA6$_L$ DNA binding to the indicated loci of CYP3A5 (D) and PLAU (E). (F) Map indicating the positions of the loci containing GATA binding sites in CYP3A5 and PLAU. ChIP-qPCR data for wildtype and S37A addback HT-29 cells stimulated with 100 ng/ml TNF for one hour are shown as the mean percent input ± SEM of n = 34 independent experiments. ChIP experiments without anti-FLAG (α-FLAG) antibody were substituted with an equivalent amount of naive mouse IgG.

If GATA6$_L$ mediates the insulin-stimulated repression of the crosstalk cluster and is stabilized by inhibition of the GSK3 pathway, then phosphorylation of Ser^{37} would provide a mechanism for TNF-insulin crosstalk. Furthermore, the S37A mutant of GATA6$_L$ should mimic the effect of insulin on TNF-stimulated gene expression for transcripts in the crosstalk cluster and dampen the crosstalk observed when insulin is added to S37A mutant cells. We tested this prediction with the wild-type and S37A addback lines (Figure 3.3E and F), inducing GATA6$_L$ and then stimulating with TNF, insulin, or both. By qRT-PCR, we identified multiple instances in which S37A addback reduced transcript abundance similar to that observed in wild-type addback cells stimulated with insulin (Figure 3.8C), green. We also found many examples in which TNF + insulin-induced transcript abundance was higher in S37A addback cells compared to wild-type addback cells (Figure 3.6C, purple) and more similar to TNF-treated cells, suggesting reduced crosstalk. Analysis of the entire Cluster#9 dataset revealed significant interactions between GATA6 and TNF or insulin (interaction P $<10^{-10}$, five-way ANOVA), indicating that the Ser^{37} genotype alters the transcriptional response to both stimuli. Taken together, our data support a model whereby TNF promotes and insulin inhibits the formation and degradation of GATA6$_L$ monophosphorylated on Ser^{37} (Figure 3.6D). This phosphoregulation is ultimately reflected by the abundance of transcripts in the crosstalk cluster.

3.4 Ser^{37} phosphorylation of different GATA6$_L$ forms is observed in diverse cell types

The model of GATA6$_L$ phosphorylation-mediated regulation (Figure 3.8D) may be specific to HT-29 cells or could occur in other cell types. We immunoblotted various cell lines with the phospho-Ser^{37} antibody in comparison to affinity-purified antisera binding the nonphosphorylated peptide surrounding Ser^{37} and to other commercial GATA6 antibodies [134, 146]. In HCT-8 and DLD-1 colorectal cancer lines, AC16 cardiomyocytes [147], and MCF10A-5E breast epithelial cells [148], we observed the 60 kD and 75 kD forms of GATA6$_L$, which were phosphorylated to variable extents according to phospho-Ser^{37} immunoreactivity (Figure 3.10A and B). Multiple GATA6 antibodies also recognized another species at ∽100 kD (Figure 3.10A to D), suggesting that an even more phosphorylated form of GATA6$_L$ may remain to be characterized. Indeed, the aggregate number of reported phosphorylation sites on GATA6$_L$ now exceeds 20 (Figure 2.11K).

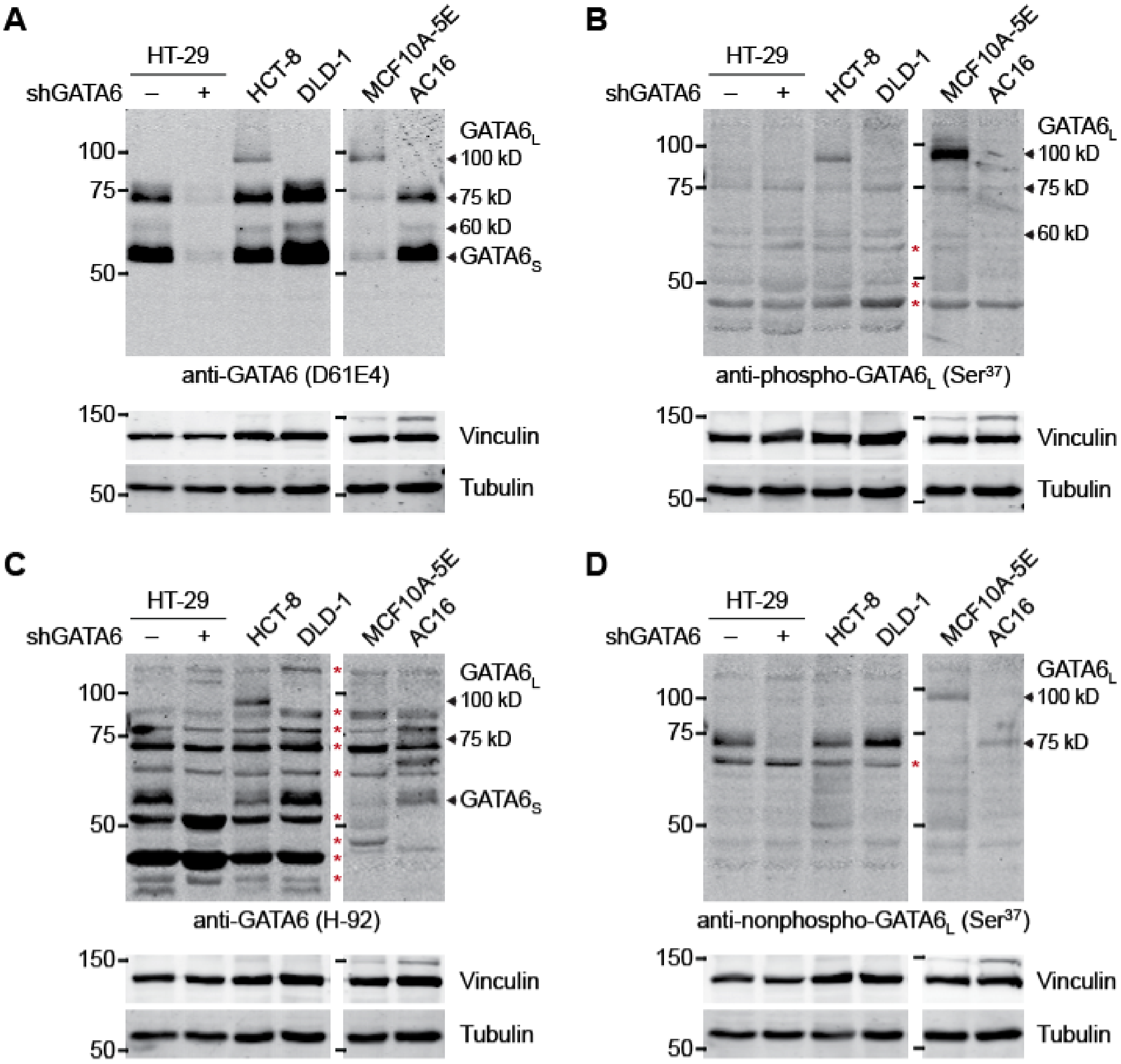


Figure 3.10: Diversity of GATA6$_L$ forms across different cell lineages. Arrows indicate the GATA6 forms confirmed earlier by knockdown or observed with multiple antibodies. Red asterisks indicate nonspecific bands. The MCF10A-5E and AC16 samples are on an immunoblot; HT-29, HCT-8, and DLD-1 are on another blot. The blots have been scaled to match. Data are representative of n 3 independent experiments.

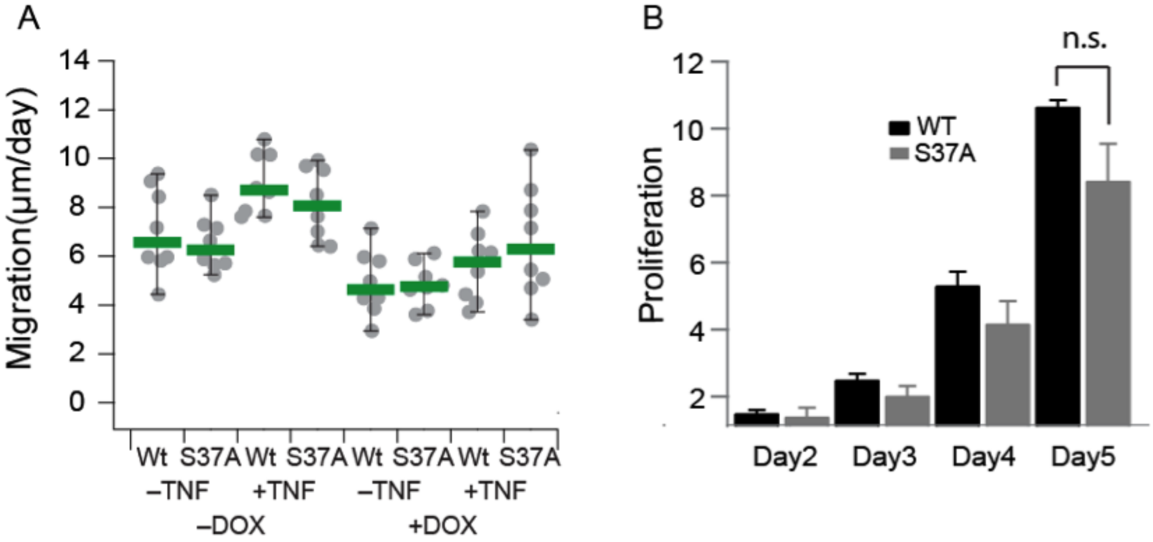


Figure 3.11: *phosphoS37GATA6*$_L$ only has marginal effect on wound healing and proliferation. Wound closure rate measured by scratch assay (A) and (B) Proliferation rate measured by counting cells both in a dox-inducible addback model of GATA6$_L$ in HT29 cells.

3.5 The effect of GATA6$_L$ on cellular phenotype

One member of cluster#9 genes that is significantly downregulated upon GATA6$_L$ overexpression is "PLAU" (Figure 3.8B). PLAU encodes for a secreted serine proteinase (uPA) that can activate other proteases capable of degrading ECM proteins. PLAU also has been implicated to have a role in cellular migration. Thus, we decided to look at cell migration as a candidate phenotype that GATA6-mediated regulation of PLAU gene would have effect on. To investigate the functional role of serine^{37} phosphorylation on epithelial sheet migration, we performed wound healing assays to assess changes in cell migration using HT29 cell inducibly expressing wt and S37A-GATA6$_L$. We observed that ectopic expression of GATA6$_L$ in a DOX-inducible addback model of HT29 cells only marginally decrease the wound healing rate in a

$GATA6_L$ and non-S37 dependent manner (Figure 3.11A). In addition, upon TNF treatment both wtGATA6 and S37AGATA6 expressing cells heal the wound slightly faster compared to the matched DOX-induced, no-treatment control. These effects can not be explained by the differences in the proliferation of wt and mutant cell lines (Figure 3.11A and B).

3.6 Summary

We found that GSK3 phosphorylation of Ser^{37} on the long form of GATA6 ($GATA6_L$) accelerated its degradation in cells. The increased turnover reduced transcriptional repression by $GATA6_L$ of genes induced by TNF. Collectively, our results showed that $GATA6_L$ integrated growth factor-induced signaling activity and inflammatory transcriptional regulation. By coupling systematic experiments with statistical modeling approaches, such as tensor PLSR, one can identify relationships that would otherwise go unnoticed. Although it remains atypical to collect transcriptomic data as tensors, we expect widespread systematization of transcriptomics as expression-profiling costs drop. A model is just the first step, however, because the most surprising data-derived connections will require the identification of previously unrecognized mechanisms to explain them. These, in turn, require hypothesis-driven experiments with the best molecular-genetic and pharmacologic perturbations available. For understanding how gene expression is controlled by complex stimuli, the integration of molecular biology and systems biology has yet to be fully exploited.

Chapter 4

Research Significance, Future Direction and Conclusion

4.1 Research significance

4.1.1 Developing proper reagents to study GATA6$_L$

The majority of GATA6 studies so far, had been focusing on GATA6 mRNA detection. Fewer GATA6 studies investigated GATA6 protein levels inside the cells and they either mostly had been studying only one proteoform of GATA6 (GATA6$_S$) or it is unclear which form they are referring to. The unavailability of proper commercial reagents including DNA constructs that encode GATA6$_L$ and antibodies that distinguish GATA6$_L$ from GATA6$_S$ have certainly played an important role in the ambiguity of GATA6 literature. Since our research hypothesis suggested a role for the N-terminal extension of GATA6$_L$, we needed to develop proper GATA6$_L$ reagents to test our hypothesis. Thus, we cloned the full length human GATA6 gene from an HT29 cDNA library into a DOX-inducible lentivector. This full length human GATA6 plasmid was engineered to bear a 3×FLAG tag in its N-terminus and a 3×AU1 tag in its C-terminus. Furthermore, to isolate the role of GATA6$_L$ from GATA6$_S$, we developed GATA6 addback cell lines in which the endogenous gene was knocked down and the ectopic expression of GATA6$_L$ or GATA6$_S$ was titered close to endogenous levels. Collectively these systems enabled us to reliably detect and distinguish different forms of GATA6 and isolate a role for GATA6$_L$ in HT29 cells.

4.1.2 Resolving longstanding GATA6 riddle suggests new interpretation of old data

Although GATA6$_L$ is less abundant than GATA6$_S$ in most cell types including HT29 cells, there is evidence that GATA6$_L$ is the more potent transcriptional regulator [38,149]. Takada K. et al based on a series of GATA6 mutation studies, provided evidence suggesting that the proline-glutamate-serine-threonine (PEST) degradation motif (Glu^{31}-Cys^{46}) within the 146 amino acid N-terminal extension of the Long form GATA6, plays a crucial role in the transcriptional potency of GATA6$_L$. However the authors did not provide further evidence supporting a mechanism for their observation. Interestingly, Ser^{37} is located within this serine-rich N-terminal region (Glu^{31}-Cys^{46}). In addition, the same authors reported an unusual electrophoretic mobility for GATA6$_L$ on SDS PAGE. This observation is in line with our primarily puzzling observations detecting several GATA6-reactive bands on the immuno-blots with a protein mass profile including but not limited to the reported size. I believe the molecular mechanism that we found and provided evidence for in previous chapters can explain the unusual size profile of GATA6 proteoforms. This mechanism also can explain the previously documented higher transcriptional potency of GATA6$_L$ based on the accepted notion that the stability of a transcription factor and its transcriptional activity are inversely related [150].

In a recent publication [145], Wamaitha et al. demonstrated that GATA6 (or GATA6$_L$, not clear based on their plasmid vendor info) can simultaneously act as a repressor and an activator in pluripotent cells. In the light of our findings about GATA6 proteoform diversity and regulation together with previous work I could envision at least two alternative mechanisms to explain the authors' observations. One possible mechanism through which the differential effect of GATA6 on different promoters can be explained is the choice of GATA6 cofactor in each case. It remains to be studied whether the PEST sequence within the evolutionary conserved N-terminal extension of GATA6 plays a role in the choice of GATA6$_L$ co-interactor. Another possible mechanism is that GATA6$_L$ and GATA6$_S$ may act antagonistically. This is a plausible hypothesis supported by the evidence that GATA6$_L$ and GATA6$_S$ can form dimers [149] added to the fact that both forms have intact DNA binding domains (Fig 2.12). The heterodimers and homodimers of GATA6$_L$ and GATA6$_S$ can have different affinities to GATA binding sites in the genome which has been previously observed for translational isoforms of (C/EBP)β [151, 152].

4.1.3 The significance of findings

Collectively, I believe our findings can potentially change the interpretation of previous published observations regarding GATA6 function (specific examples mentioned in Section 4.1.2) and change the future pursuit of research on this important regulator. In conducting our research questions, we found it necessary to develop our own reagents which proved to be crucial in unravelling part of GATA6 function. These reagents can be used to leverage research on the role of GATA6 and the importance or interchangeability of GATA6_L and GATA6_S. In lieu of our findings, I believe researchers working on GATA6 need to be more cautious both in using the available GATA6 reagents as well as making interpretations of their observations.

Transcription factors other than nuclear receptors are well recognized as difficult drug targets. Thus, identifying either their downstream target genes or upstream regulators provide possible alternative and easily druggable targets. Here in this thesis, we found that the GSK3 mediated phosphorylation of GATA6 transcription factor alters the turnover of GATA6_L affecting the expression of a group of genes. However, changes in GATA6_L turnover does not significantly alter the occupancy of GATA binding sites in our hands as assessed by ChIP assay(Figure 3.9). These findings not only shed light on GATA6 biology but also suggest potential new drug targets with sensitivity to small molecules for pathologic states in which GATA6 plays a central role.

4.1.4 The significance of approach: making new discoveries by combining holistic and reductionist approaches

Our study here introduces and implements tensor PLSR as an approach for structured biological datasets. Considering that signaling dynamics often occur in discrete temporal phases [91,114,153,154], tensor PLSR provides an attractive means to deconstruct time-course data in a systematic manner. Although the mathematics have been established for decades [111], data types that can exploit the tensor framework are relatively new to cell signaling. For tensor generation, a multiplex technique that simply measures many genes or proteins is insufficient. The method must also be cost-effective, reproducible, and scalable for repeated use across multiple treatments, time points, and perturbations. The newest technologies rarely meet these criteria, prompting our use of long-established methods at a scale not typically considered.

We applied tensor PLSR with the goal of discovering molecular mechanisms that connect signaling to transcriptional regulation. Ideally, the mechanisms would involve proteins not originally included in the systematic dataset. Systems-level studies rarely uncover these “hidden nodes” and validate them experimentally like we achieved here for GATA6$_L$ [106,155]. Testing model- or bioinformatics-derived predictions requires a skill set entirely different from the one needed to perform the analysis. Our findings argue for the benefits of dual training, where computationalists work at the bench and experimentalists use quantitative models, gaining an appreciation for the thematic similarities in each approach. For example, just as modeling assumptions should be subject to falsification [88], we sought to challenge the prevailing biological assumptions about GATA6 and its different forms.

The deceptive electrophoretic mobilities of GATA6$_S$ and GATA6$_L$ have important implications for biological function. Although GATA6$_L$ is generally less abundant than GATA6$_S$ in most cell types, there is evidence that GATA6$_L$ is the more potent transcriptional regulator [38]. GATA6 promotes the expression of the stem cell marker LGR5 in colorectal cancer [146, 156]. Neither paper clarified whether the regulation occurs through GATA6$_L$, GATA6$_S$, or both. However, insulin-like growth factor inhibits GSK3 and promotes expansion of Lgr^{5+} stem cells in mice [115, 157]. Our results indicate that one mechanism for this expansion is the stabilization of GATA6$_L$.

The phosphorylation of Ser^{37} adds a GATA6$_L$-specific mode of regulation to reports of posttranslational modifications that would presumably target both long and short forms [47, 49, 138]. Although we were unable to reproduce the mechanism exactly [138], modification of Ser^{436} by AKT2 should coincide with loss of Ser^{37} phosphorylation to stabilize GATA6$_L$ synergistically in contexts where both pathways operate. GATA6$_L$ phosphorylation-mediated regulation could prove important in endothelial cells, a TNF- and insulin-responsive cell type in which GSK3 and GATA6 interact as a complex [158]. Our mass spectrometry study also uncovered other GATA6$_L$-specific proline-directed modification sites (Thr^{62} and Ser^{137}) that could function with the Ser^{266} site phosphorylated by ERK [47, 49]. Such complex layers of regulation should be expected of a transcription factor that is central to embryonic development and cell specification [128].

4.2 Future Direction

4.2.1 A brief summary of alternative hypothesis testing

As previously mentioned, phosphorylation of a protein can change its function through different mechanisms. To study the functional consequences that phosphorylation of GATA6 on Ser^{37} entails, we tested several alternative hypotheses in parallel that will be discussed briefly in the following paragraphs. For example, phosphorylation of Ser^{37} on GATA6$_L$ can hypothetically lead to a change in the localization of GATA6$_L$. Our preliminary studies of human GATA6$_L$ localization by immunofluorescent staining using a couple of different antibodies did not suggest any obvious change in the localization of the GATA6$_L$ in a phosphoSer^{37}-dependent manner. Another possibility is that the phosphorylation of Ser^{37} changes the conformation of GATA6$_L$ affecting its affinity for DNA. In order to test this hypothesis, we performed Chromatin Immuno-precipitation (ChIP)(Figure 3.9). We observed enrichment of both wt-GATA6$_L$ and S37A-GATA6$_L$ on GATA binding sites in the promoters of genes within the crosstalk cluster. Again, this promoter enrichment was not phosphoSer^{37}–dependent in our hands. Lastly, we moved on to test another hypothesis that can further explain the observed repressory function of hyperphsophorylated-75kD-GATA6 (pGATA6$_L$). We proposed that preferred interaction of pGATA6 with co-interactors can lead to a decrease in the transcription of downstream genes. We first investigated

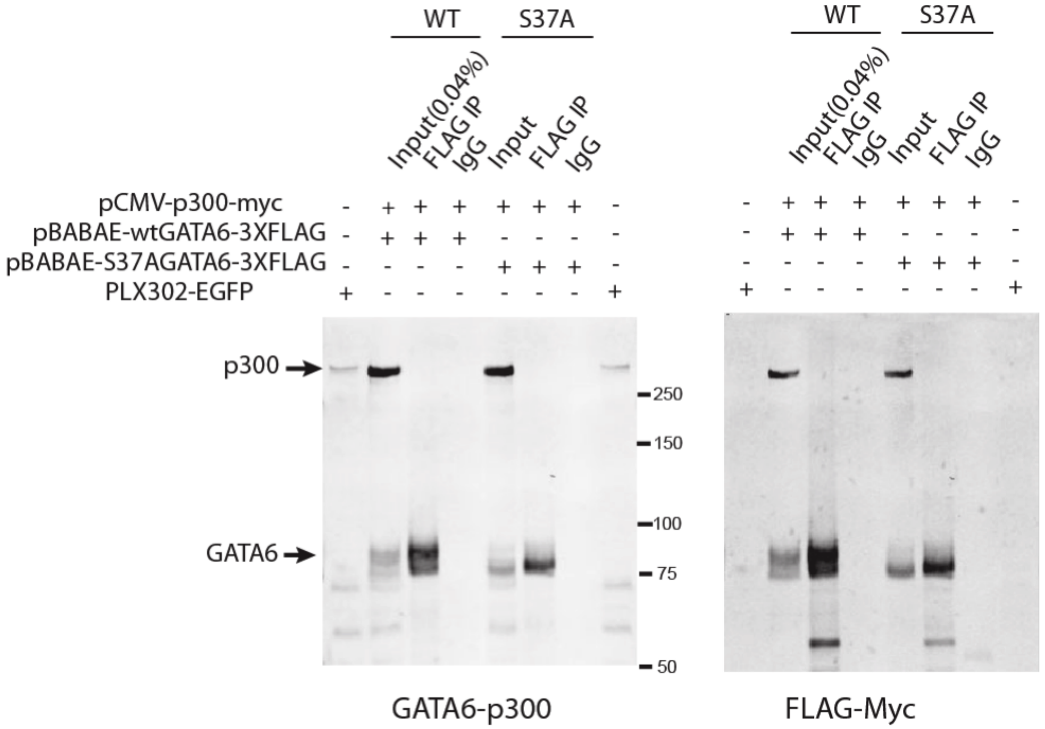

WT
S37A
Input(0.04%)
FLAG IP
IgG
Input
FLAG IP
IgG
pCMV-p300-myc - + + + + + + -
pBABAE-wtGATA6-3XFLAG - + + + - - - -
pBABAE-S37AGATA6-3XFLAG - - - - + + + -
PLX302-EGFP + - - - - - - +
p300
GATA6
250
150
100
75
50
GATA6-p300
FLAG-Myc

Figure 4.1: P300 does not co-immunoprecipitate with GATA6 in a 293T overexpression system. 293T cells were co-transfected with either pBABE-3XFLAG-wtGATA6 or S37AGATA6 and pCMV-p300-myc followed by FLAG immunoprecipitation and western blot analysis

this hypothesis by performing co-immuno precipitations (co-IPs) in a GATA6 overexpression system with or without DSP cross-linker followed by commassie brilliant blue staining of SDS PAGE gels. We did not detect any band that could correspond to and suggest the presence of a GATA6 co-interactor in a phosphoS^{37}-dependent manner. Alternatively we co-expressed GATA6 and p300 a histone acetyl-transferase and a suggested interactor of GATA6 [?] in HEK293T cells. In my hands, p300 did not co-IP with GATA6 in our overexpression system (Figure 4.2.1). However, our lab is currently developing new reagents to assess GATA6$_L$ co-interactors in a more systematic way.

4.2.2 New Research Directions

4.2.2.1 The functional impact of GATA6$_L$/GATA6$_S$ ratio

As mentioned earlier, GATA6$_S$ in most cell types is more abundant than GATA6$_L$ which is the more potent transcriptional regulator. One logical question that follows would be why then evolution supports the constant production of GATA6$_S$ and does the GATA6$_S$/GATA6$_L$ balance in cells have any biological consequence. One good documented example regarding the functional significance of the ratio of translational isoforms is the transcription factor CCAAT/enhancer binding protein (C/EBP)β [151, 152]. Similar to GATA6, (C/EBP)β mRNA can be translated from two different in frame methionine resulting in the production of (C/EBP)β_L (LAP) and (C/EBP)β_S(LIP). In these studies, the authors provided evidence that (C/EBP)β_L/(C/EBP)β_S ratio is important and changes during liver terminal differentiation and mammary gland morphogenesis, respectively. Interestingly they showed that the two translational isoforms are mutually antagonistic using functional assays in different cell contexts. Based on our observations regarding the cell type dependent expression of different proteoforms of GATA6, it would be interesting to investigate how these proteoform heterogeneities map into GATA6 function in different cell types. To this end, a starting point would be to assess whether we can observe changes in GATA6$_L$ forms in response to insulin, TNF or other relevant stimuli.

Given the documented critical role of GATA6 in early development [37], an interesting line of investigation beyond the scope of this thesis would be to assess whether

$\text{GATA6}_L/\text{GATA6}_S$ ratio changes early in development or during intestinal differentiation. This can elucidate the mechanism by which this critical master regulator is regulated early in development as well as adult life. One platform for the preliminary assessment of this research question could be the Caco-2 cell line. This human colon carcinoma cell line can be induced to differentiation by growing cells into confluency in a 2D culture dish. Then the $\text{GATA6}_L/\text{GATA6}_S$ ratio can be quantified by western blotting before and after differentiation to assess a change in GATA6 levels that coinside with a certain phenotype.

4.2.2.2 GATA6 regulated gene expression: GATA6 and CCRL2

Chemokine (CC motif) receptor-like 2 (CCRL2) is a member of the crosstalk gene cluster. This 7-transmembrane receptor has recently been deorphaned by evidence of interacting with ligands such as CCL5, CCL19 and chemerin. Among others, chemerin, a natural non-chemokine chemoattractant ligand and an adipokine, binds to CCRL2 and changes the bioavailability of this ligand. However, it does not induce intracellular calcium flux, ligand internalization or cell migration unlike chemokine-like receptor 1 (CMKLR1) and other chemerin signaling receptors [159,160]. CCRL2's expression has been studied in heamatopoitic cells, immune cells, endothelial cells, airway epithelium and astrocytes. Interestingly, the expression of this receptor is known to be induced by inflammatory stimuli including TNF [159]. CCRL2 is also a member of the group of crosstalk transcripts which we demonstrated to be induced by

TNF and attenuated by insulin through GATA6$_L$-dependent mechanism. Moreover, we observed enrichment of GATA6 transcription factor in the proximal promoter of CCRL2 providing additional evidence supporting the GATA6 regulated expression of this gene. Collectively, we identified a TNF-insulin-GATA6-CCRL2 circuit. Our findings regarding this recently de-orphaned receptor suggest that CCRL2 transcript is modulated in colonic epithelial cells in a GATA6-dependent manner in the context of TNF and insulin. Moreover, chemerine is secreted from colonic epithelial cells and that its bioavailability can be changed once bound to CCRL2. In addition, CCRL2$^{-/-}$ mice display enhanced tissue inflammation and immune cell infiltration [161].Taken together, the role of this GATA6 regulated, TNF-insulin crosstalk gene in disease contexts that both stimuli are involved such as inflammatory bowl disease and diabetes is an interesting topic of investigation. For example, one can investigate the changes in chemerine bioavailability upon single or cotreatment of TNF and insulin and whether this can affect immune cell infiltration. Colonic epithelium or endothelial cells are good candidate systems in which these questions can be pursued due to their documented TNF and insulin responsiveness and GSK3-GATA6 as well as chemerine-CCRL2 relevance [158, 162].

4.3 Summary

Many common diseases are caused by malfunction of more than one factor in the signalling network inside the cells. After close to a century of cell signalling research [15], we have a wealth of information about many signalling pathways in isolation. However, the number of developed efficacious therapeutics compared to our knowledge in the postgenomics era remains low. We believe that in order to take our understanding of cellular signalling to the next level, we need to study the pathways in the context of other pathways. Systems biology is well suited to handle this challenge by studying multiple pathways simultaneously. This network-based approach results in the accumulation of big data sets that need to be analysed. This challenge can be handled by computation, a tool that systems biologist are armed with.

With an interest in the cross-communication between signaling pathways within the signaling network, here we sought to study two pathways that malfunction in a number of common diseases. Taking a top-down approach here, we started with a signaling data compendium and a condition-matched transcriptomics data set. Then a data-derived tensor PLSR model was built. In addition, a careful inspection of the model complemented by bioinformatics enabled us to generate a crosstalk hypothesis (chapter 2). Notably, the hypothesis involved molecules with no direct quantitative measurement within the original data sets. The hypothesis was then tested in vitro using a collection of molecular biology techniques. Taken together, this work demonstrate that combinatorial approaches which combine systems data collection and computational methods with reductionist molecular biology techniques are advantageous in making exciting, new discoveries in the big data era.

Appendix A

Methods

A.1 Cell culture

HT-29, 293T, DLD-1, and HCT-8 cells were obtained from the American Type Culture Collection and cultured according to their recommendations. The 5E clone of MCF10A cells was cultured as described previously [148]. AC16 cells [147] were purchased from M. Davidson (Columbia University) and cultured in Dulbecco's modified Eagle's medium/F-12 medium (Life Technologies) with 12.5% tetracycline-free fetal bovine serum (Clontech) and penicillin-streptomycin (Gibco).

A.2 Cell stimulation

HT-29 cells were plated at 50,000 cells/cm^2 for 24 hours, sensitized with 200 U/ml human IFNγ for 24 hours (Roche), and then treated with 100 ng/ml TNF (Peprotech), 500 ng/ml insulin (Sigma), or both for the indicated times. HT-29 cells engineered to express GATA6$_L$ inducibly were treated with 1 μg/ml DOX for 24 hours (overexpression) or 48 hours (addback) before cytokine stimulation.

A.3 Plasmids

Wild-type GATA6$_L$ was amplified by PCR from HT-29 RNA that had been reverse transcribed with a GATA6-specific primer (CAAAAGCAGACACGAGTGGA). An N-terminal 3×FLAG tag and a C-terminal 3×AU1 tag were added by PCR before cloning into the BamHI and SalI sites of pBabe puro [163] the MfeI and SpeI sites of pEN_TTmiRc2 [137]. The pEN_TT donor vector containing GATA6$_L$ was then recombined with the pSLIKneo destination vector [137] by using LR clonase (Invitrogen). The shGATA6 sequence (CCCAGACCACTTGCTATGAAA, #TRCN0000005390 from The RNAi Consortium) was cloned into tet-pLKO puro [164] as described previously [130]. S33A, T34A, S37A, and 3SA point mutants were prepared by site-directed mutagenesis (QuikChange XL II, Agilent). RNAi-resistant mutants of wild-type and S37A GATA6$_L$ were prepared by introducing four silent mutations into the sequence targeted by shGATA6, which replace with rare mammalian codons that would minimize ectopic expression. The phosphorylation- and degradation-resistant IκBα super-repressor plasmid has been previously described [165]. All DNA constructs were verified by sequencing.

A.4 Production and purification of phospho-GATA6$_L$ (Ser^{37}) antibody

The peptide sequence Ac-CREPSTPPpSPIS-amide was conjugated to keyhole limpet hemocyanin and used to immunize rabbits according to the manufacturer's recommendations (Covance). Serum samples were tested by immunoblotting with positive and negative controls for phospho-GATA6$_L$ (Ser^{37}). Serum pooled from the production and terminal bleeds was negatively selected on a CREPSTPP$\underline{\text{S}}$PIS peptide-conjugated N-hydroxysuccinimide (NHS) sepharose column. The bound IgG was eluted as the nonphospho-GATA6$_L$ custom antibody while the flow through was exposed to a second CREPSTPP$\underline{\text{pS}}$PIS peptide-conjugated NHS sepharose column. The bound IgG was eluted as the phospho-GATA6$_L$ (Ser^{37}) antibody and used for detection by immunoblotting.

A.5 Lentiviral packaging and transduction

Lentiviruses were prepared in HEK293T cells (ATCC) by calcium phosphate transfection of the lentivector together with psPAX2 and pMD.2G (Addgene). Lentiviral transduction of HT-29 cells was performed as described previously [129]. Transduced cells were selected in growth medium containing 2 μg/ml puromycin or 600 μg/ml G418 until control plates had cleared. For addback experiments, viral titers were reduced to ensure single-virion transductants that matched the endogenous protein abundance as closely as possible.

A.6 Microarray profiling

HT-29 cells were plated at 50,000 cells/cm^2 for 24 hours and sensitized with 200 U/ml IFNγ (Roche) for 24 hours before stimulation with 0, 5, or 100 ng/ml TNF; 0, 1, or 100 ng/ml EGF; and 0, 5, or 500 ng/ml insulin for 4, 8, or 16 hours. RNA isolation was performed with the RNeasy Mini Kit (Qiagen), and integrity of purified RNA was confirmed on a Bioanalyzer (Agilent). Preparation of labeled complementary RNA, hybridization to GeneChip Human Genome U133A Arrays (Affymetrix), microarray scanning, and microarray processing were performed as previously described [166].

For inducible GATA6$_L$ overexpression, stably transduced HT-29 cells were plated at 50,000 cells/cm^2 for 24 hours, induced with 1 μg/ml DOX and sensitized with 200 U/ml IFNγ (Roche) for 24 hours before stimulation with 100 ng/ml TNF for two hours. RNA was purified as described above and amplified with the Illumina Total Prep-96 RNA Amplification Kit (Life Technologies) before hybridization to a HumanHT-12 v4 Expression BeadChip.

A.7 Hierarchical and CLICK clustering

One-way hierarchical clustering of the signaling and transcriptomic compendia was performed in MATLAB with the clustergram function using Euclidean distance and

Ward's linkage after row standardization. CLICK clustering was performed as described [112] with the default homogeneity parameter.

A.8 Tensor PLSR

Tensor PLSR was performed in MATLAB with Version 2.02 of the NPLS Toolbox [167]. The signaling compendium was structured by stimulus condition (mode 1), time point (mode 2), and measured signal (mode 3). The transcriptomic profiles were structured by stimulus condition (mode 1), time point (mode 2), and CLICK gene cluster (mode 3). Both data tensors were mean centered along mode 1 and variance scaled along modes 2 and 3 before calculation of latent variables [100]. The scores and time weights of the fourth latent variable of the signaling tensor were both multiplied by -1 to improve model interpretability. Randomized models were constructed in MATLAB with the shufflematrix function applied within each stimulus condition before preprocessing and calculation of latent variables.

A.9 Bioinformatic analyses of crosstalk cluster

The 20 transcripts from Cluster #9 confirmed present by qRT-PCR were submitted to three promoter-analysis algorithms. First, the proximal promoter of each transcript (defined as 2000 bp upstream and 500 bp downstream of the transcription start site) was collected from NCBI and used as an input set for MEME, which uses expectation maximization to define recurrent motifs in a set of sequences [168]. The top five enriched motifs were searched against a database of 843 binding specificities [169] using TOMTOM [170] to identify known transcription factor recognition sequences. A GATA motif was also enriched when using 2000 bp of upstream sequence alone or 1500 bp of upstream sequence and 500 bp of downstream sequence.

Expression-verified Cluster #9 transcripts were additionally analyzed with DiRE [120], which uses interspecies sequence conservation to define motifs that are searched against the TRANSFAC 10.2 database of roughly 400 transcription factor binding motifs. In DiRE, the occurrence metric reflects the overall frequency of a conserved binding motif in the input dataset, whereas the importance metric reflects the specificity of the binding motif to the input dataset compared to a background dataset of 5000 randomly selected genes. The top 20 motifs based on occurrence were used as the DiRE predictions.

Last, X2K [132] was used to identify bioinformatic connections between Cluster #9 transcripts and signaling pathways. X2K integrates the ChEA database [171] of transcription factor binding sites detected by chromatin immunoprecipitation, the JASPAR and TRANSFAC position weight matrices, as well as various protein-protein interaction and kinase-substrate databases to connect kinase signaling events to gene expression patterns. The top 20 transcription factors linked to signaling and Cluster #9 transcripts in a 2011 analysis were used as the X2K predictions.

A.10 Quantitative RT-PCR

RNA from cultured cells was isolated with the RNeasy Plus Mini kit (Qiagen) according to the manufacturer's protocol. First-strand cDNA synthesis and qRT-PCR were performed as described [77]. Parental HT-29 samples were normalized to the geometric mean of *GAPDH, HINT1, PPIA*, and *PRDX6*. GATA6$_L$ addback samples were normalized to the geometric mean of *GAPDH, HINT1, PPIA, PRDX6, B2M*, and *GUSB*. Primer sequences are available in table A.1.

A.11 Mass spectrometry

HT-29 cells stably expressing doxycycline-inducible 3×FLAG-GATA6$_L$ were induced with 1 μg/ml doxycycline for 24 hours and lysed in Nonidet P-40 (NP-40) lysis buffer plus protease and phosphatase inhibitors [93]. 60 mg of protein extract in 6 ml volume was first cleared with 50 μl of mouse IgG-agarose beads (Sigma) for one hour at 4°C on a nutator.

The cleared lysates were subjected to immunopurification using 80 μl of anti-FLAG M2 affinity gel (Sigma) for 34 hours followed by two washes with NP-40 lysis buffer, one wash with 500 mM NaCl, and one wash with Tris-buffered saline. Immunoprecipitates were eluted with 500 ng/ml 3×FLAG peptide (Sigma) in 100 μl for 30 minutes at 4°C on a nutator. The eluate was concentrated using an Amicon ultra centrifugal filter (Millipore), and samples were prepared in dithiothreitol-containing Laemmli sample buffer and separated by SDS PAGE on an 8% polyacrylamide gel

Table A.1: qRT-PCR primer sequences.

Human transcript	Forward sequence	Reverse sequence
ARL14	GCAATGATCACTGGGAATGA	CATAGGAGCACGGTGGAAAT
B2M	TTAGAGGTGGGGAGCAGAGA	TCCCCCAAATTCTAAGCAGA
BHLHE40	TCTCAAAGGCCTAACCAAGC	TGCATTTTGAAAAGCTGCTG
C3ORF52	ATTGCCTGTTGTGAGGGAAC	AGTCTGAGGCAGACACTCTGG
CCNG2	AGAGGCCATGTGGAGAGAGA	CCAAAACCTCGTGGCTTAAA
CCNL1	AAAAACCATGGTCAGGTTCAA	CACATGCAGACAAACCAGTGT
CCRL2	CTTCGCCTCCTACCACTTGT	CCAGGGTTTGGAGTTTGATG
CHMP2	GGTGAAATTGCCTTGGTATCT	AAAATGTCCACCATCCCAAA
CYP3A5	AAGAAAAGTCGCCTCAACGA	GAACACTGCTGGTGGTTTCA
DUSP6	CCAAATCATGGGCTCACTTT	CCATGCTCACACACACACAC
F3	GGGCTGACTTCAATCCATGT	GAAGGTGCCCAGAATACCAA
GAPDH	AACGTGTCAGTGGTGGACCT	TCGCTGTTGAAGTCAGAGGA
GATA1	ACTGCCCATCTCTACCAAGG	CAGAGACTTGGGTTGTCCAG
GATA2	AGCCTGTCTTGTCAGGTGGA	GCAAAGCTCCAACCTTGTGT
GATA3	ATTCAGTTGGCCTAAGGTGGT	GCACGCTGGTAGCTCATACA
GATA4	TCCTAGCCCTTGGTCAGATG	CAGCAATGTGCAGAGGAGAA
GATA5	GTCTGGTAGCCCTGTCCTGA	CTGTGGCCTCCTGACATACA
GATA6	GCCTTGCCTGCTATGGAATA	TGTGTACCAAATGGCCTCAA
GUSB	AACCAAAAAGTGCAGCGTTC	TTGTTCTGCTGCTGTGGAAG
HINT1	GCCTTGCTTTCCATGACATT	CCTTATTCAGGCCCAGATCA
INPP1	AATGGGACTCTTGTGCTGCT	CCGCTTCCTGGATCTGTATG
KLK10	ATCAGCTGGGTTTGTCATCC	CAGCTCCTTGAGGGTAGTGC
KRT20	GAACGCCAGAACAACGAATA	CGACCTTGCCATCCACTACT
LGALS8	TTGGGCACATTTTCACAGAA	TCCCTTTCAGGTTCGATTTG
MBNL2	CAAAAATCACCAATATCCAAGACA	CCAGTGGTATGGCTGAAAGAA
MCL1	GCCCATCTCAGAGCCATAAG	GATTTGGCAGACAGGCTTTT
MET	GGAAACATCCCATCAACAGG	GCTGCAGGTATAGGCAGTGA
NR4A2	TTGAAGAAGGCAAAGGCTTG	GCAAGTGCCTATAACATTTTC
PLAU	GGGTGGTCCTGACTCAACAT	TTGGCTAAGCTCCCTCAAGA
PPIA2	TACAGTGCTTGCTGGCAGTT	AATTGCCCAACACACCAAAT
PRDX6	CGTGTGGTGTTTGTTTTTGG	CTTCTTCAGGGATGGTGGGA
SNX16	AGTGAGTTTGCCTCTTTCAGAAT	TGCAATGTCAGGAGTCAAGC

followed by coomassie brilliant blue staining. The stained bands were cut and subsequently reduced, alkylated and digested with trypsin, chymotrypsin, or pepsin. Peptides from each enzymatic digestion were acrylamide extracted and subjected to LC-MS on a Thermo Electron Orbitrap Velos ETD mass spectrometer system. The data were analyzed using the Sequest search algorithm against the IPI Human Proteome Database and the predicted GATA6 protein sequence. Full mass spectrometry details are available in the Supplementary text.

A.12 Immunoblotting

Quantitative immunoblotting was performed as described previously in detail [93] with primary antibodies recognizing the following proteins or epitopes: p-GATA6$_L$ (Ser^{37}, Covance, 1:1000 for crude antiserum and 1:500 after affinity purification), nonphospho-GATA6$_L$ (Ser^{37}, Covance, 1:500), GATA6 (D61E4, Cell Signaling Technology #5851, 1:2000), GATA6 (H-92, Santa Cruz Biotechnology #9055, 1:600), p-GS (Ser^{641}, Cell Signaling Technology #3891, 1:1000), GS (Cell Signaling Technology #3893, 1:1000), GSK3α (Cell Signaling Technology #9338, 1:1000), p-GSK3α (Ser^{21}, Cell Signaling Technology #9316, 1:1000), p-Akt (Ser^{473}, Cell Signaling Technology #4060, 1:1000), Akt (Cell Signaling Technology #9272, 1:1000), p-S6 (Ser$^{240/244}$, Cell Signaling Technology #5364, 1:1000), S6 (54D2, Cell Signaling Technology #2317, 1:1000), FLAG (M2, Sigma #F1804, 1:10,000), β-actin (Ambion #4302, 1:5000), vinculin (Millipore #05-386, 1:10,000), GAPDH (Ambion #4300, 1:20,000), tubulin

(Abcam #89984, 1:20,000), and p38 (C-20, Santa Cruz Biotechnology #535, 1:5000). Membrane blocking, antibody probing, and near-infrared fluorescence detection were performed as described [93], except for phospho-GATA6 (Ser^{37}) immunoblotting, where blocking with 5% nonfat skim milk and use of Tris-buffered saline buffers was required.

Phos-tag immunoblotting was performed on a 6% polyacrylamide gel containing 10 μM Phos-tag acrylamide AAL-107 (Wako Chemical) and 0.1 μM MnCl2. Gels were run with Wide-view prestained protein markers under constant current (40 mA) for 170 minutes. Before electrophoretic transfer, gels were incubated with 1 mM EDTA in modified Towbin's transfer buffer [93] for 15 minutes. Membrane blocking, antibody probing, and near-infrared fluorescence detection were then performed as described [93].

For Gaussian mixture modeling of GATA6$_L$ forms, raw 16-bit pixel intensities were integrated horizontally across each lane and then plotted along the vertical dimension. Using the fit function in MATLAB, the vertical trace was fit by nonlinear least-squares to the following function:

$$f(x) = b + w_1 \exp\left(-\frac{x-\mu_1}{2\sigma^2}\right) + w_2 \exp\left(-\frac{x-\mu_2}{2\sigma^2}\right)$$

where f(x) is height of the vertical trace as a function of the vertical position x, b is a fixed background, w$_1$ and w$_2$ are the relative weights of the two bands, μ1 and μ2 are the mean vertical positions of the two bands, and σ2 is a shared variance for the two bands. Normalized versions of w_1 and w_2 were taken as the relative band densities for the two forms.

A.13 p-GATA6$_L$ (Ser^{37}) immunoprecipitation

HT-29 cells were plated at 75,000 cells/cm^2, pretreated with human IFNγ for 24 hours (Roche), and then treated with 100 ng/ml TNF (Peprotech), 500 ng/ml insulin (Sigma), or both for one hour. Cells were lysed in NP-40 lysis buffer (51) supplemented with 10 mM sodium pyrophosphate and 30 mM sodium fluoride, and $\backsim$4 mg of cellular extract (adjusted according to total GATA6$_L$ abundance based on immunoblotting) was incubated with 10 μl of p-GATA6$_L$ (Ser^{37}) antiserum overnight on a nutator at 4°C. The following day, 30 μl Protein A/G Plus UltraLink resin (Thermo) was added to the immune complexes for 1 hour on a nutator at 4°C. Beads were washed twice with ice-cold supplemented NP-40 lysis buffer and twice with ice-cold PBS before elution in Laemmli sample buffer [172].

A.14 Chromatin immunoprecipitation

Five million wild-type and S37A GATA6$_L$-addback HT-29 cells were seeded in 10-cm culture plates for 24 hours before inducing knockdown-addback with 1 μg/ml doxycycline for 48 hours. Cells were fixed for 710 minutes by adding a 37% formaldehyde stock to the culture medium to a final concentration of 1%. Fixation was quenched with 1/20 volume of 2.5 M glycine for 710 minutes at room temperature. Cells were washed twice with cold PBS, scraped into 1 ml PBS, and centrifuged at 400 relative centrifugal force (rcf) for 3 minutes. The cell pellets from four 10-cm plates were combined and lysed in ChIP lysis buffer [129] to a final volume of 1.5 ml. Lysates were incubated on ice for 10 minutes and then sonicated using a Branson digital sonifier for 5 minutes at 40% amplitude with 0.7 sec "on" and 1.3 sec "off" pulse cycles. After centrifugation at 14,000 rcf for 20 minutes, the supernatant was collected, and 20 μl of the soluble chromatin was retained as the input fraction. Soluble chromatin was diluted 10-fold in dilution buffer [129], precleared with 100 μl mouse IgG-conjugated agarose beads (Sigma) for four hours at 4°C with constant agitation, and then incubated with 100 μl anti-FLAG M2 affinity gel (Sigma) or mouse IgG-conjugated agarose beads overnight at 4°C with constant agitation. Agarose beads were collected and washed as previously described [129]. DNA from the beads and the input fraction was eluted by reversing methylene cross-links with 500 μl elution buffer [129] at 65°C for five hours. Samples were then treated with 100 μg/ml RNase for 30 minutes at 37°C and 200 μg/ml proteinase K for 90 minutes at 50°C, followed by phenol-chloroform extraction. The aqueous fraction was ethanol-precipitated, washed once in 70% ethanol, air dried, and dissolved in nuclease-free water. The samples were diluted tenfold in nuclease-free water and quantified by PCR with primers designed for proximal promoter regions of selected crosstalk genes [156]. Primer sequences are available in table A.2.

Table A.2: ChIP primer sequences.

Genomic locus	Forward sequence	Reverse sequence
CCRL2 locus #1	AGGTCACAGGGAAATCAAAGG	GAGGGGCACGAGAAACCA
CCRL2 locus #2	GACATCACTTCCTTGCTACCAC	GTCTGAAAGCAACCGGGAAG
CCRL2 locus #3	CACACAGTGCCCCTCAAAA	AGCAGGGAGGAAGGTATGTG
PLAU locus #1	CAGTTTATGCCCTAGGACTTTGT	GTTTGTTTGATGGTGCTATCAGA
PLAU locus #2	TCTGATAGCACCATCAAACAAAC	CCCAGAAGCACAGACAGAAAA
PLAU locus #3	CGGACTTTAAACCAAGCTGC	TGGCAGTCCCCAGATATCAC
ARL14 locus #1	TTCAAGGAAGCAAGCGTGG	AAGCCTGGTTCTCTTCCTGT
ARL14 locus #2	TGGTCCGCAGAATCCTTAGA	TCCACATCATCCCGTTGTCA
ARL14 locus #3	CAATGTGGCGTCTTGGATGG	ACTCTGCACCTAAGTTCCCC
ARL14 locus #4	TGGACAAGTTGCTAGCTTGC	TCCTTGTGAAATTTATGCCTGCT
ARL14 locus #5	TCAAAACATGTAATCCCAGTCCT	CCATCCAAGACGCCACATTG
CYP3A5 locus #1	CCCCATTTATCACTAACCCACC	CATCCATTCACAATCCCAGCA
CYP3A5 locus #2	CCCAGGAGAAACAAAATGGCT	TCATGAACTCCTGAACTCAAACA
CYP3A5 locus #3	GGTGGGAGCGAGACTTTGTC	TCATATGCAATTGAGCCAAAACT
CYP3A5 locus #4	GGAAAACAGCTGTAGACGCT	GGAGACAGATGAGAAAACAGGC
CYP3A5 locus #7	ACCTAGCCTTCTGAGAGGATTT	GATTTTCCATAGGCGTTACTCTT

A.15 Statistical analysis

Microarray data were analyzed by four-way analysis of variance (factors: TNF, EGF, insulin, and time) in MATLAB with the anovan function at a false-discovery rate of 5%. qRT-PCR data of Cluster #9 genes were analyzed by four-way analysis of variance (factors: transcript, TNF, insulin, and time) or, for GATA6$_L$ addback, by five-way analysis of variance (factors: GATA6$_L$ genotype, transcript, TNF, insulin, and time) in MATLAB with the anovan function after log transformation. Half-lives of GATA6$_L$ forms were estimated by nonlinear least-squares curve fitting to the following function:

$$g(t) = c\exp\left(-\frac{\ln(2)}{\tau_{1/2}}t\right) + b$$

where g(t) is the relative band intensity as a function of time t, c is the scaling coefficient, b is a fixed background, and $\tau 1/2$ is the half-life. Differences in means were assessed by Welch's t test, and differences in geometric means were assessed by Welch's t test after log transformation. One- or two-sidedness was based on prior evidence or expectation for a directional change. Tests for enrichment were performed by binomial test. Differences between immunoblotting time courses were assessed by two-way analysis of variance.

Bibliography

[1] Yong Li. Pathway Crosstalk Network. In Sangdun Choi, editor, *Systems Biology for Signaling Networks*, Systems Biology, pages 491–504. Springer New York, 2010. DOI: 10.1007/978-1-4419-5797-9_20.

[2] Alfred G. Gilman, Melvin I. Simon, Henry R. Bourne, Bruce A. Harris, Rochelle Long, Elliott M. Ross, James T. Stull, Ronald Taussig, Adam P. Arkin, Melanie H. Cobb, Jason G. Cyster, Peter N. Devreotes, James E. Ferrell, David Fruman, Michael Gold, Arthur Weiss, Michael J. Berridge, Lewis C. Cantley, William A. Catterall, Shaun R. Coughlin, Eric N. Olson, Temple F. Smith, Joan S. Brugge, David Botstein, Jack E. Dixon, Tony Hunter, Robert J. Lefkowitz, Anthony J. Pawson, Paul W. Sternberg, Harold Varmus, Shankar Subramaniam, Robert S. Sinkovits, Joshua Li, Dennis Mock, Yuhong Ning, Brian Saunders, Paul C. Sternweis, Donald Hilgemann, Richard H. Scheuermann, Dianne DeCamp, Robert Hsueh, Keng-Mean Lin, Yan Ni, William E. Seaman, Paul C. Simpson, Timothy D. O'Connell, Tamara Roach, Sangdun Choi, Pamela Eversole-Cire, Iain Fraser, Marc C. Mumby, Yingming Zhao, Deirdre Brekken, Hongjun Shu, Tobias Meyer, Grischa Chandy, Won Do Heo, Jen Liou, Nancy O'Rourke, Mary Verghese, Susanne M. Mumby, Heping Han, H. Alex Brown, Jeffrey S. Forrester, Pavlina Ivanova, Stephen B. Milne, Patrick J. Casey, T. Kendall Harden, John Doyle, Martha L. Gray, Stephen

Michnick, Martin A. Schmidt, Mehmet Toner, Roger Y. Tsien, Madhusudan Natarajan, Rama Ranganathan, and Gilberto R. Sambrano. Overview of the Alliance for Cellular Signaling. *Nature*, 420(6916):703–706, December 2002.

[3] Kanae Oda, Yukiko Matsuoka, Akira Funahashi, and Hiroaki Kitano. A comprehensive pathway map of epidermal growth factor receptor signaling. *Molecular Systems Biology*, 1:2005.0010, 2005.

[4] Kanae Oda and Hiroaki Kitano. A comprehensive map of the toll-like receptor signaling network. *Molecular Systems Biology*, 2, April 2006.

[5] Madhusudan Natarajan, Keng-Mean Lin, Robert C. Hsueh, Paul C. Sternweis, and Rama Ranganathan. A global analysis of cross-talk in a mammalian cellular signalling network. *Nature Cell Biology*, 8(6):571–580, June 2006.

[6] Neil D. Perkins. Integrating cell-signalling pathways with NF-B and IKK function. *Nature Reviews Molecular Cell Biology*, 8(1):49–62, January 2007.

[7] Sumito Ogawa, Jean Lozach, Chris Benner, Gabriel Pascual, Rajendra K. Tangirala, Stefan Westin, Alexander Hoffmann, Shankar Subramaniam, Michael David, Michael G. Rosenfeld, and Christopher K. Glass. Molecular Determinants of Crosstalk between Nuclear Receptors and Toll-like Receptors. *Cell*, 122(5):707–721, September 2005.

[8] A. F. Candia, T. Watabe, S. H. Hawley, D. Onichtchouk, Y. Zhang, R. Derynck, C. Niehrs, and K. W. Cho. Cellular interpretation of multiple TGF-beta signals: intracellular antagonism between activin/BVg1 and BMP-2/4 signaling mediated by Smads. *Development*, 124(22):4467–4480, November 1997.

[9] Manash S. Chatterjee, Jeremy E. Purvis, Lawrence F. Brass, and Scott L. Diamond. Pairwise agonist scanning predicts cellular signaling responses to combinatorial stimuli. *Nature Biotechnology*, 28(7):727–732, July 2010.

[10] Naama Geva-Zatorsky, Erez Dekel, Ariel A. Cohen, Tamar Danon, Lydia Cohen, and Uri Alon. Protein Dynamics in Drug Combinations: a Linear Superposition of Individual-Drug Responses. *Cell*, 140(5):643–651, March 2010.

[11] Robert C. Hsueh, Madhusudan Natarajan, Iain Fraser, Blake Pond, Jamie Liu, Susanne Mumby, Heping Han, Lily I. Jiang, Melvin I. Simon, Ronald Taussig, and Paul C. Sternweis. Deciphering signaling outcomes from a system of complex networks. *Science Signaling*, 2(71):ra22, 2009.

[12] Kevin A. Janes. Paring down signaling complexity. *Nature Biotechnology*, 28(7):681–682, July 2010.

[13] Robin Donaldson and Muffy Calder. Modelling and analysis of biochemical signalling pathway cross-talk. *EPTCS*, 19:40–54, February 2010.

[14] Xing Guo and Xiao-Fan Wang. Signaling cross-talk between TGF-/BMP and other pathways. *Cell Research*, 19(1):71–88, January 2009.

[15] A. Harvey. *Cancer Cell Signaling*. John Wiley & Sons, 2013.

[16] Dirk Brenner, Heiko Blaser, and Tak W. Mak. Regulation of tumour necrosis factor signalling: live or let die. *Nature Reviews Immunology*, 15(6):362–374, June 2015.

[17] Herbert Fluhr, Stefanie Krenzer, Gerburg M. Stein, Bjrn Stork, Margarita Deperschmidt, Diethelm Wallwiener, Sebastian Wesselborg, Marek Zygmunt, and Peter Licht. Interferon-λ and tumor necrosis factor-α sensitize primarily

resistant human endometrial stromal cells to fas-mediated apoptosis. *Journal of Cell Science*, 120, December 2007.

[18] Linda F. Watkins, Laurie R. Lewis, and Alan E. Levine. Characterization of the synergistic effect of insulin and transferrin and the regulation of their receptors on a human colon carcinoma cell line. *International Journal of Cancer*, 45(2):372–375, February 1990.

[19] J. Bjrk, J. Nilsson, R. Hultcrantz, and C. Johansson. Growth-Regulatory Effects of Sensory Neuropeptides, Epidermal Growth Factor, Insulin, and Somatostatin on the Non-Transformed Intestinal Epithelial Cell Line IEC-6 and the Colon Cancer Cell Line HT 29. *Scandinavian Journal of Gastroenterology*, 28(10):879–884, January 1993.

[20] Hong Ruan, Philip D. G. Miles, Christine M. Ladd, Kenneth Ross, Todd R. Golub, Jerrold M. Olefsky, and Harvey F. Lodish. Profiling gene transcription in vivo reveals adipose tissue as an immediate target of tumor necrosis factor-alpha: implications for insulin resistance. *Diabetes*, 51(11):3176–3188, November 2002.

[21] J. M. Stephens and P. H. Pekala. Transcriptional repression of the GLUT4 and C/EBP genes in 3t3-L1 adipocytes by tumor necrosis factor-alpha. *Journal of Biological Chemistry*, 266(32):21839–21845, November 1991.

[22] Jeffery D. Molkentin. The Zinc Finger-containing Transcription Factors GATA-4, -5, and -6 UBIQUITOUSLY EXPRESSED REGULATORS OF TISSUE-SPECIFIC GENE EXPRESSION. *Journal of Biological Chemistry*, 275(50):38949–38952, December 2000.

[23] Jonathan Chou, Sylvain Provot, and Zena Werb. GATA3 in development and cancer differentiation: cells GATA have it! *Journal of Cellular Physiology*, 222(1):42–49, January 2010.

[24] Marcela Rosas, Luke C. Davies, Peter J. Giles, Chia-Te Liao, Bashar Kharfan, Timothy C. Stone, Valerie B. O'Donnell, Donald J. Fraser, Simon A. Jones, and Philip R. Taylor. The transcription factor Gata6 links tissue macrophage phenotype and proliferative renewal. *Science (New York, N.Y.)*, 344(6184):645–648, May 2014.

[25] Larysa Pevny, M. Celeste Simon, Elizabeth Robertson, William H. Klein, Shih-Feng Tsai, Vivette D'Agati, Stuart H. Orkin, and Frank Costantini. Erythroid differentiation in chimaeric mice blocked by a targeted mutation in the gene for transcription factor GATA-1. *Nature*, 349(6306):257–260, January 1991.

[26] Fong-Ying Tsai, Gordon Keller, Frank C. Kuo, Mitchell Weiss, Jianzhou Chen, Margery Rosenblatt, Frederick W. Alt, and Stuart H. Orkin. An early haematopoietic defect in mice lacking the transcription factor GATA-2. *Nature*, 371(6494):221–226, September 1994.

[27] Kim-Chew Lim, Ganesh Lakshmanan, Susan E. Crawford, Yi Gu, Frank Grosveld, and James Douglas Engel. Gata3 loss leads to embryonic lethality due to noradrenaline deficiency of the sympathetic nervous system. *Nature Genetics*, 25(2):209–212, June 2000.

[28] J. D. Molkentin, Q. Lin, S. A. Duncan, and E. N. Olson. Requirement of the transcription factor GATA4 for heart tube formation and ventral morphogenesis. *Genes & Development*, 11(8):1061–1072, April 1997.

[29] C. T. Kuo, E. E. Morrisey, R. Anandappa, K. Sigrist, M. M. Lu, M. S. Parmacek, C. Soudais, and J. M. Leiden. GATA4 transcription factor is required for ventral morphogenesis and heart tube formation. *Genes & Development*, 11(8):1048–1060, April 1997.

[30] Edward E. Morrisey, Zhihua Tang, Kirsten Sigrist, Min Min Lu, Fang Jiang, Hon S. Ip, and Michael S. Parmacek. GATA6 regulates HNF4 and is required for differentiation of visceral endoderm in the mouse embryo. *Genes & Development*, 12(22):3579–3590, November 1998.

[31] M. Koutsourakis, A. Langeveld, R. Patient, R. Beddington, and F. Grosveld. The transcription factor GATA6 is essential for early extraembryonic development. *Development*, 126(9):723–732, May 1999.

[32] T. Yoshida, R. Sato, S. Mahmood, S. Kawasaki, M. Futai, and M. Maeda. GATA-6 DNA binding protein expressed in human gastric adenocarcinoma MKN45 cells. *FEBS letters*, 414(2):333–337, September 1997.

[33] E. Suzuki, T. Evans, J. Lowry, L. Truong, D. W. Bell, J. R. Testa, and K. Walsh. The human GATA-6 gene: structure, chromosomal location, and regulation of expression by tissue-specific and mitogen-responsive signals. *Genomics*, 38(3):283–290, December 1996.

[34] I. C. Huggon, A. Davies, C. Gove, G. Moscoso, C. Moniz, Y. Foss, F. Farzaneh, and P. Towner. Molecular cloning of human GATA-6 DNA binding protein: high levels of expression in heart and gut. *Biochimica Et Biophysica Acta*, 1353(2):98–102, August 1997.

[35] Yongmei Jiang and Todd Evans. TheXenopusGATA-4/5/6 Genes Are Associated with Cardiac Specification and Can Regulate Cardiac-Specific Transcrip-

tion during Embryogenesis. *Developmental Biology*, 174(2):258–270, March 1996.

[36] A. C. Laverriere, C. MacNeill, C. Mueller, R. E. Poelmann, J. B. Burch, and T. Evans. GATA-4/5/6, a subfamily of three transcription factors transcribed in developing heart and gut. *Journal of Biological Chemistry*, 269(37):23177–23184, September 1994.

[37] Gove Brewer, McNulty Davies, Koutsourakis Barrow, Pizzey Farzaneh, and Patient R. Bomford. The human and mouse gata-6 genes utilize two promoters and two initiation codons. *J. Biol. Chem.*, 274(53):38004–38016, December 1999.

[38] Mika Takeda, Kanako Obayashi, Ayako Kobayashi, and Masatomo Maeda. A unique role of an amino terminal 16-residue region of long-type GATA-6. *Journal of Biochemistry*, 135(5):639–650, May 2004.

[39] Nelly Khidekel and Linda C. Hsieh-Wilson. A 'molecular switchboard'–covalent modifications to proteins and their impact on transcription. *Organic & Biomolecular Chemistry*, 2(1):1–7, January 2004.

[40] A. J. Whitmarsh and R. J. Davis. Regulation of transcription factor function by phosphorylation. *Cellular and molecular life sciences: CMLS*, 57(8-9):1172–1183, August 2000.

[41] Jesper V. Olsen, Michiel Vermeulen, Anna Santamaria, Chanchal Kumar, Martin L. Miller, Lars J. Jensen, Florian Gnad, Jrgen Cox, Thomas S. Jensen, Erich A. Nigg, Sren Brunak, and Matthias Mann. Quantitative phosphoproteomics reveals widespread full phosphorylation site occupancy during mitosis. *Science Signaling*, 3(104):ra3, 2010.

[42] Yonghao Yu, Sang-Oh Yoon, George Poulogiannis, Qian Yang, Xiaoju Max Ma, Judit Villn, Neil Kubica, Gregory R. Hoffman, Lewis C. Cantley, Steven P. Gygi, and John Blenis. Phosphoproteomic Analysis Identifies Grb10 as an mTORC1 Substrate That Negatively Regulates Insulin Signaling. *Science*, 332(6035):1322–1326, June 2011.

[43] Mirita Franz-Wachtel, Stephan A. Eisler, Karsten Krug, Silke Wahl, Alejandro Carpy, Alfred Nordheim, Klaus Pfizenmaier, Angelika Hausser, and Boris Macek. Global detection of protein kinase D-dependent phosphorylation events in nocodazole-treated human cells. *Molecular & cellular proteomics: MCP*, 11(5):160–170, May 2012.

[44] Tingfang Yi, Bo Zhai, Yonghao Yu, Yoshikawa Kiyotsugu, Thomas Raschle, Manuel Etzkorn, Hee-Chan Seo, Michal Nagiec, Rafael E. Luna, Ellis L. Reinherz, John Blenis, Steven P. Gygi, and Gerhard Wagner. Quantitative phosphoproteomic analysis reveals system-wide signaling pathways downstream of SDF-1/CXCR4 in breast cancer stem cells. *Proceedings of the National Academy of Sciences*, 111(21):E2182–E2190, May 2014.

[45] Arminja N. Kettenbach, Devin K. Schweppe, Brendan K. Faherty, Dov Pechenick, Alexandre A. Pletnev, and Scott A. Gerber. Quantitative Phosphoproteomics Identifies Substrates and Functional Modules of Aurora and Polo-Like Kinase Activities in Mitotic Cells. *Sci. Signal.*, 4(179):rs5–rs5, June 2011.

[46] Noah Dephoure, Chunshui Zhou, Judit Villn, Sean A. Beausoleil, Corey E. Bakalarski, Stephen J. Elledge, and Steven P. Gygi. A quantitative atlas of

mitotic phosphorylation. *Proceedings of the National Academy of Sciences of the United States of America*, 105(31):10762–10767, August 2008.

[47] Y. Adachi, Y. Shibai, J. Mitsushita, W. H. Shang, K. Hirose, and T. Kamata. Oncogenic Ras upregulates NADPH oxidase 1 gene expression through MEK-ERK-dependent phosphorylation of GATA-6. *Oncogene*, 27(36):4921–4932, August 2008.

[48] Hironori Ushijima and Masatomo Maeda. cAMP-dependent proteolysis of GATA-6 is linked to JNK-signaling pathway. *Biochemical and Biophysical Research Communications*, 423(4):679–683, July 2012.

[49] Mollie L. Kelly, Artyom Astsaturov, Jennifer Rhodes, and Jonathan Chernoff. A Pak1/Erk Signaling Module Acts through Gata6 to Regulate Cardiovascular Development in Zebrafish. *Developmental Cell*, 29(3):350–359, May 2014.

[50] Kelly F. Benedict, Feilim Mac Gabhann, Robert K. Amanfu, Arvind K. Chavali, Erwin P. Gianchandani, Lydia S. Glaw, Matthew A. Oberhardt, Bryan C. Thorne, Jason H. Yang, Jason A. Papin, Shayn M. Peirce, Jeffrey J. Saucerman, and Thomas C. Skalak. Systems analysis of small signaling modules relevant to eight human diseases. *Ann Biomed Eng*, 39, 2011.

[51] H. Steven Wiley, Stanislav Y. Shvartsman, and Douglas A. Lauffenburger. Computational modeling of the EGF-receptor system: a paradigm for systems biology. *Trends in Cell Biology*, 13(1):43–50, January 2003.

[52] Takashi Nakakuki, Marc R. Birtwistle, Yuko Saeki, Noriko Yumoto, Kaori Ide, Takeshi Nagashima, Lutz Brusch, Babatunde A. Ogunnaike, Mariko Okada-Hatakeyama, and Boris N. Kholodenko. Ligand-specific c-Fos expression

emerges from the spatiotemporal control of ErbB network dynamics. *Cell*, 141(5):884–896, May 2010.

[53] Bree B. Aldridge, Julio Saez-Rodriguez, Jeremy L. Muhlich, Peter K. Sorger, and Douglas A. Lauffenburger. Fuzzy logic analysis of kinase pathway crosstalk in TNF/EGF/insulin-induced signaling. *PLoS computational biology*, 5(4):e1000340, April 2009.

[54] Julio Saez-Rodriguez, Leonidas G. Alexopoulos, Jonathan Epperlein, Regina Samaga, Douglas A. Lauffenburger, Steffen Klamt, and Peter K. Sorger. Discrete logic modelling as a means to link protein signalling networks with functional analysis of mammalian signal transduction. *Molecular Systems Biology*, 5:331, 2009.

[55] Kevin A. Janes and Douglas A. Lauffenburger. A biological approach to computational models of proteomic networks. *Current Opinion in Chemical Biology*, 10(1):73–80, February 2006.

[56] Kevin A. Janes and Michael B. Yaffe. Data-driven modelling of signal-transduction networks. *Nature Reviews. Molecular Cell Biology*, 7(11):820–828, November 2006.

[57] John G. Albeck, Gavin MacBeath, Forest M. White, Peter K. Sorger, Douglas A. Lauffenburger, and Suzanne Gaudet. Collecting and organizing systematic sets of protein data. *Nature Reviews Molecular Cell Biology*, 7(11):803–812, November 2006.

[58] Suzanne Gaudet, Kevin A. Janes, John G. Albeck, Emily A. Pace, Douglas A. Lauffenburger, and Peter K. Sorger. A compendium of signals and responses

triggered by prodeath and prosurvival cytokines. *Molecular & cellular proteomics: MCP*, 4(10):1569–1590, October 2005.

[59] Marco Vilela and Gaudenz Danuser. What's wrong with correlative experiments? *Nature Cell Biology*, 13(9):1011, September 2011.

[60] Karin J. Jensen and Kevin A. Janes. Modeling the latent dimensions of multivariate signaling datasets. *Physical Biology*, 9(4):045004, August 2012.

[61] Orly Alter. Genomic signal processing: from matrix algebra to genetic networks. *Methods in Molecular Biology (Clifton, N.J.)*, 377:17–60, 2007.

[62] P. T. Spellman, G. Sherlock, M. Q. Zhang, V. R. Iyer, K. Anders, M. B. Eisen, P. O. Brown, D. Botstein, and B. Futcher. Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. *Molecular Biology of the Cell*, 9(12):3273–3297, December 1998.

[63] Kevin A. Janes, John G. Albeck, Suzanne Gaudet, Peter K. Sorger, Douglas A. Lauffenburger, and Michael B. Yaffe. A systems model of signaling identifies a molecular basis set for cytokine-induced apoptosis. *Science (New York, N.Y.)*, 310(5754):1646–1653, December 2005.

[64] I. T. Jolliffe. *Principal Component Analysis.* 2002.

[65] Karl Pearson. LIII. *On lines and planes of closest fit to systems of points in space. Philosophical Magazine Series 6*, 2(11):559–572, November 1901.

[66] Harald Martens and M. Martens. *Multivariate Analysis of Quality: An Introduction.* John Wiley & Sons, February 2001.

[67] J. N. R. Jeffers. Two Case Studies in the Application of Principal Component Analysis. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 16(3):225–236, 1967.

[68] M. Crescenzi and A. Giuliani. The main biological determinants of tumor line taxonomy elucidated by a principal component analysis of microarray data. *FEBS letters*, 507(1):114–118, October 2001.

[69] Jie Hu, Jason W. Locasale, Jason H. Bielas, Jacintha O'Sullivan, Kieran Sheahan, Lewis C. Cantley, Matthew G. Vander Heiden, and Dennis Vitkup. Heterogeneity of tumor-induced gene expression changes in the human metabolic network. *Nature Biotechnology*, 31(6):522–529, June 2013.

[70] Jatin Misra, William Schmitt, Daehee Hwang, Li-Li Hsiao, Steve Gullans, George Stephanopoulos, and Gregory Stephanopoulos. Interactive exploration of microarray gene expression patterns in a reduced dimensional space. *Genome Research*, 12(7):1112–1120, July 2002.

[71] J. Schlens. A tutorial on principal component analysis: Derivation, discussion, and singular value decomposition, 2003.

[72] K. Pearson. On lines and planes of closest fit to systems of points in space. *Philosphical Magazine*, 2:559–572, 1901.

[73] O. Alter, P. O. Brown, and D. Botstein. Singular value decomposition for genome-wide expression data processing and modeling. *Proceedings of the National Academy of Sciences of the United States of America*, 97(18):10101–10106, August 2000.

[74] Pamela K. Kreeger, Roli Mandhana, Shannon K. Alford, Kevin M. Haigis, and Douglas A. Lauffenburger. RAS mutations affect tumor necrosis factor-induced apoptosis in colon carcinoma cells via ERK-modulatory negative and positive feedback circuits along with non-ERK pathway effects. *Cancer Research*, 69(20):8191–8199, October 2009.

[75] Ken S. Lau, Alwin M. Juchheim, Kimberly R. Cavaliere, Sarah R. Philips, Douglas A. Lauffenburger, and Kevin M. Haigis. In vivo systems analysis identifies spatial and temporal aspects of the modulation of TNF--induced apoptosis and proliferation by MAPKs. *Science Signaling*, 4(165):ra16, 2011.

[76] Michael J. Lee, Albert S. Ye, Alexandra K. Gardino, Anne Margriet Heijink, Peter K. Sorger, Gavin MacBeath, and Michael B. Yaffe. Sequential Application of Anticancer Drugs Enhances Cell Death by Rewiring Apoptotic Signaling Networks. *Cell*, 149(4):780–794, May 2012.

[77] Kathryn Miller-Jensen, Kevin A. Janes, Joan S. Brugge, and Douglas A. Lauffenburger. Common effector processing mediates cell-specific responses to stimuli. *Nature*, 448(7153):604–608, August 2007.

[78] Andrea R. Tentner, Michael J. Lee, Gerry J. Ostheimer, Leona D. Samson, Douglas A. Lauffenburger, and Michael B. Yaffe. Combined experimental and computational analysis of DNA damage signaling reveals context-dependent roles for Erk in apoptosis and G1/S arrest after genotoxic stress. *Molecular Systems Biology*, 8:568, 2012.

[79] Elsa M. Beyer and Gavin MacBeath. Cross-talk between receptor tyrosine kinase and tumor necrosis factor- signaling networks regulates apoptosis but

not proliferation. *Molecular & cellular proteomics: MCP*, 11(6):M111.013292, June 2012.

[80] Dhiraj Kumar, Ravichandran Srikanth, Helena Ahlfors, Riitta Lahesmaa, and Kanury V. S. Rao. Capturing cell-fate decisions from the molecular signatures of a receptor-dependent signaling response. *Molecular Systems Biology*, 3:150, 2007.

[81] Neil Kumar, Alejandro Wolf-Yadlin, Forest M. White, and Douglas A. Lauffenburger. Modeling HER2 effects on cell behavior from mass spectrometry phosphotyrosine data. *PLoS computational biology*, 3(1):e4, January 2007.

[82] Melissa L. Kemp, Lucia Wille, Christina L. Lewis, Lindsay B. Nicholson, and Douglas A. Lauffenburger. Quantitative network signal combinations downstream of TCR activation can predict IL-2 production response. *Journal of Immunology (Baltimore, Md.: 1950)*, 178(8):4984–4992, April 2007.

[83] Michael Dworkin, Sayak Mukherjee, Ciriyam Jayaprakash, and Jayajit Das. Dramatic reduction of dimensionality in large biochemical networks owing to strong pair correlations. *Journal of the Royal Society, Interface / the Royal Society*, 9(73):1824–1835, August 2012.

[84] Pamela K. Kreeger. Using partial least squares regression to analyze cellular response data. *Science Signaling*, 6(271):tr7, 2013.

[85] Paul Geladi and Bruce R. Kowalski. Partial least-squares regression: a tutorial. *Analytica Chimica Acta*, 185:1–17, January 1986.

[86] Anders Krogh. What are artificial neural networks? *Nature Viotechnology*, 26(2):195–197, 2006.

[87] William S. noble. What is a support vector machine? *Nature Viotechnology*, 24(12):1565–1567, 2006.

[88] Kevin A. Janes and Douglas A. Lauffenburger. Models of signalling networks what cell biologists can gain from them and give to them. *J Cell Sci*, 126:1913–21, 2013.

[89] Kevin A. Janes, Suzanne Gaudet, John G. Albeck, Ulrik B. Nielsen, Douglas A. Lauffenburger, and Peter K. Sorger. The response of human epithelial cells to tnf involves an inducible autocrine cascade. *cell*, 124(6):1225–39, March 2006.

[90] M. T. Abreu-Martin, A. Vidrich, D. H. Lynch, and S. R. Targan. Divergent induction of apoptosis and IL-8 secretion in HT-29 cells in response to TNF-alpha and ligation of Fas antigen. *Journal of Immunology (Baltimore, Md.: 1950)*, 155(9):4147–4154, November 1995.

[91] Kevin A. Janes, John G. Albeck, Lili X. Peng, Peter K. Sorger, Douglas A. Lauffenburger, and Michael B. Yaffe. A high-throughput quantitative multiplex kinase assay for monitoring information flow in signaling networks: application to sepsis-apoptosis. *Molecular & cellular proteomics: MCP*, 2(7):463–473, July 2003.

[92] Ulrik B. Nielsen, Mike H. Cardone, Anthony J. Sinskey, Gavin MacBeath, and Peter K. Sorger. Profiling receptor tyrosine kinase activation by using Ab microarrays. *Proceedings of the National Academy of Sciences of the United States of America*, 100(16):9330–9335, August 2003.

[93] Kevin A. Janes. An analysis of critical factors for quantitative immunoblotting. *Science Signaling*, 8(371):rs2, April 2015.

[94] Ido Amit, Ami Citri, Tal Shay, Yiling Lu, Menachem Katz, Fan Zhang, Gabi Tarcic, Doris Siwak, John Lahad, Jasmine Jacob-Hirsch, Ninette Amariglio, Nora Vaisman, Eran Segal, Gideon Rechavi, Uri Alon, Gordon B. Mills, Eytan Domany, and Yosef Yarden. A module of negative feedback regulators defines growth factor signaling. *Nature Genetics*, 39(4):503–512, April 2007.

[95] Sophie Rome, Karine Clment, Rmi Rabasa-Lhoret, Emmanuelle Loizon, Christine Poitou, Greg S. Barsh, Jean-Paul Riou, Martine Laville, and Hubert Vidal. Microarray profiling of human skeletal muscle reveals that insulin regulates approximately 800 genes during a hyperinsulinemic clamp. *The Journal of Biological Chemistry*, 278(20):18063–18068, May 2003.

[96] D. Fambrough, K. McClure, A. Kazlauskas, and E. S. Lander. Diverse signaling pathways activated by growth factor receptors induce broadly overlapping, rather than independent, sets of genes. *Cell*, 97(6):727–741, June 1999.

[97] Taesung Park, Sung-Gon Yi, Seungmook Lee, Seung Yeoun Lee, Dong-Hyun Yoo, Jun-Ik Ahn, and Yong-Sung Lee. Statistical tests for identifying differentially expressed genes in time-course microarray experiments. *Bioinformatics (Oxford, England)*, 19(6):694–703, April 2003.

[98] John D. Storey, Wenzhong Xiao, Jeffrey T. Leek, Ronald G. Tompkins, and Ronald W. Davis. Significance analysis of time course microarray experiments. *Proceedings of the National Academy of Sciences of the United States of America*, 102(36):12837–12842, September 2005.

[99] Bjorn L. Millard, Mario Niepel, Michael P. Menden, Jeremy L. Muhlich, and Peter K. Sorger. Adaptive informatics for multifactorial and high-content biological data. *Nature Methods*, 8(6):487–493, June 2011.

[100] R. Bro and A. K. Smilde. Centering and scaling in component analysis. *J. Chemometr*, 17:16–33, 2003.

[101] Rasmus Bro. PARAFAC. Tutorial and applications. *Chemometrics and Intelligent Laboratory Systems*, 38(2):149–171, October 1997.

[102] Tamara G. Kolda and Brett W. Bader. Tensor Decompositions and Applications. *SIAM Rev.*, 51(3):455–500, August 2009.

[103] Larsson Omberg, Gene H. Golub, and Orly Alter. A tensor higher-order singular value decomposition for integrative analysis of DNA microarray data from different studies. *Proceedings of the National Academy of Sciences of the United States of America*, 104(47):18371–18376, November 2007.

[104] Larsson Omberg, Joel R. Meyerson, Kayta Kobayashi, Lucy S. Drury, John F. X. Diffley, and Orly Alter. Global effects of DNA replication and DNA replication origin activity on eukaryotic gene expression. *Molecular Systems Biology*, 5:312, 2009.

[105] Preethi Sankaranarayanan, Theodore E. Schomay, Katherine A. Aiello, and Orly Alter. Tensor GSVD of patient- and platform-matched tumor and normal DNA copy-number profiles uncovers chromosome arm-wide patterns of tumor-exclusive platform-consistent alterations encoding for cell transformation and predicting ovarian cancer survival. *PloS One*, 10(4):e0121396, 2015.

[106] Karin J. Jensen, Farshid S. Garmaroudi, Jingchun Zhang, Jun Lin, Seti Boroomand, Mary Zhang, Zongshu Luo, Decheng Yang, Honglin Luo, Bruce M. McManus, and Kevin A. Janes. An ERK-p38 subnetwork coordinates host cell apoptosis and necrosis during coxsackievirus B3 infection. *Cell Host & Microbe*, 13(1):67–76, January 2013.

[107] Kevin A. Janes, Christian H. Reinhardt, and Michael B. Yaffe. Cytokine-induced signaling networks prioritize dynamic range over signal strength. *Cell*, 135(2):343–354, October 2008.

[108] Andrew Gordus, Jordan A. Krall, Elsa M. Beyer, Alexis Kaushansky, Alejandro Wolf-Yadlin, Mark Sevecka, Bryan H. Chang, John Rush, and Gavin MacBeath. Linear combinations of docking affinities explain quantitative differences in RTK signaling. *Molecular Systems Biology*, 5:235, 2009.

[109] Mario Niepel, Marc Hafner, Emily A. Pace, Mirra Chung, Diana H. Chai, Lili Zhou, Birgit Schoeberl, and Peter K. Sorger. Profiles of Basal and stimulated receptor signaling networks predict drug response in breast cancer lines. *Science Signaling*, 6(294):ra84, September 2013.

[110] Kevin A. Janes, Jason R. Kelly, Suzanne Gaudet, John G. Albeck, Peter K. Sorger, and Douglas A. Lauffenburger. Cue-signal-response analysis of TNF-induced apoptosis by partial least squares regression of dynamic multivariate data. *Journal of Computational Biology: A Journal of Computational Molecular Cell Biology*, 11(4):544–561, 2004.

[111] Rasmus Bro. Multiway calibration. multilinear pls. *J Chemometr*, 10(1):47–61, 1996.

[112] Roded Sharan, Adi Maron-Katz, and Ron Shamir. CLICK and EXPANDER: a system for clustering and visualizing gene expression data. *Bioinformatics (Oxford, England)*, 19(14):1787–1799, September 2003.

[113] Najeeb Halabi, Olivier Rivoire, Stanislas Leibler, and Rama Ranganathan. Protein sectors: evolutionary units of three-dimensional structure. *Cell*, 138(4):774–786, August 2009.

[114] John G. Albeck, John M. Burke, Bree B. Aldridge, Mingsheng Zhang, Douglas A. Lauffenburger, and Peter K. Sorger. Quantitative analysis of pathways controlling extrinsic apoptosis in single cells. *Molecular Cell*, 30(1):11–25, April 2008.

[115] D. A. Cross, D. R. Alessi, P. Cohen, M. Andjelkovich, and B. A. Hemmings. Inhibition of glycogen synthase kinase-3 by insulin mediated by protein kinase B. *Nature*, 378(6559):785–789, December 1995.

[116] Janice M. Knowlden, Helen E. Jones, Denise Barrow, Julia M. W. Gee, Robert I. Nicholson, and Iain R. Hutcheson. Insulin receptor substrate-1 involvement in epidermal growth factor receptor and insulin-like growth factor receptor signalling: implication for Gefitinib ('Iressa') response and resistance. *Breast Cancer Research and Treatment*, 111(1):79–91, September 2008.

[117] O. N. Ozes, L. D. Mayo, J. A. Gustin, S. R. Pfeffer, L. M. Pfeffer, and D. B. Donner. NF-kappaB activation by tumour necrosis factor requires the Akt serine-threonine kinase. *Nature*, 401(6748):82–85, September 1999.

[118] Jason A. Gustin, Osman N. Ozes, Hakan Akca, Roxana Pincheira, Lindsey D. Mayo, Qiutang Li, Javier Rivera Guzman, Chandrashekhar K. Korgaonkar, and David B. Donner. Cell type-specific expression of the IkappaB kinases determines the significance of phosphatidylinositol 3-kinase/Akt signaling to NF-kappa B activation. *The Journal of Biological Chemistry*, 279(3):1615–1620, January 2004.

[119] Wan-Nan U. Chen, Ronald L. Woodbury, Loel E. Kathmann, Lee K. Opresko, Richard C. Zangar, H. Steven Wiley, and Brian D. Thrall. Induced autocrine

signaling through the epidermal growth factor receptor contributes to the response of mammary epithelial cells to tumor necrosis factor alpha. *The Journal of Biological Chemistry*, 279(18):18488–18496, April 2004.

[120] Valer Gotea and Ivan Ovcharenko. DiRE: identifying distant regulatory elements of co-expressed genes. *Nucleic Acids Research*, 36(Web Server issue):W133–139, July 2008.

[121] Jiao-Yang Zheng, Jun-Jie Zou, Wen-Zhao Wang, Xiao-Yun Feng, Yong-Yuan Shi, Ying Zhao, Gang Jin, and Zhi-Min Liu. Tumor necrosis factor- increases angiopoietin-like protein 2 gene expression by activating Foxo1 in 3t3-L1 adipocytes. *Molecular and Cellular Endocrinology*, 339(1-2):120–129, June 2011.

[122] Matthew S. Hayden and Sankar Ghosh. Signaling to nf-kappab. *Genes Dev*, 18(18):2195–2224, September 2004.

[123] N. Mukaida, Y. Mahe, and K. Matsushima. Cooperative interaction of nuclear factor-kappa B- and cis-regulatory enhancer binding protein-like factor binding elements in activating the interleukin-8 gene by pro-inflammatory cytokines. *The Journal of Biological Chemistry*, 265(34):21128–21133, December 1990.

[124] S. K. Hansen, C. Nerlov, U. Zabel, P. Verde, M. Johnsen, P. A. Baeuerle, and F. Blasi. A novel complex between the p65 subunit of NF-kappa B and c-Rel binds to a DNA element involved in the phorbol ester induction of the human urokinase gene. *The EMBO journal*, 11(1):205–213, January 1992.

[125] S. R. Datta, A. Brunet, and M. E. Greenberg. Cellular survival: a play in three Akts. *Genes & Development*, 13(22):2905–2927, November 1999.

[126] Timothy L. Bailey, James Johnson, Charles E. Grant, and William S. Noble. The MEME Suite. *Nucleic Acids Research*, 43(W1):W39–49, July 2015.

[127] V. Korinek, N. Barker, P. J. Morin, D. van Wichen, R. de Weger, K. W. Kinzler, B. Vogelstein, and H. Clevers. Constitutive transcriptional activation by a beta-catenin-Tcf complex in APC-/- colon carcinoma. *Science (New York, N.Y.)*, 275(5307):1784–1787, March 1997.

[128] Roger K. Patient and James D. McGhee. The GATA family (vertebrates and invertebrates). *Current Opinion in Genetics & Development*, 12(4):416–422, August 2002.

[129] Lixin Wang, Joan S. Brugge, and Kevin A. Janes. Intersection of FOXO- and RUNX1-mediated gene expression programs in single breast epithelial cells during morphogenesis and tumor progression. *Proceedings of the National Academy of Sciences of the United States of America*, 108(40):E803–812, October 2011.

[130] Sameer S. Bajikar, Christiane Fuchs, Andreas Roller, Fabian J. Theis, and Kevin A. Janes. Parameterizing cell-to-cell regulatory heterogeneities via stochastic transcriptional profiles. *Proceedings of the National Academy of Sciences of the United States of America*, 111(5):E626–635, February 2014.

[131] Tobias Ehrenberger, Lewis C. Cantley, and Michael B. Yaffe. Computational prediction of protein-protein interactions. *Methods in Molecular Biology (Clifton, N.J.)*, 1278:57–75, 2015.

[132] Edward Y. Chen, Huilei Xu, Simon Gordonov, Maribel P. Lim, Matthew H. Perkins, and Avi Ma'ayan. Expression2kinases: mRNA profiling linked to multiple upstream regulatory layers. *Bioinformatics (Oxford, England)*, 28(1):105–111, January 2012.

[133] Peter V. Hornbeck, Indy Chabra, Jon M. Kornhauser, Elzbieta Skrzypek, and Bin Zhang. PhosphoSite: A bioinformatics resource dedicated to physiological protein phosphorylation. *Proteomics*, 4(6):1551–1561, June 2004.

[134] Q. Liang, L. J. De Windt, S. A. Witt, T. R. Kimball, B. E. Markham, and J. D. Molkentin. The transcription factors GATA4 and GATA6 regulate cardiomyocyte hypertrophy in vitro and in vivo. *The Journal of Biological Chemistry*, 276(32):30245–30253, August 2001.

[135] Xiaoping Yang, Jesse S. Boehm, Xinping Yang, Kourosh Salehi-Ashtiani, Tong Hao, Yun Shen, Rakela Lubonja, Sapana R. Thomas, Ozan Alkan, Tashfeen Bhimdi, Thomas M. Green, Cory M. Johannessen, Serena J. Silver, Cindy Nguyen, Ryan R. Murray, Haley Hieronymus, Dawit Balcha, Changyu Fan, Chenwei Lin, Lila Ghamsari, Marc Vidal, William C. Hahn, David E. Hill, and David E. Root. A public genome-scale lentiviral expression library of human ORFs. *Nature Methods*, 8(8):659–661, August 2011.

[136] Robert L. Strausberg, Elise A. Feingold, Lynette H. Grouse, Jeffery G. Derge, Richard D. Klausner, Francis S. Collins, Lukas Wagner, Carolyn M. Shenmen, Gregory D. Schuler, Stephen F. Altschul, Barry Zeeberg, Kenneth H. Buetow, Carl F. Schaefer, Narayan K. Bhat, Ralph F. Hopkins, Heather Jordan, Troy Moore, Steve I. Max, Jun Wang, Florence Hsieh, Luda Diatchenko, Kate Marusina, Andrew A. Farmer, Gerald M. Rubin, Ling Hong, Mark Stapleton, M. Bento Soares, Maria F. Bonaldo, Tom L. Casavant, Todd E. Scheetz, Michael J. Brownstein, Ted B. Usdin, Shiraki Toshiyuki, Piero Carninci, Christa Prange, Sam S. Raha, Naomi A. Loquellano, Garrick J. Peters, Rick D. Abramson, Sara J. Mullahy, Stephanie A. Bosak, Paul J. McEwan, Kevin J. McKernan, Joel A. Malek, Preethi H. Gunaratne, Stephen Richards,

Kim C. Worley, Sarah Hale, Angela M. Garcia, Laura J. Gay, Stephen W. Hulyk, Debbie K. Villalon, Donna M. Muzny, Erica J. Sodergren, Xiuhua Lu, Richard A. Gibbs, Jessica Fahey, Erin Helton, Mark Ketteman, Anuradha Madan, Stephanie Rodrigues, Amy Sanchez, Michelle Whiting, Anup Madan, Alice C. Young, Yuriy Shevchenko, Gerard G. Bouffard, Robert W. Blakesley, Jeffrey W. Touchman, Eric D. Green, Mark C. Dickson, Alex C. Rodriguez, Jane Grimwood, Jeremy Schmutz, Richard M. Myers, Yaron S. N. Butterfield, Martin I. Krzywinski, Ursula Skalska, Duane E. Smailus, Angelique Schnerch, Jacqueline E. Schein, Steven J. M. Jones, Marco A. Marra, and Mammalian Gene Collection Program Team. Generation and initial analysis of more than 15,000 full-length human and mouse cDNA sequences. *Proceedings of the National Academy of Sciences of the United States of America*, 99(26):16899–16903, December 2002.

[137] Kum-Joo Shin, Estelle A. Wall, Joelle R. Zavzavadjian, Leah A. Santat, Jamie Liu, Jong-Ik Hwang, Robert Rebres, Tamara Roach, William Seaman, Melvin I. Simon, and Iain D. C. Fraser. A single lentiviral vector platform for microRNA-based conditional RNA interference and coordinated transgene expression. *Proceedings of the National Academy of Sciences of the United States of America*, 103(37):13759–13764, September 2006.

[138] Yi Xie, Yu Jin, Bethany L. Merenick, Min Ding, Kristina M. Fetalvero, Robert J. Wagner, Alice Mai, Scott Gleim, David F. Tucker, Morris J. Birnbaum, Bryan A. Ballif, Amelia K. Luciano, William C. Sessa, Eva M. Rzucidlo, Richard J. Powell, Lin Hou, Hongyu Zhao, John Hwa, Jun Yu, and Kathleen A. Martin. Phosphorylation of GATA-6 is required for vascular smooth muscle cell

differentiation after mTORC1 inhibition. *Science Signaling*, 8(376):ra44, May 2015.

[139] P. Cohen and S. Frame. The renaissance of GSK3. *Nature Reviews. Molecular Cell Biology*, 2(10):769–776, October 2001.

[140] Eiji Kinoshita, Emiko Kinoshita-Kikuta, Kei Takiyama, and Tohru Koike. Phosphate-binding tag, a new tool to visualize phosphorylated proteins. *Molecular & cellular proteomics: MCP*, 5(4):749–757, April 2006.

[141] David B. Ring, Kirk W. Johnson, Erik J. Henriksen, John M. Nuss, Dane Goff, Tyson R. Kinnick, Sylvia T. Ma, John W. Reeder, Isa Samuels, Trina Slabiak, Allan S. Wagman, Mary-Ellen Wernette Hammond, and Stephen D. Harrison. Selective glycogen synthase kinase 3 inhibitors potentiate insulin activation of glucose transport and utilization in vitro and in vivo. *Diabetes*, 52(3):588–595, March 2003.

[142] M. J. Hart, R. de los Santos, I. N. Albert, B. Rubinfeld, and P. Polakis. Downregulation of beta-catenin by human Axin and its association with the APC tumor suppressor, beta-catenin and GSK3 beta. *Current biology: CB*, 8(10):573–581, May 1998.

[143] M. Rechsteiner and S. W. Rogers. PEST sequences and regulation by proteolysis. *Trends in Biochemical Sciences*, 21(7):267–271, July 1996.

[144] P. Rice, I. Longden, and A. Bleasby. EMBOSS: the European Molecular Biology Open Software Suite. *Trends in genetics: TIG*, 16(6):276–277, June 2000.

[145] Sissy E. Wamaitha, Ignacio del Valle, Lily T. Y. Cho, Yingying Wei, Norah M. E. Fogarty, Paul Blakeley, Richard I. Sherwood, Hongkai Ji, and Kathy K.

Niakan. Gata6 potently initiates reprograming of pluripotent and differentiated cells to extraembryonic endoderm stem cells. *Genes & Development*, 29(12):1239–1255, June 2015.

[146] Shinnosuke Tsuji, Yoshihiro Kawasaki, Shiori Furukawa, Kenzui Taniue, Tomoatsu Hayashi, Masumi Okuno, Masaya Hiyoshi, Joji Kitayama, and Tetsu Akiyama. The miR-363-GATA6-Lgr5 pathway is critical for colorectal tumourigenesis. *Nature Communications*, 5:3150, 2014.

[147] Mercy M. Davidson, Claudia Nesti, Lluis Palenzuela, Winsome F. Walker, Evelyn Hernandez, Lev Protas, Michio Hirano, and Nithila D. Isaac. Novel cell lines derived from adult human ventricular cardiomyocytes. *Journal of Molecular and Cellular Cardiology*, 39(1):133–147, July 2005.

[148] Kevin A. Janes, Chun-Chao Wang, Karin J. Holmberg, Kristin Cabral, and Joan S. Brugge. Identifying single-cell molecular programs by stochastic profiling. *Nature Methods*, 7(4):311–317, April 2010.

[149] Kayoko Takada, Kanako Obayashi, Kazuaki Ohashi, Ayako Ohashi-Kobayashi, Mayumi Nakanishi-Matsui, and Masatomo Maeda. Amino-terminal extension of 146 residues of L-type GATA-6 is required for transcriptional activation but not for self-association. *Biochemical and Biophysical Research Communications*, 452(4):962–966, October 2014.

[150] Masafumi Muratani and William P. Tansey. How the ubiquitin-proteasome system controls transcription. *Nature Reviews. Molecular Cell Biology*, 4(3):192–201, March 2003.

[151] P. Descombes and U. Schibler. A liver-enriched transcriptional activator protein, LAP, and a transcriptional inhibitory protein, LIP, are translated from the same mRNA. *Cell*, 67(3):569–579, November 1991.

[152] Y. Hirai, D. Radisky, R. Boudreau, M. Simian, M. E. Stevens, Y. Oka, K. Takebe, S. Niwa, and M. J. Bissell. Epimorphin mediates mammary luminal morphogenesis through control of C/EBPbeta. *The Journal of Cell Biology*, 153(4):785–794, May 2001.

[153] Satoru Sasagawa, Yu-ichi Ozaki, Kazuhiro Fujita, and Shinya Kuroda. Prediction and validation of the distinct dynamics of transient and sustained ERK activation. *Nature Cell Biology*, 7(4):365–373, April 2005.

[154] A. Roulston, C. Reinhard, P. Amiri, and L. T. Williams. Early activation of c-Jun N-terminal kinase and p38 kinase regulate cell survival in response to tumor necrosis factor alpha. *The Journal of Biological Chemistry*, 273(17):10232–10239, April 1998.

[155] Shao-shan Carol Huang, David C. Clarke, Sara J. C. Gosline, Adam Labadorf, Candace R. Chouinard, William Gordon, Douglas A. Lauffenburger, and Ernest Fraenkel. Linking proteomic and transcriptional data through the interactome and epigenome reveals a map of oncogene-induced signaling. *PLoS computational biology*, 9(2):e1002887, 2013.

[156] Gavin Whissell, Elisa Montagni, Paola Martinelli, Xavier Hernando-Momblona, Marta Sevillano, Peter Jung, Carme Cortina, Alexandre Calon, Anna Abuli, Antoni Castells, Sergi Castellvi-Bel, Ana Silvina Nacht, Elena Sancho, Camille Stephan-Otto Attolini, Guillermo P. Vicent, Francisco X. Real, and Eduard Batlle. The transcription factor GATA6 enables self-renewal of colon adenoma

stem cells by repressing BMP gene expression. *Nature Cell Biology*, 16(7):695–707, July 2014.

[157] Laurianne Van Landeghem, M. Agostina Santoro, Amanda T. Mah, Adrienne E. Krebs, Jeffrey J. Dehmer, Kirk K. McNaughton, Michael A. Helmrath, Scott T. Magness, and P. Kay Lund. IGF1 stimulates crypt expansion via differential activation of 2 intestinal stem cell populations. *FASEB journal: official publication of the Federation of American Societies for Experimental Biology*, 29(7):2828–2842, July 2015.

[158] Konstantin Tsoyi, Hwa Jin Jang, Irina Tsoy Nizamutdinova, Kyungok Park, Young Min Kim, Hye Jung Kim, Han Geuk Seo, Jae Heun Lee, and Ki Churl Chang. PTEN differentially regulates expressions of ICAM-1 and VCAM-1 through PI3k/Akt/GSK-3/GATA-6 signaling pathways in TNF--activated human endothelial cells. *Atherosclerosis*, 213(1):115–121, November 2010.

[159] Justin Monnier, Susanna Lewn, Edward O'Hara, Kexin Huang, Hua Tu, Eugene C. Butcher, and Brian A. Zabel. Expression, regulation, and function of atypical chemerin receptor CCRL2 on endothelial cells. *Journal of Immunology (Baltimore, Md.: 1950)*, 189(2):956–967, July 2012.

[160] Francoise Bachelerie, Adit Ben-Baruch, Amanda M. Burkhardt, Christophe Combadiere, Joshua M. Farber, Gerard J. Graham, Richard Horuk, Alexander Hovard Sparre-Ulrich, Massimo Locati, Andrew D. Luster, Alberto Mantovani, Kouji Matsushima, Philip M. Murphy, Robert Nibbs, Hisayuki Nomiyama, Christine A. Power, Amanda E. I. Proudfoot, Mette M. Rosenkilde, Antal Rot, Silvano Sozzani, Marcus Thelen, Osamu Yoshie, and Albert Zlotnik. International Union of Basic and Clinical Pharmacology. [corrected].

LXXXIX. Update on the extended family of chemokine receptors and introducing a new nomenclature for atypical chemokine receptors. *Pharmacological Reviews*, 66(1):1–79, 2014.

[161] Brian A. Zabel, Susumu Nakae, Luis Ziga, Ji-Yun Kim, Takao Ohyama, Carsten Alt, Junliang Pan, Hajime Suto, Dulce Soler, Samantha J. Allen, Tracy M. Handel, Chang Ho Song, Stephen J. Galli, and Eugene C. Butcher. Mast cell-expressed orphan receptor CCRL2 binds chemerin and is required for optimal induction of IgE-mediated passive cutaneous anaphylaxis. *The Journal of Experimental Medicine*, 205(10):2207–2220, September 2008.

[162] Safiye Gonzalvo-Feo, Annalisa D. Prete, Monika Pruenster, Valentina Salvi, Li Wang, Marina Sironi, Susanne Bierschenk, Markus Sperandio, Annunciata Vecchi, and Silvano Sozzani. Endothelial cellderived chemerin promotes dendritic cell transmigration. *J immunol*, 192, March 2014.

[163] J. P. Morgenstern and H. Land. Advanced mammalian gene transfer: high titre retroviral vectors with multiple drug selection markers and a complementary helper-free packaging cell line. *Nucleic Acids Research*, 18(12):3587–3596, June 1990.

[164] Dmitri Wiederschain, Susan Wee, Lin Chen, Alice Loo, Guizhi Yang, Alan Huang, Yan Chen, Giordano Caponigro, Yung-Mae Yao, Christoph Lengauer, William R. Sellers, and John D. Benson. Single-vector inducible lentiviral RNAi system for oncology target validation. *Cell Cycle (Georgetown, Tex.)*, 8(3):498–504, February 2009.

[165] Jesse S. Boehm, Jean J. Zhao, Jun Yao, So Young Kim, Ron Firestein, Ian F. Dunn, Sarah K. Sjostrom, Levi A. Garraway, Stanislawa Weremow-

icz, Andrea L. Richardson, Heidi Greulich, Carly J. Stewart, Laura A. Mulvey, Rhine R. Shen, Lauren Ambrogio, Tomoko Hirozane-Kishikawa, David E. Hill, Marc Vidal, Matthew Meyerson, Jennifer K. Grenier, Greg Hinkle, David E. Root, Thomas M. Roberts, Eric S. Lander, Kornelia Polyak, and William C. Hahn. Integrative genomic approaches identify IKBKE as a breast cancer oncogene. *Cell*, 129(6):1065–1079, June 2007.

[166] Byong H. Kang, Karin J. Jensen, Jaime A. Hatch, and Kevin A. Janes. Simultaneous profiling of 194 distinct receptor transcripts in human cells. *Science Signaling*, 6(287):rs13, August 2013.

[167] C. A. Andersson and R. Bro. The n-way toolbox for matlab. *Chemometr. Intell. Lab.*, 52:1–4, 2000.

[168] T. L. Bailey and C. Elkan. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proceedings / ... International Conference on Intelligent Systems for Molecular Biology ; ISMB. International Conference on Intelligent Systems for Molecular Biology*, 2:28–36, 1994.

[169] Arttu Jolma, Jian Yan, Thomas Whitington, Jarkko Toivonen, Kazuhiro R. Nitta, Pasi Rastas, Ekaterina Morgunova, Martin Enge, Mikko Taipale, Gonghong Wei, Kimmo Palin, Juan M. Vaquerizas, Renaud Vincentelli, Nicholas M. Luscombe, Timothy R. Hughes, Patrick Lemaire, Esko Ukkonen, Teemu Kivioja, and Jussi Taipale. DNA-binding specificities of human transcription factors. *Cell*, 152(1-2):327–339, January 2013.

[170] Shobhit Gupta, John A. Stamatoyannopoulos, Timothy L. Bailey, and William Stafford Noble. Quantifying similarity between motifs. *Genome Biology*, 8(2):R24, 2007.

[171] Alexander Lachmann, Huilei Xu, Jayanth Krishnan, Seth I. Berger, Amin R. Mazloom, and Avi Ma'ayan. ChEA: transcription factor regulation inferred from integrating genome-wide ChIP-X experiments. *Bioinformatics (Oxford, England)*, 26(19):2438–2444, October 2010.

[172] U. K. Laemmli. Cleavage of structural proteins during the assembly of the head of bacteriophage T4. *Nature*, 227(5259):680–685, August 1970.