## Title

Integrative Approach to Targeting Chromatin Remodeling in Breast Cancer Therapy

## Author

Mollah, Shamim

## Publication Date

2019

UNIVERSITY OF CALIFORNIA SAN DIEGO

# Integrative Approaches to Targeting Chromatin Remodeling in Breast Cancer Therapy

A dissertation submitted in partial satisfaction of the
requirements for the degree Doctor of Philosophy

in

Bioinformatics and Systems Biology

by

Shamim Ara Mollah

Committee in charge:

Professor Shankar Subramaniam, Chair
Professor Vineet Bafna, Co-Chair
Professor Hannah Carter
Professor Teressa Gaasterland
Professor Nathan Lewis

2019

The Dissertation of Shamim Ara Mollah is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

Co-Chair

Chair

University of California San Diego

2019

# DEDICATION

This dissertation is dedicated to my mom and dad.

# EPIGRAPH

*The purpose of computation is*

*insight, not numbers.*

-Richard Hamming

# TABLE OF CONTENTS

# LIST OF FIGURES

# LIST OF TABLES

# ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my advisor, Prof. Shankar Subramaniam for his support, guidance, and mentorship during my graduate study. His scientific enthusiasm and curiosity are simply infectious. I should also thank him for giving me the opportunity to participate in grant writing, mentoring undergraduate students, and guiding me toward becoming an independent scientist and a mentor. It has been my privilege to have him as my research advisor. I would like to thank my doctoral committee members: Professor Vineet Bafna, Professor Hannah Carter, Professor Nathan Lewis, and Prof. Terry Gaasterland for their support and guidance. Terry has been an important advocate of my success and a friend throughout my graduate study, providing guidance and numerous advice on my professional and personal growth.

I would like to extend my gratitude to Dr. Shakti Gupta, Dr. Andrew Caldwell, Dr. Mano R. Maurya, for stimulating discussions and all the past and present lab members who made the lab an enjoyable place: Dr. Pam Bhattacharya, Dr. Sindhu Raghunandan, Dr. Priya Nayak, Dr. Nassim Ajami, Dr. Maryam Masnadi-Shirazi, Sara Rahiminejad, Lina Aboulmuna, and Carol Kling.

I would also like to thank Dr. Elizabeth Chen, Dr. Philip Payne, Dr. Niroshana Anandasabapathy, Dr. Dinh Diep, Dr. Justin Mih, Dr. Anjana Srivatsan, Dr. Brian Tsui for being invaluable friends and for constant support during my graduate study.

Finally, I would like to thank my family for their love and support during my time in school. I want to thank my mom and dad for instilling the value of education in me when I was just a child and without whose hard work and inspiration it would not have been possible for me to come this far.

Chapter 2, is currently being prepared for submission for publication of the material. Shamim Mollah and Shankar Subramaniam. The dissertation author was a primary investigator and author of this paper.

Chapter 3, is currently being prepared for submission for publication of the material. Shamim Mollah and Shankar Subramaniam. The dissertation author was a primary investigator and author of this paper.

Chapter 4, is currently being prepared for submission for publication of the material. Shamim Mollah and Shankar Subramaniam. The dissertation author was a primary investigator and author of this paper.

# VITA

| | |
|---|---|
| 1999 | Bachelor of Science in Computer science, Indiana University, South Bend |
| 1999 | Bachelor of Arts in Mathematics, Indiana University, South Bend |
| 2003 | Master of Arts in Biomedical Informatics, Columbia University, New York |
| 2019 | Doctor of Philosophy, Bioinformatics and Systems Biology,<br>University of California San Diego |

## PUBLICATIONS

1. **Mollah SA** and Subramaniam S; "Global Chromatin Profiling Fingerprints Reveals Therapeutic Efficacy in Breast Cancer". Available at http://doi.org/10.2139/ssrn.3413902.
2. Tsui B, **Mollah S**, Skola D, Dow M, Hsu C, Carter H; "Creating a scalable deep learning based Named Entity Recognition model for biomedical textual data by repurposing biosample specimen free-text annotation", available on BioRxiv. https://doi.org/10.1101/414136.
3. **Mollah SA**, Dobrin J, Feder R, Tse S, Matos I, Cheong C, Steinman R, Anandasabapathy N; "Flt3L dependence helps define an uncharacterized subset of murine cutaneous dendritic cells", Journal of Investigative Dermatology. 2014 May;134(5):1265-1275. doi: 10.1038/jid.2013.515. Epub 2013 Nov 28. PubMed PMID:24288007.
4. Anandasabapathy N, Feder R, **Mollah S**, Tse S, Longhi M, Mehandru S, Matos I, Cheong C, Ruane D, Brane L, Teixeira A, Dobrin J, Mizenina O, Park C, Meredith M, Clausen B, Nussenzweig M, Steinman R; "Classical Flt3L-dependent dendritic cells control immunity to protein vaccine", Journal of Experimental Medicine. 2014 Aug 25;211(9):1875-91. doi: 10.1084/jem.20131397. Epub 2014 Aug 18. PubMed PMID: 25135299.
5. Elbatarny M, **Mollah S**, Grabell J, Bae S, Deforest M, Tuttle A, Hopman W, Clark DS, Mauer AC, Bowman M, Riddel J, Christopherson PA, Montgomery RR, Zimmerman Program Investigators, Rand ML, Coller B, James PD, "Normal Range of Bleeding Scores for the ISTH-BAT: Adult and Pediatric Data from The Merging Project", Haemophilia. 2014 Nov;20(6):831-5. doi: 10.1111/hae.12503. Epub 2014 Sep 6. PubMed PMID: 25196510.
6. **Mollah, SA**, PB James, Grabell J, Barbour EM, Coller B, "Diagnostic Prediction of Von Willebrand Disease using Multiple Bleeding Phenomics Datasets", American Medical

Informatics Association Joint Summits Translational Science Proceeding. 2013 Mar 18;2013:184. eCollection 2013. PubMed PMID: 24303262

7. **Mollah, SA**, James, P, Coller, B, et al, "A Machine Learning Approach to Merging and Analyzing Data from the Condensed MCMDM1 VWD Bleeding Questionnaire", 2012 International Society of Thrombosis and Hemostasis proceeding.

8. Sim I, Carini S, Tu SW, Detwiler LT, Brinkley J, **Mollah SA**, Burke K, Lehmann HP, Chakraborty S, Wittkowski KM, Pollock BH, Johnson TM, Huser V. "Ontology-Based Federated Data Access to Human Studies Information", American Medical Informatics Association Annual Symposium Proceeding. 2012;2012:856-65. Epub 2012 Nov 3. Pubmed PMID:23304360.

9. Sim I, Carini S, Tu S, Wynden R, Pollock BH, **Mollah SA**, Gabriel D, Hagler HK, Scheuermann RH, Lehmann HP, Wittkowski KM, Nahm M, Bakken S. The Human Studies Database Project: Federating Human Studies Design Data Using the Ontology of Clinical Research. American Medical Informatics Association Joint Summits Translational Science Proceeding. 2010 Mar 1;2010:51-5. Pubmed PMID:21347149.

10. AC Mauer, EM Barbour, NA Khazanov, N Levenkova, **SA Mollah**, BS Coller. "Initial Deployment of a Comprehensive, Ontology-Backed, Web-Based Bleeding History Phenotyping Instrument in Normal Individuals". Journal of Thrombosis and Haemostasis 2009,7:14.

11. Carini S, Pollock BH, Lehmann HP, Bakken S, Barbour EM, Gabriel D, Hagler HK, Harper CR, **Mollah SA**, Nahm M, Nguyen HH, Scheuermann RH, Sim I. "Development and Evaluation of a Study Design Typology for Human Subjects Research". American Medical Informatics Association Annual Symposium Proceeding. 2009 Nov 14;2009:81-5. Pubmed PMID: 20351827.

12. Mauer, A C, Barbour, E, **Mollah, SA** et al, Creating an Ontology-Based Human Phenotyping System: The Rockefeller University Bleeding History Experience. Clinical Translational Science. 2009 Oct;2(5):382-5. doi: 10.1111/j.1752-8062.2009.00147.x. PubMed PMID: 20443924.

13. **Mollah, SA**, Cimino, C, "Semi-Automatic Indexing of PostScript Files using Medical Text Indexer in Medical Education". American Medical Informatics Association Annual Symposium Proceeding. 2007 Oct 11:1053. Pubmed PMID:18694151.

14. **Mollah, SA**, Johnson, SB, "Automatic learning of the morphology of medical language using information compression", 49th American Medical Informatics Association proceeding, November 2003. **PMID:14728443**.
15. Das, A, **Mollah, SA**, Shafii-Mousavi, M, "Mathematical Model of the Hunting Strategy of a Velociraptor," Annual Indiana University Undergraduate Research Conference, November, 1998.

# ABSTRACT OF THE DISSERTATION

Integrative Approach to Targeting Chromatin Remodeling in Breast Cancer Therapy

by

Shamim Ara Mollah

Doctor of Philosophy in Bioinformatics and Systems Biology

University of California San Diego, 2019

Professor Shankar Subramaniam, Chair
Professor Vineet Bafna, Co-Chair

With the advent of high-throughput technologies, large-scale multi-omics data integration approaches have revolutionized our understanding of cancer and its progression. In this dissertation, we focus on deciphering the molecular mechanisms of various targeted cancer treatments and growth factors on chromatin remodeling through integrative analyses of multi-omics data. Chromatin remodeling is involved in the stability of genome structure, gene expression, and DNA repair, as well as, cell growth and progression; therefore, it plays an essential role in tumor suppression. The regulation of chromatin remodeling is carried out by the precise coordination of covalent histone modifications and remodeler proteins through catalytic activities. Disruption of these regulated activities confers a unique ability for healthy cells to reprogram their genome for the maintenance of oncogenic phenotypes. Hence, these histone modifications and remodeler proteins are potential targets for cancer treatments. While many drugs have the potential to target histone modifications and remodeler proteins, their precise mechanisms of action, i.e., alterations in cellular reprogramming, are not well studied. To address this gap, we utilized latent space models to integrate multi-omics data such as proteomic/phosphoproteomic, transcriptomic, and epigenomic data to understand the effects of various drugs and growth factors on specific genes, proteins, and phosphoproteins that are involved in the regulation of a wide range of cellular processes (growth, proliferation, and cell division) and gene activity states in cancer and healthy cell lines. In addition to providing mechanistically-driven targets that can impact chromatin remodeling, the chromatin fingerprints generated from our study can serve as a signature for assessing the efficacy of a given drug in treating cancer. Further, increasing evidence, supported by our study of breast cancer, indicates that this paradigm applies to various cancers, and further analysis can provide insights into more detailed chromatin-based mechanisms. Our study implies that the cancer state is one where chromatin gets remodeled, and effective drugs attempt to restore the chromatin state to that of a healthy cell. Overall, the integrative frameworks we developed reveal the mechanisms of action of specific drugs on chromatin remodeling machinery in breast cancer cells and of growth factors on cellular phenotypes in normal breast cells, which lay the foundation for improved development of chromatin-based cancer therapy.

# CHAPTER 1: Introduction

The human epigenome is made up of chemical compounds and proteins that can attach to DNA and orchestrate cellular activities by controlling gene expression. They enable the diversity of cells in multicellular organisms. Epigenetic plays an essential role in mediating environment-induced phenotypic plasticity.

Two well-studied epigenetic phenomena are DNA methylation and histone modifications. DNA methylation is a stable and reversible epigenetic mark that occurs on the fifth carbon of cytosines and chemically named “5-methylcytosine” (5mC). Histones are structural proteins that form the basic units of nucleosomes. These nucleosomes are utilized for DNA packaging within the nucleus to form chromatin. The post-translational modifications of histones such as methylation, acetylation, etc., can influence the structure of the chromatin and modulate genes expression patterns. Deregulation of chromatin leads to altered gene activation and gene silencing. Recent studies suggest that by altering chromatin structures, gene-translocations results to oncogenesis (Croce and Di Croce 2005;Donehower et al. 1992;Martens and Stunnenberg 2010;Mitelman, Johansson, and Mertens 2007;Uribesalgo and Di Croce 2011;Cairns 2007).

In this work, we will focus on post-translational modification (PTM) of histones and other signaling proteins in human cancer, specifically for profiling chromatin alterations due to various treatments and growth factors, and new approaches to integrate multi-omics data towards chromatin-based cancer therapy.

## 1.1 Organizational structure of chromatin in human epigenome

Histones are a group of relatively small proteins (~15-20 kDa) whose high arginine and lysin amino acids make them positively charged. Thus, binding of the negatively charged DNA to histones is stabilized by ionic bonds. In the human cells, genetic information is organized in a highly conserved structure polymer called chromatin (**Fig 1.1A**). This structure is attained by wrapping approximately 3 m of DNA around the histone proteins. Chromatin assembly undergoes dynamic changes actively mediating gene function and expression (Felsenfeld and Groudine 2003). It is believed that one approach by which gene regulatory mechanisms are modulated is through the interaction of the genomic DNA with these histone proteins (Arents and Moudrianakis 1995). Nucleosomes, the building blocks of chromatin, are comprised of 146 base pairs of DNA wrapped around an octamer of core histones (two of each 2HA, 2HB, H3, and H4) (**Fig 1.1A**). Linker histones of the H1 class associate between single chromosomes resulting in "beads on a string" structure, which folds into loop establishing higher organization. The nucleosome is the key player in regulation of gene expression serving as a "memory bank" by transmitting epigenetic information from one cell generation to next. Epigenetics refer to inheritable changes in gene function and expression without any changes in DNA sequences. The gene regulatory functions of nucleosomes are primarily controlled by the N-terminal tails of the four core histones. The N-terminal tails are subjected to enzymatic catalytic PTMs including phosphorylation, methylation, acetylation, and ubiquitination. The enzymes responsible for creating histone PTMs are termed as histone modifying enzymes and are broadly categorized into "Reader(s)", "Writer(s)" and "Eraser(s)", based on the ability to catalyze either by reading, addition or removal of specific PTMs, respectively (**Fig 1.1B**). These modifications act in various transcriptional activation/inactivation, chromosome packaging, and DNA damage/repair,  and can be highly reversible making them potential for effective drug targets.

## 1.2 General arithmetic of histone modifications

In this thesis we will work with H3 histones which primarily are acetylated at lysines 9, 14, 18, 23, and 56, methylated at arginine 2 and lysines 4, 9, 27, 36, and 79, and phosphorylated at ser10, ser28, Thr3, and Thr11 (**Fig 1.2**). Thus, mapping these modifications to various signaling pathways would allow us to better understand epigenetic regulation of cellular processes and the development of histone modifying enzyme-targeted drugs. A list of canonical H3 histones and their epigenetic functions are illustrated in **Table 1.1**. Kinases and phosphatases are two type of enzymes that catalyze transfer or removal of phosphate to and from a molecule from adinosen tri phosphate (ATP). Recently serine/threonine kinases have emerged as strong candidates for drug therapy and around one-third of all kinase inhibitors are currently in development target serine/threonine kinases.

## 1.3 Scope of the dissertation

The purpose of this work is to create integrative approaches to quantify the effects of specific drugs and growth factors on cellular phenotypes, and systematically assess the impact of these phenotypes on chromatin-based therapy outcome and cancer formation.

In chapter 2, we develop an integrative latent space model to generate a 3D phosphoprotein-histone-drug network (iPhDnet) using multi-omics cancer data obtained from LINCS. We identified four histones signatures in response to drugs as chromatin fingerprints and deciphered the unique phosphoprotein networks and signaling pathways that describe these fingerprints.

In chapter 3, we describe the mechanisms of action of specific drugs on chromatin remodeling machinery in breast cancer. We applied the iPhDnet from the previous chapter to implicate CDK inhibitors flavopiridol and dinaciclib as potential therapeutics mediated by BRD4, NSD3, EZH2 and MYC targeting H3K27me3K36me3 status change in breast cancer.

In chapter 4, We provide a comprehensive characterization of how normal breast cells acquire inflammation, proliferation, migration, and differentiation statuses when treated with six growth factors, namely, epidermal growth fact (EGF), hepatocyte growth factor (HGF), oncostatin M (OSM), bone morphogenetic protein 2 (BMP2), transforming growth factor beta (TGFB), and interferon-gamma (IFNG). We showed that STAT3 induction promotes BRD4 to regulate cell proliferation when treated with OSM.

Lastly, in Chapter 5, we present the conclusions of the dissertation.

## 1.4 Figures

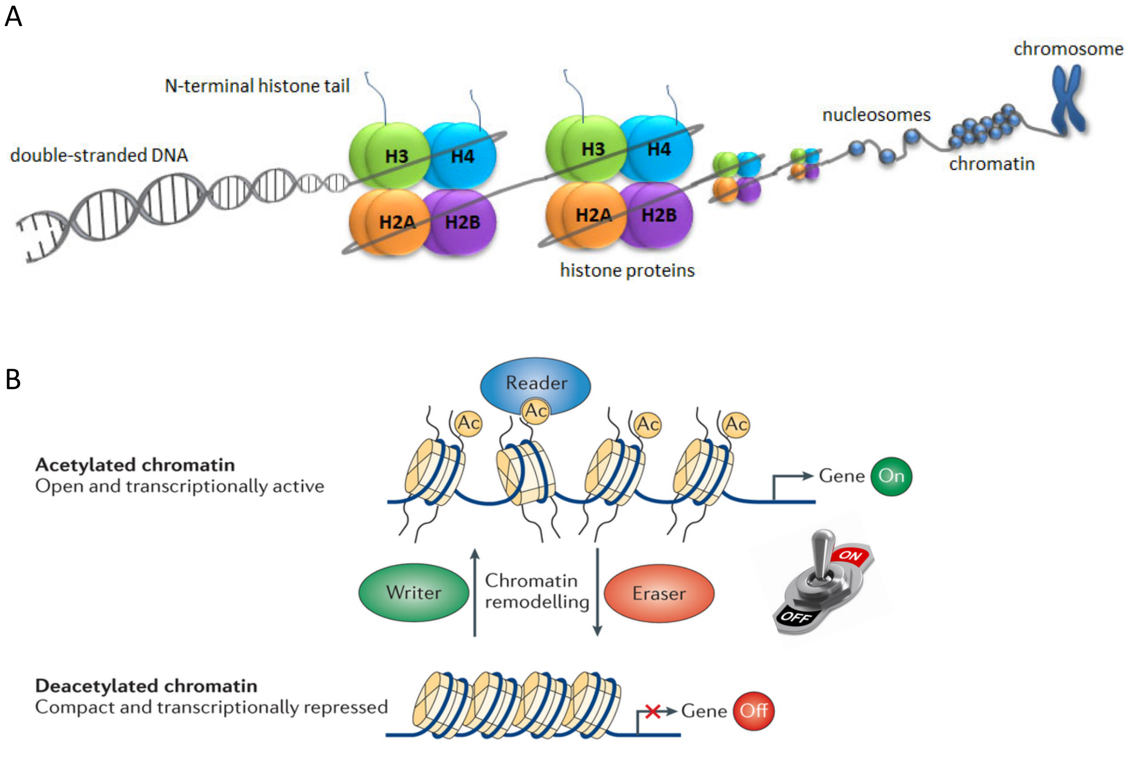


**Figure 1.1. Higher organizational structure of chromatin complex.**
(**A**) DNA wrapped around the core histone proteins (two of histone H2A, histone H2B, histone H3 and histone H4) forming an octamer nucleosome subunit. Together with H1 linker histone these nucleosomes represent beads on string like structure forming the chromatin complex. (**B**) Depicting interplay among the chromatin remodeler proteins: reader, writer, and eraser. Their precise coordination allows genes to switch from euchromatin (open and transcriptionally active) state to heterochromatin (compact and transcriptionally repressed) state and vice versa. Images are adapted from (http://www.whatisepigenetics.com/histone-modifications/).

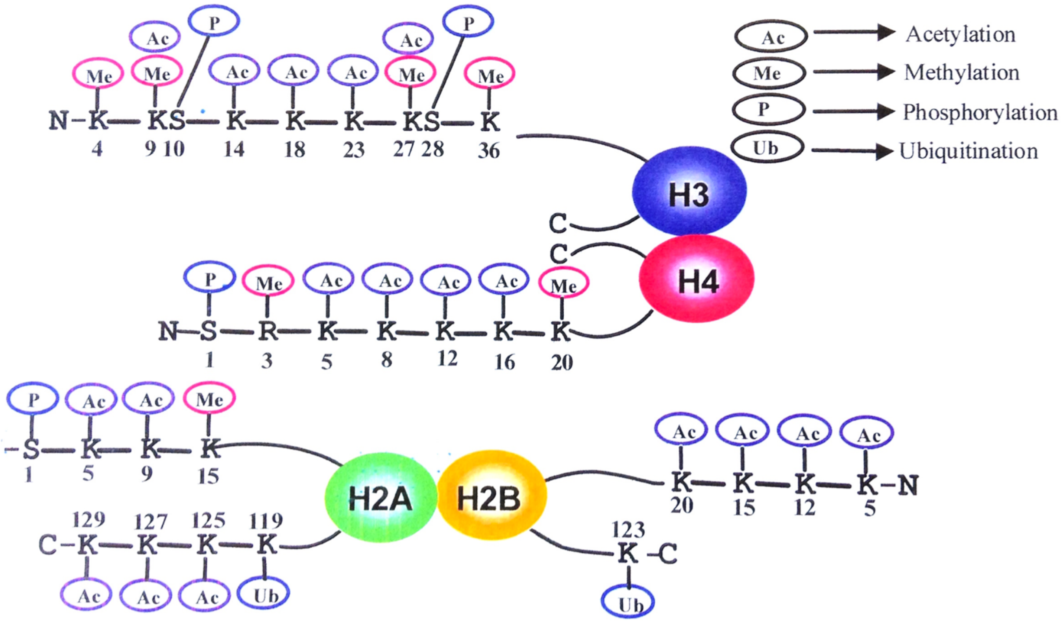


**Figure 1.2. General arithmetic of histone modifications.**
Schematic showing some common histone modifications at C and N terminal of the histone tails. Unique combinations of these modifications at various amino acid loci give rise to unique histone codes influencing unique regulatory roles in gene expressions machinery.

## 1.5 Tables

**Table 1.1. Summary of canonical histone codes and their gene expression statuses.** Showing known gene expression statuses for the canonical histone codes.

| **Type of modification** | **Histone** | | | | | |
|---|---|---|---|---|---|---|
| | H3K4 | H3K9 | H3K27 | H3K36 | H3K79 | H3K14 |
| mono-methylation | activation | Activation | activation | activation | activation | |
| di-methylation | activation | Repression | repression | activation | activation | |
| tri-methylation | activation | Repression | repression | activation | activation<br>repression | |
| acetylation | | Activation | activation | | | activation |

## 1.6 References

Arents, G., and E. N. Moudrianakis. 1995. "The Histone Fold: A Ubiquitous Architectural Motif Utilized in DNA Compaction and Protein Dimerization." *Proceedings of the National Academy of Sciences*. https://doi.org/10.1073/pnas.92.24.11170.

Cairns, Bradley R. 2007. "Chromatin Remodeling: Insights and Intrigue from Single-Molecule Studies." *Nature Structural & Molecular Biology*. https://doi.org/10.1038/nsmb1333.

Croce, Luciano Di, and Luciano Di Croce. 2005. "Chromatin Modifying Activity of Leukaemia Associated Fusion Proteins." *Human Molecular Genetics*. https://doi.org/10.1093/hmg/ddi109.

Donehower, Lawrence A., Michele Harvey, Betty L. Slagle, Mark J. McArthur, Charles A. Montgomery, Janet S. Butel, and Allan Bradley. 1992. "Mice Deficient for p53 Are Developmentally Normal but Susceptible to Spontaneous Tumours." *Nature*. https://doi.org/10.1038/356215a0.

Felsenfeld, Gary, and Mark Groudine. 2003. "Controlling the Double Helix." *Nature*. https://doi.org/10.1038/nature01411.

Martens, Joost H. A., and Henk G. Stunnenberg. 2010. "The Molecular Signature of Oncofusion Proteins in Acute Myeloid Leukemia." *FEBS Letters*. https://doi.org/10.1016/j.febslet.2010.04.002.

Mitelman, Felix, Bertil Johansson, and Fredrik Mertens. 2007. "The Impact of Translocations and Gene Fusions on Cancer Causation." *Nature Reviews Cancer*. https://doi.org/10.1038/nrc2091.

Uribesalgo, I., and L. Di Croce. 2011. "Dynamics of Epigenetic Modifications in Leukemia." *Briefings in Functional Genomics*. https://doi.org/10.1093/bfgp/elr002.

# CHAPTER 2: iPhDnet: A 3D Latent Space Model for Cancer Therapy

## 2.1 Abstract

As the molecular complexity of cancer etiology exists at multiple levels, there is a need for multi-omics data integration approaches to uncover the underlying molecular mechanisms of cancer biology. Understanding the molecular mechanisms of cancer biology is critical for designing effective anti-cancer therapies. In the past, single-level omics have contributed to our current understanding of cancer-specific drug targets; however, they lacked to establish the link between pathology and therapeutic mechanisms. In contrast, integrative multi-omics approaches offer a global view covering interconnectivity of the signaling proteins and their functional contributions to cancer phenotypes. In this study, we developed a 3D network using latent space models to infer cell-specific pohospho-histone-drug signaling network. We combined an unsupervised clustering technique with a supervised multivariate regression technique to identify signaling modules and predicted their associations with phosphoproteins. We applied our model on the P100 and GCP phosphoproteome datasets from LINCS to establish relationships among drugs, phosphoproteins, and histone modifications in MCF7 human breast cancer cell line.

## 2.2 Introduction

In recent years, molecular biology has improved our understanding of cancer, however, the use of therapeutic drugs faces major challenges (Sorger and Schoeberl 2012). This is partly due to the inadequate understanding of the causal connections between the pathology and the therapeutic mechanisms at the cellular function level. The understanding of cellular function at the molecular level involves the study of intracellular signaling, metabolic pathways and gene regulatory networks, through "omics" measurements on biological systems. Protein phosphorylation is one of the most important post-translational modifications (PTMs) in intracellular signaling (Li et al. 2013), (Sacco et al. 2012) that regulates cell cycle, cell growth, cell differentiation, and metabolism (Delom and Chevet 2006). Phosphorylation is a key reversible modification that activates and deactivates proteins via phosphorylation/dephosphorylation events due to specific kinases and phosphatases (Hunter 1995). Protein phosphorylation events occurs on serine (S), threonine (T), and tyrosine residues (Y) (*O*-phosphorylation) that can regulate enzymatic activity, subcellular localization, complex formation and protein degradation (Roskoski 2012). Thus, reconstructing phosphoprotein networks from "omics" measurements can help us understand and model cellular signaling pathways as well as uncovering the mechanisms of actions of drugs. While single-level "omics" approaches have contributed towards understanding of cancer mutations, subtypes, and epigenetic alterations based on gene/protein expressions, they lack the resolving power to establish causal relationship between molecular signatures and cellular phenotypes. In contrast, multi-omics approaches involving interrogation of proteins/drugs in multiple dimensions have the potential to uncover the intricate molecular mechanisms underlying various cellular phenotypes.

In a recent large-scale initiative, Library of Integrated Network-Based Cellular Signatures (LINCS) (http://www.lincsproject.org), has carried out multi-omics characterization of response of five cancer cells to 31 drugs, through measurement of phosphoproteins (P100) (Abelin et al. 2016) and global chromatin profiles (GCP) (Creech et al. 2015). Some of these measurements were carried out at multiple time points post-treatment of cells. The P100 and GCP are Mass Spectrometry (MS) based targeted proteomics assays that include a representative set of phosphopeptides, and different combinations of histone modifications treated by multiple drugs respectively. Identifying the treatment effects of a compound on certain cell line is of great significance of discovering new potential drugs and improving the response to therapies in clinic. This consortium has created unprecedented opportunities to reveal underlying oncogenic molecular signatures beneath phenotypes.

In the past, a number of mathematical and statistical approaches to high-throughput biological data has been used extensively to understand the relationship between different cellular components to partially reconstruct intracellular networks. With the availability of large-scale omics data, computational systems biology has made substantial progress towards modeling and reconstruction of data-driven networks using (1) iCluster (Shen et al. 2012), (2) Principal Component Regression (Pradervand, Maurya, and Subramaniam 2006), (3) probabilistic graphical models such as Bayesian network-based models (Dojer et al. 2006; Faryabi et al. 2009; Altay 2012), and (4) information theory-based methods such as integrated correlation and transfer entropy based approach and C3NET ((Dojer et al. 2006; Faryabi et al. 2009; Altay 2012)). Other approaches includes differential equations (Mestl, Plahte, and Omholt 1995) and structural equation methods (Xiong, n.d.).

However, cancer  is complex and is regulated at multiple layers which can be manifested by these assays. While theses assays offer a glimpse of the complex system, these events are interdependent (or interactive). Thus, when integrating several different omics data to discover the coherent biological signatures, it is challenging to incorporate different biological layers of information to predict accurate phenotypes. Our objective in this study is to develop a generalized 3D latent space model that can generate an integrated phosphoprotein-histone-drug network (iPhDnet) to uncover the number of distinct ways in which drugs relate to global chromatin profile and decipher the unique phosphoproteins networks and pathways that describe histone response to 31 drug treatments. In the present work, we have applied a robust version of Non-negative Matrix Factorization (NMF) to generate histone signatures for MCF7 breast cancer cell line, and Partial Least Square Regression (PLSR) to predict which of these enriched phosphoproteins were associated with specific histone signatures (**Fig 2.1**).

Due to the fact that, there were only 24-hour matching data for GCP and P100 available for MCF7 breast cancer cell line, we used these two datasets to build our model.

## 2.3 Methods and Materials

**Data Acquisition.** The experimental data were generated by the NIH LINCS Proteomic Characterization Center for Signaling and Epigenetics (PCCSE) repository. Level 3 (log 2 normalized) targeted phosphoproteomics assay (P100) against 96 phosphopeptides data, and level 3 (log 2 normalized) global chromatin profiling assay (GCP) against 60 probes that monitor combinations of post-translational modification on histones data using in MCF7 breast cancer cell line. These assays were treated with 31 serine/threonine kinase inhibitors (drugs) at various concentrations, DMSO as a negative control and consisted of three biological replicates. We used the 24-hour GCP and P100 datasets to test our model.

**Data Preprocessing.** Replicates were used to impute missing data by taking their weighted average values during the pre-processing step. Differential histone modifications and phosphorylation changes were computed by taking fold changes of each perturbed phosphopeptide and histone code with respect to DMSO. These resulted in two data matrices, i) phosphoprotein profiles consisting [96 peptides x 31 drugs], and ii) global chromatin profiles consisting [60 histone modifications x 31 drugs]. Prior to modeling, data were normalized with respect to the mean and standard deviation of the respective variables.

**Functional module extraction using NMF.** An unsupervised clustering technique, non-negative matrix factorization (NMF), was used to stratify histone signatures. Similar to vector quantization methods such as principal components analysis (PCA) and singular value decomposition (SVD), the objective of NMF is to explain the observed data using a compact number of latent features, i.e., basis components, which when combined approximate the original data as accurately as possible. In NMF both the matrix representing the basis components (histone signatures) as well as the matrix of mixture coefficients (drug prototypes) are constrained to have non-negative values, and unlike PCA and SVD, no independence or orthogonally constraints are imposed on the basis components. Because NMF assumes an additive model, non-log transformed values were used in our analysis.

Mathematically, NMF consists of finding an approximation

$$A \approx WH, \qquad \textit{(Equation 1)}$$

$$W, H \geq 0$$

where W, H are n × k and k × m non-negative matrices respectively where n are rows – samples and m are columns –the measured features in A (**Fig 2.3**). Since the objective is to reduce the dimension of the original data A, the factorization rank k is often chosen such that k << (n, m). **W** contains basis vectors and **H** contains encoding vectors that estimate the extent to which each basis vector is used to reconstruct each input vector. We used a version of NMF to minimize the divergence function (KL divergence) given by (Brunet et al. 2004). The function is related to the Poisson likelihood of generating A from W and H, more specifically, based on randomly initialized matrices **W** and **H**, NMF finds the solution of

$$\min D(\mathbf{A}||\mathbf{WH}) = \sum_{i=1}^{n} \sum_{j=1}^{m} (A_{ij} log \frac{A_{ij}}{(WH)_{ij}} - A_{ij} + (WH)_{ij})$$

where, D is a loss function, via an iterative process (Lee et al., 1999). At each step, W and H are updated by using the following coupled divergence equations:

$$H_{a\mu} \leftarrow H_{a\mu} \frac{\sum_i W_{ia} A_{i\mu}/(WH)_{i\mu}}{\sum_k W_{ka}} \qquad \textit{(Equation 2)}$$

$$W_{ia} \leftarrow HW_{ia} \frac{\sum_{\mu} WH_{a\mu} A_{i\mu}/(WH)_{i\mu}}{\sum_{v} H_{av}} \qquad \textit{(Equation 3)}$$

where $\mathbf{A}_{i,j} = [\mathbf{A}]_{i,j}$ indicates (i,j)-th element of the matrix $\mathbf{A}$.

Because (1) is non-convex optimization with respects to W and H, there is no guarantee of obtaining a local minimum. Moreover, the above iterative update rules are notorious for slow convergence (i.e., require more iterations) and have a complexity of $O(mnk^2N_iN_o)$ where $N_i$ is the number of inner iterations to solve the non-negative linear model and $N_o$ is the number of outer iterations to alternate W and H steps. As a result, the initialization of the pair of factors (W, H) is considered an important component in the design of successful NMF methods (Lee and Seung 1999). We used a robust initialization strategy using the seeding algorithm (Boutsidis and Gallopoulos 2008), that is based on a non-negative double singular value decomposition (nndSVD). The whole process then becomes deterministic and needs to run once and the complexity is reduced to $O(mnk^2N_i + N_o)$. Our NMF framework works as follows:

1. Initialize W, H ∈ $R_{m\times k}$, $R_{k\times n}$ respectively with non-negative elements using nndSVD.

2. Repeat until a convergence criterion is satisfied:

$$H_{a\mu} \leftarrow H_{a\mu} \frac{\sum_{i} W_{ia} A_{i\mu}/(WH)_{i\mu}}{\sum_{k} W_{ka}}$$

$$H \geq 0$$

where W is fixed, and

$$W_{ia} \leftarrow HW_{ia} \frac{\sum_{\mu} WH_{a\mu} A_{i\mu}/(WH)_{i\mu}}{\sum_{v} H_{av}}$$

$$W \geq 0$$

where H is fixed

3. The columns of W are normalized and the rows of H are scaled accordingly.

**Cluster Validation.** To identify the optimal rank k, we used the cophenetic correlation coefficient (Sokal and Rohlf 1962) to determine the most robust clustering as:

$$c = \frac{\sum_{i<j}(x(i,j) - \bar{x})(t(i,j) - \bar{t})}{\sqrt{[\sum_{i<j}(x(i,j) - \bar{x})^2][\sum_{i<j}(t(i,j) - \bar{t})^2]}}. \quad \textit{(Equation 4)}$$

C measures how reliably the same histone codes are assigned to the same cluster across many iterations of the clustering algorithm with random initializations. The cophenetic correlation coefficient lies between 0 and 1 and reflects the probability that samples i and j cluster together. Higher values indicate more stable cluster assignments. We selected optimal k= 4 based on the largest observed cophenetic coefficient and where the magnitude of the cophenetic correlation begins to decrease by varying values of k from 2 to 10 (**Fig 2.5**). We used the NMF package in R to implement and compute these calculations.

In eq 3, $x(i, j) = |x_i - x_j|$, the ordinary Euclidean distance between the *i*th and *j*th observations. $t(i, j)$ = the dendrogrammatic distance between the model points $t_i$ and $t_j$ (height of the node at which these two points are first joined), *x* bar is the average of the $x(i, j)$, and *t* bar is the average of the $t(i, j)$. After factorizing **A** into the basis matrix **W** and the *encoding* matrix **H**, we used the basis matrix **W** for histone stratification. Specifically, we grouped histone codes into *four groups (k=4). We assigned histone code $x_i$ to cluster k* which has the highest value based on the basis vector, as:*

$$k^* = \arg\max_k W_{i,k} \quad \textit{(Equation 5)}$$

Similarly, we assigned targeted pathways for each drug $d_j$ to cluster k* which has the highest value based on the encoding vector, as:

$$k^* = \arg\max_k H_{j,k} \quad \textit{(Equation 6)}$$

**Histone Prediction Model using PLSR.** Histone-peptide interaction network was generated using partial least square regression (PLSR) method based on Kraemer et al. formulation (Krämer and Sugiyama 2011). PLSR is a multivariate regression method for constructing predictive model when the number of factors/predictor variables (in our case phosphopeptides) exceeds the number of responses / dependent variables (histone marks), and collinearity exists (phosphopeptides are correlated with one another). A past study (Gupta et.al, 2010) had shown the effectiveness of partial least square (PLS) application in understanding crosstalk between phosphoprotein signaling in macrophage cells, thus, prompting us to consider a PLS-based regression model. The general idea behind PLSR is to try to extract latent factors, accounting for as much of the observed variation as possible while modeling the responses well. For each sample n, the value $y_{nj}$ is defined as:

$$y_{ni} = \sum_{i=0}^{k} b_i \, x_{ni} + \varepsilon_{ni} \qquad \textit{(Equation 7)}$$

Where $y_{ni}$ is a response, $b_i$is the coefficient, $x_{ni}$is an explanatory variable and $\varepsilon_{ni}$ is an error term. This model is similar to linear regression; however, the way these $\beta_i$ are found is different. To see this, a matrix format of equation (7) can be expressed as ***Y*=*XB*+*E*** where ***Y*** is an *n* cases by *m* variables response matrix (in our case it is drugs x histone data), ***X*** is an *n* cases by *p* variables predictor matrix (in our case it is drugs x phosphopeptides data), ***B*** is a *p* by *m* regression coefficient matrix, and ***E*** is a noise term for the model which has the same dimensions as ***Y***. For our ***X*** predictor matrix, we first normalized all the phosphosignal values to their corresponding z-scores and centered ***Y*** response matrix (histone values). Intuitively, partial least squares regression produces a *p* by *c* weight matrix ***W*** for ***X*** such that ***T*=*XW***, i.e., the columns of $\boldsymbol{W}$ are weight vectors for the $\boldsymbol{X}$ columns producing the corresponding $n$ by $c$ factor score matrix $\boldsymbol{T}$. These weights are computed so that each of them maximizes the covariance between responses and the corresponding factor scores. Ordinary least squares procedures for the regression of $\boldsymbol{Y}$ on $\boldsymbol{T}$ are then performed to produce $\boldsymbol{Q}$, the loadings for $\boldsymbol{Y}$ (or weights for $\boldsymbol{Y}$) such that $\boldsymbol{Y}=\boldsymbol{TQ}+\boldsymbol{E}$. Once $\boldsymbol{Q}$ is computed where $\boldsymbol{B}=\boldsymbol{WQ}$, we have $\boldsymbol{Y}=\boldsymbol{XB}+\boldsymbol{E}$, and the prediction model is complete. To provide a complete description of PLSR, we also need a $p$ by $c$ factor loading matrix $\boldsymbol{P}$ which gives a factor model $\boldsymbol{X}=\boldsymbol{TP}+\boldsymbol{F}$, where $\boldsymbol{F}$ is the unexplained part of the $\boldsymbol{X}$ scores. On the training data, we calculated the optimal model parameter using 10-fold cross-validation. We assessed the predictive performance by computing the residual sum of square (RSS) error of prediction on the test set.

We identified the optimal number of components (principal component, PC) that could be used to predict the model accurately using residual square sum (RSS) value < 0.05. Once the coefficients ($\beta_i$) are generated, we retained only the significant peptides (p_value < 0.001) using the statistical significant t-test where the degree of freedom DOF was computed as:

DOF = min (column of X, row of X) – PC – 1.

## 2.4 Results and Discussions

**Four Pathway-based Histone Signatures Identified by Clustering Method Constitute "Global Chromatin Fingerprint Profiles."** In order to identify fingerprint histone profiles, we investigated the relationships between the 31 drugs targeting serine-threonine kinases in various cell lines including the breast cancer line (MCF7), and the resulting GCP response at 24 hours. We calculated the histone code fold changes by accounting for their differential modifications i.e., changes in histone levels from pre-treatment (MCF7 treated with DMSO) to post-treatment (MCF7 treated with a specific drug) state. Using a non-negative matrix factorization (NMF) clustering method on these histone code fold changes, we identified four pathway-based functional histone modules c1, c2, c3 and c4 (**Fig 2.2, Table 2.1**) and refer to them as "histone signatures" that characterize the response to drugs. Briefly, the objective of NMF is to explain the observed data using a compact number of latent features, i.e., basis components, which abstract the original data as accurately as possible. No independence or orthogonality constraints are imposed on the basis components leading to a simple and intuitive interpretation of the factors that allows the basis components to overlap. This unique feature is particularly interesting in histone modules, where overlapping basis components identify combinatorial histone codes resulting from multiple signaling pathways and indicating a specific signature (method section).

To provide a comprehensive mapping of these histone signatures to drugs with respect to their shared signaling pathways, we then generated a molecular network consisting of 91 nodes (comprising histone codes and drugs) and 554 edges (node interactions). Coefficients generated from the assignments of each histone signature profile to the drug prototypes (see method section) are used to represent the strength of the interactions between a histone code and a drug (**Fig 2.2**).

Edge thickness represents the strength of the contributions of drugs to histone codes belonging to the same histone signature. In c1 histone signature, we found 46 histone codes are strongly associated with ten drugs  showing mostly inhibitory effects shared by nine common pathways (**Fig 2.2**). We observed all cyclin-dependent kinase (CDK) inhibitors (flavopiridol, dinaciclib, and PD-0332991) and replication stress inhibitors (VX-970 and SCH 90076) were grouped with the same histone signature. We observed similar groupings for the c2 signature associating 3 histone codes with ten drugs where all AkT/PI3K variants of Ras inhibitors (IPI145, afuresertib, BYL719, dactolisib) grouped together (**Table 2.1**). Similarly, in c3, 2 repressive histone marks, H3K56me1 and H3K56me2, are associated with 5 drugs targeting Jak3, Mek1/2, Jnk, Hippo and multikinase pathways. In addition, H3K56me2 is activated by all drugs while H3K56me1 had inhibitory effects on Jnk, Mek1/2 and multikinase pathways, and activation effects on Jak3 and Hippo pathways. The involvement of these pathways regulating these two histones is less clearly established in breast cancer and merits further experimental investigation. In the c4 signature module, we observed nine histone codes associated with IKkKB, Mek1/2, Notch/Wnt/Hedgehog, Gamma secretase, mTOR, and p38 MAPK pathways. We observed a reduction of monomethylation at lysine 9 and phosphorylation at serine 10 (H3K9me1S10ph1K14ac0), a repressive histone code, in all the pathways in this module suggesting its potential as a therapeutic marker mediated by these shared pathways in breast cancer. In addition, we observed overlaps in c2, c3 and c4 for drugs targeting IkKB, Notch, Mek1/2 and Jnk pathways suggesting crosstalk among histone signatures. Collectively, our results suggest strong selective preferences of histone codes towards specific pathway-based therapeutic effects as well as possible crosstalk among the pathways which may lead to off-target effects.

**Histone Prediction Model Provides Quantitative Contributions of Enriched Phosphoproteins Toward Histone Codes.** Next, we sought to identify the phosphoprotein networks representing various interactions among the enriched phosphoproteins and histone codes. Using the P100 phosphoproteins and GCP responses at 24 hours after treatment with the 31 drugs, we developed a quantitative-qualitative estimate of significant phosphoproteins contributing to the alterations of histone codes using a partial least square regression (PLSR) method. PLSR is a multivariate method for constructing a predictive model when the number of factors (covariates, e.g., 96 phosphoproteins) exceeds the number of responses (e.g., 60 histones) and the factors are highly correlated. PLSR attempts to extract latent factors across data, accounting for as much of the observed variation as possible while modeling the output/responses accurately. A past study (Gupta et.al, 2010) had shown the effectiveness of partial least square (PLS) application in understanding crosstalk between phosphoprotein signaling in macrophage cells, thus, prompting us to consider a PLS-based regression model. Using PLSR we generated a system model where each histone code is considered as an outcome/response to combined influences (i.e., coefficients) of all phosphoproteins. Each coefficient represents the contribution of individual phosphoprotein towards the level of a histone code (see method section). We evaluated our PLSR model using 10-fold cross-validation. Using a p-value <1.0e-4 (see method section), our model generated a histone-phosphoprotein network comprised of 113 nodes, representing histone codes and phosphoproteins and 230 edges (interactions between them) (**Fig 2.3**). Our results showed H3K27me3K36me3, H3K9ac1S10ph1K14ac0, H3K56me2, and H3K18ac0K23ub1 as highly connected histone codes (hubs with the highest degree), influenced by the statistically significant phosphoproteins: BRD4 (Bromodomain Containing 4), ATAD2 (ATPase Family, AAA Domain

Containing 2), NOLC1 (Nucleolar and Coiled-Body Phosphoprotein 1), SRRM2 (Serine/Arginine Repetitive Matrix 2), and CASC3 (Cancer Susceptibility Candidate 3).

Briefly, BRD4 and ATAD2 are bromodomain proteins. BRD4 is an epigenetic "reader" and belongs to BET family protein that maintains epigenetic memory and regulates cell cycle progression; BRD4 has been shown to have an intrinsic binding specificity for transcription factors such as c-MYC and p53 which are known to promote cancer (Delmore et al. 2011), making it a promising drug target. Similarly, ATAD2 is a novel cofactor for MYC, overexpressed and amplified in aggressive tumors. It has been shown that downregulation of ATAD2 via siRNA results in increased apoptotic activity, suggesting a role for inhibitors of ATAD2 in cancer cell death and tumor regression (Caron et al. 2010). NOLC1 is a nucleolar protein that regulates RNA polymerase I by connecting RNA polymerase I to ribosomal processing and remodeling enzymes, resulting in translational remodeling. It has a high binding affinity to c-MYC and Max transcription factors which play an important role in cancer. Although NOLC1 has not been studied extensively, a previous study (Hwang et al. 2009) found NOLC1 to have transcription factor-like activity in nasopharyngeal cancer progression suggesting its possible role in other cancers. SRRM2 protein is known to be involved in pre-mRNA splicing and has binding specificity for p53. SRRM2 has been detected as a 5'-3' Exoribonuclease 2 (Xrn2)-interacting protein that is involved in premature termination of RNA polymerase II transcription (Sansó et al. 2016; Brannan et al. 2012) thus affecting cell cycle progression. CASC3, also known as MLN51 is a component of the exon junction complex (EJC) whose expression has been shown to be elevated in some breast cancer cell lines (Tomasetto et al. 1995). The EJC is known to be involved in a surveillance mechanism that degrades mRNAs with premature translation termination codons through a nonsense-mediated mRNA decay (NMD) function, thereby, promoting cell cycle arrest. Currently, the implications of

ATAD2, NOLC1, SRRM2 and CASC3 proteins in cancer regulation is poorly understood, and our results suggest that these regulators serve as potential novel oncogenic drivers mediating histone modifications in breast cancer.

Taken together, the histone-peptide network reveals candidate phosphoproteins that serve as potential therapeutic targets; these candidate phosphoproteins alter histone codes resulting in alterations in cell cycle progression in breast cancer.

**Integrated Approach Provides a Three-dimensional View of Molecular Interactions among Drugs-Phosphoproteins-Histones.** To further elucidate the influence of specific drugs on phosphoproteins and downstream histone codes, we then developed a 3D view of the molecular interactions (phosphoproteins-drugs-histones) by integrating histone signatures with the drug-phosphoprotein interaction network resulting in an integrated phosphoproteins-histones-drugs network (iPhDnet). The network consisted of 144 nodes, 742 interactions (**Fig 2.4**) and 2157 unique maps (interaction profiles) which are stored into a relational database. Using this database, we developed a web-based tool called chromatin reader eraser writer database (CREWdb) to further characterize these enriched phosphoprotein (**Fig 2.7**). The iPhDnet serves as a quantitative atlas of global chromatin profile fingerprints that can be used to generate hypotheses linking drugs, pathways, phosphoproteins, and histones, to understand drug response pathways in cancer.

Our chromatin profile fingerprints revealed an overall reduction in histone levels in active marks such as methylation of H3K36, H3K4, and acetylation of H3K9 when treated with drugs, which are consistent with previous studies (Lewis et al. 2013; Leroy et al. 2013; Zhu et al. 2015). We observed that reduction of phosphoprotein level in SRRM2 was positively correlated (p-Val < 2.7e-04) with H3K4me1 and H3K4me3 when treated with drugs that belonged to c1 signature histone module. While trimethylation of H3K27 is a repressive histone mark associated with transcriptionally silenced chromatin in most cancers (Lewis et al. 2013; Leroy et al. 2013; Zhu et al. 2015), our analysis revealed inhibitory effects of differential modification levels of trimethylation of H3K27. These findings are consistent with the results of prior studies in breast cancer (Holm et al. 2012; Ren et al. 2012). Likewise, Abelson interactor protein-1 (ABI1), an adaptor protein involved in cell migration, along with its downstream effector phospho-Akt (p-Akt) has been implicated in the spread of breast cancer (Wang et al. 2011); ABl1 is positively correlated with reduced H3K27me3K36me3 histone mark (p-Val < 7.02e-05) when inhibited by CDK inhibitor flavopiridol. Additionally, we observed significant associations (p-val < 7.5e-05) of BRD4, NOLC1, ATAD2 and SRRM2 with H3K27me3K36me3 when treated with CDK inhibitors flavopiridol, dinaciclib, and palbociclib (PD-033291) (**Table 2.1**). Taken together, iPhDnet shows that inhibiting ABI1, BRD4, NOLC1, ATAD2, and SRRM2 with the help of CDK inhibitors may be sufficient to induce the heterochromatin state where the repressive mark H3K27me3 colocalizes with the active mark H3K36me3. This finding suggests that for a stable reversion of epigenetic silencing state in breast cancer, a reversal from the malignant euchromatin to normal heterochromatin may be dictated by H3K36me3. In addition, we observed a decrease of H3K27me3 when treated with all CDK inhibitors supporting the observation made by a prior study where induction of a CDK inhibitor was associated with a lower level of H3K27me3 in breast cancer (Yang et al. 2009). Most likely the reduced level of H3K27me3 is associated with CDK inhibitors in breast cancer. Therefore, we postulate that ABI1, BRD4, NOLC1, ATAD2, and SRRM2 can mediate H3K27me3K36me3 for reversion of epigenetic silencing in breast cancer using CDK inhibitors. We also postulate that the reduction of H3K27me3 may be a compensatory effect of other cofactors of these phosphoproteins working together to induce cell growth arrest, suggesting the potential for a combinatorial treatment strategy in breast cancer.

## 2.5 Evaluation

**Benchmarking and performance analysis.** We evaluated the performance of NMF clustering using a cophenetic coefficient score and known canonical signaling pathways associated with drug treatment. Cophenetic coefficient measures how reliably the same histone codes are assigned to the same cluster across many iterations of the clustering algorithm with random initializations. We found that 100 iterations were enough to attain cluster stability for our data. The cophenetic correlation coefficient lies between 0 and 1 and reflects the probability that samples i and j cluster together. Higher values indicate more stable cluster assignments. We selected optimal k= 4 based on the largest observed cophenetic coefficient and where the magnitude of the cophenetic correlation begins to decrease by varying values of k from 2 - 10 (**Fig. 2.5**). See Online Methods for details. In addition, the optimal k = 4 is reflected by the groupings of the of the drugs to their respective canonical signaling pathways.

We evaluated PLSR model using cross validation. On the training data, we calculated the optimal model parameter using a 10-fold cross-validation. We assessed the predictive performance by computing the residual sum of square (RSS) error of prediction on the test sets. We identified the optimal number of components (principal component, PC) that could be used to predict the model accurately using residual square sum (RSS) value < 0.05 (**Fig. 2.6**). Once the coefficients are generated, we retained only the significant peptides (p_value < 0.0001) using a t-test and degree of freedom (DOF). See Methods for details.

## 2.6 Conclusions and Future Work

The iPhDnet serves as a quantitative atlas of global chromatin profiles thereby, providing a detailed view of histone modifications by various drugs and phophoproteins in breast cancer. Our study reveals strong correlation between flavopiridol and dinaciclib with strongest selective inhibitory potential against BRD4 that impacts H3K27me3K36me3 histone mark. Thus, implicating these drugs as potential BRD4 mediated therapeutics targeting H3K27me3K36me3 in breast cancer. All of these findings warrant experimental validation.

In conclusion, iPhDnet can serve as a valuable method for integrative analysis of multi-omics datasets. The framework can also be extended to accommodate other types of omics data, for instance scRNASeq, ATAC-seq and Hi-C data. In the future, we want to incorporate nonlinear models in the framework to capture the dynamics of cellular mechanisms. Finally, the observations made by applying iPhDnet to analyze GCP and phosphoprotein data have implications for discovery of cancer therapeutics suggesting mechanistically-associated targets.

## 2.7 Figures

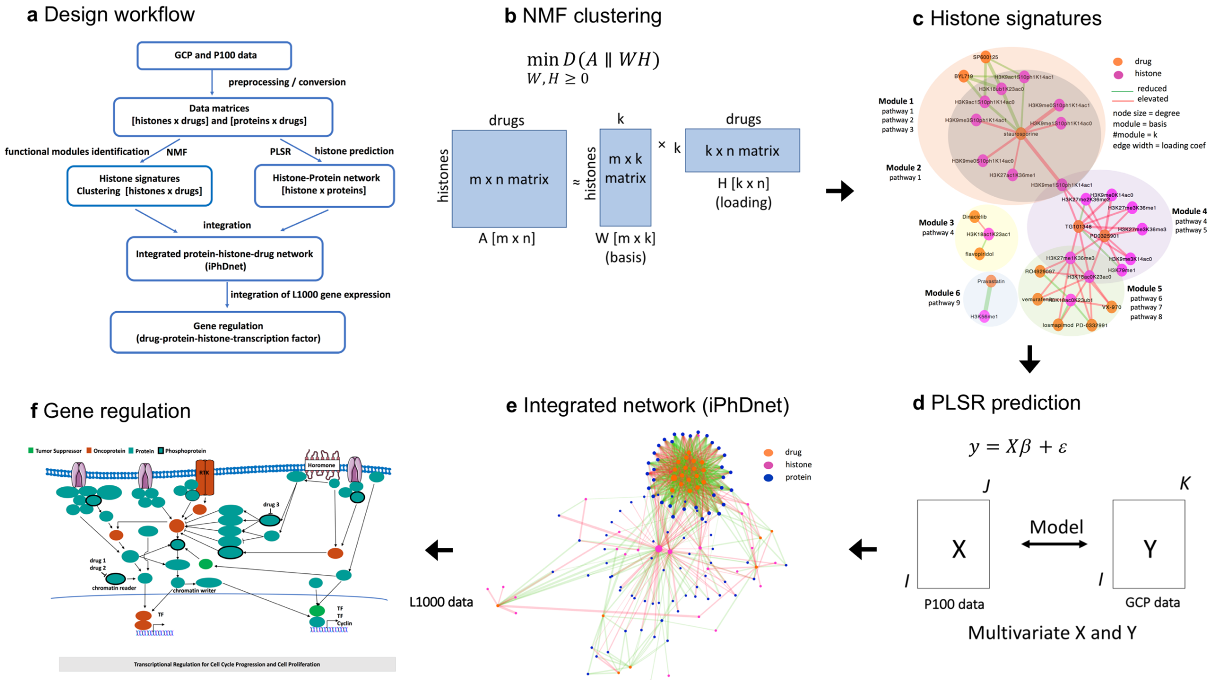


**Figure 2.1. Schematic view of the 3D latent space model to generate global chromatin fingerprints in breast cancer.**
(**a**) Flowchart of the model steps. (**b**) GCP data is clustered using consensus-based non-negative factorization (NMF). (**c**) Pathway-based "histone signatures" are recovered by NMF clustering and linking drugs to the specific histones. (**d**) A quantitative estimate of significant phosphoproteins contributing to histone modifications is assessed by generating a histone-protein interaction network using partial least square regression (PLSR) method. (**e**) A 3-dimensional view of molecular interactions of drugs-phosphoproteins-histones is generated by integrating the histone signatures with the histone-protein interaction network. (**f**) Showing the utility of the iPhDnet.

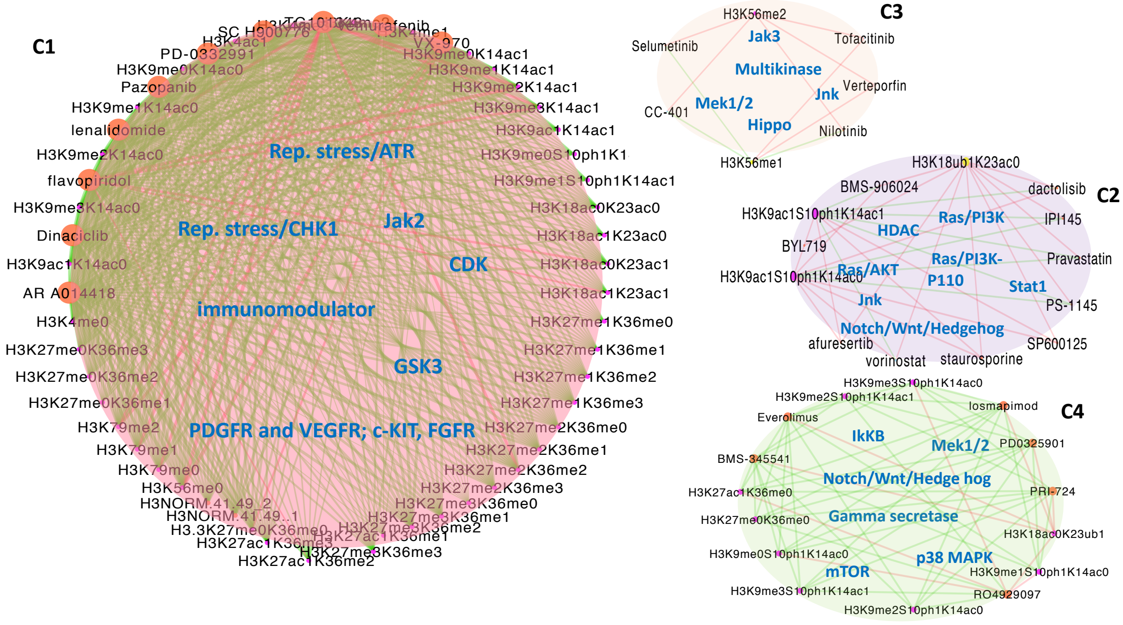


**Figure 2.2. Histone-drug interaction map using NMF model.**

Four "histone signatures" are obtained by NMF clustering of GCP at 24-hour post-treatment. Drugs and histones are depicted by orange and magenta nodes respectively, the color of edges signifies whether the interaction between a drug and a histone resulted in elevated (red) or reduced (green) histone level.

.

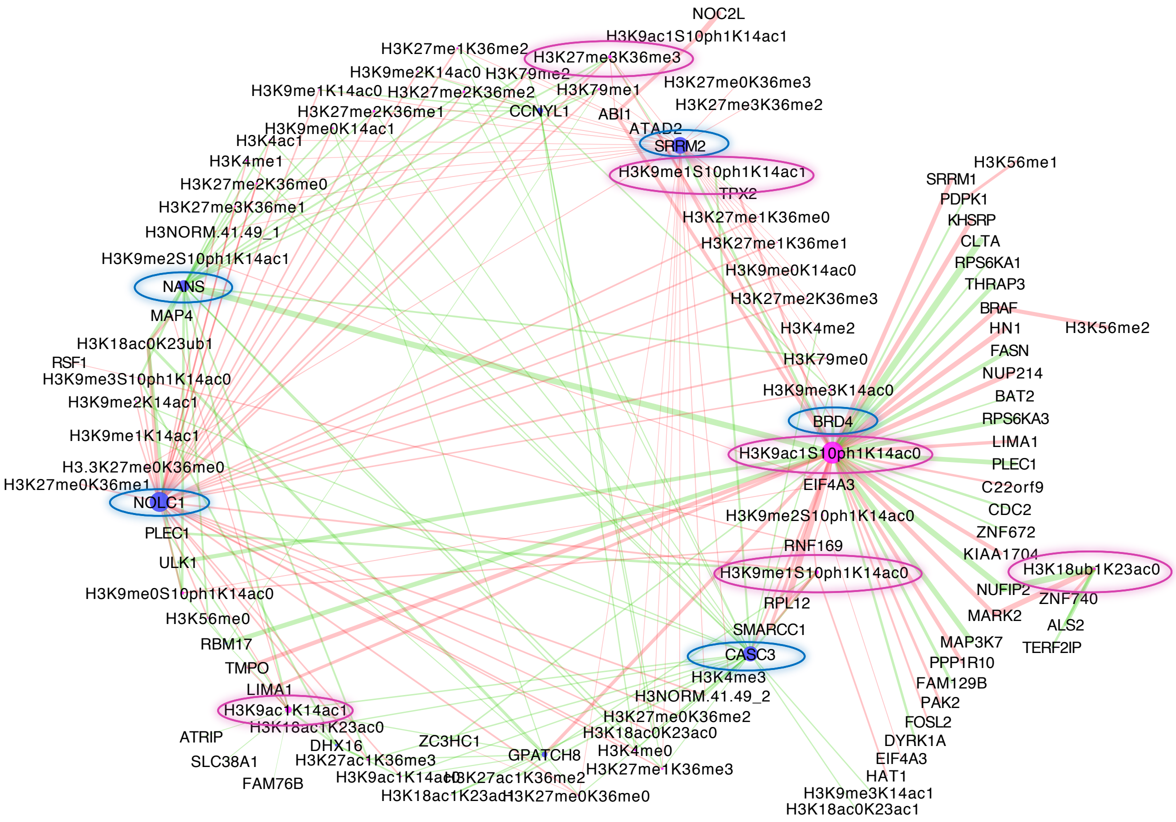


**Figure 2.3. Histone-phosphopeptide interaction map using PLSR model.**
Histone-phosphoprotein interaction network using a PLSR prediction model. Histones and phosphoproteins are depicted by magenta and blue nodes respectively, the color of edges depicts whether the interaction between a phosphoprotein and a histone is positively (red) or negatively (green) correlated.

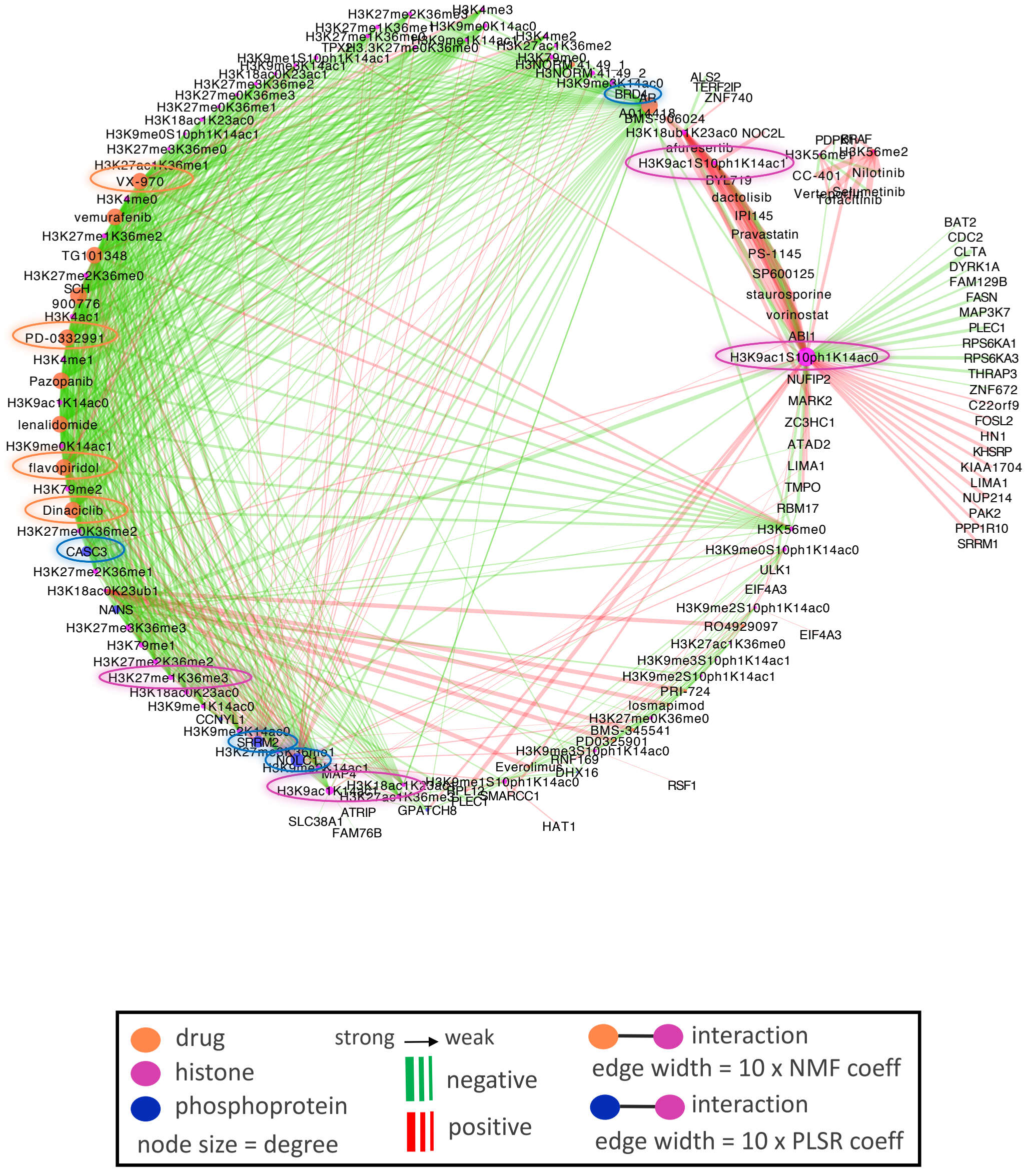


**Figure 2.4. Integrated phosphoprotein-drug-histone network (iPhDnet).**
iPhDnet shows enrichment of modification levels on H3K9ac1S10ph1K14ac0, H3K56me2, H3K27me3K36me3, H3K18ac0K23ub1 histone codes, acting as highly connected nodes (hubs) and positively induced by various drugs affecting enriched phosphoproteins including BRD4, ATAD2, and NOLC1. The strength of an interaction is captured by the width of an edge

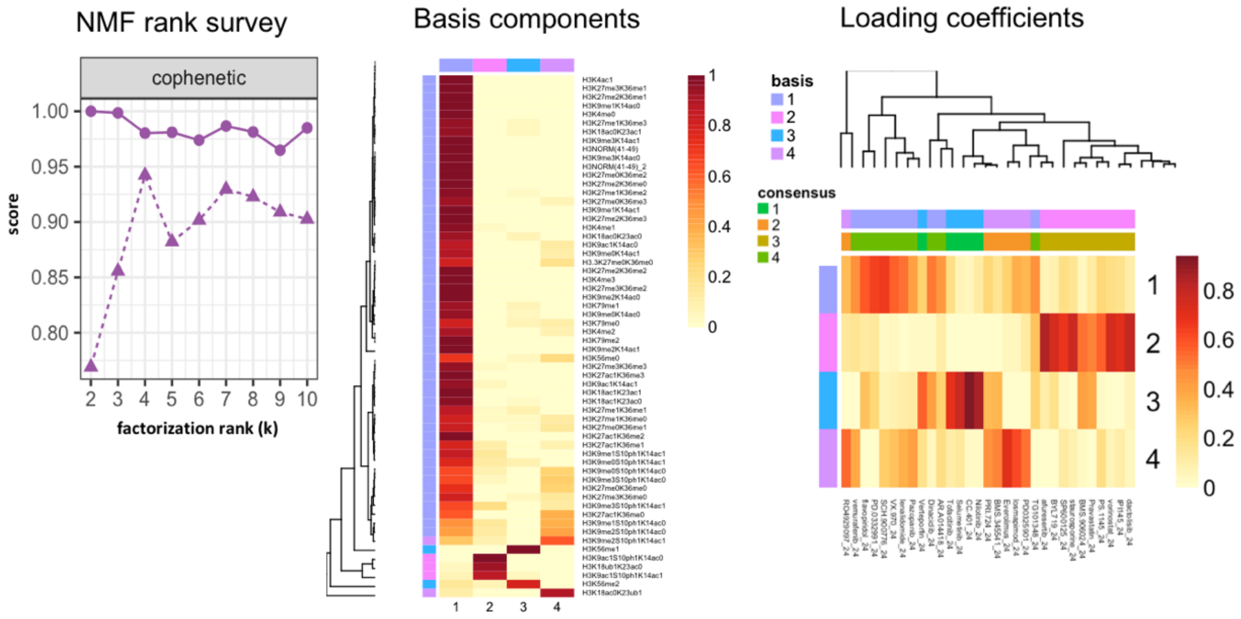


**Figure 2.5. Benchmarking and Performance analysis.**

Estimation of the factorization rank of NMF and its cluster component. Cophenetic score is computed from 100 runs for each value of rank k, by varying k= 2, 3...10 on 24-hour GCP data. Rank k represents the number of clusters or basis components. The solid line represents the original data and the dotted line represents simulated data using 100 runs. The cophenetic scores identify optimal k to be 4. Heatmap of the basis components (histones and their cluster membership). The heatmap of the basis components shows likelihood of each histone mark belonging to each signature module. Mixture (loading) coefficients (quantitative contribution of drugs to histone cluster membership) are shown for k=4.

$$RSS = \sum_{i=1}^{n}(\varepsilon_i)^2 = \sum_{i=1}^{n}(y_i - (\alpha + \beta x_i))^2,$$

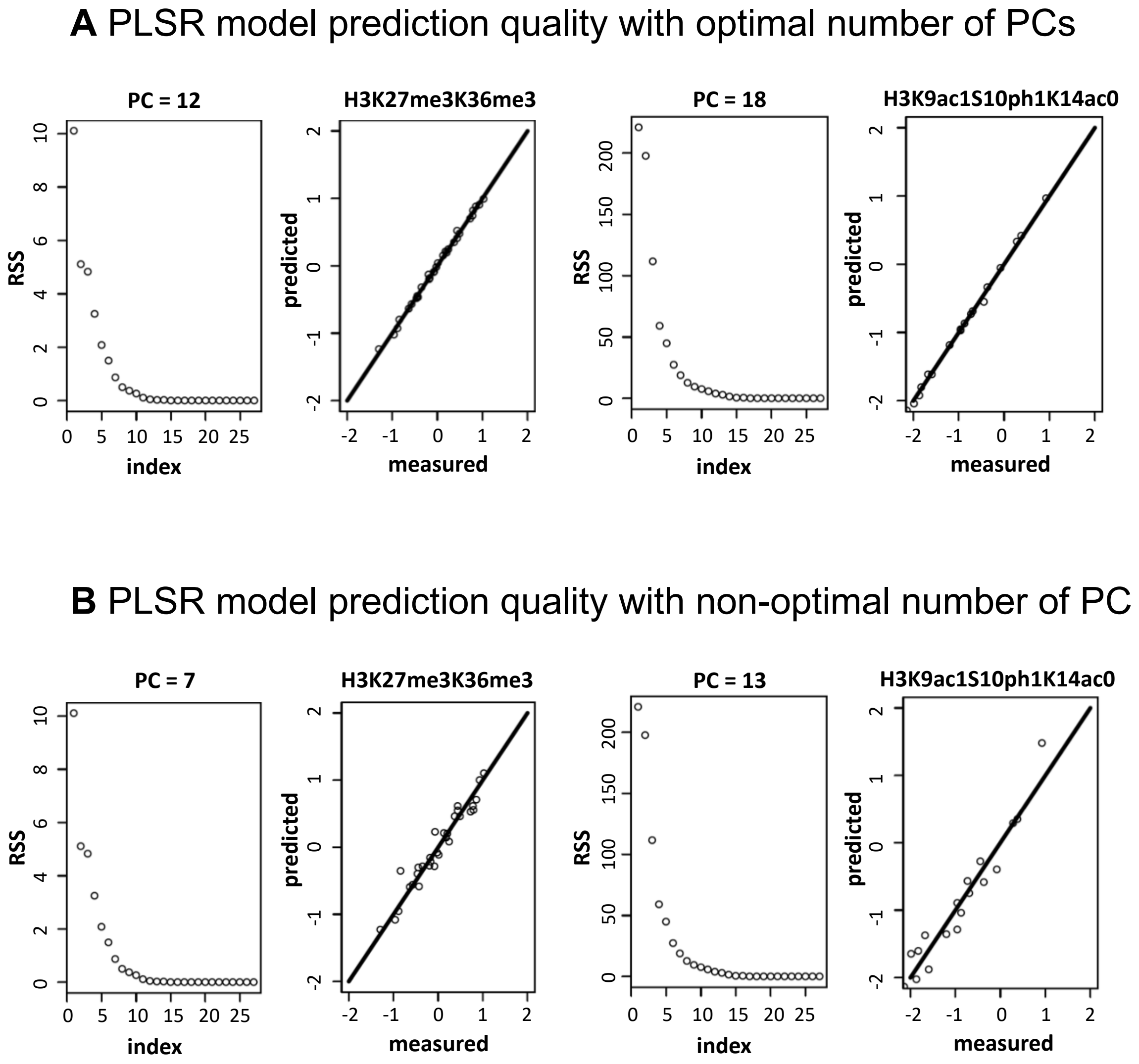
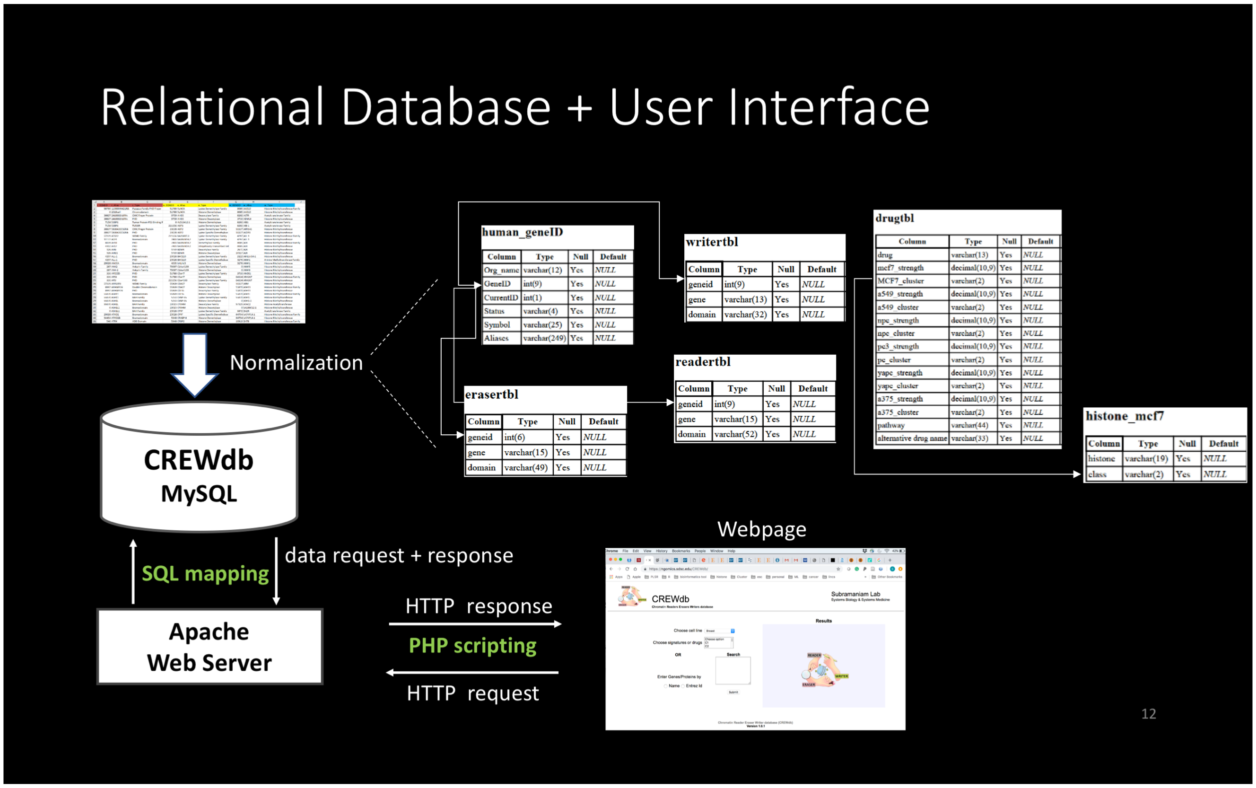


**Figure 2.6. Performance of PLSR model.**
Showing two examples of histone codes (H3K27me3K36me3, H3K9ac1S10ph1K14ac0 and H3K18ub1K23ac0). (**A**) Showing model performance using optimal number of components. The optimal number of components (principal component, PC) is used to predict the model accurately using residual sum square (RSS) value < 0.05. (**B**) Depicting model performance using sub optimal number of components.



Relational Database + User Interface

Normalization

**human_geneID**

| Column | Type | Null | Default |
|---|---|---|---|
| Org_name | varchar(12) | Yes | *NULL* |
| GeneID | int(9) | Yes | *NULL* |
| CurrentID | int(1) | Yes | *NULL* |
| Status | varchar(4) | Yes | *NULL* |
| Symbol | varchar(25) | Yes | *NULL* |
| Aliases | varchar(249) | Yes | *NULL* |

**writertbl**

| Column | Type | Null | Default |
|---|---|---|---|
| geneid | int(9) | Yes | *NULL* |
| gene | varchar(13) | Yes | *NULL* |
| domain | varchar(32) | Yes | *NULL* |

**erasertbl**

| Column | Type | Null | Default |
|---|---|---|---|
| geneid | int(6) | Yes | *NULL* |
| gene | varchar(15) | Yes | *NULL* |
| domain | varchar(49) | Yes | *NULL* |

**readertbl**

| Column | Type | Null | Default |
|---|---|---|---|
| geneid | int(9) | Yes | *NULL* |
| gene | varchar(15) | Yes | *NULL* |
| domain | varchar(52) | Yes | *NULL* |

**drugtbl**

| Column | Type | Null | Default |
|---|---|---|---|
| drug | varchar(13) | Yes | *NULL* |
| mcf7_strength | decimal(10,9) | Yes | *NULL* |
| MCF7_cluster | varchar(2) | Yes | *NULL* |
| a549_strength | decimal(10,9) | Yes | *NULL* |
| a549_cluster | varchar(2) | Yes | *NULL* |
| npc_strength | decimal(10,9) | Yes | *NULL* |
| npc_cluster | varchar(2) | Yes | *NULL* |
| pc3_strength | decimal(10,9) | Yes | *NULL* |
| pc_cluster | varchar(2) | Yes | *NULL* |
| yapc_strength | decimal(10,9) | Yes | *NULL* |
| yapc_cluster | varchar(2) | Yes | *NULL* |
| a375_strength | decimal(10,9) | Yes | *NULL* |
| a375_cluster | varchar(2) | Yes | *NULL* |
| pathway | varchar(44) | Yes | *NULL* |
| alternative drug name | varchar(33) | Yes | *NULL* |

**histone_mcf7**

| Column | Type | Null | Default |
|---|---|---|---|
| histone | varchar(19) | Yes | *NULL* |
| class | varchar(2) | Yes | *NULL* |

**Figure 2.7. Communication between CREWdb database and GUI.**
Schematic view of MySQL database construction and integration with Apache webserver via PhP scripting.

## 2.8 Table

**Table 2.1. Characteristics of the NMF based four histone signatures in MCF7.** Membership of histone codes and drugs in their respective histone signatures. These signatures influence specific canonical signaling pathways shown in the right most column (represented visually in Figure 3.1).

| Histone signature | # of Histone codes | Histone codes | # of Drugs | Drugs | Canonical pathways |
|---|---|---|---|---|---|
| c1 | 46 | H3K4me0,<br>H3K4me1,<br>H3K4me2,<br>H3K4me3,<br>H3K4ac1,<br>H3K9me0K14ac0,<br>H3K9me1K14ac0,<br>H3K9me2K14ac0,<br>H3K9me3K14ac0,<br>H3K9ac1K14ac0,<br>H3K9me0K14ac1,<br>H3K9me1K14ac1,<br>H3K9me2K14ac1,<br>H3K9me3K14ac1,<br>H3K9ac1K14ac1,<br>H3K9me0S10ph1K14ac1,<br>H3K9me1S10ph1K14ac1,<br>H3K18ac0K23ac0,<br>H3K18ac1K23ac0,<br>H3K18ac0K23ac1,<br>H3K18ac1K23ac1,<br>H3K27me1K36me0,<br>H3K27me1K36me1,<br>H3K27me1K36me2,<br>H3K27me1K36me3,<br>H3K27me2K36me0 | 10 | AR-A014418,<br>Dinaciclib,<br>Flavopiridol,<br>Lenalidomide,<br>Pazopanib,<br>PD-0332991,<br>SCH900776,<br>TG101348,<br>Vemurafenib,<br>VX970 | GSK3 inhibitor,<br>CDK/1,2,4,5,6,9 inhibitor,<br>immunomodulator,<br>PDGFR and VEGFR; Also c-KIT, FGFR, inhibitor,<br>Rep. stress/CHK1 inhibitor,<br>Jak2 inhibitor,<br>Raf inhibitor,<br>Rep. stress/ATR inhibitor |
| c2 | 3 | H3K9ac1S10ph1K14ac0,<br>H3K9ac1S10ph1K14ac1,<br>H3K18ub1K23ac0 | 10 | afuresertib,<br>BMS906024,<br>BYL719,<br>dactolisib,<br>IPI145,<br>Pravastatin,<br>PS-1145,<br>SP600125,<br>staurosporine,<br>vorinostat | Ras/AKT inhibitor,<br>Notch/other inhibitor,<br>Ras/PI3K-P110a inhibitor,<br>Ras/PI3K inhibitor,<br>Ras/PI3K-P110g,d inhibitor,<br>Stat1 inhibitor,<br>IkKB inhibitor,<br>Jnk inhibitor,<br>Kinase inhibitor; general,<br>HDAC inhibitor; general |
| c3 | 2 | H3K56me1,<br>H3K56me2 | 5 | CC-401,<br>Nilotinib,<br>Selumetinib,<br>Tofacitinib,<br>Verteporfin | Jnk inhibitor,<br>Multikinase inhibitor,<br>Mek1/2 inhibitor,<br>Jak3 inhibitor,<br>Hippo inhibitor |
| c4 | 9 | H3K9me0S10ph1K14ac0,<br>H3K9me1S10ph1K14ac0,<br>H3K9me2S10ph1K14ac0,<br>H3K9me3S10ph1K14ac0,<br>H3K9me2S10ph1K14ac1,<br>H3K9me3S10ph1K14ac1,<br>H3K18ac0K23ub1,<br>H3K27me0K36me0,<br>H3K27ac1K36me0 | 6 | BMS-345541,<br>Everolimus,<br>losmapimod,<br>PD0325901,<br>PRI-724,<br>RO4929097 | IkKB inhibitor,<br>mTOR inhibitor,<br>p38 MAPK inhibitor,<br>Mek1/2 inhibitor,<br>Notch/Wnt/Hedgehog inhibitor,<br>Notch/gamma secretase inhibitor |

## 2.9 Author Contributions

**Original Concept:** Shamim Ara Mollah

**Project Supervision:** Shankar Subramaniam

**Project Planning and Experimental Design:** Shamim Ara Mollah and Shankar Subramaniam

**Method Implementation, Processing and Analysis:** Shamim Ara Mollah

**Preparation of Manuscript:** Shamim Ara Mollah and Shankar Subramaniam

## 2.10 Acknowledgements

We would like to thank Drs. Shakti Gupta and Mano for helpful discussion and comments and. This work was funded by R01 LM012595, U01 DK097430, U01 CA200147, U01 CA198941, U19 AI090023, R01 HL106579, R01HL108735, R01 HD084633, R01 DK109365, NSF CCF-0939370, and a Wellcome-Trust Award to S.S. S.A.M. received first-year support from National Institutes of Health Graduate Training Grant (T15LM011271).

Chapter 2, is currently being prepared for submission for publication of the material. Shamim Mollah, Shankar Subramaniam. The dissertation author was a primary investigator and author of this paper.

## 2.11 References

Abelin, Jennifer G., Jinal Patel, Xiaodong Lu, Caitlin M. Feeney, Lola Fagbami, Amanda L. Creech, Roger Hu, et al. 2016. "Reduced-Representation Phosphosignatures Measured by Quantitative Targeted MS Capture Cellular States and Enable Large-Scale Comparison of Drug-Induced Phenotypes." *Molecular & Cellular Proteomics: MCP* 15 (5): 1622–41.

Altay, G. 2012. "Empirically Determining the Sample Size for Large-Scale Gene Network Inference Algorithms." *IET Systems Biology* 6 (2): 35–43.

Boutsidis, C., and E. Gallopoulos. 2008. "SVD Based Initialization: A Head Start for Nonnegative Matrix Factorization." *Pattern Recognition* 41 (4): 1350–62.

Brunet, Jean-Philippe, Pablo Tamayo, Todd R. Golub, and Jill P. Mesirov. 2004. "Metagenes and Molecular Pattern Discovery Using Matrix Factorization." *Proceedings of the National Academy of Sciences of the United States of America* 101 (12): 4164–69.

Creech, Amanda L., Jordan E. Taylor, Verena K. Maier, Xiaoyun Wu, Caitlin M. Feeney, Namrata D. Udeshi, Sally E. Peach, et al. 2015. "Building the Connectivity Map of Epigenetics: Chromatin Profiling by Quantitative Targeted Mass Spectrometry." *Methods*. https://doi.org/10.1016/j.ymeth.2014.10.033.

Delom, Frederic, and Eric Chevet. 2006. "Proteome Science." https://doi.org/10.1186/1477-5956-4-15.

Dojer, Norbert, Anna Gambin, Andrzej Mizera, Bartek Wilczyński, and Jerzy Tiuryn. 2006. "Applying Dynamic Bayesian Networks to Perturbed Gene Expression Data." *BMC Bioinformatics* 7 (May): 249.

Faryabi, Babak, Golnaz Vahedi, Jean-Francois Chamberland, Aniruddha Datta, and Edward R. Dougherty. 2009. "Intervention in Context-Sensitive Probabilistic Boolean Networks Revisited." *EURASIP Journal on Bioinformatics & Systems Biology*, April, 360864.

Gupta, Shakti, Mano Ram Maurya, and Shankar Subramaniam. 2010. "Identification of Crosstalk between Phosphoprotein Signaling Pathways in RAW 264.7 Macrophage Cells." *PLoS Computational Biology* 6 (1): e1000654.

Hunter, T. 1995. "Protein Kinases and Phosphatases: The Yin and Yang of Protein Phosphorylation and Signaling." *Cell* 80 (2): 225–36.

Krämer, Nicole, and Masashi Sugiyama. 2011. "The Degrees of Freedom of Partial Least Squares Regression." *Journal of the American Statistical Association* 106 (494): 697–705.

Lee, D. D., and H. S. Seung. 1999. "Learning the Parts of Objects by Non-Negative Matrix Factorization." *Nature* 401 (6755): 788–91.

Li, Xun, Matthias Wilmanns, Janet Thornton, and Maja Köhn. 2013. “Elucidating Human Phosphatase-Substrate Networks.” *Science Signaling* 6 (275): rs10.

Mestl, Thomas, Erik Plahte, and Stig W. Omholt. 1995. “A Mathematical Framework for Describing and Analysing Gene Regulatory Networks.” *Journal of Theoretical Biology*. https://doi.org/10.1006/jtbi.1995.0199.

Pradervand, Sylvain, Mano R. Maurya, and Shankar Subramaniam. 2006. “Identification of Signaling Components Required for the Prediction of Cytokine Release in RAW 264.7 Macrophages.” *Genome Biology* 7 (2): R11.

Roskoski, Robert. 2012. “ERK1/2 MAP Kinases: Structure, Function, and Regulation.” *Pharmacological Research*. https://doi.org/10.1016/j.phrs.2012.04.005.

Sacco, Francesca, Livia Perfetto, Luisa Castagnoli, and Gianni Cesareni. 2012. “The Human Phosphatase Interactome: An Intricate Family Portrait.” *FEBS Letters* 586 (17): 2732–39.

Shen, Ronglai, Qianxing Mo, Nikolaus Schultz, Venkatraman E. Seshan, Adam B. Olshen, Jason Huse, Marc Ladanyi, and Chris Sander. 2012. “Integrative Subtype Discovery in Glioblastoma Using iCluster.” *PloS One* 7 (4): e35236.

Sokal, Robert R., and F. James Rohlf. 1962. “The Comparison of Dendrograms by Objective Methods.” *Taxon* 11 (2): 33–40.

Sorger, Peter K., and Birgit Schoeberl. 2012. “An Expanding Role for Cell Biologists in Drug Discovery and Pharmacology.” *Molecular Biology of the Cell* 23 (21): 4162–64.

Xiong, Momiao. n.d. “Structural Equation for Identification of Genetic Networks.” *Analysis of Microarray Data*. https://doi.org/10.1002/9783527622818.ch9.

# CHAPTER 3: Network-based Global Chromatin Profiling Fingerprints Reveal Therapeutic Efficacy in Breast Cancer

## 3.1 Abstract

Regulatory abnormalities caused by epigenetic changes due to chromatin modifications are being increasingly recognized as contributors to cancer. While many molecularly targeted drugs have the potential to revert these modifications, their precise mechanism of action in cellular reprogramming is yet to be deciphered. We generated "network-based global chromatin profiling fingerprints" by integrating proteomic/phosphoproteomic, transcriptomic and regulatory genomic data to understand how unique chromatin alterations by post-translational histone modifications regulate cell state changes when treated with drugs in breast cancer. We find H3K27me3K36me3 as a key fingerprint, mediated by chromatin remodelers BRD4, NSD3, EZH2, and a proto-oncogene MYC. We show CDK inhibitors flavopiridol and dinaciclib display selective inhibitory potential toward BRD4/MYC, implicating them as potential therapeutic targets to restitute H3K27me3K36me3 status in breast cancer.

## 3.2 Introduction

Breast cancer is one of the few tumor types in which effective therapies have led to significant improvements in patient survival (Perez 2011). However, the molecular and clinical heterogeneity of breast cancer makes the identification of the most specific and effective therapies challenging (Rugo et al. 2016). To address this problem, recent studies (Mertins et al. 2016; Ellis et al. 2012; Yamamoto et al. 2014; Cancer Genome Atlas Network 2012; Curtis et al. 2012) have employed high-throughput genomic and proteomic technologies to discover the molecular events and critical pathways involved in breast cancer, leading to contextually targeted therapies as well as the development of novel therapeutic targets. Furthermore, analysis and identification of the most appropriate targets have the ability to provide insights into tumor progression and drug resistance mechanisms caused by cellular reprogramming in disease (Dravis et al. 2018; Wahl and Spike 2017).

Cancer arises due to aberrant genetic (Hanahan and Weinberg 2011) and epigenetic dysregulation (Baylin and Jones 2011; Sandoval and Esteller 2012), causing normal cells to proliferate. There is increasing evidence that these regulatory abnormalities are caused by epigenetic alterations through post-translational modification (PTM) of histones (Leroy et al. 2013) resulting from a variety of covalent modifications including, phosphorylation, methylation, acetylation, and ubiquitination at the N-terminal tails of histones. A single or combinatorial set of these modifications on one or more histone tail comprises a 'histone code' (Strahl and Allis 2000) which greatly influences the control of the chromatin structure, function, and interactions and leads to altered downstream cellular processes. The chromatin-associating proteins also known as chromatin remodelers (readers, writers, erasers) recognize, add and remove specific histone modifications through their specialized protein-binding domains (Strahl and Allis 2000; Jenuwein and Allis 2001; Schreiber and Bernstein 2002; Fischle et al. 2003). The histone codes play key roles in the regulation of gene expressions activities, acting as cellular regulators switching genes on and off by making the DNA accessible/inaccessible to transcriptional machinery through remodeling euchromatin (active) and heterochromatin (silenced) states respectively. In contrast to the irreversible genomic mutations that activate oncogenes or inactivate tumor suppressor genes in cancer, histone modifications are reversible and can be used as potential biomarkers for normal or cancer state of cells, and as markers of drug response. Furthermore, chromatin remodelers themselves can be targets of therapy if their specific roles in histone modifications are understood. Epigenetic therapeutics such as vorinostat (SAHA) and romidepsin, inhibitors of histone deacetylases (HDAC), are approved for the treatment of refractory cutaneous T-cell lymphoma (Foss et al. 2011; Khan and La Thangue 2012). Although these drugs have been successful in treatment, their precise mechanism of action, i.e., their action in cellular reprogramming is poorly understood due to the lack of reliable biomarkers for the prediction of their clinical activity.

In a recent large-scale initiative, Library of Integrated Network-Based Cellular Signatures (LINCS) (http://www.lincsproject.org), has carried out multi-omics characterization of response of five cancer cells to 31 drugs, through measurement of phosphoproteins (P100) (Abelin et al. 2016), transcripts (L1000) (Subramanian et al. 2017), and global chromatin profiles (GCP) (Creech et al. 2015). Some of these measurements were carried out at multiple time points post-treatment of cells. The P100 and GCP are Mass Spectrometry (MS) based targeted proteomics assays that include a representative set of phosphopeptides, and different combinations of histone modifications treated by multiple drugs respectively. L1000 data was generated using a microarray platform containing landmark transcript probes obtained from a Connectivity Map (Subramanian et al. 2017) of genes which were invariant across cell states. We report here the relationship between the treatments and cellular response to the treatments. Our primary objective was to explore and identify epigenetic fingerprints uniquely characterizing effective drug response. The combinatorial histone marks measured in response to each treatment represent the epigenetic changes associated with the remodeled chromatin topology and serve as fingerprints of the cellular state.

For initial characterization of the epigenetic fingerprint responses, we used the MCF7 cell line from LINCS study that profiled 96 phosphopeptides at three time points and 60 histone marks profiled 24 hours after treatment with 31 established drugs. This type of high-dimensional data represents significant challenges in analyzing the pattern of drug responses affecting GCP, and deciphering pathways that are causally involved in responses leading to specific GCP. We approached this from the perspective of data and dimension reduction in order to develop mechanistic models of drug response through GCP fingerprints. In the following sections, we describe the integrated network we developed for analyzing the LINCS breast cancer data to 1) uncover the number of distinct ways in which drugs relate to GCP; 2) decipher the unique phosphoproteins networks and pathways that describe histone response to 31 drug treatments; and 3) identify mechanisms involving phosphoproteins regulating a wide range of cellular processes (growth, proliferation and cell division) and gene activity states. Our results demonstrate fingerprints of GCP that comprehensively describe the drug response in cancer cells and further help elucidate the detailed causal mechanisms that lead to these epigenetic profiles.

## 3.3 Results

**Four Pathway-based Histone Signatures Identified by Clustering Method Constitute "Global Chromatin Fingerprint Profiles."** In order to identify fingerprint histone profiles, we investigated the relationships between the 31 drugs targeting serine-threonine kinases in various cell lines including the breast cancer line (MCF7), and the resulting GCP response at 24 hours. We calculated the histone code fold changes by accounting for their differential modifications i.e., changes in histone levels from pre-treatment (MCF7 treated with DMSO) to post-treatment (MCF7 treated with a specific drug) state. Using a non-negative matrix factorization (NMF) clustering method on these histone code fold changes (supplement figure S1), we identified four pathway-based functional histone modules c1, c2, c3 and c4 (**Fig 3.1A, 3.S1D**) and refer to them as "histone signatures" that characterize the response to drugs. Briefly, the objective of NMF is to explain the observed data using a compact number of latent features, i.e., basis components, which abstract the original data as accurately as possible. No independence or orthogonality constraints are imposed on the basis components leading to a simple and intuitive interpretation of the factors that allows the basis components to overlap. This unique feature is particularly interesting in histone modules, where overlapping basis components identify combinatorial histone codes resulting from multiple signaling pathways and indicating a specific signature (**Fig 3.S1**, method section).

To provide a comprehensive mapping of these histone signatures to drugs with respect to their shared signaling pathways, we then generated a molecular network consisting of 91 nodes (comprising histone codes and drugs) and 554 edges (node interactions). Coefficients generated from the assignments of each histone signature profile to the drug prototypes (see method section) are used to represent the strength of the interactions between a histone code and a drug (**Fig 3.S1**). Edge thickness represents the strength of the contributions of drugs to histone codes belonging to the same histone signature. In c1 histone signature, we found 46 histone codes are strongly associated with ten drugs (**Fig 3.S1**) showing mostly inhibitory effects shared by nine common pathways (**Fig 3.1A**). We observed all cyclin-dependent kinase (CDK) inhibitors (flavopiridol, dinaciclib, and PD-0332991) and replication stress inhibitors (VX-970 and SCH 90076) were grouped with the same histone signature. We observed similar groupings for the c2 signature associating 3 histone codes with ten drugs where all AkT/PI3K variants of Ras inhibitors (IPI145, afuresertib, BYL719, dactolisib) grouped together (**Fig S1**). Similarly, in c3, 2 repressive histone marks, H3K56me1 and H3K56me2, are associated with 5 drugs targeting Jak3, Mek1/2, Jnk, Hippo and multikinase pathways. In addition, H3K56me2 is activated by all drugs while H3K56me1 had inhibitory effects on Jnk, Mek1/2 and multikinase pathways, and activation effects on Jak3 and Hippo pathways. The involvement of these pathways regulating these two histones is less clearly established in breast cancer and merits further experimental investigation. In the c4 signature module, we observed nine histone codes associated with IKkKB, Mek1/2, Notch/Wnt/Hedgehog, Gamma secretase, mTOR, and p38 MAPK pathways. We observed a reduction of monomethylation at lysine 9 and phosphorylation at serine 10 (H3K9me1S10ph1K14ac0), a repressive histone code, in all the pathways in this module suggesting its potential as a therapeutic marker mediated by these shared pathways in breast cancer. In addition, we observed overlaps in c2, c3 and c4 for drugs targeting IkKB, Notch, Mek1/2 and Jnk pathways (**Fig S1**) suggesting crosstalk among histone signatures. Collectively, our results suggest strong selective preferences of histone codes towards specific pathway-based therapeutic effects as well as possible crosstalk among the pathways which may lead to off-target effects.

**<u>Histone Prediction Model Provides Quantitative Contributions of Enriched Phosphoproteins Toward Histone Codes</u>.** Next, we sought to identify the phosphoprotein networks representing various interactions among the enriched phosphoproteins and histone codes. Using the P100 phosphoproteins and GCP responses at 24 hours after treatment with the 31 drugs, we developed a combined quantitative and qualitative estimate of significant phosphoproteins contributing to the alterations of histone codes using a partial least square regression (PLSR) method. PLSR is a multivariate method for constructing a predictive model when the number of factors (covariates, e.g., 96 phosphoproteins) exceeds the number of responses (e.g., 60 histones) and the factors are highly correlated. PLSR attempts to extract latent factors across data, accounting for as much of the observed variation as possible while modeling the output/responses accurately. A past study (Gupta et.al, 2010) had shown the effectiveness of partial least square (PLS) application in understanding crosstalk between phosphoprotein signaling in macrophage cells, thus, prompting us to consider a PLS-based regression model. Using PLSR we generated a system model where each histone code is considered as an outcome/response to combined influences (i.e., coefficients) of all phosphoproteins. Each coefficient represents the contribution of individual phosphoprotein towards the level of a histone code (see method section). We evaluated our PLSR model using 10-fold cross-validation. Using a p-value <1.0e-4 (see method section, **Fig 3.S2**), our model generated a histone-phosphoprotein network comprised of 113 nodes, representing histone codes and phosphoproteins and 230 edges (interactions between them) (**Fig 3.1B**). Our results showed H3K27me3K36me3, H3K9ac1S10ph1K14ac0, H3K56me2, and H3K18ac0K23ub1 as highly connected histone codes (hubs with the highest degree), influenced by the statistically significant phosphoproteins: BRD4 (Bromodomain Containing 4), ATAD2 (ATPase Family, AAA Domain Containing 2), NOLC1 (Nucleolar and Coiled-Body

Phosphoprotein 1), SRRM2 (Serine/Arginine Repetitive Matrix 2), and CASC3 (Cancer Susceptibility Candidate 3).

Briefly, BRD4 and ATAD2 are bromodomain proteins. BRD4 is an epigenetic "reader" and belongs to BET family protein that maintains epigenetic memory and regulates cell cycle progression; BRD4 has been shown to have an intrinsic binding specificity for transcription factors such as c-MYC and p53 which are known to promote cancer (Delmore et al. 2011), making it a promising drug target. Similarly, ATAD2 is a novel cofactor for MYC, overexpressed and amplified in aggressive tumors. It has been shown that downregulation of ATAD2 via siRNA results in increased apoptotic activity, suggesting a role for inhibitors of ATAD2 in cancer cell death and tumor regression (Caron et al. 2010). NOLC1 is a nucleolar protein that regulates RNA polymerase I by connecting RNA polymerase I to ribosomal processing and remodeling enzymes, resulting in translational remodeling. It has a high binding affinity to c-MYC and Max transcription factors which play an important role in cancer. Although NOLC1 has not been studied extensively, a previous study (Hwang et al. 2009) found NOLC1 to have transcription factor-like activity in nasopharyngeal cancer progression suggesting its possible role in other cancers. SRRM2 protein is known to be involved in pre-mRNA splicing and has binding specificity for p53. SRRM2 has been detected as a 5'-3' Exoribonuclease 2 (Xrn2)-interacting protein that is involved in premature termination of RNA polymerase II transcription (Sansó et al. 2016; Brannan et al. 2012) thus affecting cell cycle progression. CASC3, also known as MLN51 is a component of the exon junction complex (EJC) whose expression has been shown to be elevated in some breast cancer cell lines (Tomasetto et al. 1995). The EJC is known to be involved in a surveillance mechanism that degrades mRNAs with premature translation termination codons through a nonsense-mediated mRNA decay (NMD) function, thereby, promoting cell cycle arrest. Currently, the implications of

ATAD2, NOLC1, SRRM2 and CASC3 proteins in cancer regulation is poorly understood, and our results suggest that these regulators serve as potential novel oncogenic drivers mediating histone modifications in breast cancer.

Taken together, the histone-peptide network reveals candidate phosphoproteins that serve as potential therapeutic targets; these candidate phosphoproteins alter histone codes resulting in alterations in cell cycle progression in breast cancer.

**Integrated Approach Provides a Three-dimensional View of Molecular Interactions among Drugs-Phosphoproteins-Histones.** To further elucidate the influence of specific drugs on phosphoproteins and downstream histone codes, we then developed a 3D view of the molecular interactions (phosphoproteins-drugs-histones) by integrating histone signatures with the drug-phosphoprotein interaction network resulting in an integrated phosphoproteins-histones-drugs network (iPhDnet). The network consisted of 144 nodes, 742 interactions (**Fig 3.1C**) and 2157 unique maps (interaction profiles) which are stored into a relational database. The iPhDnet serves as a quantitative atlas of global chromatin profile fingerprints that can be used to generate hypotheses linking drugs, pathways, phosphoproteins, and histones, to understand drug response pathways in cancer.

Our chromatin profile fingerprints revealed an overall reduction in histone levels in active marks such as methylation of H3K36, H3K4, and acetylation of H3K9 when treated with drugs, which are consistent with previous studies (Lewis et al. 2013; Leroy et al. 2013; Zhu et al. 2015). We observed that reduction of phosphoprotein level in SRRM2 was positively correlated ($p\text{-val} < 2.7e\text{-}04$) with H3K4me1 and H3K4me3 when treated with drugs that belonged to c1 signature histone module. While trimethylation of H3K27 is a repressive histone mark associated with transcriptionally silenced chromatin in most cancers (Lewis et al. 2013; Leroy et al. 2013; Zhu et al. 2015), our analysis revealed inhibitory effects of differential modification levels of trimethylation of H3K27. These findings are consistent with the results of prior studies in breast cancer (Holm et al. 2012; Ren et al. 2012). Likewise, Abelson interactor protein-1 (ABI1), an adaptor protein involved in cell migration, along with its downstream effector phospho-Akt (p-Akt) has been implicated in the spread of breast cancer (Wang et al. 2011); ABl1 is positively correlated with reduced H3K27me3K36me3 histone mark (p-val < 7.02e-05) when inhibited by CDK inhibitor flavopiridol. Additionally, we observed significant associations (p-val < 7.5e-05) of BRD4, NOLC1, ATAD2 and SRRM2 with H3K27me3K36me3 when treated with CDK inhibitors flavopiridol, dinaciclib, and palbociclib (PD-033291) (**Fig 3.1C, table 3.S1**). Taken together, iPhDnet shows that inhibiting ABI1, BRD4, NOLC1, ATAD2, and SRRM2 with the help of CDK inhibitors may be sufficient to induce the heterochromatin state where the repressive mark H3K27me3 colocalizes with the active mark H3K36me3. This finding suggests that for a stable reversion of epigenetic silencing state in breast cancer, a reversal from the malignant euchromatin to normal heterochromatin may be dictated by H3K36me3. In addition, we observed a decrease of H3K27me3 when treated with all CDK inhibitors supporting the observation made by a prior study where induction of a CDK inhibitor was associated with a lower level of H3K27me3 in breast cancer (Yang et al. 2009). Most likely the reduced level of H3K27me3 is associated with CDK inhibitors in breast cancer. Therefore, we postulate that ABI1, BRD4, NOLC1, ATAD2, and SRRM2 can mediate H3K27me3K36me3 for reversion of epigenetic silencing in breast cancer using CDK inhibitors. We also postulate that the reduction of H3K27me3 may be a compensatory effect of other cofactors of these phosphoproteins working together to induce cell growth arrest, suggesting the potential for a combinatorial treatment strategy in breast cancer.

**Flavopiridol and Dinaciclib Emerge as Potential CDK mediated Therapeutics Modulating H3K27me3K36me3 in Breast Cancer.** To examine the validity of the identified enriched phosphoproteins mediated by specific drugs, we first compared our findings with prior experiments on identification of various histone codes in breast cancer. Our findings are consistent with other reports that interrogated specific PTMs in breast cancers. We observed the reduction in H3K4me1, H3K4me3, H3K9ac, H3K27me3K36me0, and elevation in H3K18ac0K23ub1 histone codes. We are unaware of any studies that examined modifications of H3K56 methylation or other combinatorial histone codes in breast cancer. A summary of our PTM findings is provided in a table (**Fig 3.2A**). While the aforementioned studies have investigated the modification of H3K27me3 and H3K36me3, combinatorial assembly of repressive H3K27me3 and active H3K36me3 marks (H3K27me3K36me3), have not been previously studied in breast cancer. Hence, we further analyzed the molecular mechanisms associated with H3K27me3K36me3 modulation to identify potential targets for therapeutic interventions in breast cancer.

Since H3K27me3K36me3 was assigned to the C1 histone signature, we considered evaluating the effect of various CDK inhibitors and the kinase inhibitors that belonged to C1 histone signature on the enriched phosphoproteins. Together with the phosphorylation profiles (phosphorylation status at 3, 6 and 24 hours) of these enriched phosphoproteins and their molecular interactions, we observed anti-tumorigenic effects of flavopiridol on ABI1, BRD4, NOLC1 and SRRM2; anti-tumorigenic effects of dinaciclib on BRD4, NOLC1, and SRRM2; and anti-tumorigenic effects of PD-0332991 on NOLC1 modulating H3K27me3K36me3 (**Fig 3.2B**). Similarly, we observed an anti-tumorigenic effect of drugs in c1 histone signature modulating H3K4me1 and H3K4me3 (**Fig 3.2C**) through CASC3 and SRRM2. From these findings, we observe that flavopiridol and dinaciclib induce similar regulatory pathways suggesting similar therapeutic responses.

Flavopiridol is a CDK inhibitor with high selectivity for CDK9 (Bosken et al., 2014). It has been used in a phase II clinical trial for the treatment of relapsed/refractory lymphoma or multiple myeloma (Dispenzieri et al., 2006). Similarly, dinaciclib is a highly potent CDK inhibitor with selectivity for CDK1, CDK2, CDK5, and CDK9 (Paruch et al., 2010). It is in phase III clinical trials for the treatment of refractory chronic lymphocytic leukemia. To further evaluate the concordance of these two inhibitors at a global level, we performed NMF clustering analysis on LINCS data from four other cancer cell lines namely, pancreas (YAPC), skin (A375), lung (A549), prostate (PC3) and as a control a neural progenitor cell line (NPC). This resulted in histone signature modules for each of these cell lines (**Fig 3.2D**). Using these signature modules, we then computed the Rand Index (RI) between each paired cell lines to measure the similarity between two data groupings (see method for details). The RI value range from 0 (completely dissimilar group assignment) to 1 (exactly same group assignment). We observed a high concordance (RI=0.65) between MCF7 breast cancer and YAPC pancreas cancer cell lines (**Fig 3.2D**). Similarly, we computed RI on drug assignments to see how many drugs were grouped in the same histone signatures across all six cell lines. Interestingly, the analysis assigned flavopiridol and dinaciclib in the same histone modules across all six cell lines with an RI score of 1 (**Fig 3.2E**). To further corroborate these findings, we performed a Pearson correlation analysis on the phosphoprotein data. The results provide further support for the concordance between flavopiridol and dinaciclib showing a strong correlation between the two drug responses at 3 to 6 hour (r = 0.59) and 6 to 24 hour (r = 0.69) (**Fig 3.S3A and 3.S3B**). Furthermore, a linear regression analysis of histone expressions at 24 hours showed similar treatment effects between flavopiridol and dinaciclib (positive slope, p-value = 1.85e-11, adjusted r-squared =0.536 ) (**Fig 3.S3D**). Additional support is provided by a comparative structural analysis study (Ember et al., 2014) that indicates a similar affinity of flavopiridol and dinaciclib for acetylated lysine (KAc) binding site of bromodomain (BRD). Collectively, our findings show the similarity and the efficacy of flavopiridol and dinaciclib as potential candidates for BRD mediated CDK therapeutics in breast cancer.

**Mechanistic Causal Network (MCN) Reconstruction Supports BRD4 Mediated Cell Cycle Arrest caused by Impaired Transcriptional Elongation when Treated with Flavopiridol and Dinaciclib in Breast Cancer.** To gain mechanistic insights into H3K27me3K36me3 mediated regulation by flavopiridol and dinaciclib, we reconstructed mechanistic causal networks (MCN) that demonstrated the dynamics of the regulatory machinery involving the enriched phosphoproteins measured at varying time points. To construct a dynamic signaling network we first generated protein-protein interactions (PPI) for the enriched phosphoproteins, at 3, 6 and 24 hours, using the STRING database (http://string-db.org/). We only considered experimentally validated proteins that had a moderate to a high confidence score (see methods section). Using one-way analysis of variance (ANOVA) with astringent statistical criterion (p-value < 1.0e-4), we then generated significant phosphoproteins that were enriched at 3 and 6 hours when treated with flavopiridol. Next, we extracted a list of inferred proteins by carrying out PPI on the enriched phosphoproteins that were previously generated from iPhDnet at 24-hour time post-treatment. Using these inferred proteins, we then performed back propagation PPI by linking them to earlier time points, 6 and 3 hour enriched phosphoproteins. We repeated this process for dinaciclib. The resulting networks were then visualized using the Cytoscape software (Shannon et al. 2003)

(**Fig 3.3A, 3.3C**).

Our results showed four phosphoproteins in flavopiridol (BRD4, TMPO, FAM76B, and RBM14) and three phosphoproteins (TMPO, FAM76B, and TPX2) in dinaciclib remained enriched across 3, 6, and 24-hour time points. Moreover, the network showed binding of BRD4/NSD3 which is consistent with a previous study (Rahman et al., 2011) where they found that reduced H3K36 methylation was a result of depletion of BRD4 or NSD3. It has been reported that Nuclear receptor SET domain-containing 3 (NSD3), also known as WHSCL1, is a methyltransferase that binds to BRD4 complexes at the promoter region to regulate levels of H3K36me3, affecting DNA repair, transcription initiation and elongation/termination process (Wen et al., 2014; Li et al., 2013). To further investigate the mechanisms by which this BRD4/NSD3 complex contributes to mediating cell cycle progression through the recruitment of H3K36me3 and binding to upstream regulators/cofactors, we performed enrichment analyses on the genes representing these phosphoproteins using the Enrichr tool (Chen et al., 2013; Kuleshov et al., 2016). The enrichment analysis identified MYC, POU5F1 (OCT4), ESR2, UPF1, SMARCA4 and BRCA1 as commonly enriched upstream/core regulators of phosphoproteins for flavopiridol and dinaciclib. These core regulatory factors have been shown to interact with super-enhancers which are master transcription factors that control cell identity by exhibiting higher sensitivity to transcription activities (Whyte et al., 2013). Additionally, POU5F1 (OCT4) is a pioneer transcription factor whose expression has been shown to have an association with a high level of H3K36me3 active mark (Musselman et al., 2012). Moreover, Loven et al. postulated that the heightened sensitivity of super-enhancer genes to reduced levels of BRD4 may lead genes associated with super-enhancers to a greater transcriptional reduction than genes with average enhancers when BRD4 is inhibited. This further implicates the efficacy of flavopiridol and dinaciclib targeting BRD4 in breast cancer.

Furthermore, our enrichment analysis showed interactions between spliceosome mediated activities through the core regulators: E2F4, UPF1, ILF3, and SMARCA4, and the components of exon junction complex (EJC) comprised of enriched phosphoproteins namely, NOLC1, SRRM2, CASC3, EIF4A3 and RBM8A (Le et al., 2016). Interestingly, EJC has been shown to have an association with Wnt/Notch signaling activity in the cancer signaling pathway (Liu, et al., 2016) suggesting crosstalk among pathways with possible off-target effects. The enrichment analysis showed interactions among the mitotic regulators (TPX2, AURKA) with TP53 activity and ATAD2 that formed a cluster, regulating cell cycle through alternative splicing. These regulators: TPX2, AURKA and EJC complex are known substrates of positive transcription elongation factor's (P-TEFb), which bind indirectly with BRD4. From these findings, we postulate that these regulators may serve as potential novel targets towards breast cancer therapy. Further experimental validation is warranted.

We further observed enrichment of estrogen receptors ESR1 and ESR2 as upstream regulators for SRRM2 and NOLC1 supporting possible MYC mediated endocrine activities. A recent study showed high MYC transcription mediated by CDK9 as a critical determinant of endocrine-therapy resistance breast cancers (Sengupta et.al, 2014). Therefore, it is reasonable to postulate that inhibition of SRRM2 and NOLC1 which interact with BET proteins may prove to be efficacious for endocrine therapy refractory breast cancers in a clinical setting.

Next, we investigated how flavopiridol and dinaciclib lead to preferential loss of BRD4/NSD3 impacting the super-enhancer-associated oncogene MYC, thereby, promoting cell cycle arrest in breast cancer. Based on our results and the evidence from previous studies

(Horiuchi et al., 2012; Kwak et al., 2013; Li, et al., 2013; Lu, et al., 2015), we postulate that cell cycle arrest  associated with the reduced H3K27me3K36me3 phenotype occurs through the following mechanism: 1) flavopiridol and dinaciclib inhibit BRD4, 2) as a result, H3K36me3 level is reduced through BRD4's interacting partner NSD3, 3) reduction of BRD4 then impairs the catalytic activity of CDK9's ability to bind to positive transcription elongation factor b (P-TEFb), which is sequestered by 7SK snRNP to acetylated chromatin at the MYC locus, 4) this suppresses P-TEFb's phosphorylation at serine 2 of the Pol II carboxyl-terminal domain (CTD) and the DRB Sensitivity Inducing Factor (DSIF) subunit SPT5, causes widespread RNA polymerase II to pause at gene promoters, thereby promoting cell cycle arrest. As a functional consequence of the loss of CDK9 activity, MYC expression is elevated as a compensatory effect, which activates EZH2, a subunit of the PRC2 complex, resulting in methyltransferase activity leading to H3K27me3 reduction. This is a plausible rationale for the global reduction of H3K27me3K36me3 in breast cancer when treated with flavopiridol and dinaciclib. However, the reduction of H3K27me3 and H3K36me3 levels alone were markedly pronounced than the reduction level of the 'bivalent domains' H3K27me3K36me3, further suggesting that targeting H3K36me3 or H3K27me3K36me3 by these drugs is more efficacious than targeting H3K27me3 alone in breast cancer.

In conjunction with the proteomic analyses, we performed transcriptomic analyses using L1000 data on genes representing the enriched phosphoproteins to capture in vitro gene activity levels. We examined the efficacy of flavopiridol and dinaciclib in breast cancer at transcriptomic level, by looking at the CDK inhibitor genes, down-regulated genes that represented cell cycle genes and parent proteins of enriched phosphoproteins associated with H3K27me3K36me3 mark in the MCF7 cell line. We isolated 31 functionally significant genes with $p < 0.05$:  example includes, CDKNA2, BRCA1, AURKA, MELK, EZH2, CCNA2, BRD4, NOLC1, SRRM2, MYC, CASC3 (**Fig 3.3B, 3.3D**). From the results, we observed an increase in CDK inhibitor gene CDKNA2 expression and decrease in expressions for all cell cycle genes across 3, 6, and 24-hour time points in flavopiridol and 6 and 24-hour time points in dinaciclib. Together, the results from these transcriptomic analyses further validate our proteomics analyses.

**Functional analysis of transcriptomic data associate gene regulators response to cell cycle.** To perform functional analyses on the differential expressions (DE) of these 31 genes after treated with flavopiridol and dinaciclib, we then performed DAVID annotation analysis (D. W. Huang, Sherman, and Lempicki 2009) and KEGG pathway analysis. KEGG pathway analysis of these genes showed their involvement in the cell cycle, p53 signaling pathway, ErbB signaling, and pathways in cancer across 3, 6, and 24 hours for flavopiridol and 6 and 24 hours for dinaciclib (**Fig 4A, 4B**). Furthermore, these genes are associated with the following GO categories: nuclear lumen, regulation of cell death, cell cycle, programmed cell death, and protein kinase activities (**Fig 3.4C**). We then identified marker genes (MYC, CCNA2, EZH2, MELK, and AURKA) whose DE were statistically significant across normal, pre-treated and post-treated breast cancer tissues at 3, 24 hours. The DE results showed downregulation of MYC, CCNA2, EZH2, MELK, and AURKA marker genes when normal breast tissue (MCF10A) treated with DMSO is compared against breast cancer tissue (MCF7) treated with flavopiridol or dinaciclib at 24 hour. We observed similar effects when MCF7 treated with DMSO (pre-treatment) is compared against MCF7 treated with flavopiridol or dinaciclib (post-treatment) at 24 hour (**Fig 4SA**). The comparisons between normal vs cancer: MCF10A(DMSO) vs MCF7(DMSO), showed upregulation of EZH2 and AURKA and downregulation of MYC, CCNA2, and MELK (**Fig 4SA**). In addition to the vitro analyses, we performed a systematic analysis on in vivo gene activities of these marker genes using the molecular taxonomy of breast cancer international consortium (METABRIC) study (Curtis C, et al. 2012) dataset from TCGA data repository. These samples consisted of 113 normal patients breast tissue and 303 ER+/HR+/HER2- cancer patients breast tissue. Consistent with literature and our pre-treatment vs post-treatment, these marker genes are downregulated in normal patients. Corresponding survival z-scores for these marker genes were obtained from the prediction of clinical outcomes from genomic profiles (PRECOG) datasets where higher z-scores of these genes have shown to be prognostics for longer patient survival in breast cancer making them potential biomarker candidates (**Fig 3.4SB**). Taken together, the transcriptomic profile shows strong associations of MYC, AURKA, MELK, EZH2, CCNA2, BRD4, NOLC1, SRRM2, CASC3 with the regulation of cell cycle and programmed cell death activities in breast cancer when treated with flavopiridol and dinaciclib.

**Fingerprint global chromatin profiling reveals crosstalk among "regulators" in breast cancer signaling pathways**. Finally, to highlight potential BRD4 mediated off-target effects of flavopiridol and dinaciclib using our global chromatin profiling fingerprints, we constructed a detailed view of the crosstalk among the various regulators (phosphoproteins, protein complexes, transcription factors) (**Fig 5**). We accomplished this by further generating PPI using STRINGdb to incorporate inferred proteins/protein complexes for other signaling pathways that may interact with the CDK pathway. The detailed view of the of the breast cancer signaling landscape revealed various regulators associated with specific signaling pathways mediating cellular activities such as cell cycle regulation, apoptosis and transcriptional regulation for cell cycle progression and cell proliferation. As part of the cell cycle, inhibition of CDK by flavopiridol and dinaciclib showed molecular cascades of interactions among BRD4, NSD3, SRRM2, NOLC1, MYC with the P-TEFb complex and its recruitment to the proximal promoter region of MYC to block transcriptional elongation of RNA Pol II. In addition, our results showed the presence of crosstalk among CDK, IkK, AKT, PI3K, and Map3K7 pathways when P-TEFb binds to AURKA, TPX2, and other proteins. For example, IkB, an enzyme complex that is part of the NF-κB signaling pathway, interacts with P-TEFb via AURKA to activate the CDK pathway. P-TEFb targets the intrinsic kinase activity directed towards RNA Pol II essential for transcriptional initiation, elongation, and inhibition. Furthermore, BRD4 has been implicated in activating NF-κB pathway by recruiting P-TEFb to acetylated RELA (Huang et al. 2009). As a consequence of the CDK and BRD4 inhibition, we observed a reduction of Map3K7 phosphorylation, which inhibited JNK expression resulting in an increase of H3K56me2 level. Hyperactive RAS then acts as a signaling switch that converts JNK's role from pro- to anti-tumor signaling through the regulation of Hippo signaling activity by inhibiting the PDPK1 phosphoprotein. A recent study has shown that the combined effect of PI3K and BET inhibition in a wide range of cancer cell lines resulted in apoptosis, tumor regression, and clamped inhibition of PI3K signaling (Stratikopoulos et al., 2015). While EJC regulators indirectly bind to P-TEFb recruited by BRD4 via RBM8A and ZC3H18, they have a secondary binding effect with Wnt/Notch signaling pathway components. From our analysis, we observed NOLC1 interacted with the EJC regulators: SRRM2, CASC3, EIF4A3 and RBM8A proteins; Thus, we postulate that NOLC1 mediates Wnt/Notch signaling activity through Notch intracellular domain (NICD) and monoubiquitylation of H3K23 (H3K18ac0K23ub1) by translocating to RNA Pol I. Collectively, these results indicate that BRD4 is an atypical kinase that could interact with a diverse group of kinases resulting in pleiotropic effects when treated with flavopiridol and dinaciclib. However, a number of small molecules such as JQ1, i-CDK9 have shown high selectivity and potent inhibitory activity against CDK9, thereby, demonstrating the efficacy of BET bromodomain inhibitors for treating cancers (Filippakopoulos et al., 2010).

### 3.4 Discussion

Global chromatin profiling fingerprints represent a new way to identify response of tumor cells to drug treatments. Further, the GCP also serve as endpoints of mechanisms responding to drugs and has the potential to provide insights into the detailed networks and their perturbations. Through our integrative network analysis, we were able to identify four distinct histone signatures and enriched phosphoproteins that contributed to specific histone codes. While our network shows multiple drugs and phosphoproteins regulating H3K27me3K36me3, flavopiridol, and dinaciclib convincingly demonstrated selective inhibitory effects on chromatin reader BRD4 modulating H3K27me3K36me3. In particular, our study shows that BRD4 mediates cell cycle regulation that impacts H3K27me3K36me3 histone mark. This implicates H3K27me3K36me3 as a potential biomarker that can be targeted by flavopiridol and dinaciclib to induce cell cycle alterations by BRD4.

An important objective of our study is to understand the dynamics of the regulatory machinery of H3K27me3K36me3 modulation by flavopiridol and dinaciclib. Our MCN reconstructions identify NSD3, AURKA, CCNA2, EZH2, MYC as interacting partners of BRD4. Our transcriptomic results show overexpression of a CDK inhibitor gene, CDKN2A, which acts as a tumor suppressor gene by inducing cell cycle arrest in G1 and G2-M phase. The results also show reduced expression of several cell cycle genes, AURKA, BRCA1, CCNA2, MELK, TP53, EZH2, suggesting their roles in the response to these drug treatments. We show that these drugs inhibit BRD4 and reduce H3K36me3 level through BRD4’s interaction partner NSD3. They displace the P-TEFb complex from acetylated chromatin to MYC locus to inhibit transcription. These cascading effects in turn induce growth arrest of breast cancer cells. We observed a global loss in H3K27me3, an initial decrease of MYC at 3 and 6 hours, and subsequent induction of

BRD4, NSD3 and MYC expressions at 24 hour by flavopiridol and dinaciclib. We postulate that the global loss of H3K27me3 due to the compensatory effect of CDK9 loss is a general phenomenon likely caused by a common mechanism, independent of BRD4, NSD3 and MYC expression levels (figure 3B, 3D). Because of this compensatory mechanism, our results show that the inhibition of both CDK9 catalytic activity and MYC expression, mediated by BRD4, cause synergistic induction of growth arrest of cancer cells. This suggests possible dual roles of BRD4 on H3K27me3K36me3 simultaneously inhibiting CDK9 and inducing MYC to effectively induce cell cycle arrest. Therefore, it is tempting to speculate a possible feedback loop present between BRD4, NSD3, CDK9, and MYC regulation. Although we provided a plausible mechanistic view of the molecular machinery of BRD4 mediated H3K27me3K36me3 modulation, the mechanisms underlying the loss of H3K27me3 in tumors remains somewhat unclear. A number of previous studies reported that overexpression of EZH2 results in a different PRC complex, namely, PRC4 showing histone substrate specificities (Kuzmichev et al., 2005); therefore, loss of H3K27me3 may relate to a new PRC complex formation (Cao et al., 2004), or protein modification in components of PRC complexes (Cha et al., 2005). Experimental verification is needed to gain further insights into the underlying loss of H3K27me3.

From the MCN analyses, we infer that BRD4 enrichment is maintained across 3, 6, 24-hour time points in flavopiridol, which lead us to conclude that the cell cycle arrest is induced via direct BRD4 mediation during G1/S phase as well as during late mitosis, G2/M transition. In dinaciclib, however, we observed enrichment of the EJC regulators involved in spliceosome related activities at 3 and 6-hour time points, which indicates cell cycle arrest most likely occurred due to nonsense-mediated mRNA decay during G1/S phase, and indirect BRD4 mediation during G2/M phase (figure 3E). From these analyses, we can link cell cycle control to cell cycle arrest through the presence of an alternative splicing network. When treated with dinaciclib, we observed EJC members SRRM2, CASC3 and EIF4A3 interacted with both upstream transcription factors UPF1, a known regulator of nonsense-mediated mRNA decay (NMD), and Interleukin Enhancer Binding Factor 3 (ILF3). These transcription factors are known to modulate the Wnt/Notch signaling pathway through NMD and are highly active in pluripotent cells (Lou, et al., 2016), suggesting possible influences in cellular state remodeling. We also observed enrichment of AURKA and TPX2 regulators which are known to modulate cell program death via Bcl-x, a BCL2 family apoptosis regulator (Moore et al., 2010). All of these findings warrant further experimental investigation.

From our MCN analysis, we further observe the presence of a super-enhancer binding gene POU5F1 (OCT4) upstream of BRD4 suggesting the possible role of BRD4 in regulating pluripotency gene expression by exhibiting a “stemness” behavior. Previous studies have shown positive correlation between BRD4 and the level of H3K36me3 with OCT4 (Liu et al., 2014; Barrand et al.,2010). Depletion of BRD4 has been shown to decrease the pluripotency of OCT4 by changing the cellular fate through disruption of signaling pathways controlling differentiation (Wu et al., 2015). Our analysis shows that, BRD4 interacts with the transcription factor SMARCA4 (SWI/SNF-related, matrix-associated, actin-dependent regulator of chromatin, subfamily a, member 4), a key regulator of ESC self-renewal and pluripotency, known to regulate NANOG homeobox (NANOG) expression in the NANOG regulatory regions (Liu et al., 2014). NANOG interacts directly with OCT4, and SRY-Box 2 (SOX2) genes, which function as pluripotent transcription factors contributing to the reprogramming of somatic cells into an ESC-like pluripotent state (Liu et al., 2013). As a result, they have a profound impact in cancer biology as they provide great promise for clinical applications where reducing their expression or blocking their pathways, may inhibit tumor growth and turn-off the cancer "switch" (Liu et al., 2013). This makes H3K36me3 a potential biomarker, regulated by the super enhancer-mediated BRD4 to study tumor transformation, tumorigenesis, and metastasis in breast cancers as well as in other cancers. From our in vitro transcriptomic analysis, we found OCT4 expression was slightly upregulated (not shown), and SOX2 and NANOG were downregulated (not shown). These findings raise the possibility that targeting NANOG, OCT4, and SOX2 in determining chromatin marks mediated by enhancer-binding proteins, and their precise regulatory mechanisms will identify new components of the transcriptional regulatory networks that may be relevant to tumor progression.

Finally, the global chromatin profiling fingerprints of breast cancer landscape reveal crosstalk among various signaling pathways belonging to specific histone signatures, suggesting possible combinatorial targeted therapies to address off-target effects. In particular, these fingerprints show crosstalk between CDK and IkB signaling pathways involving interactions between P-TEFb and AURKA. Overexpression of AURKA is linked to many cancers. Our transcriptomic results show down-regulation of AURKA when treated with flavopiridol and dinaciclib, further suggesting the efficacy and ability of these drugs to minimize off-target effects in breast cancer. These observations have implications for the discovery of cancer therapeutics directed at global chromatin fingerprinting in diverse cancer types.

## 3.5 Materials and Methods

**Data Acquisition.** The experimental data were generated by the NIH LINCS Proteomic Characterization Center for Signaling and Epigenetics (PCCSE) repository. Level 3 (log 2 normalized) targeted phosphoproteomics assay (P100) against 96 phosphopeptides data, and level 3 (log 2 normalized) global chromatin profiling assay (GCP) against 60 probes that monitor combinations of post-translational modification on histones data using various cancer cell lines including MCF7 (breast), YAPC (pancreas), A375 (melanoma), PC3 (prostate), A549 (lung) and NPC (Neural Progenitor) were downloaded. These assays were treated with 31 serine/threonine kinase inhibitors (drugs) at various concentrations, DMSO as a negative control and consisted of three biological replicates. Three time points (3, 6, 24 hour) were available for P100 data while a single time point (24 hour) was available for GCP data in MCF7 cells. Single time point (3 hour) was available for P100 data and a single time point (24 hour) was available for GCP data in YAPC, A375, PC3, A549, and NPC cell lines.

The experimental transcriptomic data was generated by the NIH LINCS Connectivity Map (CMap) using a microarray-based platform. This assay, which is known as L1000, contained 978 landmark transcripts whose expressions were invariant across cell states. In addition, 11350 inferred genes were also obtained using the L1000 inference algorithm (Subramanian et al. 2017). Level 3 (log 2 normalized) L1000 data for two breast cell lines: MCF10A (normal tissue) and MCF7 (cancer tissue) were obtained. For MCF10A, 3 and 24-hour pre-treatment (DMSO, 150, 116 respectively biological replicates) and post-treatment (flavopiridol, 4 biological replicates) were available. For MCF7 pre-treatment (DMSO) 100 replicates at 3, 24 hour, 3 replicates at 6 hour and post-treatment (flavopiridol) 4 replicates at 3, 24 hour and 3 replicates at 6 hour were available. Additionally, pre-treatment (DMSO) 11 replicates, and post-treatment (dinaciclib) 3 replicates at 24 hour were available.

Patient-level data were obtained from The Cancer Genome Atlas (TCGA) where breast tissue samples were obtained from 113 normal patients and breast cancer tissue samples were obtained from 303 ER+/HR+/HER2- cancer patients from molecular taxonomy of breast cancer international consortium (METABRIC) study (Curtis C, et al. 2012). Integrated cancer gene expression and clinical outcome data were obtained from the prediction of clinical outcomes from gene profiles (PRECOG), encompassing 166  cancer expression datasets from ~18000 patients diagnosed with 39 malignancies with their overall survival data (Gentles et al., 2015).

**Data Pre-processing.** Replicates were used to impute missing data by taking their weighted average values during the pre-processing step. Differential histone modifications and phosphorylation changes were computed by taking fold changes of each perturbed phosphopeptide and histone code with respect to DMSO.  These resulted in two data matrices, i) phosphoprotein profiles consisting of [96 peptides x 31 drugs], and ii) global chromatin profiles consisting of [60 histone modifications x 31 drugs]. Prior to modeling, data were normalized with respect to the mean and standard deviation of the respective variables. A log2 transformation was performed on TCGA data.

**Experimental Validation.** L1000 genes expressions were used to validate differential gene expressions of the 31 functionally significant genes (cell cycle genes, CDK inhibitor gene CDKN2A, transcription factor MYC and genes representing the enriched phosphoproteins) to capture in vitro gene activity levels in normal (MCF10A) vs cancer (MCF7) cell lines. In addition, TCGA and METABRIC datasets were used to validate in vivo gene activities of the marker genes. Corresponding survival z-scores for these marker genes were obtained from the PRECOG datasets where lower expressions of these genes were used as prognostics for longer patient survival in breast cancer.

**Histone Signature Identification.** An unsupervised clustering technique, non-negative matrix factorization (NMF) outlined in chapter 2, was used to stratify histone signatures. R Statistics package was used for the calculation and Cytoscape was used to generate network graphs.

**Histone Prediction Model.** Histone-peptide interaction network was generated using partial least square regression (PLSR) method based on Kraemer et al. formulation (Krämer and Sugiyama 2011) (outlined in chapter 2).

**Integrated Phosphoprotein-Histone-Drug Network (iPhDnet).** Using the coefficients from the histone signatures (c1, c2, c3, and c4) and the drug prototypes using NMF and model coefficients of phosphoproteins towards histone model prediction using PLSR, an integrated 3D network file is constructed connecting drugs to phosphoproteins and phosphoproteins to histones (iPhDnet). This is described in chapter 2.

**Mechanistic Causal Network (MCN) Reconstruction.** A time-varying mechanistic causal network was constructed by back propagating iPhDnet, previously generated for 24 hour from P100 data. We first used a one-way ANOVA with a p-value of 1.0e-4 to populate enriched (statistically significant) phosphoproteins at 6 and 3-hour time points. We then inferred protein-protein interactions for the phosphoproteins enriched in 24 hour by mapping them to the STRING database. An interaction score of 0.8 and above, experimentally validated PPIs, and gene fusions criteria were used to obtain these inferred proteins. Our final MCN was constructed by back propagating our mapping of the inferred proteins from 24 hour to enriched phosphoproteins in 6 and to 3 hour. Additional protein-coding genes were generated and added to the final MCN using the EnrichR tool (http://amp.pharm.mssm.edu/Enrichr/enrich). We then validated our MCN by matching them against significant differentially expressed (DE) genes in L1000. Cytoscape was used to view the final reconstructed MCN.

**Differential Expression and Functional Analyses of L1000 and TCGA Data.** Differential Expression and Functional Analyses of L1000 and TCGA Data Differential expression analyses for 978 landmark genes from L1000 assay treated with flavopiridol and dinaciclib were performed using the unpaired t-test implemented in CyberT. Cyber-T is based on a regularized Bayesian framework that addresses technology biases and low replication levels in high throughput data (Baldi and Long 2001). These analyses were performed on 3, 6 and 24-hour datasets. Multiple corrections were applied to p-values using Benjamini Hochberg. Similarly, differential expression analyses of TCGA matched normal vs cancer patients were performed using unpaired t-tests. Cyber-T web server (Kayala and Baldi 2012) was used to generate these analyses.

To examine the in vitro effectiveness of flavopiridol and dinaciclib, we compared differential gene expressions between i) DMSO treated normal breast tissue (MCF10A) and DMSO treated breast cancer tissue (MCF7), ii) DMSO treated MCF10A and flavopiridol/dinaciclib treated MCF7, and finally iii) DMSO treated MCF7 and flavopiridol/dinaciclib treated MCF7. To account for between-group differences for specific genes, a one-way ANOVA with Tukey's post hoc test was performed. P-values for overall difference between groups were corrected using the Benjamini Hochberg multiple corrections. We considered genes to be potential drug targets only if they satisfied the following criteria: 1) statistically different in normal vs cancerous tissue, i.e., p-val <0.05 in DMSO treated MCF10A vs DMSO treated MCF7 and 2) statistically different in pre-treatment and post-treatment, i.e., p-val <0.05 in DMSO treated MCF7 vs flavopiridol/dinaciclib treated MCF7.

To examine the functional enrichment of significant DE genes in L1000 data treated with flavopiridol and dinaciclib, Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway analysis and Gene Ontology (GO) enrichment were performed using the DAVID Functional Annotation Tool (D. W. Huang, Sherman, and Lempicki 2009) for each treatment.

**Cluster Similarity Evaluation.** We used the Rand Index (RI) to evaluate the similarity of cluster assignments between every paired treatment in breast cancer and between paired cell lines. RI computes the percentage of pairs of objects for which both classification methods, the computed and the ideal one, agree. It is computed using False Positives (FP), False Negatives (FN), True Positives (TP) and True Negatives (TN) as follows:

$$RI = \frac{(TP+TN)}{(TP+TN+FP+FN)} \qquad \textit{(Equation 8)}$$

The RI value ranges from 0 (completely dissimilar group assignment) to 1 (exactly same group assignment).

**Data and Software availability.** Genomic, transcriptomic, epigenetic, and proteomic data files are available from the public online portal (https://panoramaweb.org/project/LINCS/GCP/begin.view?). Source codes are implemented in R 3.3.1 and are freely available for download at (https://github.com/smollah/iPhDnet).

## 3.6 Figures

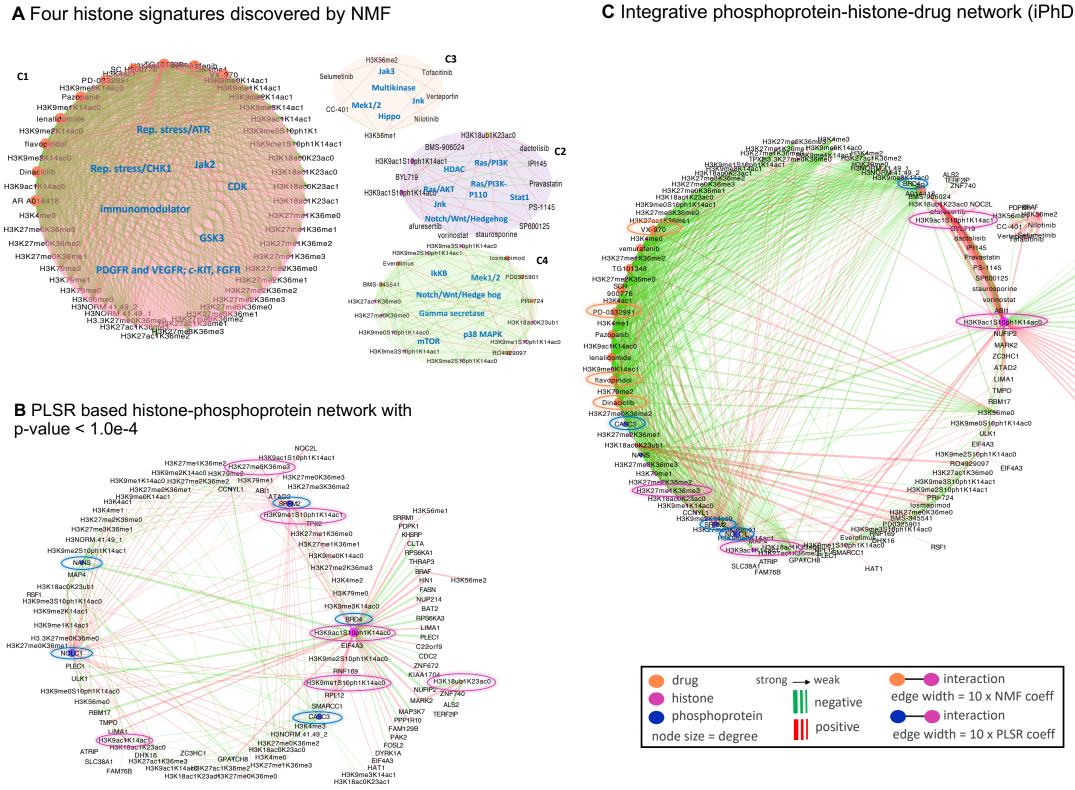


**Figure 3.1. A 3-dimensional view of molecular interactions of phosphoproteins-histones-drugs generated by integrating the histone signatures with the histone-protein interaction network.**

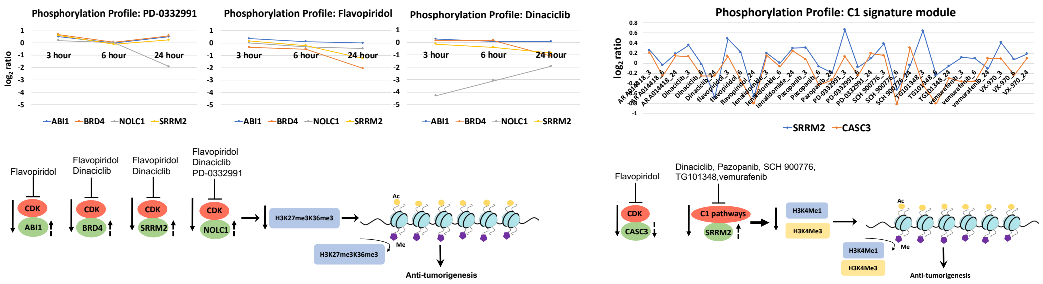
(**A**) Four "histone signatures" are obtained by NMF clustering of GCP at 24-hour post-treatment. Drugs and histones are depicted by orange and magenta nodes respectively, the color of edges signifies whether the interaction between a drug and a histone resulted in elevated (red) or reduced (green) histone level. (**B**) Histone-phosphoprotein interaction network using a PLSR prediction model. Histones and phosphoproteins are depicted by magenta and blue nodes respectively, the color of edges depicts whether the interaction between a phosphoprotein and a histone is positively (red) or negatively (green) correlated. (**C**) Integrated phosphoproteins-histones-drug network (iPhDnet). iPhDnet shows enrichment of modification levels on H3K9ac1S10ph1K14ac0, H3K56me2, H3K27me3K36me3, H3K18ac0K23ub1 histone codes, acting as highly connected nodes (hubs) and positively induced by various drugs affecting enriched phosphoproteins including BRD4, ATAD2, and NOLC1. The strength of an interaction is captured by the width of an edge. See also Figures S1 and S2.

**A** PTMs findings in our study and other published studies in breast cancer

| Histone modification | Regulators | Pathways | Observation in our study | Observation in other studies |
|---|---|---|---|---|
| H3K27me3K36me3 | ABI1, BRD4, NOLC1,SRRM2 | CDK (Flavopiridol, Dinaciclib, PD-0332991) | reduced | |
| H3K9ac1S10ph1K14ac1 | ABI1, NOC2L | c2 | reduced | |
| H3K56me2 | BRAF | c3 | elevated | |
| H3K18ac0K23ub1 | GPATCH8, NANS, PLEC1, MAP4, NOLC1 | c4 | elevated | |
| H3K27ac1K36me2 | CASC3, GPATCH8 | c1 | elevated | |
| H3K27me3K36me0 | DHX16, GPATCH8 | c1 | reduced | reduced ( Yang et al) |
| H3K9me3K14ac0, H3K9me3K14ac1 | BRD4, CASC3, CCNYL1 | c1 | reduced | reduced(Leszenski et al) |
| H3K9ac | NOLC1, CASC3,DHX16, GPATCH8 | c1 | reduced | reduced (Choe et al) |
| H3K4Me1 | CASC3, SRRM2, NOLC1,NANS | c1 | reduced | reduced (Choe et al) |
| H3K4Me3 | CASC3, SRRM2 | c1 | reduced | reduced (Choe et al) |

**B** CDK mediated regulation

**C** C1 pathway mediated regulation



**D** Comparison of histone signatures discovered in other cancer and neural progenitor cell lines

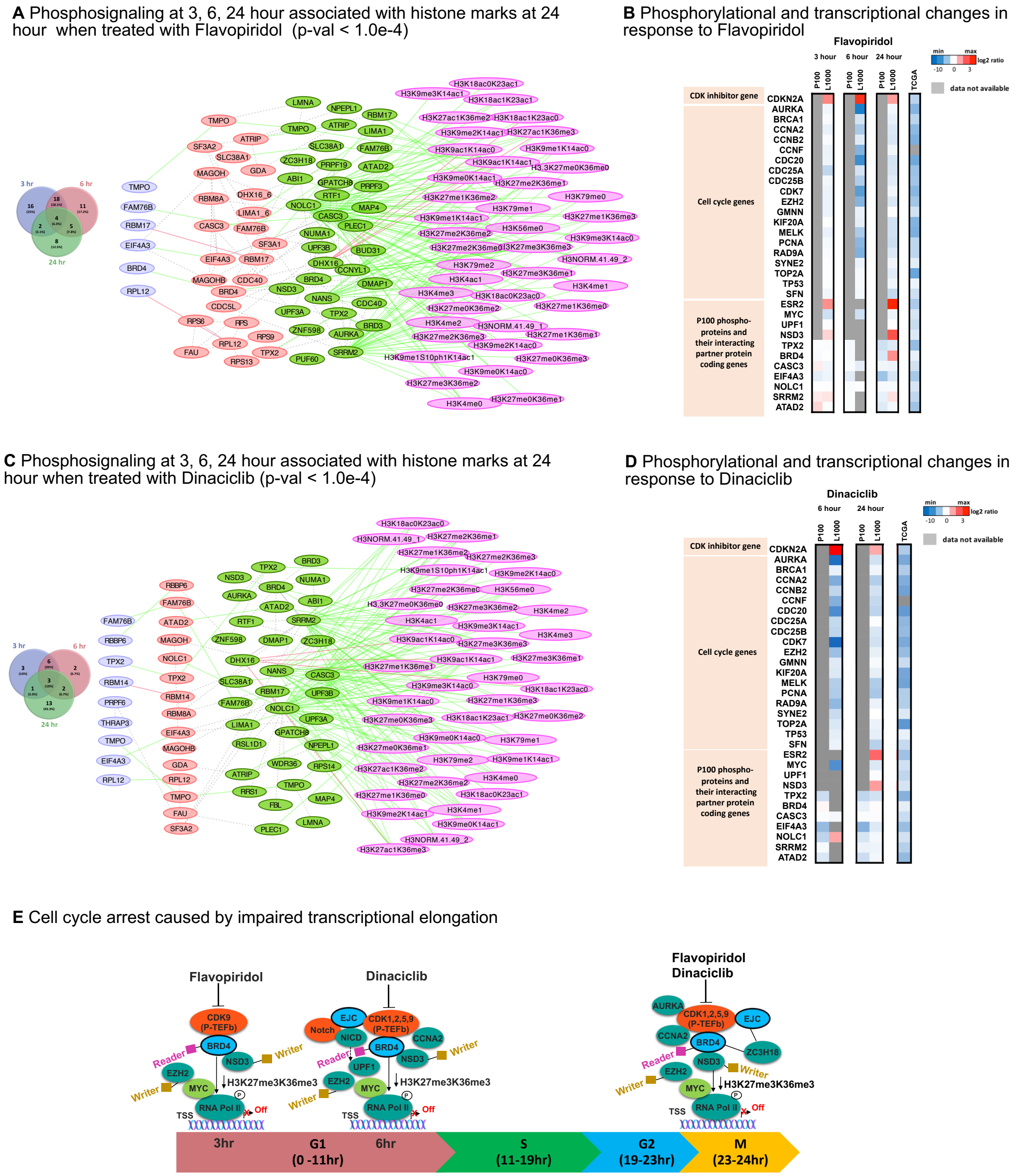
| Cell line | # of Histone signatures | Rand Index (RI) | | | | | |
|---|---|---|---|---|---|---|---|
| | | MCF7 | YAPC | A375 | PC3 | A549 | NPC |
| MCF7<br>(Breast cancer) | 4 | 1 | 0.65 | 0.54 | 0.55 | 0.54 | 0.6 |
| YAPC<br>(Pancreas cancer) | 4 | | 1 | 0.53 | 0.53 | 0.65 | 0.58 |
| A375<br>(Melanoma) | 4 | | | 1 | 0.51 | 0.59 | 0.59 |
| PC3<br>(Prostate cancer) | 3 | | | | 1 | 0.54 | 0.58 |
| A549<br>(Lung cancer) | 4 | | | | | 1 | 0.6 |
| NPC (Neural<br>Progenitor Cells) | 3 | | | | | | 1 |

**E** Comparison of drugs in other cancer and neural progenitor cell lines

| Rand Index (RI) | | | | | | | |
|---|---|---|---|---|---|---|---|
| Drugs | AR A014418 | Dinaciclib | Flavopiridol | Lenalidomide | Pazopanib | PD 0332991 | TG101348 |
| AR A014418 | 1.00 | 0.67 | 0.60 | 0.73 | 0.53 | 0.60 | 0.6 |
| Dinaciclib | | 1.00 | **1.00** | 0.67 | 0.60 | 0.80 | 0.93 |
| Flavopiridol | | | 1.00 | 0.60 | 0.67 | 0.73 | 0.87 |
| Lenalidomide | | | | 1.00 | 0.80 | 0.73 | 0.73 |
| Pazopanib | | | | | 1.00 | 0.67 | 0.67 |
| PD 0332991 | | | | | | 1.00 | 0.87 |
| TG101348 | | | | | | | 1.00 |
| .<br>.<br>. | | | | | | | |

**Figure 3.2. Flavopiridol and dinaciclib emerge as potential CDK mediated therapeutics in breast cancer.**

(**A**) Summary of PTM results showing consistency of our findings with other reports that interrogated specific PTMs in breast cancers. (**B**) CDK mediated regulation in flavopiridol, dinaciclib, and PD-0332991. (**C**) Showing multiple phosphosignaling pathways regulated by the specific drugs in C1 "histone signature". (**D**) Comparison of histone signatures in six cell lines. (**E**) Group assignments of drugs. See also Figure S3.


A Phosphosignaling at 3, 6, 24 hour associated with histone marks at 24 hour when treated with Flavopiridol (p-val < 1.0e-4)

B Phosphorylational and transcriptional changes in response to Flavopiridol

C Phosphosignaling at 3, 6, 24 hour associated with histone marks at 24 hour when treated with Dinaciclib (p-val < 1.0e-4)

D Phosphorylational and transcriptional changes in response to Dinaciclib

E Cell cycle arrest caused by impaired transcriptional elongation

**Figure 3.3. Mechanistic causal network (MCN) reconstruction supports BRD4 mediated cell cycle arrest when treated with flavopiridol and dinaciclib in breast cancer.**

(**A**) MCN reconstruction for enriched phosphoproteins (p-val < 1.0e-4) upon flavopiridol treatment. (**B**) Phosphorylation changes of proteins and transcription changes of 31 functionally significant genes in response to flavopiridol. (**C**) A similar mechanistic causal network reconstruction for enriched phosphoproteins after dinaciclib treatment is obtained using the protocol described in A. (**D**) Similarly, phosphorylation and transcription changes of the same phosphoproteins and genes in B, in response to dinaciclib. (**E**) Demonstrating possible cell cycle arrest mechanisms caused by transcriptional elongation of the participating regulators over various time points corresponds to cell cycle stages.

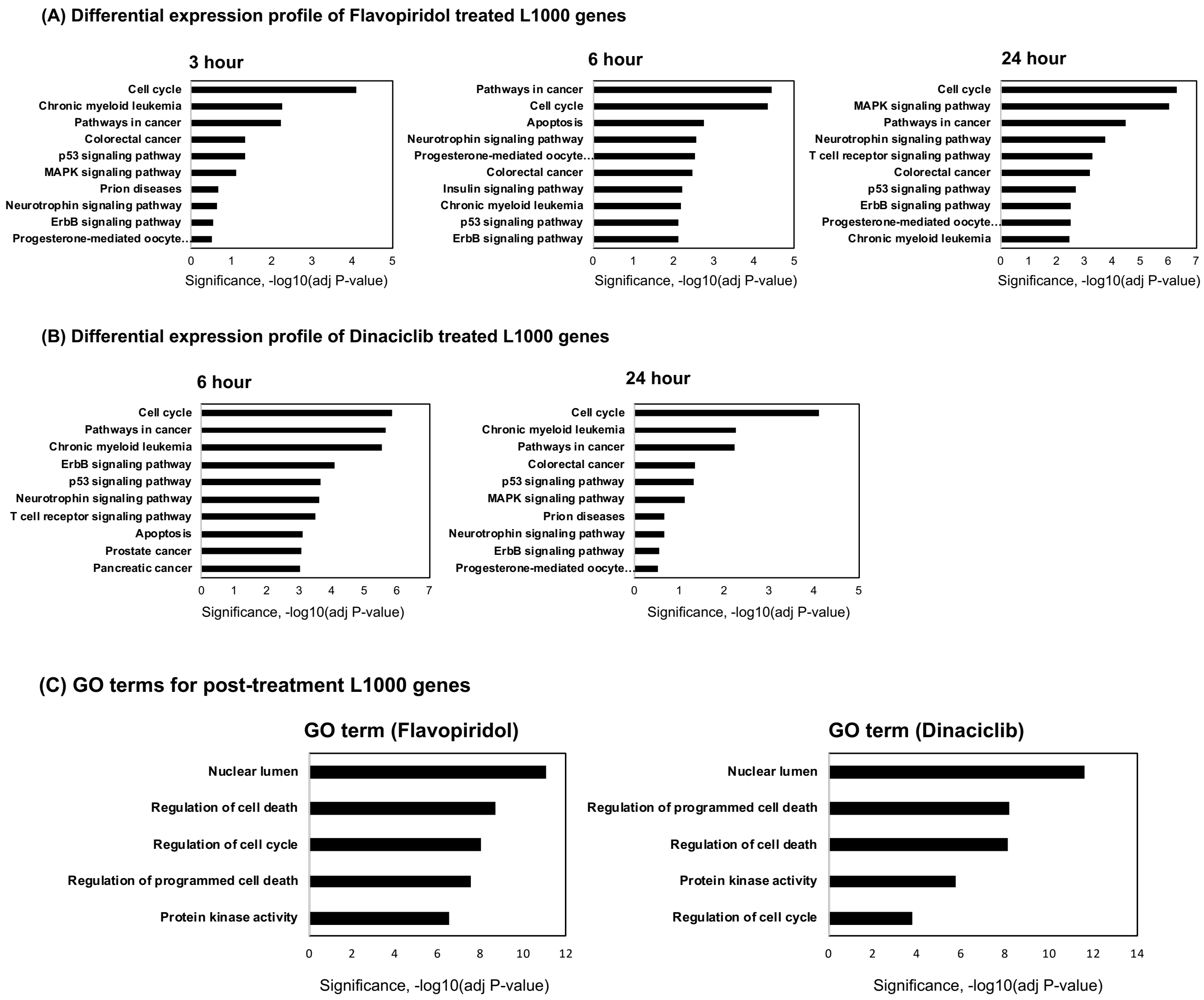


**Figure 3.4 Functional analyses on differentially expressed L1000 genes with drug treatments.**

(**A**) KEGG pathway analysis showing the involvement of L1000 differentially expressed (DE) genes significantly related to cell cycle, p53 signaling pathway, ErbB signaling, and pathways in cancer across 3, 6, and 24-hour time points after treatment with flavopiridol. (**B**) The same pathways are significantly related when considering the differentially expressed genes at 6, and 24-hour time points post dinaciclib treatment. (**C**) Shown here are statistically significant DE genes associated with gene ontology (GO) categories: nuclear lumen, regulation of cell death, cell cycle, programmed cell death and protein kinase activities.

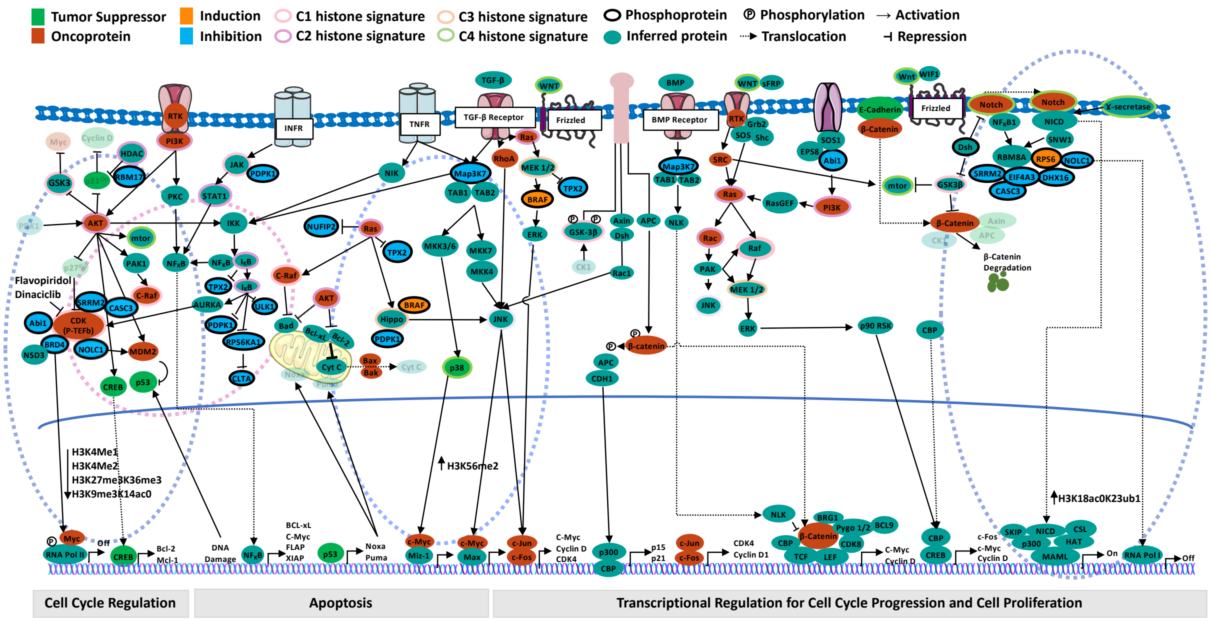


**Figure 3.5. Fingerprint global chromatin profiling in breast cancer signaling.** The crosstalk among histone signature pathways is depicted by linking inferred proteins/protein complexes generated from STRINGDB PPI for signaling pathways that may interact with the CDK pathway in cancer signaling. As part of cell cycle regulation, inhibition of CDK by flavopiridol and dinaciclib is highlighted showing molecular cascades of interactions among BRD4, NSD3, SRRM2, NOLC1, MYC with the P-TEFb complex and its recruitment to promoter region to block transcriptional elongation of RNA Pol II (the first blue dashed oval). As a consequence, reduced levels of H3K27me3K36me3, H3K4me1, and H3K4me2 are observed. Examples of crosstalk includes: Ikkb inhibiting TPX2 which binds to AURKA to activate CDK targeting the intrinsic kinase activity directed towards RNA Pol II (pink dashed oval); reduction of Map3K7 brings JNK level down resulting in an increase of H3K56me2 level; hyperactive RAS acts as a signaling switch to convert JNK's role from pro- to anti-tumor signaling through the regulation of Hippo signaling activity by inhibiting PDPK1 protein (the second blue dashed oval). NOLC1 interacts with the EJC junction formed by SRRM2, CASC3, EIF4A3, and RBM8A proteins. It mediates Wnt/Notch signaling activity through the Notch intracellular domain (NICD) and monoubiquitylation of H3K23 (H3K18ac0K23ub1) by translocating to RNA Pol I (the third blue dashed oval).The color of molecules represents tumor suppressors (green), oncoproteins (red), inferred proteins (teal) and phosphoproteins (black ring). The color of protein molecule indicates whether the protein was induced (orange) or inhibited (blue). Each histone signature (C1, C2, C3, and C4) is highlighted using a distinct outer ring color. Phosphorylation, activation, and repression are indicated by Ⓟ, arrowheads (→), and cross-bars (⊣), respectively.

## 3.7 Tables

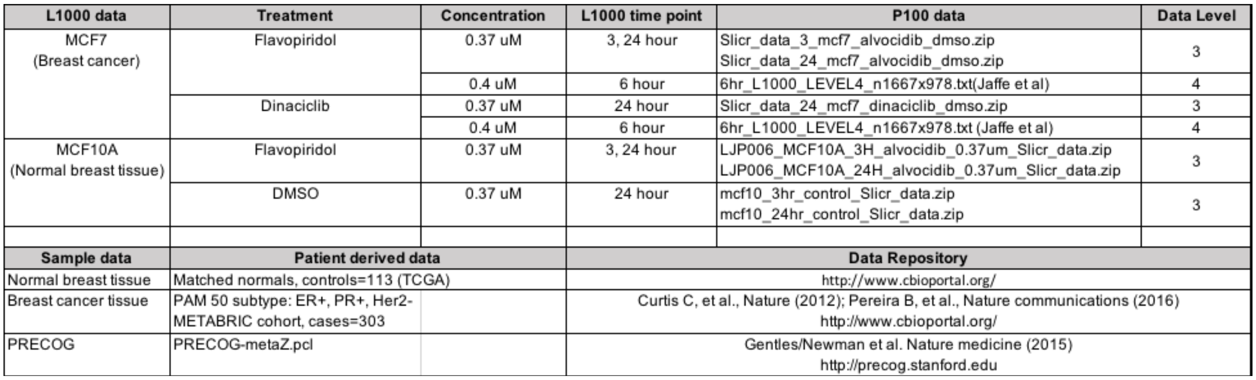
**Table S3.1. LINCS proteomics dataset used in the study.**

Listed here are the global chromatin profiles (GCP) and phosphoprotein (P100) data in six cancer cell lines from LINCS  at various time points.

| Cell line | GCP data | GCP time point | P100 data | P100 time |
|---|---|---|---|---|
| MCF7 (Breast cancer) | LINCS_GCP_Plate29_annotated_minimized_2016-01-08_09-27-26_unprocessed.gct | 24 hour | LINCS_P100_PRM_Plate29_03H_annotated_minimized_2016-01-28_11-00-43.gct<br>LINCS_P100_PRM_Plate29_06H_annotated_minimized_2016-01-28_17-11-17.gct<br>LINCS_P100_PRM_Plate29_24H_annotated_minimized_2016-01-28_17-11-22.gct | 3, 6, 24 hour |
| YAPC (Pancreas cancer) | LINCS_P100_PRM_Plate32_annotated_minimized_2016-07-22_11-29-42.gct | 24 hour | LINCS_GCP_Plate32_annotated_minimized_2016-07-22_11-29-04.gct | 3 hour |
| A375 (Melanoma) | LINCS_GCP_Plate28_annotated_minimized_2016-04-14_14-24-24.gct | 24 hour | LINCS_P100_PRM_Plate28_annotated_minimized_2016-04-08_15-41-04.gct | 3 hour |
| PC3 (Prostate cancer) | LINCS_GCP_Plate34_annotated_minimized_2016-07-07_14-16-01.gct | 24 hour | LINCS_P100_DIA_Plate34_annotated_minimized_2016-08-15_10-38-52.gct | 3 hour |
| A549 (Lung cancer) | LINCS_GCP_Plate33_annotated_minimized_2016-06-03_14-58-02.gct | 24 hour | LINCS_P100_DIA_Plate33_annotated_minimized_2016-06-29_12-24-03.gct | 3 hour |
| NPC (Neural Progenitor Cells) | LINCS_GCP_Plate27_annotated_minimized_2016-04-14_14-24-09.gct | 24 hour | LINCS_P100_DIA_Plate27_annotated_minimized_2016-02-01_15-53-14.gct | 3 hour |

**Table S3.2. Transcriptomic datasets used in the study.**

Listed here are the L1000 data for two cell lines (normal like breast tissue MCF10A, and breast cancer tissue MCF7) from LINCS and case-control patient level normal and breast cancer data from TCGA, METABRIC and PRECOG.



| L1000 data | Treatment | Concentration | L1000 time point | P100 data | Data Level |
|---|---|---|---|---|---|
| MCF7<br>(Breast cancer) | Flavopiridol | 0.37 uM | 3, 24 hour | Slicr_data_3_mcf7_alvocidib_dmso.zip<br>Slicr_data_24_mcf7_alvocidib_dmso.zip | 3 |
| | | 0.4 uM | 6 hour | 6hr_L1000_LEVEL4_n1667x978.txt(Jaffe et al) | 4 |
| | Dinaciclib | 0.37 uM | 24 hour | Slicr_data_24_mcf7_dinaciclib_dmso.zip | 3 |
| | | 0.4 uM | 6 hour | 6hr_L1000_LEVEL4_n1667x978.txt (Jaffe et al) | 4 |
| MCF10A<br>(Normal breast tissue) | Flavopiridol | 0.37 uM | 3, 24 hour | LJP006_MCF10A_3H_alvocidib_0.37um_Slicr_data.zip<br>LJP006_MCF10A_24H_alvocidib_0.37um_Slicr_data.zip | 3 |
| | DMSO | 0.37 uM | 24 hour | mcf10_3hr_control_Slicr_data.zip<br>mcf10_24hr_control_Slicr_data.zip | 3 |
| | | | | | |

| Sample data | Patient derived data | Data Repository |
|---|---|---|
| Normal breast tissue | Matched normals, controls=113 (TCGA) | http://www.cbioportal.org/ |
| Breast cancer tissue | PAM 50 subtype: ER+, PR+, Her2-<br>METABRIC cohort, cases=303 | Curtis C, et al., Nature (2012); Pereira B, et al., Nature communications (2016)<br>http://www.cbioportal.org/ |
| PRECOG | PRECOG-metaZ.pcl | Gentles/Newman et al. Nature medicine (2015)<br>http://precog.stanford.edu |

## 3.8 Supplemental Figures

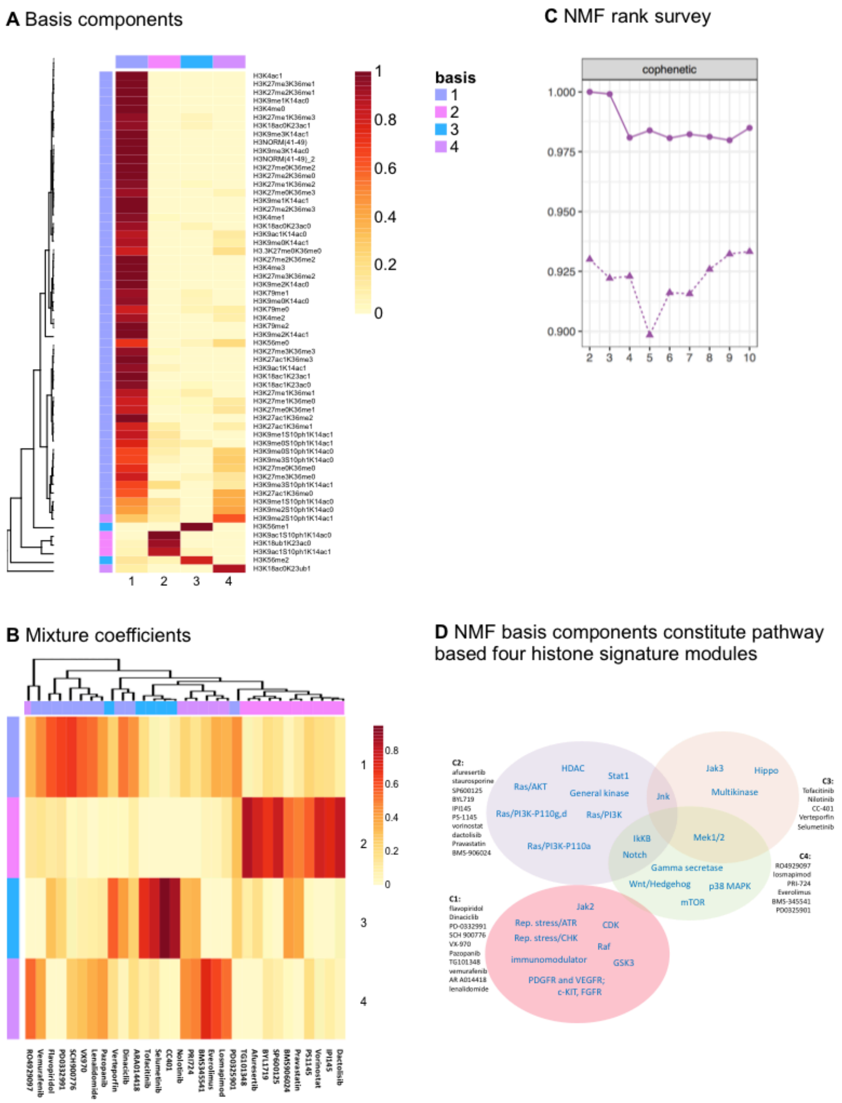
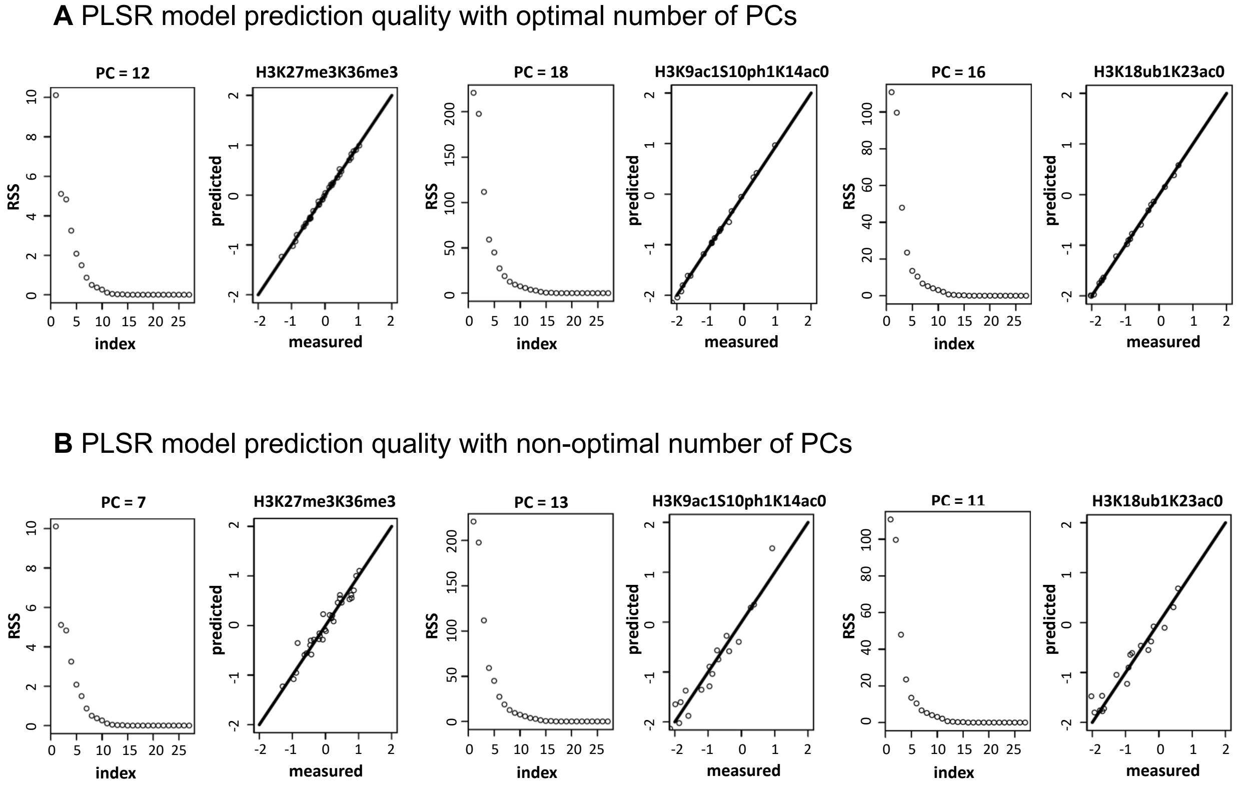


**Figure 3.S1. Estimation of the factorization rank of NMF and its cluster components.** (**A**) Heatmap of the basis components (histones and their cluster memberships). Showing likelihood of each histone code belonging to a specific signature module. (**B**) Showing membership contributions of each drug toward 4 signature modules (k=4). (**C**) Cophenetic score is computed from 100 runs for each value of rank k by varying k= 2, 3...10 on 24-hour GCP data. Rank k represents the number of clusters or basis components. The solid line represents the original data and the dotted line represents random data. (**D**) Showing these 4 basis components corresponds to 4 pathway-based functional modules (c1, c2, c3 and c4). These functional modules constitute histone signatures.

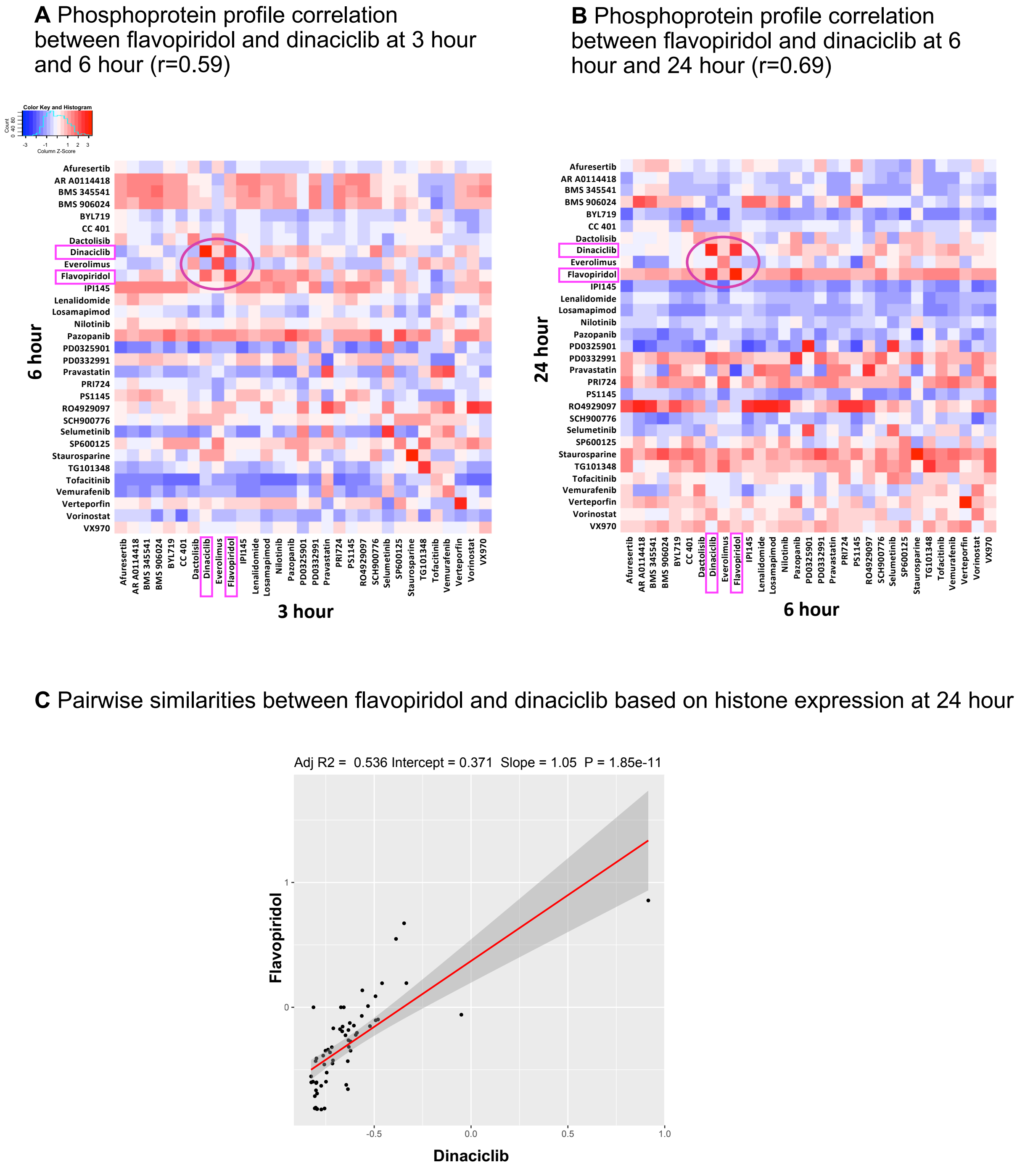


**A** PLSR model prediction quality with optimal number of PCs

**B** PLSR model prediction quality with non-optimal number of PCs

**Figure 3.S2. Performance of PLRS model.**

Showing three examples of histone codes (H3K27me3K36me3, H3K9ac1S10ph1K14ac0 and H3K18ub1K23ac0). (**A**) Showing model performance using optimal number of components. The optimal number of components (principal component, PC) is used to predict the model accurately using residual sum square (RSS) value < 0.05. (**B**) Depicting model performance using sub optimal number of components.



**A** Phosphoprotein profile correlation between flavopiridol and dinaciclib at 3 hour and 6 hour (r=0.59)

**B** Phosphoprotein profile correlation between flavopiridol and dinaciclib at 6 hour and 24 hour (r=0.69)

**C** Pairwise similarities between flavopiridol and dinaciclib based on histone expression at 24 hour

**Figure 3.S3. Phosphoprotein and global chromatin correlation profiles between drug pairs.**

Pearson correlation between paired drugs at 3 and 6 hours. Showing a strong positive correlation (r=0.59) between flavopiridol and dinaciclib (circled) at 3 and 6 hours. (**B**) Strong positive correlation (r=0.69) is sustained between flavopiridol and dinaciclib (circled) at 6 and 24 hours. (**C**) Pairwise correlation between flavopiridol and dinaciclib based on 24 hour GCP data. Showing positive correlations between flavopiridol and dinaciclib (positive slope), using a linear regression line on 24 hour normalized global chromatin data (p-value = 1.85e-11, adjusted r-squared =0.536).

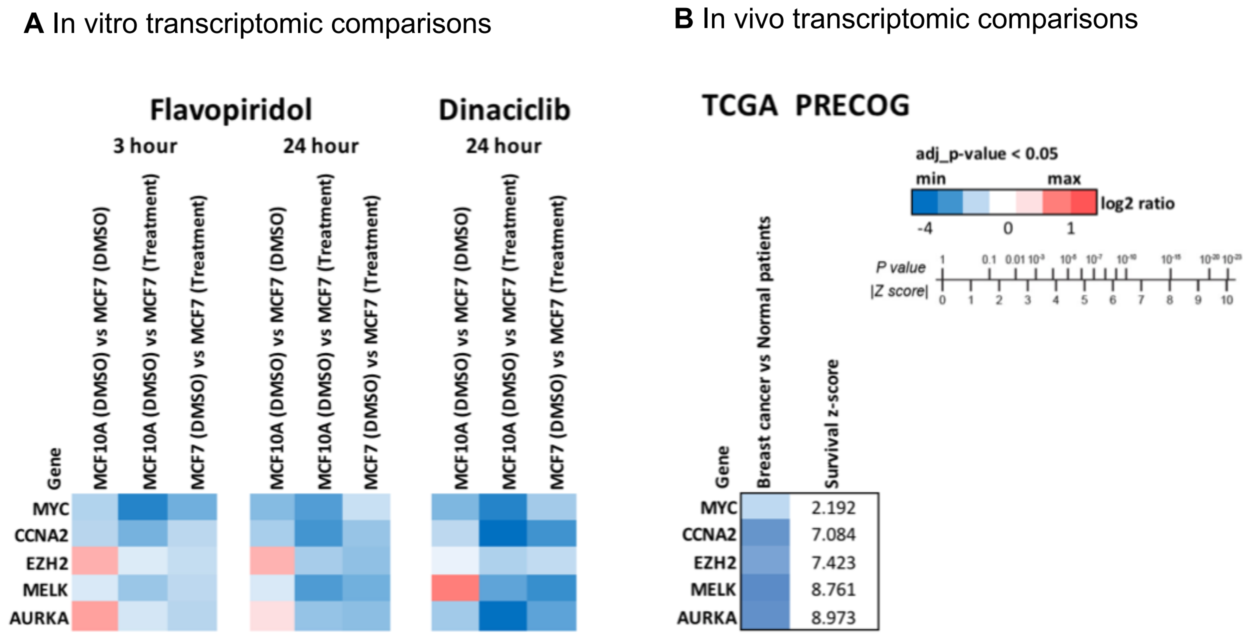


**Figure 3.S4. In vitro and in vivo transcriptomic analysis  of marker genes across normal, pre-treated and post-treated breast cancer tissues.**

(A) Gene expressions of L1000 genes in normal, pre-treated and post-treated cancer tissues treated with flavopiridol at 3, 24 hour and dinaciclib at 24 hour. Shown here are in vitro comparisons  of the marker genes in these two treatments across these time points. These comparisons include i) normal vs cancer: MCF10A (treated with DMSO) vs MCF7 (treated with DMSO), showing upregulation (red) of EZH2 and AURKA and downregulation (blue) of MYC, CCNA2 and MELK; ii) normal vs cancer post-treatment: MCF10A (treated with DMSO) vs MCF7 (treated with flavopiridol/dinaciclib), showing downregulation for all of these genes; and iii) cancer pre-treatment vs post-treatment: MCF7 (treated with DMSO) vs MCF7 (treated with flavopiridol/dinaciclib) showing downregulation of all these genes. The heatmaps highlight enhanced potency of flavopiridol and dinaciclib against normal and cancer tissues. (B) In vivo comparisons between cancer vs normal patients for the same marker genes in (A); differential expression analysis was performed on TCGA matched normal breast datasets (n=113 normals) and METABRIC (n=303 cases) . Consistent with literature and our pre-treatment vs post-treatment results, these markers were downregulated in normal patients. Showing corresponding survival z-scores for these marker genes obtained from the prediction of PRECOG dataset.

## 3.9 Author Contributions

**Original Concept:** Shamim Ara Mollah and Shankar Subramaniam

**Project Supervision:** Shankar Subramaniam

**Project Planning and Experiemental Design:** Shamim Ara Mollah and Shankar Subramaniam

**Method Implementation, Data Acquisition, Processing and Analysis:** Shamim Ara Mollah

**Preparation of Manuscript:** Shamim Ara Mollah and Shankar Subramaniam

### 3.10 Acknowledgements

We are grateful to Dr. Jacob D. Jaffe, LINCS, and TCGA consortium for making the data publicly available. We thank Drs. Mano Maurya and Shakti Gupta for helpful discussion and comments and Marissa Elliot for helping with figures. This work was funded by R01 LM012595, U01 DK097430, U01 CA200147, U01 CA198941, U19 AI090023, R01 HL106579, R01HL108735, R01 HD084633, R01 DK109365, NSF CCF-0939370, and a Wellcome-Trust Award to S.S. S.A.M. received first-year support from National Institutes of Health Graduate Training Grant (T15LM011271).

Chapter 3, is currently being prepared for submission for publication of the material. Shamim Mollah and Shankar Subramaniam. The dissertation author was a primary investigator and author of this paper.

## 3.11 References

Abelin, Jennifer G., Jinal Patel, Xiaodong Lu, Caitlin M. Feeney, Lola Fagbami, Amanda L. Creech, Roger Hu, Daniel Lam, Desiree Davison, Lindsay Pino, Jana W. Qiao, Eric Kuhn, Adam Officer, Jianxue Li, Susan Abbatiello, Aravind Subramanian, Richard Sidman, Evan Snyder, Steven A. Carr, and Jacob D. Jaffe. 2016. "Reduced-Representation Phosphosignatures Measured by Quantitative Targeted MS Capture Cellular States and Enable Large-Scale Comparison of Drug-Induced Phenotypes." *Molecular & Cellular Proteomics: MCP* 15 (5): 1622–41.

Baldi, Pierre, and Anthony D. Long. 2001. "A Bayesian Framework for the Analysis of Microarray Expression Data: Regularized T -Test and Statistical Inferences of Gene Changes." *Bioinformatics* 17 (6). Oxford University Press: 509–19.

Barrand, Sanna, Ingrid S. Andersen, and Philippe Collas. 2010. "Promoter-Exon Relationship of H3 Lysine 9, 27, 36 and 79 Methylation on Pluripotency-Associated Genes." *Biochemical and Biophysical Research Communications* 401 (4): 611–17.

Baylin, Stephen B., and Peter A. Jones. 2011. "A Decade of Exploring the Cancer Epigenome - Biological and Translational Implications." *Nature Reviews. Cancer* 11 (10): 726–34.

Bösken, Christian A., Lucas Farnung, Corinna Hintermair, Miriam Merzel Schachter, Karin Vogel-Bachmayr, Dalibor Blazek, Kanchan Anand, Robert P. Fisher, Dirk Eick, and Matthias Geyer. 2014. "The Structure and Substrate Specificity of Human Cdk12/Cyclin K." *Nature Communications* 5 (March): 3505.

Boutsidis, C., and E. Gallopoulos. 2008. "SVD Based Initialization: A Head Start for Nonnegative Matrix Factorization." *Pattern Recognition* 41 (4): 1350–62.

Brannan, Kris, Hyunmin Kim, Benjamin Erickson, Kira Glover-Cutter, Soojin Kim, Nova Fong, Lauren Kiemele, Kirk Hansen, Richard Davis, Jens Lykke-Andersen, and David L. Bentley. 2012. "mRNA Decapping Factors and the Exonuclease Xrn2 Function in Widespread Premature Termination of RNA Polymerase II Transcription." *Molecular Cell* 46 (3): 311–24.

Brunet, Jean-Philippe, Pablo Tamayo, Todd R. Golub, and Jill P. Mesirov. 2004. "Metagenes and Molecular Pattern Discovery Using Matrix Factorization." *Proceedings of the National Academy of Sciences of the United States of America* 101 (12): 4164–69.

Cancer Genome Atlas Network. 2012. "Comprehensive Molecular Portraits of Human Breast Tumours." *Nature* 490 (7418): 61–70.

Cao, Ru, and Yi Zhang. 2004. "The Functions of E(Z)/EZH2-Mediated Methylation of Lysine 27 in Histone H3." *Current Opinion in Genetics & Development* 14 (2): 155–64.

Caron, C., C. Lestrat, S. Marsal, E. Escoffier, S. Curtet, V. Virolle, P. Barbry, A. Debernardi, C. Brambilla, E. Brambilla, S. Rousseaux, and S. Khochbin. 2010. "Functional Characterization of ATAD2 as a New Cancer/testis Factor and a Predictor of Poor Prognosis in Breast and Lung Cancers." *Oncogene* 29 (37): 5171–81.

Cha, Tai-Lung, Binhua P. Zhou, Weiya Xia, Yadi Wu, Cheng-Chieh Yang, Chun-Te Chen, Bo Ping, Arie P. Otte, and Mien-Chie Hung. 2005. "Akt-Mediated Phosphorylation of EZH2 Suppresses Methylation of Lysine 27 in Histone H3." *Science* 310 (5746): 306–10.

Chen, Edward Y., Christopher M. Tan, Yan Kou, Qiaonan Duan, Zichen Wang, Gabriela Vaz Meirelles, Neil R. Clark, and Avi Ma'ayan. 2013. "Enrichr: Interactive and Collaborative HTML5 Gene List Enrichment Analysis Tool." *BMC Bioinformatics* 14 (April): 128.

Creech, Amanda L., Jordan E. Taylor, Verena K. Maier, Xiaoyun Wu, Caitlin M. Feeney, Namrata D. Udeshi, Sally E. Peach, Jesse S. Boehm, Jeannie T. Lee, Steven A. Carr, and Jacob D. Jaffe. 2015. "Building the Connectivity Map of Epigenetics: Chromatin Profiling by Quantitative Targeted Mass Spectrometry." *Methods* 72 (January): 57–64.

Curtis, Christina, Sohrab P. Shah, Suet-Feung Chin, Gulisa Turashvili, Oscar M. Rueda, Mark J. Dunning, Doug Speed, Andy G. Lynch, Shamith Samarajiwa, Yinyin Yuan, Stefan Gräf, Gavin Ha, Gholamreza Haffari, Ali Bashashati, Roslin Russell, Steven McKinney, METABRIC Group, Anita Langerød, Andrew Green, Elena Provenzano, Gordon Wishart, Sarah Pinder, Peter Watson, Florian Markowetz, Leigh Murphy, Ian Ellis, Arnie Purushotham, Anne-Lise Børresen-Dale, James D. Brenton, Simon Tavaré, Carlos Caldas, and Samuel Aparicio. 2012. "The Genomic and Transcriptomic Architecture of 2,000 Breast Tumours Reveals Novel Subgroups." *Nature* 486 (7403): 346–52.

Delmore, Jake E., Ghayas C. Issa, Madeleine E. Lemieux, Peter B. Rahl, Junwei Shi, Hannah M. Jacobs, Efstathios Kastritis, Timothy Gilpatrick, Ronald M. Paranal, Jun Qi, Marta Chesi, Anna C. Schinzel, Michael R. McKeown, Timothy P. Heffernan, Christopher R. Vakoc, P. Leif Bergsagel, Irene M. Ghobrial, Paul G. Richardson, Richard A. Young, William C. Hahn, Kenneth C. Anderson, Andrew L. Kung, James E. Bradner, and Constantine S. Mitsiades. 2011. "BET Bromodomain Inhibition as a Therapeutic Strategy to Target c-Myc." *Cell* 146 (6): 904–17.

Dispenzieri, Angela, Morie A. Gertz, Martha Q. Lacy, Susan M. Geyer, Tom R. Fitch, Robert G. Fenton, Rafael Fonseca, Crescent R. Isham, Steven C. Ziesmer, Charles Erlichman, and Keith C. Bible. 2006. "Flavopiridol in Patients with Relapsed or Refractory Multiple Myeloma: A Phase 2 Trial with Clinical and Pharmacodynamic End-Points." *Haematologica* 91 (3): 390–93.

Dravis, Christopher, Chi-Yeh Chung, Nikki K. Lytle, Jaslem Herrera-Valdez, Gidsela Luna, Christy L. Trejo, Tannishtha Reya, and Geoffrey M. Wahl. 2018. "Epigenetic and Transcriptomic Profiling of Mammary Gland Development and Tumor Models Disclose Regulators of Cell State Plasticity." *Cancer Cell*, August. doi:10.1016/j.ccell.2018.08.001.

Ellis, Matthew J., Li Ding, Dong Shen, Jingqin Luo, Vera J. Suman, John W. Wallis, Brian A. Van Tine, Jeremy Hoog, Reece J. Goiffon, Theodore C. Goldstein, Sam Ng, Li Lin, Robert

Crowder, Jacqueline Snider, Karla Ballman, Jason Weber, Ken Chen, Daniel C. Koboldt, Cyriac Kandoth, William S. Schierding, Joshua F. McMichael, Christopher A. Miller, Charles Lu, Christopher C. Harris, Michael D. McLellan, Michael C. Wendl, Katherine DeSchryver, D. Craig Allred, Laura Esserman, Gary Unzeitig, Julie Margenthaler, G. V. Babiera, P. Kelly Marcom, J. M. Guenther, Marilyn Leitch, Kelly Hunt, John Olson, Yu Tao, Christopher A. Maher, Lucinda L. Fulton, Robert S. Fulton, Michelle Harrison, Ben Oberkfell, Feiyu Du, Ryan Demeter, Tammi L. Vickery, Adnan Elhammali, Helen Piwnica-Worms, Sandra McDonald, Mark Watson, David J. Dooling, David Ota, Li-Wei Chang, Ron Bose, Timothy J. Ley, David Piwnica-Worms, Joshua M. Stuart, Richard K. Wilson, and Elaine R. Mardis. 2012. "Whole-Genome Analysis Informs Breast Cancer Response to Aromatase Inhibition." *Nature* 486 (7403): 353–60.

Ember, Stuart W. J., Jin-Yi Zhu, Sanne H. Olesen, Mathew P. Martin, Andreas Becker, Norbert Berndt, Gunda I. Georg, and Ernst Schönbrunn. 2014. "Acetyl-Lysine Binding Site of Bromodomain-Containing Protein 4 (BRD4) Interacts with Diverse Kinase Inhibitors." *ACS Chemical Biology* 9 (5): 1160–71.

Filippakopoulos, Panagis, Jun Qi, Sarah Picaud, Yao Shen, William B. Smith, Oleg Fedorov, Elizabeth M. Morse, Tracey Keates, Tyler T. Hickman, Ildiko Felletar, Martin Philpott, Shonagh Munro, Michael R. McKeown, Yuchuan Wang, Amanda L. Christie, Nathan West, Michael J. Cameron, Brian Schwartz, Tom D. Heightman, Nicholas La Thangue, Christopher A. French, Olaf Wiest, Andrew L. Kung, Stefan Knapp, and James E. Bradner. 2010. "Selective Inhibition of BET Bromodomains." *Nature* 468 (September). Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved.: 1067.

Fischle, Wolfgang, Yanming Wang, and C. David Allis. 2003. "Histone and Chromatin Cross-Talk." *Current Opinion in Cell Biology* 15 (2): 172–83.

Foss, Francine M., Pier Luigi Zinzani, Julie M. Vose, Randy D. Gascoyne, Steven T. Rosen, and Kensei Tobinai. 2011. "Peripheral T-Cell Lymphoma." *Blood* 117 (25): 6756–67.

Gentles, Andrew J., Aaron M. Newman, Chih Long Liu, Scott V. Bratman, Weiguo Feng, Dongkyoon Kim, Viswam S. Nair, Yue Xu, Amanda Khuong, Chuong D. Hoang, Maximilian Diehn, Robert B. West, Sylvia K. Plevritis, and Ash A. Alizadeh. 2015. "The Prognostic Landscape of Genes and Infiltrating Immune Cells across Human Cancers." *Nature Medicine* 21 (8): 938–45.

Gupta, Shakti, Mano Ram Maurya, and Shankar Subramaniam. 2010. "Identification of Crosstalk between Phosphoprotein Signaling Pathways in RAW 264.7 Macrophage Cells." *PLoS Computational Biology* 6 (1): e1000654.

Hanahan, Douglas, and Robert A. Weinberg. 2011. "Hallmarks of Cancer: The next Generation." *Cell* 144 (5): 646–74.

Holm, Karolina, Dorthe Grabau, Kristina Lövgren, Steina Aradottir, Sofia Gruvberger-Saal, Jillian Howlin, Lao H. Saal, Stephen P. Ethier, Pär-Ola Bendahl, Olle Stål, Per Malmström, Mårten

Fernö, Lisa Rydén, Cecilia Hegardt, Åke Borg, and Markus Ringnér. 2012. "Global H3K27 Trimethylation and EZH2 Abundance in Breast Tumor Subtypes." *Molecular Oncology* 6 (5): 494–506.

Horiuchi, Dai, Leonard Kusdra, Noelle E. Huskey, Sanjay Chandriani, Marc E. Lenburg, Ana Maria Gonzalez-Angulo, Katelyn J. Creasman, Alexey V. Bazarov, James W. Smyth, Sarah E. Davis, Paul Yaswen, Gordon B. Mills, Laura J. Esserman, and Andrei Goga. 2012. "MYC Pathway Activation in Triple-Negative Breast Cancer Is Synthetic Lethal with CDK Inhibition." *The Journal of Experimental Medicine* 209 (4): 679–96.

Huang, Bo, Xiao-Dong Yang, Ming-Ming Zhou, Keiko Ozato, and Lin-Feng Chen. 2009. "Brd4 Coactivates Transcriptional Activation of NF-kappaB via Specific Binding to Acetylated RelA." *Molecular and Cellular Biology* 29 (5): 1375–87.

Huang, Da Wei, Brad T. Sherman, and Richard A. Lempicki. 2009. "Systematic and Integrative Analysis of Large Gene Lists Using DAVID Bioinformatics Resources." *Nature Protocols* 4 (1): 44–57.

Hwang, Yu-Chyi, Tung-Ying Lu, Dah-Yeou Huang, Yuan-Sung Kuo, Cheng-Fu Kao, Ning-Hsing Yeh, Han-Chung Wu, and Chin-Tarng Lin. 2009. "NOLC1, an Enhancer of Nasopharyngeal Carcinoma Progression, Is Essential for TP53 to Regulate MDM2 Expression." *The American Journal of Pathology* 175 (1): 342–54.

Jenuwein, T., and C. D. Allis. 2001. "Translating the Histone Code." *Science* 293 (5532): 1074–80.

Kayala, Matthew A., and Pierre Baldi. 2012. "Cyber-T Web Server: Differential Analysis of High-Throughput Data." *Nucleic Acids Research* 40 (Web Server issue): W553–59.

Khan, Omar, and Nicholas B. La Thangue. 2012. "HDAC Inhibitors in Cancer Biology: Emerging Mechanisms and Clinical Applications." *Immunology and Cell Biology* 90 (1): 85–94.

Krämer, Nicole, and Masashi Sugiyama. 2011. "The Degrees of Freedom of Partial Least Squares Regression." *Journal of the American Statistical Association* 106 (494). Taylor & Francis: 697–705.

Kuleshov, Maxim V., Matthew R. Jones, Andrew D. Rouillard, Nicolas F. Fernandez, Qiaonan Duan, Zichen Wang, Simon Koplev, Sherry L. Jenkins, Kathleen M. Jagodnik, Alexander Lachmann, Michael G. McDermott, Caroline D. Monteiro, Gregory W. Gundersen, and Avi Ma'ayan. 2016. "Enrichr: A Comprehensive Gene Set Enrichment Analysis Web Server 2016 Update." *Nucleic Acids Research* 44 (W1): W90–97.

Kuzmichev, Andrei, Raphael Margueron, Alejandro Vaquero, Tanja S. Preissner, Michael Scher, Antonis Kirmizis, Xuesong Ouyang, Neil Brockdorff, Cory Abate-Shen, Peggy Farnham, and Danny Reinberg. 2005. "Composition and Histone Substrates of Polycomb Repressive

Group Complexes Change during Cellular Differentiation." *Proceedings of the National Academy of Sciences of the United States of America* 102 (6): 1859–64.

Kwak, Hojoong, and John T. Lis. 2013. "Control of Transcriptional Elongation." *Annual Review of Genetics* 47 (September): 483–508.

Lee, D. D., and H. S. Seung. 1999. "Learning the Parts of Objects by Non-Negative Matrix Factorization." *Nature* 401 (6755): 788–91.

Le Hir, Hervé, Jérôme Saulière, and Zhen Wang. 2016. "The Exon Junction Complex as a Node of Post-Transcriptional Networks." *Nature Reviews. Molecular Cell Biology* 17 (1): 41–54.

Leroy, Gary, Peter A. Dimaggio, Eric Y. Chan, Barry M. Zee, M. Andres Blanco, Barbara Bryant, Ian Z. Flaniken, Sherry Liu, Yibin Kang, Patrick Trojer, and Benjamin A. Garcia. 2013. "A Quantitative Atlas of Histone Modification Signatures from Human Cancer Cells." *Epigenetics & Chromatin* 6 (1): 20.

Lewis, Peter W., Manuel M. Müller, Matthew S. Koletsky, Francisco Cordero, Shu Lin, Laura A. Banaszynski, Benjamin A. Garcia, Tom W. Muir, Oren J. Becher, and C. David Allis. 2013. "Inhibition of PRC2 Activity by a Gain-of-Function H3 Mutation Found in Pediatric Glioblastoma." *Science* 340 (6134): 857–61.

Li, Feng, Guogen Mao, Dan Tong, Jian Huang, Liya Gu, Wei Yang, and Guo-Min Li. 2013. "The Histone Mark H3K36me3 Regulates Human DNA Mismatch Repair through Its Interaction with MutSα." *Cell* 153 (3): 590–600.

Liu, Anfei, Xiya Yu, and Shanrong Liu. 2013. "Pluripotency Transcription Factors and Cancer Stem Cells: Small Genes Make a Big Difference." *Chinese Journal of Cancer* 32 (9): 483–87.

Liu, Min, Yajuan Li, Aiguo Liu, Ruifeng Li, Ying Su, Juan Du, Cheng Li, and Alan Jian Zhu. 2016. "The Exon Junction Complex Regulates the Splicing of Cell Polarity Gene dlg1 to Control Wingless Signaling in Development." *eLife* 5 (August). doi:10.7554/eLife.17200.

Liu, W., P. Stein, X. Cheng, W. Yang, N-Y Shao, E. E. Morrisey, R. M. Schultz, and J. You. 2014. "BRD4 Regulates Nanog Expression in Mouse Embryonic Stem Cells and Preimplantation Embryos." *Cell Death and Differentiation* 21 (12): 1950–60.

Lou, Chih-Hong, Jennifer Dumdie, Alexandra Goetz, Eleen Y. Shum, David Brafman, Xiaoyan Liao, Sergio Mora-Castilla, Madhuvanthi Ramaiah, Heidi Cook-Andersen, Louise Laurent, and Miles F. Wilkinson. 2016. "Nonsense-Mediated RNA Decay Influences Human Embryonic Stem Cell Fate." *Stem Cell Reports* 6 (6): 844–57.

Lovén, Jakob, Heather A. Hoke, Charles Y. Lin, Ashley Lau, David A. Orlando, Christopher R. Vakoc, James E. Bradner, Tong Ihn Lee, and Richard A. Young. 2013. "Selective Inhibition of Tumor Oncogenes by Disruption of Super-Enhancers." *Cell* 153 (2): 320–34.

Lu, Huasong, Yuhua Xue, Guoying K. Yu, Carolina Arias, Julie Lin, Susan Fong, Michel Faure, Ben Weisburd, Xiaodan Ji, Alexandre Mercier, James Sutton, Kunxin Luo, Zhenhai Gao, and Qiang Zhou. 2015. "Correction: Compensatory Induction of MYC Expression by Sustained CDK9 Inhibition via a BRD4-Dependent Mechanism." *eLife* 4 (July): e09993.

Mertins, Philipp, D. R. Mani, Kelly V. Ruggles, Michael A. Gillette, Karl R. Clauser, Pei Wang, Xianlong Wang, Jana W. Qiao, Song Cao, Francesca Petralia, Emily Kawaler, Filip Mundt, Karsten Krug, Zhidong Tu, Jonathan T. Lei, Michael L. Gatza, Matthew Wilkerson, Charles M. Perou, Venkata Yellapantula, Kuan-Lin Huang, Chenwei Lin, Michael D. McLellan, Ping Yan, Sherri R. Davies, R. Reid Townsend, Steven J. Skates, Jing Wang, Bing Zhang, Christopher R. Kinsinger, Mehdi Mesri, Henry Rodriguez, Li Ding, Amanda G. Paulovich, David Fenyö, Matthew J. Ellis, Steven A. Carr, and NCI CPTAC. 2016. "Proteogenomics Connects Somatic Mutations to Signalling in Breast Cancer." *Nature* 534 (7605): 55–62.

Moore, Michael J., Qingqing Wang, Caleb J. Kennedy, and Pamela A. Silver. 2010. "An Alternative Splicing Network Links Cell-Cycle Control to Apoptosis." *Cell* 142 (4): 625–36.

Musselman, Catherine A., Nikita Avvakumov, Reiko Watanabe, Christopher G. Abraham, Marie-Eve Lalonde, Zehui Hong, Christopher Allen, Siddhartha Roy, James K. Nuñez, Jac Nickoloff, Caroline A. Kulesza, Akira Yasui, Jacques Côté, and Tatiana G. Kutateladze. 2012. "Molecular Basis for H3K36me3 Recognition by the Tudor Domain of PHF1." *Nature Structural & Molecular Biology* 19 (12): 1266–72.

Paruch, Kamil, Michael P. Dwyer, Carmen Alvarez, Courtney Brown, Tin-Yau Chan, Ronald J. Doll, Kerry Keertikar, Chad Knutson, Brian McKittrick, Jocelyn Rivera, Randall Rossman, Greg Tucker, Thierry Fischmann, Alan Hruza, Vincent Madison, Amin A. Nomeir, Yaolin Wang, Paul Kirschmeier, Emma Lees, David Parry, Nicole Sgambellone, Wolfgang Seghezzi, Lesley Schultz, Frances Shanahan, Derek Wiswell, Xiaoying Xu, Quiao Zhou, Ray A. James, Vidyadhar M. Paradkar, Haengsoon Park, Laura R. Rokosz, Tara M. Stauffer, and Timothy J. Guzi. 2010. "Discovery of Dinaciclib (SCH 727965): A Potent and Selective Inhibitor of Cyclin-Dependent Kinases." *ACS Medicinal Chemistry Letters* 1 (5): 204–8.

Perez, Edith A. 2011. "Breast Cancer Management: Opportunities and Barriers to an Individualized Approach." *The Oncologist* 16 Suppl 1: 20–22.

Rahman, Shaila, Mathew E. Sowa, Matthias Ottinger, Jennifer A. Smith, Yang Shi, J. Wade Harper, and Peter M. Howley. 2011. "The Brd4 Extraterminal Domain Confers Transcription Activation Independent of pTEFb by Recruiting Multiple Proteins, Including NSD3." *Molecular and Cellular Biology* 31 (13): 2641–52.

Ren, Gang, Stavroula Baritaki, Himangi Marathe, Jingwei Feng, Sungdae Park, Sandy Beach, Peter S. Bazeley, Anwar B. Beshir, Gabriel Fenteany, Rohit Mehra, Stephanie Daignault, Fahd Al-Mulla, Evan Keller, Ben Bonavida, Ivana de la Serna, and Kam C. Yeung. 2012. "Polycomb Protein EZH2 Regulates Tumor Invasion via the Transcriptional Repression of the Metastasis Suppressor RKIP in Breast and Prostate Cancer." *Cancer Research* 72 (12): 3091–3104.

Rugo, Hope S., Olufunmilayo I. Olopade, Angela DeMichele, Christina Yau, Laura J. van 't Veer, Meredith B. Buxton, Michael Hogarth, Nola M. Hylton, Melissa Paoloni, Jane Perlmutter, W. Fraser Symmans, Douglas Yee, A. Jo Chien, Anne M. Wallace, Henry G. Kaplan, Judy C. Boughey, Tufia C. Haddad, Kathy S. Albain, Minetta C. Liu, Claudine Isaacs, Qamar J. Khan, Julie E. Lang, Rebecca K. Viscusi, Lajos Pusztai, Stacy L. Moulder, Stephen Y. Chui, Kathleen A. Kemmer, Anthony D. Elias, Kirsten K. Edmiston, David M. Euhus, Barbara B. Haley, Rita Nanda, Donald W. Northfelt, Debasish Tripathy, William C. Wood, Cheryl Ewing, Richard Schwab, Julia Lyandres, Sarah E. Davis, Gillian L. Hirst, Ashish Sanil, Donald A. Berry, Laura J. Esserman, and I-SPY 2 Investigators. 2016. "Adaptive Randomization of Veliparib-Carboplatin Treatment in Breast Cancer." *The New England Journal of Medicine* 375 (1): 23–34.

Sandoval, Juan, and Manel Esteller. 2012. "Cancer Epigenomics: Beyond Genomics." *Current Opinion in Genetics & Development* 22 (1): 50–55.

Sansó, Miriam, Rebecca S. Levin, Jesse J. Lipp, Vivien Ya-Fan Wang, Ann Katrin Greifenberg, Elizabeth M. Quezada, Akbar Ali, Animesh Ghosh, Stéphane Larochelle, Tariq M. Rana, Matthias Geyer, Liang Tong, Kevan M. Shokat, and Robert P. Fisher. 2016. "P-TEFb Regulation of Transcription Termination Factor Xrn2 Revealed by a Chemical Genetic Screen for Cdk9 Substrates." *Genes & Development* 30 (1): 117–31.

Schreiber, Stuart L., and Bradley E. Bernstein. 2002. "Signaling Network Model of Chromatin." *Cell* 111 (6): 771–78.

Sengupta, Surojeet, Michael C. Biarnes, and V. Craig Jordan. 2014. "Cyclin Dependent Kinase-9 Mediated Transcriptional de-Regulation of cMYC as a Critical Determinant of Endocrine-Therapy Resistance in Breast Cancers." *Breast Cancer Research and Treatment* 143 (1): 113–24.

Shannon, Paul, Andrew Markiel, Owen Ozier, Nitin S. Baliga, Jonathan T. Wang, Daniel Ramage, Nada Amin, Benno Schwikowski, and Trey Ideker. 2003. "Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks." *Genome Research* 13 (11): 2498–2504.

Sokal, Robert R., and F. James Rohlf. 1962. "The Comparison of Dendrograms by Objective Methods." *Taxon* 11 (2). International Association for Plant Taxonomy (IAPT): 33–40.

Strahl, B. D., and C. D. Allis. 2000. "The Language of Covalent Histone Modifications." *Nature* 403 (6765): 41–45.

Stratikopoulos, Elias E., Meaghan Dendy, Matthias Szabolcs, Alan J. Khaykin, Celine Lefebvre, Ming-Ming Zhou, and Ramon Parsons. 2015. "Kinase and BET Inhibitors Together Clamp

Inhibition of PI3K Signaling and Overcome Resistance to Therapy." *Cancer Cell* 27 (6): 837–51.

Subramanian, Aravind, Rajiv Narayan, Steven M. Corsello, David D. Peck, Ted E. Natoli, Xiaodong Lu, Joshua Gould, John F. Davis, Andrew A. Tubelli, Jacob K. Asiedu, David L. Lahr, Jodi E. Hirschman, Zihan Liu, Melanie Donahue, Bina Julian, Mariya Khan, David Wadden, Ian C. Smith, Daniel Lam, Arthur Liberzon, Courtney Toder, Mukta Bagul, Marek Orzechowski, Oana M. Enache, Federica Piccioni, Sarah A. Johnson, Nicholas J. Lyons, Alice H. Berger, Alykhan F. Shamji, Angela N. Brooks, Anita Vrcic, Corey Flynn, Jacqueline Rosains, David Y. Takeda, Roger Hu, Desiree Davison, Justin Lamb, Kristin Ardlie, Larson Hogstrom, Peyton Greenside, Nathanael S. Gray, Paul A. Clemons, Serena Silver, Xiaoyun Wu, Wen-Ning Zhao, Willis Read-Button, Xiaohua Wu, Stephen J. Haggarty, Lucienne V. Ronco, Jesse S. Boehm, Stuart L. Schreiber, John G. Doench, Joshua A. Bittker, David E. Root, Bang Wong, and Todd R. Golub. 2017. "A Next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles." *Cell* 171 (6): 1437–52.e17.

Tomasetto, C., C. Régnier, C. Moog-Lutz, M. G. Mattei, M. P. Chenard, R. Lidereau, P. Basset, and M. C. Rio. 1995. "Identification of Four Novel Human Genes Amplified and Overexpressed in Breast Carcinoma and Localized to the q11-q21.3 Region of Chromosome 17." *Genomics* 28 (3): 367–76.

Wahl, Geoffrey M., and Benjamin T. Spike. 2017. "Cell State Plasticity, Stem Cells, EMT, and the Generation of Intra-Tumoral Heterogeneity." *NPJ Breast Cancer* 3 (April): 14.

Wang, Chunjie, Danh Tran-Thanh, Juan C. Moreno, Thomas R. Cawthorn, Lindsay M. Jacks, Dong-Yu Wang, David R. McCready, and Susan J. Done. 2011. "Expression of Abl Interactor 1 and Its Prognostic Significance in Breast Cancer: A Tissue-Array-Based Investigation." *Breast Cancer Research and Treatment* 129 (2): 373–86.

Wen, Hong, Yuanyuan Li, Yuanxin Xi, Shiming Jiang, Sabrina Stratton, Danni Peng, Kaori Tanaka, Yongfeng Ren, Zheng Xia, Jun Wu, Bing Li, Michelle C. Barton, Wei Li, Haitao Li, and Xiaobing Shi. 2014. "ZMYND11 Links Histone H3.3K36me3 to Transcription Elongation and Tumour Suppression." *Nature* 508 (7495): 263–68.

Whyte, Warren A., David A. Orlando, Denes Hnisz, Brian J. Abraham, Charles Y. Lin, Michael H. Kagey, Peter B. Rahl, Tong Ihn Lee, and Richard A. Young. 2013. "Master Transcription Factors and Mediator Establish Super-Enhancers at Key Cell Identity Genes." *Cell* 153 (2): 307–19.

Wu, Tao, Hugo Borges Pinto, Yasunao F. Kamikawa, and Mary E. Donohoe. 2015. "The BET Family Member BRD4 Interacts with OCT4 and Regulates Pluripotency Gene Expression." *Stem Cell Reports* 4 (3): 390–403.

Yamamoto, Shoji, Zhenhua Wu, Hege G. Russnes, Shinji Takagi, Guillermo Peluffo, Charles Vaske, Xi Zhao, Hans Kristian Moen Vollan, Reo Maruyama, Muhammad B. Ekram, Hanfei

Sun, Jee Hyun Kim, Kristopher Carver, Mattia Zucca, Jianxing Feng, Vanessa Almendro, Marina Bessarabova, Oscar M. Rueda, Yuri Nikolsky, Carlos Caldas, X. Shirley Liu, and Kornelia Polyak. 2014. "JARID1B Is a Luminal Lineage-Driving Oncogene in Breast Cancer." *Cancer Cell* 25 (6): 762–77.

Yang, Xiaojing, R. K. Murthy Karuturi, Feng Sun, Meiyee Aau, Kun Yu, Rongguang Shao, Lance D. Miller, Patrick Boon Ooi Tan, and Qiang Yu. 2009. "CDKN1C (p57KIP2) Is a Direct Target of EZH2 and Suppressed by Multiple Epigenetic Mechanisms in Breast Cancer Cells." *PloS One* 4 (4). Public Library of Science: e5011.

Zhu, Jiajun, Morgan A. Sammons, Greg Donahue, Zhixun Dou, Masoud Vedadi, Matthäus Getlik, Dalia Barsyte-Lovejoy, Rima Al-awar, Bryson W. Katona, Ali Shilatifard, Jing Huang, Xianxin Hua, Cheryl H. Arrowsmith, and Shelley L. Berger. 2015. "Gain-of-Function p53 Mutants Co-Opt Chromatin Pathways to Drive Cancer Growth." *Nature* 525 (7568): 206–11.

# CHAPTER 4: Functional ATLAS of the Epithelial Breast Cells

## 4.1 Abstract

An in-depth molecular characterization of epithelial breast cell responses to the growth-promoting ligands is required to elucidate how the microenvironment (ME) signals affect cell-intrinsic regulatory networks and the cellular phenotypes they control, such as cell growth, progression, and differentiation. This is particularly important towards understanding the mechanisms of breast cancer initiation and progression. However, the current mechanisms by which the ME signals influence these cellular phenotypes are not well established. Using multi-omics data, we developed a functional ATLAS of the epithelial breast cell response to distinct microenvironmental perturbations in MCF10 cell line. Using responses from six growth ligands at various time points, we were able to decipher their specific effects on cellular phenotypes. Our results suggest that Oncostatin M (OSM), a growth factor, exerts its regulatory influence on STAT3 induction to promote BRD4, a chromatin reader, to regulate cell proliferation in breast epithelial cells.

## 4.2 Introduction

Our current understanding of biological actions of growth factors come from examining a growth factor in isolation. However, growth factors are known to interact with the extracellular matrix (ECM), and these associations influence cell behavior. ECM plays a vital role in the healthy mammary tissue and tumor formation. Details about these interactions within the complex and continuously changing microenvironment are not well understood and are likely to be very important during the tumor formation.

The goal of this project is to contribute to further development of the NIH Library of Integrated Network-based cellular signatures (LINCS) program by developing a dataset and computational strategy to elucidate how microenvironment (ME) signals affect cell intrinsic intracellular transcriptional- and protein-defined molecular networks to generate experimentally observable cellular phenotypes (**Fig 4.1**). We will infer these regulatory relationships by combining measurements of ME perturbagen-induced changes in multiple cellular phenotypes, RNA expression and regulatory protein expression in a core set of cell lines with measurements of responses of the same lines to chemical and genomic perturbagens generated by LINCS sites. Integrative analysis of these data will enable us to address three key questions:

How are ME peturbagen-induced cellular phenotypes regulated by specific molecular networks?

Do subsets of ME-induced perturbations elicit common molecular network changes and phenotypic responses?

Do specific molecular network signatures influence multiple cellular phenotypes?

A project workflow is depicted in **Fig 4.2.**

## 4.3 Results

**Deep profiling of signaling data shows canonical footprints of each ligand.** In order to determine the signaling effects of each growth ligand on breast epithelial phenotypes, we needed to generate a comprehensive signaling profile for each individual ligand EGF, HGF, OSM, BMP2, TGFB, and IFNG. We constructed a global heatmap for all the significant signaling proteins/phosphoproteins by analyzing their protein expressions obtained from the RPPA dataset. First, we compared the signaling effects of each ligand measured across all time points (1, 4, 8, 24, 48 hours) against PBS 0 hour. Using a t-test, we generated 201 differentially expressed (DE) proteins where at least one protein was shown to be statistically significant (p_value < 0.05) in at least one ligand across all time points. Out of 201 DE proteins, we found 117 were functionally enriched in the canonical pathways (cellular phenotypes of interest) that were associated with cell growth, migration, apoptosis, cell adhesion, focal adhesion, differentiation, etc. (**Fig 4.6, Fig 4.7**). The heatmap of these proteins showed similar signaling profiles for all six ligands across all time points. This is further corroborated by the high correlation values ($R^2$ >0.8) we observed in the pairwise regression plots between each paired of ligands across all time points (**Fig 4.5**). We particularly observed increased protein concentrations for cell cycle genes CCNB1, CDK1, PLK1, RB1, ribosomal genes RPS6 and its phosphorylation and decreased values for CDKN1A, ERBB3, BCL2L11, CASP7, histone HIST3H3, and PAR in all ligands. Their concentrations were more pronounced at 24 and 48 hours, suggesting a possible mechanistic involvement at 24 hours (**Fig 4.6**). We further observed elevated protein concentrations of GJA1, NDRG1, HIF1A and STAT3 phosphorylation unique to OSM at all time points; CD274, IRF1, TRIM25, and phosphorylated PTK2 unique to IFNG, TGM1; ITGA2 unique to TGFB; SERPINE1 unique to BMP2 and TGFB across all time points (**Fig 4.7**). We observed that nearly half of the canonical genes in cell cycle pathways,  were up and half were down  ( **Fig 4.9A**).

Similarly, to tease out the individual effects of BMP2, TGFB, and IFNG on the various cellular phenotypes, we performed differential analysis of these ligands with respect to EGF. We applied the same t-test (with p-value < 0.05) to generate a list of 113 DE proteins of which 81 belonged to our canonical pathways of interest (**Fig 4.8**). From the heatmap we found that canonical cell cycle genes were mostly down ( **Fig 4.9B**), indicating, the additive effect of EGF on these ligands in promoting cell cycle induction.

**Mapping of signaling proteins to their transcriptional profiles identified distinct cellular functions and chromatin accessibility.** To find out how the signaling proteins affect the transcriptional machinery, we first identified 35 TFs from the 201 DE proteins, ( **Fig 4.10**) and performed a functional enrichment analyses on these 35 TFs using the DAVID annotation tool (Huang, Sherman, and Lempicki 2009). We found 3 distinct functional modules associated with the responses, namely, cell cycle, proteoglycans in cancer pathways, and signaling pathways regulating pluripotency of stem cells ( **Fig 4.11**). We  repeated this analysis for transcriptomic data and identified 158 TFs from 4879 DE gene list (**Fig 4.12**). We then generated  unique TFs for each ligand using a Venn program ( **Fig 4.13**) and mapped these unique TFs to their target genes. We then performed functional enrichment analyses, on these target genes. The results showed many non-specific functional associations with various pathway activities including p53 signaling, regulation of metabolism, and cancer pathways. Thus, we focused our investigation on integrating the signaling profiles with the transcriptomic data. We generated a PPI network using 201 DE proteins (see method) and mapped them to 158 DE TFs and their target genes. We then performed a functional analyses on the subnetwork containing only DE TFs and their targets (top 25 out degrees) (**Fig. 4.14**). Our results showed functional enrichment of specific phenotypes to each ligand, such as ABC transporters for OSM,  NOD-like receptor signaling pathways, serotonergic synapse,  argenine and proline metabolism for BMP2, cell cycle and breast cancer for TGFB, and pathways in cancer, breast cancers, PI3K-AKT signaling for IFNG. However, our results did not show any unique functional phenotypes associated with EGF or HGF which is consistent with the prior observations from the signaling data (**Fig 4.14** ).

**Signaling network profiles identified BRD4 as a mediator of OSM induced cell proliferation.** To find out how each ligand's distinct signaling profile influences gene expression at the transcriptomic level,  we projected each ligand's DE protein list on to the 201 DE proteins' PPI profiles using Cytoscape Software (Shannon et al. 2003) (see PPI generation in method section). This allowed us to identify the signaling proteins that were unique to a ligand or combination of ligands at 24- and 48-hour time points (**Fig. 4.15)**. We observed BRD4 was common to EGF, HGF, and OSM, however, its gene expression was only elevated in OSM at both 24 and 48 hours. BRD4 is a chromatin reader that plays an essential role as a structural scaffold with kinase activity that influences transcription by recruiting the positive transcription elongation factor complex-b (P-TEFb) to the transcription pre-initiation complex of inducible genes.  In our previous chapter, we have shown that BRD4 is an atypical signaling protein that could interact with a diverse group of proteins resulting in various cellular changes. Therefore, we postulated that BRD4 may play an important role in mediating OSM's growth activities. We then ask the question what other genes were significantly upregulated in OSM and not in others that also interacted with BRD4's ? Our result showed that STAT3 and its targets were all significantly upregulated in OSM compared to the other ligands, confirming its enhanced signaling and transcriptional activities for OSM (**Fig 4.16**). STAT3 is a transcription factor that regulates cellular processes involving cell-cycle progression and the anti-apoptotic program (Bromberg 2001). STAT3 is activated by tyrosine and serine phosphorylation in response to various growth factors including cytokines such as IL-6, leading to the formation of dimers that rapidly translocate into the nucleus and activate the promoters of target genes (Darnell, Kerr, and Stark 1994), (Jr. and Darnell 1997). Previous work has shown that IL-6 promotes a nuclear complex of STAT3 and CDK9 (a cell cycle gene), a major component of active P-TEFb complex in activation and deactivation of RNA Pol II (Hou, Ray, and Brasier 2007; Ray et al. 2014; Hou, Ray, and Brasier 2007). These studies, as well as our previous chapter, have shown that within the P-TEFb complex, BRD4 uniquely interacted with P-TEFb complex through CDK9. Hence, the focus of our section was to understand how STAT3 may recruit BRD4 to promote a proliferative and anti-apoptotic cellular state.

**STAT3 induction promotes BRD4 to regulate cell proliferation when treated with OSM.** To determine the mechanism of enhanced transcriptional activity of STAT3 and BRD4, we sought to identify their interacting genes. So, we looked at the expressions of all the targets genes of STAT3 that may directly interacted with BRD4, and led to increased expressions of proliferation and decreased expressions of anti-apoptotic events. We observed known genes who's over expressions lead to proliferation/ angiogenesis, such as HIF1A, CDK9, IRF1, SOD2, SOCS3, and MUC1, were upregulated and had open chromatin conformations from ATACSeq, indicating their active transcribed states (**Fig 4.16**). Interestingly, we also observed, HMOX1, a liver microsomal protein with activity to degrade heme to bilirubin (Jr. and Darnell 1997; Maines 1988) to be downregulated when treated with OSM. We found that HMOX1 was a target gene of BRD4 and was differentially downregulated at 24 hour. While overexpression of HMOX1 in cancer cell promotes cancer proliferation and survival, its deficiency in normal cells enhances DNA damages and carcinogenesis (Gueron et al. 2014; Hill et al. 2005). Furthermore, a previous study has shown Src/STAT3-dependent HMOX1 induction to mediate chemo resistance of breast cancer cells to doxorubicin by promoting autophagy (Tan et al. 2015). These are all consistent with our observations of decreased expressions of HMOX1 in normal MCF10A breast cell. However, the mechanisms by which HMOX1 regulates BRD4 is not very clear and will require experimental validations. Because in the previous chapter we have provided a detailed mechanism that induced cell cycle arrest involving CDK9, BRD4, and P-TEFb, we postulate that, STAT3 induction promotes BRD4 to regulate cell proliferation when treated with OSM. We can hypothesize that BRD4 uses the same transcriptional machinery to promote cell proliferation through PTEF-b when induced by STAT3 during OSM treatment.

### 4.4 Discussion

In this study, we have generated a comprehensive functional ATLAS of breast  cells to study the effects of six growth factors, namely EGF, HGF, OSM, BMP2, TGFB, and IFNG on ECM of normal breast tissue. Using an integrative multi-omics approach, we are able to reveal unique footprints for these growth factors. It appears from our analysis that ECM and HGF exert common cellular phenotypes, this is evident by the their signaling and transcriptomic profiles of canonical cell cycle genes. We observed reduced cell cycle effects for BMP2, IFNG and TGFB when compared to EGF.  OSM demonstrated more proliferative effects by inducing STAT3 protein. Using the combined profiles, we were able to reveal  a potential mechanism for promoting cell proliferation via BRD4.

Deciphering the relationship between growth factors and ECM is essential, and can be used to growth factor-based therapies or lead to the development of novel treatment strategies for cancer. A better understanding of the relationship between these classes of molecules and how it can be exploited to treat cancer is needed.

## 4.5 Materials and Method

**Data acquisition.** All data were obtained from LINCS sponsored Microenvironment Perturbagen (MEP) project. We developed a pipeline that integrates protein measurements from RPPA data, gene expressions from data Ranse data, and chromatin conformation from ATACSeq data that provides a unified view of their profiles.

**RPPA data processing.** For proteomics data, Level1, log2 PBS normalized protein measurement data were downloaded from ( https://www.synapse.org/#!Synapse:syn12555331) on June 6, 2019. The data consisted of highly sensitive and selective 295 antibody-based protein and phosphoprotein measurements of six growth ligands, namely, EGF, HGF, OSM, BMP2, TGFB, IFNG. The data contained three replicates for each of these ligands and a control (PBS). These samples were measured at five time points (1, 4, 8, 24, and 48 hours). The quality of data was assessed using the tsn-e (Dale, Cain, and Zell 2009) plot (**Fig 4.3A**).

**RNASeq data processing.** For transcriptomic data, ~200 GB raw fastq files were downloaded from (https://www.synapse.org/#!Synapse:syn18518040) on April 23, 2019. The data contained three replicates for each ligand (EGF, HGF, OSM, BMP2, TGFB, IFNG) and 4 replicates for control (PBS) measured at two time points (24 and 48 hours). The quality of data was assessed using the tsn-e plot (**Fig 4.3B**).

**Differential expression analysis.** We used Omicsoft sequence aligner (OSA) (Hu et al. 2012) to process raw fastq files to generate Fragments Per Kilobase of transcript per Million (FPKM) mapped read counts. We retained only those genes with ten or more counts in at least one replicate sample to be considered for the subsequent differential expression analysis. We then performed differential expression (DE) analysis using  EdgeR (Hu et al. 2012; “edgeR” n.d.) package from R library to generate the DE gene list. We normalized data using the trimmed mean of M-values (TMM) (Hu et al. 2012; “edgeR” n.d.; Robinson and Oshlack 2010) between each pair of samples (control vs condition)  with a p-value <0.01. (**Table 4.1B**).

**ATACSeq data processing.** For epigenomic data (ATACseq), 87.4 MB of level1, log2(TMM+1) normalized peaks files (bed) were downloaded from (https://www.synapse.org/#!Synapse:syn18352471) on February 27, 2019. The data contained three replicates for each ligand  (EGF, HGF, OSM, BMP2, TGFB, IFNG) and three replicates for control (PBS) for two time points (24 and 48 hours).

**Library preparation and sequencing.** Data ATAC-seq libraries were prepared from 0-, 24-, and 48-hour Collection 1 MCF10A samples following the Omni-ATAC protocol (Corces, 2017). 100bp PE reads were sequenced on an Illumina HiSeq 2500 Sequencer by the OHSU MPSSR. Technical replicates of samples TGFB_48_B and PBS_0_C were added by OHSU MPSSR.

Sequence alignment, preprocessing, and peak calling. ATAC-seq files were processed and aligned using the "ATACseq (1 -> 3)" workflow on the AnswerALS Galaxy server (answer.csbi.mit.edu). 100-bp PE reads were trimmed of adapter sequences and low-quality bases using Trimmomatic (Galaxy version 0.36.5). Reads were trimmed of low-quality bases (Phred score < 15) at the read start or end, and Nextera adapter sequences (CTGTCTCTTATA) were trimmed from read ends (minimum of a 2-bp overlap required for trimming). Reads were aligned to the human genome (hg38) using Bowtie2 (Galaxy version 2.3.4.1) in paired-end mode with otherwise default settings. BAM files were filtered to remove secondary alignments, unmapped reads, and mitochondiral

DNA alignments using ngsutils bam filter (Galaxy version 0.5.9). PCR duplicates were detected and removed using Picard MarkDuplicates (Galaxy version 2.7.1.2).

The de-duplicated, filtered BAM file was used for peak calling and quantification.

Peaks were called using MACS2 (Galaxy Version 2.1.1.20160309.5) using the following parameters: -format BAMPE -nomodel -extsize 200 -shift -100 -qvalue 0.01.

**Sample quality assessment.** The fraction of reads in peaks (FRiP) was calculated to assess successful chromatin enrichment in each sample. FRiP was initially calculated from each sample's de-duplicated, filtered BAM file and MACS2 narrowPeaks output using R (3.5.1) and R package ChIPQC (1.18.1). Samples with a FRiP < 0.175 were excluded from further analysis. 26/50 samples had a FRiP score > 0.175; these samples were flagged as "PASS" in the frip_filter metadata field. 24/50 samples had a FRiP score < 0.175; these low-quality samples were marked as "FAIL" in the frip_filter field in the metadata.

**Differential peak analysis.** We performed differential peak calling R libraries DiffBind ("DiffBind" n.d.), on each paired samples (PBS vs ligand) bed files. We then ran EdgeR using the default threshold of FDR <= 0.05 to identify significantly differentially bound sites (peaks). We annotated the sites that were 1kb away from transcription start site (TSS) using human TxDb.Hsapiens.UCSC.hg38.knownGene annotation database ("TxDb.Hsapiens.UCSC.hg38.knownGene" n.d.). (**Fig 4.4**). These DE peaks denote relative open or close regions compare to the controls (PBS at 0 hour).

**Correlation analysis.** We computed pairwise linear correlations and their corresponding $R^2$ scores between each pair of ligands using linear regression for all five time points for RPPA data (**Fig 4.5**).

**Signaling footprint generation**. We used PBS as control and an unpaired (t-test) with p-value < 0.05 on the 295 signaling proteins expression measurements to generate a list of differential (DE) proteins for all ligands. We further filtered the list by taking a 1.2 fold change (FC) value on these DE proteins to generate a signaling profile for these 201 proteins.

**PPI generation.** We generated a protein-protein interaction (PPI) network by mapping 201 DE proteins and their first neighbors to Stringdb database with a 0.7 confidence score. We generated a network of 6113 nodes (proteins) and 23255 edges (interactions between nodes). We then generated individual DE protein list for each ligand (**Table 4.1A**) and projected them onto the network (**Fig 4.6**).

**TF target mapping and functional enrichment analysis**. We performed TF target analysis by mapping DE gene / protein lists for individual ligand with our local Transfac database. We then used a custom R script to perform enrichment analysis by mapping these target gene lists to KEGG annotation database. We identified 35 and 158 transcription factors (TF) in 201 DE protein list and DE gene list respectively.

**Motif analysis for ATACSeq.** We used HOMER's motif finding tool ("Homer Software and Data Download" n.d.) to identify putative TFs and their motifs from our differential peaks. We used these TFs and annotMotif.pl script to further identify their target genes and corresponding binding regions (promoter or exonic regions). Peaks detected at the promoter regions of a specific gene typically signals for transcription initiation process, whereas peaks bound at the exonic regions signals for transcribed expressions.

## 4.6 Figures

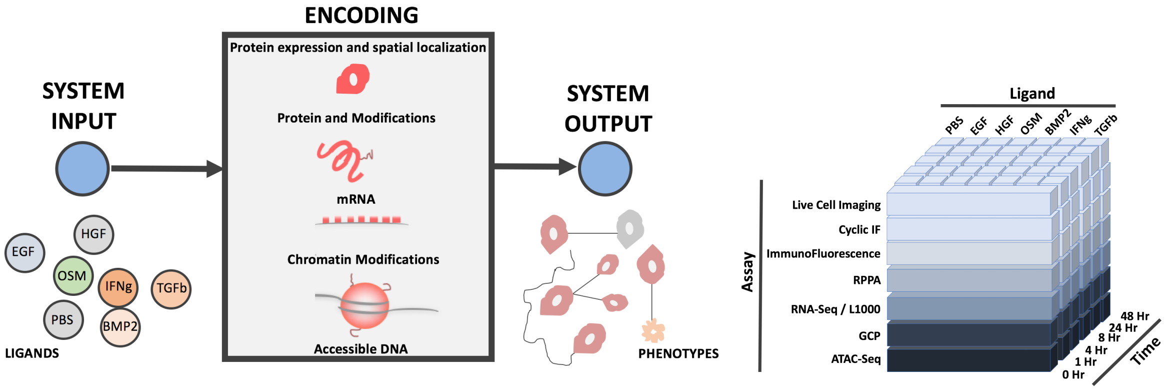


**Figure 4.1. Overview of the ECM MCF10A project.**
Showing various assays used in the study for characterization of ECM in normal breast MCF10A cell line. The objective is to characterize effects of various growth ligands on ECM structure that influences specific cellular phenotypes. These relationships are depicted as system input and output. The figure is adapted from LINCS ECM MCF10A project.

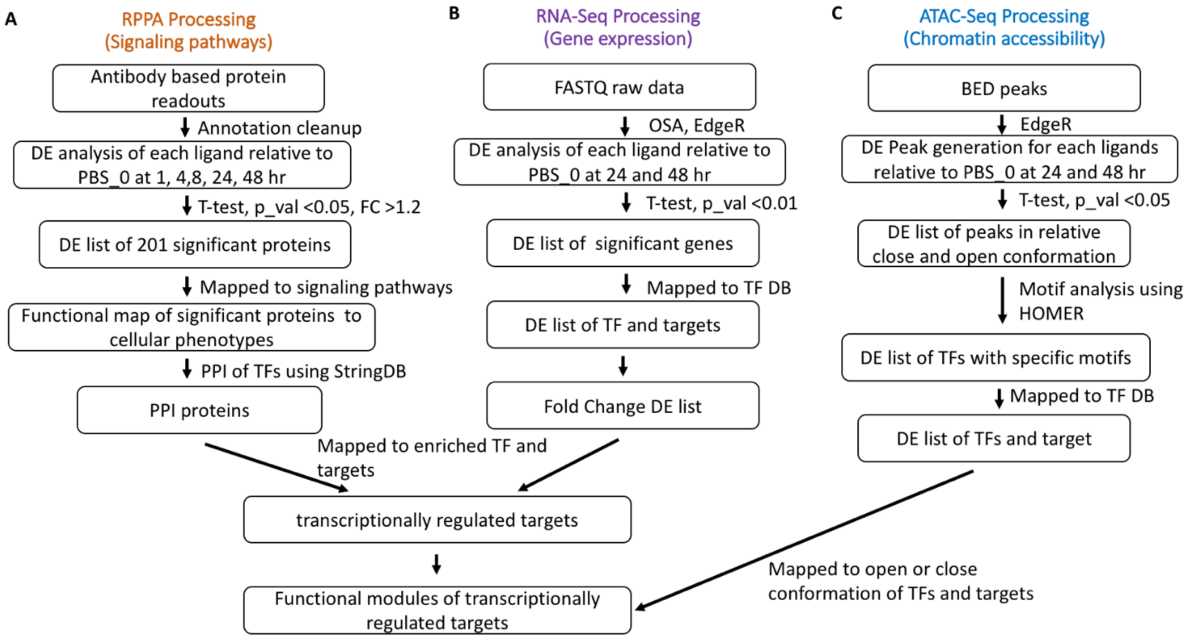

A
RPPA Processing
(Signaling pathways)
Antibody based protein readouts
Annotation cleanup
DE analysis of each ligand relative to PBS_0 at 1, 4,8, 24, 48 hr
T-test, p_val <0.05, FC >1.2
DE list of 201 significant proteins
Mapped to signaling pathways
Functional map of significant proteins to cellular phenotypes
PPI of TFs using StringDB
PPI proteins
B
RNA-Seq Processing
(Gene expression)
FASTQ raw data
OSA, EdgeR
DE analysis of each ligand relative to PBS_0 at 24 and 48 hr
T-test, p_val <0.01
DE list of significant genes
Mapped to TF DB
DE list of TF and targets
Fold Change DE list
C
ATAC-Seq Processing
(Chromatin accessibility)
BED peaks
EdgeR
DE Peak generation for each ligands relative to PBS_0 at 24 and 48 hr
T-test, p_val <0.05
DE list of peaks in relative close and open conformation
Motif analysis using HOMER
DE list of TFs with specific motifs
Mapped to TF DB
DE list of TFs and target
Mapped to enriched TF and targets
transcriptionally regulated targets
Functional modules of transcriptionally regulated targets
Mapped to open or close conformation of TFs and targets

**Figure 4.2. Design workflow of the integrative framework.**
(A) Showing processing and analyses steps for RPPA data, (B) RNASeq, and (C) ATACSeq data.

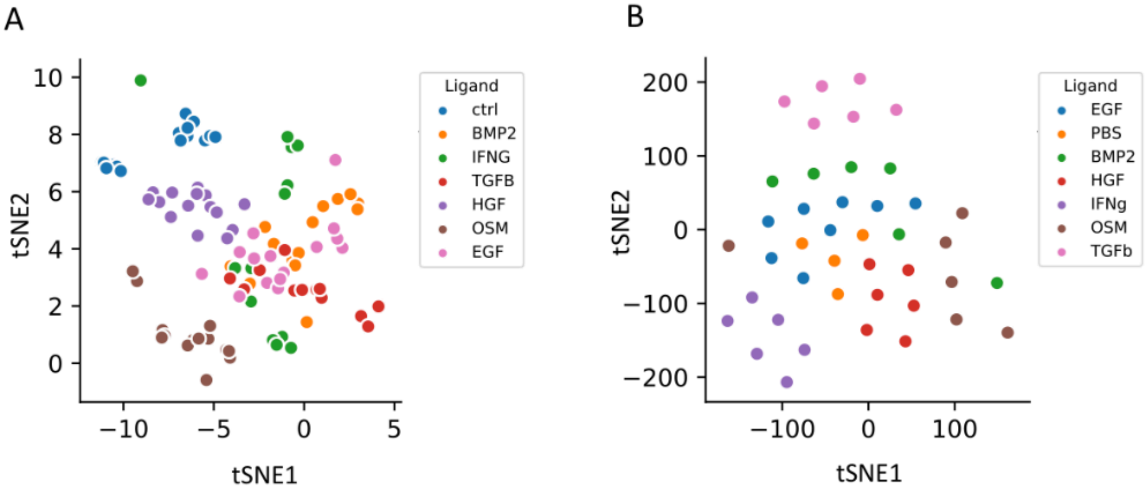
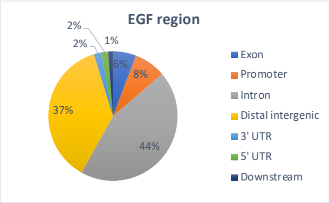
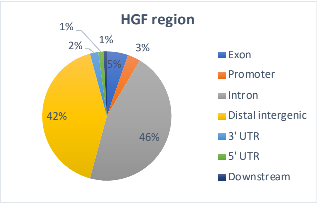


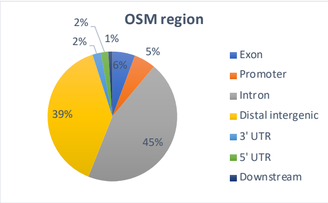
**Figure 4.3. Data quality of RPPA and RNASeq data**
(A) tSNE plots representing log2 normalized RPPA data with of six ligands, a control (PBS) and their replicates. (B) tSNE plot representing log2 normalized RNASeq data with six ligands, a control (PBS) and their replicates.

A: 24

| region | EGF region | close | open | HGF region | close | open | OSM region | close | open |
|---|---|---|---|---|---|---|---|---|---|
| **Exon** | 372 | 254 | 118 | 415 | 281 | 134 | 1082 | 550 | 532 |
| **Promoter** | 508 | 121 | 387 | 249 | 150 | 99 | 1035 | 623 | 412 |
| **Intron** | 2835 | 2025 | 810 | 3560 | 2229 | 1331 | 8593 | 4344 | 4249 |
| **Distal intergenic** | 2374 | 1835 | 539 | 3239 | 2125 | 1114 | 7451 | 3596 | 3855 |
| **3' UTR** | 120 | 84 | 36 | 162 | 112 | 50 | 400 | 185 | 215 |
| **5' UTR** | 108 | 45 | 63 | 101 | 62 | 39 | 355 | 210 | 145 |
| **Downstream** | 78 | 55 | 23 | 65 | 44 | 21 | 194 | 73 | 121 |
| **Total** | 6395 | | | 7791 | | | 19110 | | |

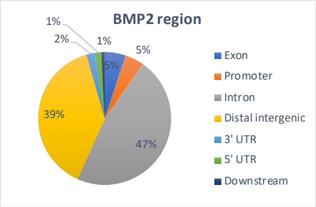


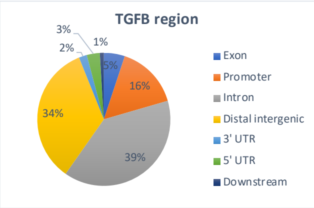




B: 24

| BMP2 region | close | open | TGFB region | close | open | IFNG region | close | open |
|---|---|---|---|---|---|---|---|---|
| 1150 | 650 | 500 | 1385 | 738 | 647 | 1549 | 802 | 747 |
| 1036 | 850 | 186 | 4181 | 3763 | 418 | 1872 | 974 | 898 |
| 10594 | 4422 | 6172 | 10647 | 4836 | 5811 | 12018 | 6707 | 5311 |
| 8756 | 3733 | 5023 | 9278 | 3962 | 5316 | 10363 | 5738 | 4625 |
| 467 | 251 | 216 | 529 | 254 | 275 | 603 | 227 | 376 |
| 354 | 281 | 73 | 858 | 701 | 157 | 563 | 297 | 266 |
| 183 | 102 | 81 | 262 | 116 | 146 | 255 | 106 | 149 |
| 22540 | | | 27140 | | | 27223 | | |





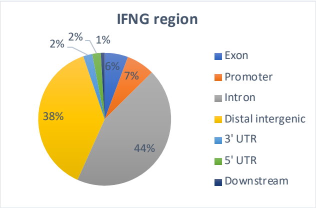
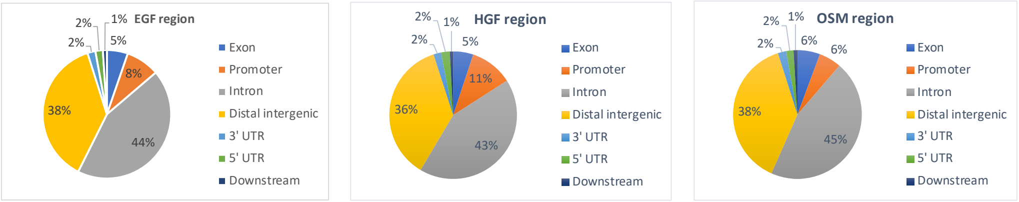


**Figure 4.4. Differential peak distributions of ATACSeq in genomic annotation at 24 and 28 hours.**

(A-D) Pie charts denoting the differential peak distribution (p-val < 0.01) of genomic regions at 24 hour for (A) EGF, HGF, OSM and (B) BMP2, TGFB, and IFNG. (C) Differential peak distribution of genomic regions at 48 hour for EGF, HGF, OSM, and for  (B) BMP2, TGFB, and IFNG.

## C: 48

| region | EGF region | close | open | HGF region | close | open | OSM region | close | open |
|---|---|---|---|---|---|---|---|---|---|
| Exon | 1223 | 741 | 482 | 1079 | 620 | 459 | 1117 | 633 | 484 |
| Promoter | 1965 | 302 | 1663 | 2293 | 312 | 1981 | 1106 | 729 | 377 |
| Intron | 10110 | 7141 | 2969 | 9016 | 5883 | 3133 | 8921 | 5144 | 3777 |
| Distal intergenic | 8745 | 6476 | 2269 | 7721 | 5067 | 2654 | 7568 | 4179 | 3389 |
| 3' UTR | 465 | 281 | 184 | 410 | 262 | 148 | 410 | 209 | 201 |
| 5' UTR | 435 | 124 | 311 | 466 | 143 | 323 | 362 | 245 | 117 |
| Downstream | 231 | 138 | 93 | 177 | 100 | 77 | 194 | 84 | 110 |
| Total | 23174 | | | 21162 | | | 19678 | | |



## D: 48

| BMP2 region | close | open | TGFB region | close | open | IFNG region | close | open |
|---|---|---|---|---|---|---|---|---|
| 614 | 383 | 231 | 1547 | 828 | 719 | 2329 | 1143 | 1186 |
| 530 | 176 | 354 | 5080 | 4511 | 569 | 4479 | 659 | 3820 |
| 5543 | 3026 | 2517 | 10646 | 5208 | 5438 | 19247 | 11169 | 8078 |
| 4704 | 2610 | 2094 | 9478 | 4523 | 4955 | 15997 | 9947 | 6050 |
| 250 | 143 | 107 | 600 | 280 | 320 | 878 | 382 | 496 |
| 166 | 106 | 60 | 998 | 820 | 178 | 1040 | 244 | 796 |
| 101 | 58 | 43 | 273 | 112 | 161 | 370 | 176 | 194 |
| 11908 | | | 28622 | | | 44340 | | |

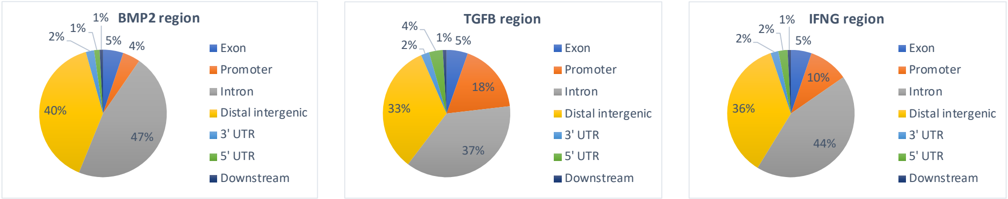
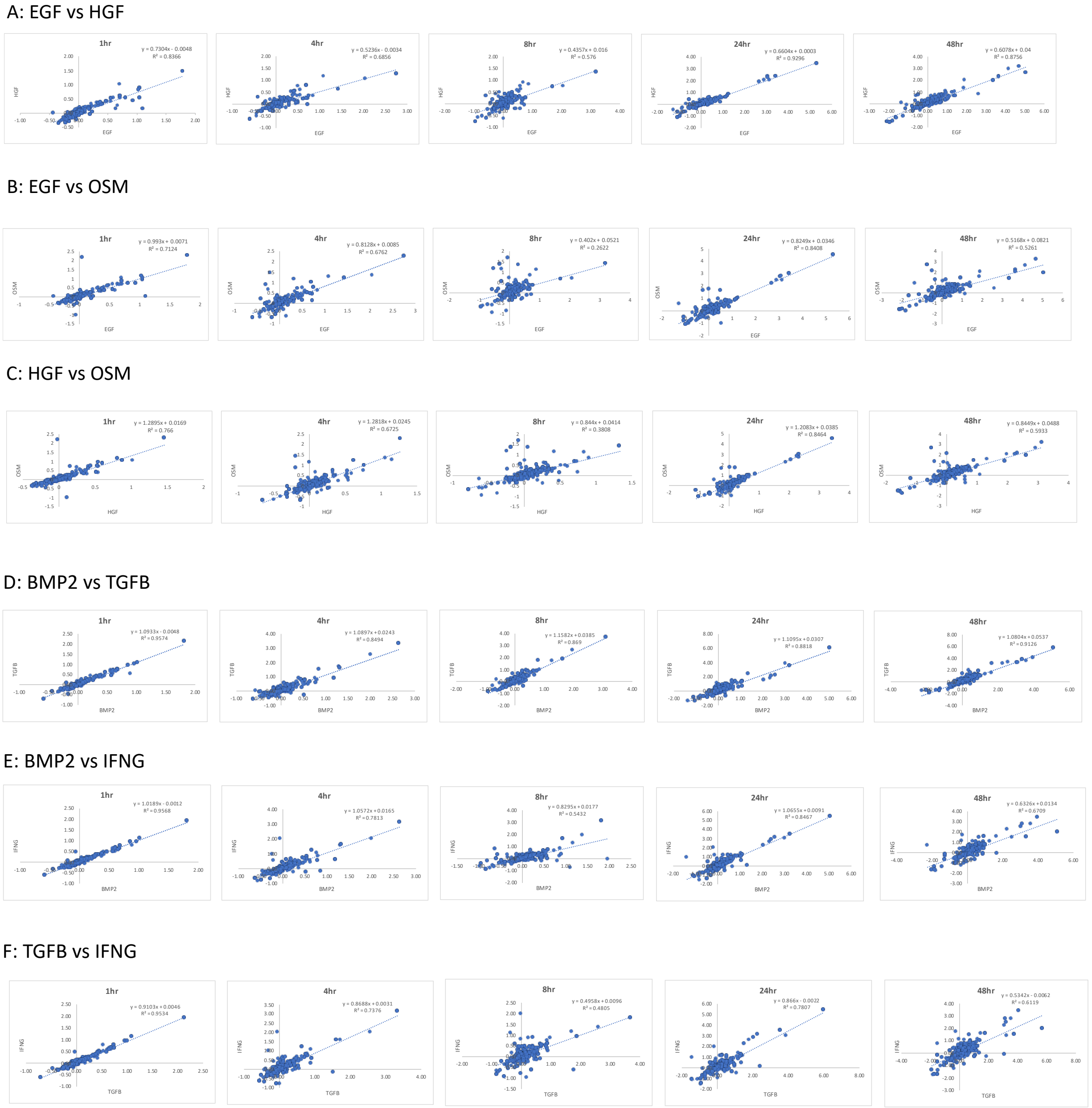


**Figure 4.4. Differential peak distributions of ATACSeq in genomic annotation at 24 and 28 hours.**

(A-D) Pie charts denoting the differential peak distribution (p-val < 0.01) of genomic regions at 24 hour for (A) EGF, HGF, OSM and (B) BMP2, TGFB, and IFNG. (C) Differential peak distribution of genomic regions at 48 hour for EGF, HGF, OSM, and for (B) BMP2, TGFB, and IFNG (continued).



A: EGF vs HGF

B: EGF vs OSM

C: HGF vs OSM

D: BMP2 vs TGFB

E: BMP2 vs IFNG

F: TGFB vs IFNG

**Figure 4.5. Correlations between pairwise ligands across time points.**
(A-D) Depicting the pairwise correlations and $R^2$ between two growth factors across all time points with respects to PBS (1, 4, 8, 24, and 48 hours) for (A) EGF vs HGF, (B) EGF vs OSM, (C) HGF vs OSM, (D) BMP2 VS TGFB, (E) BMP2 VS IFNG, and (F) TGFB VS IFNG using RPPA data.

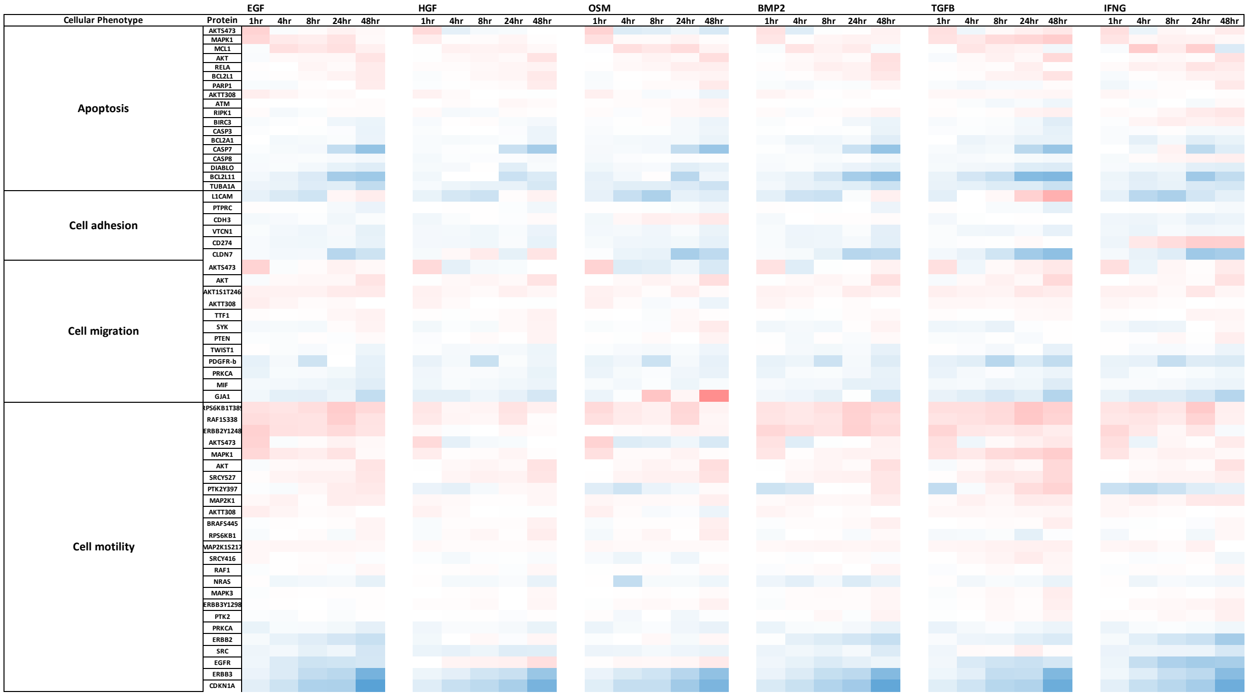


**Figure 4.6. Canonical  genes in their respective cellular pathways.**
Showing genes in their respective pathways with PBS as a control in all 6 ligands.

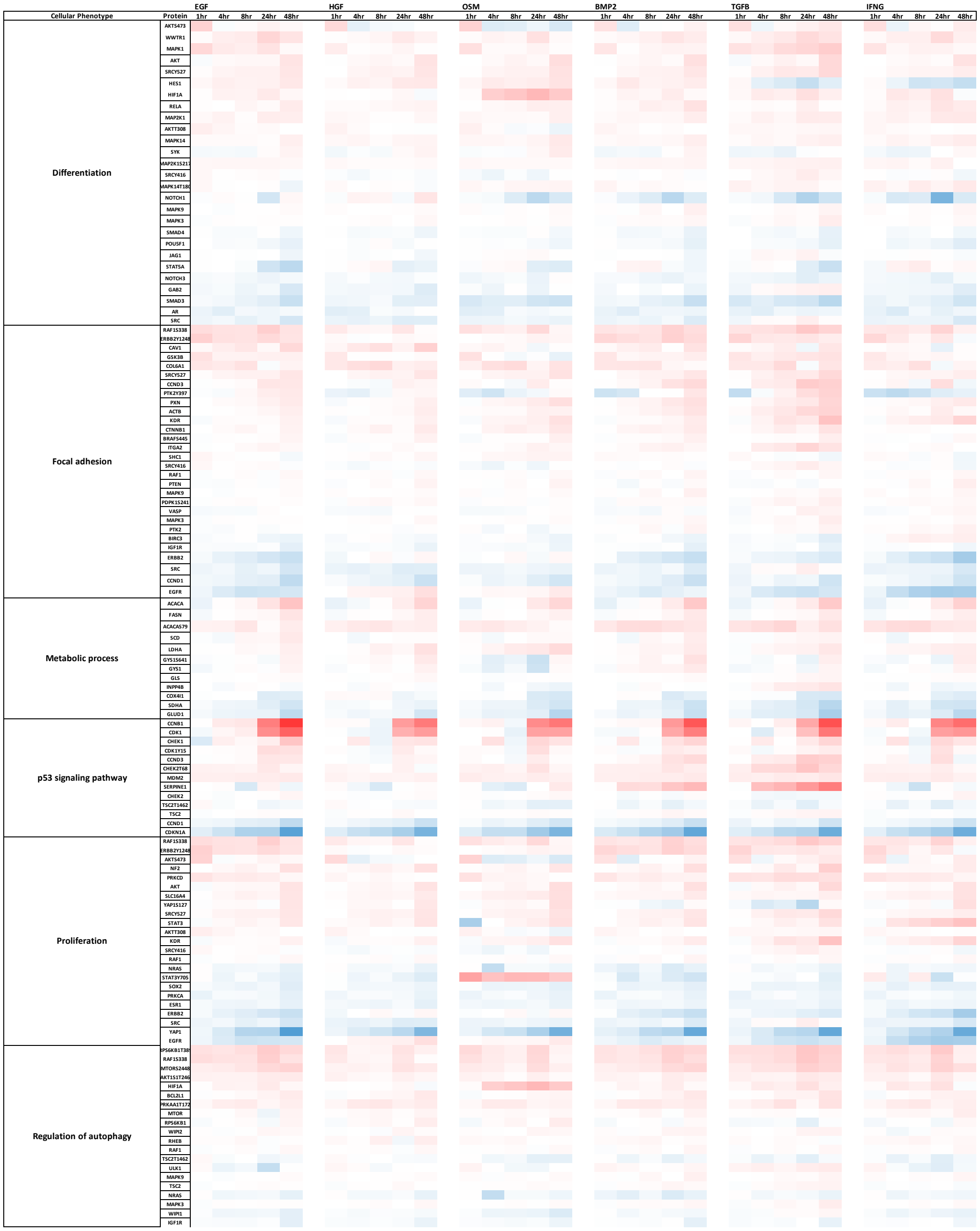


**Figure 4.7. Canonical  genes in their respective cellular pathways.**
Showing genes in their respective pathways with PBS as a control in all 6 ligands

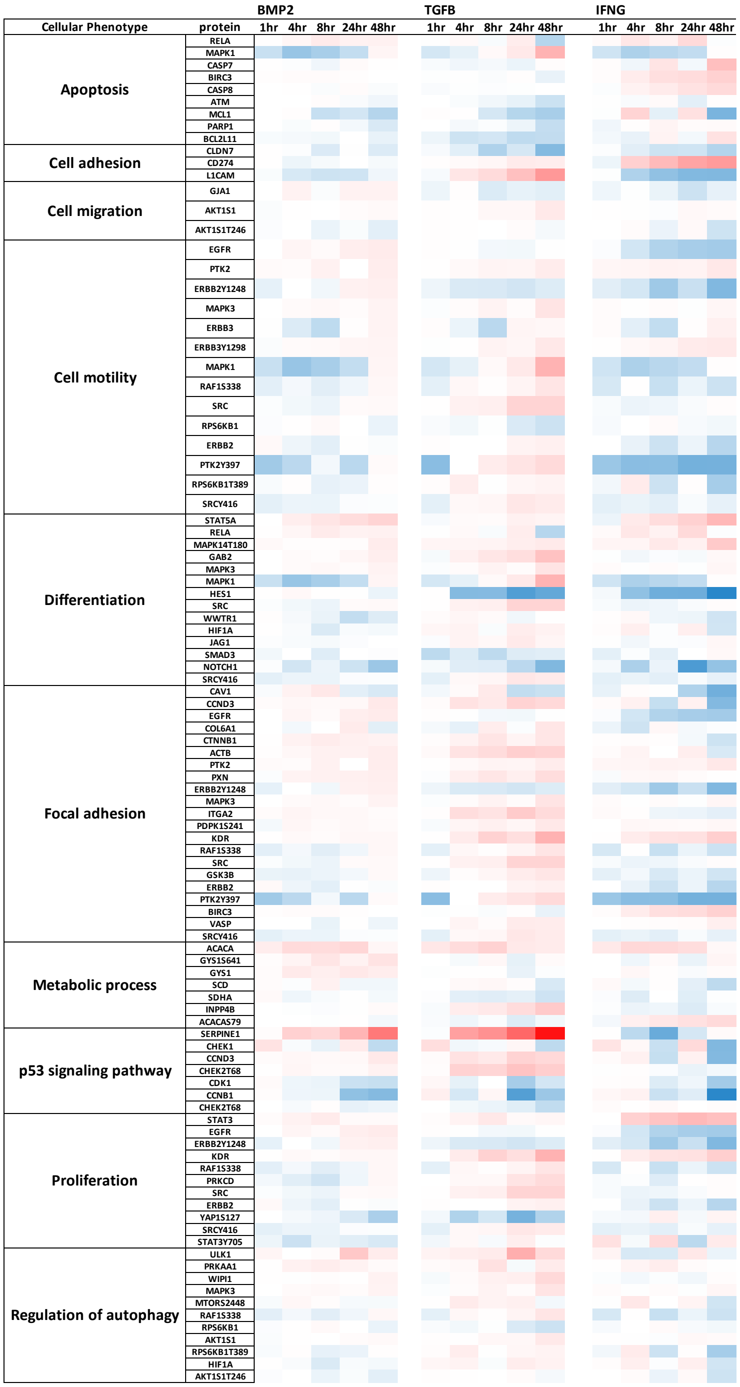


**Figure 4.8. Canonical  genes in their respective cellular pathways.**
Showing genes in their respective pathways with EGF as a control in BMP2, TGFB, and IFNG.

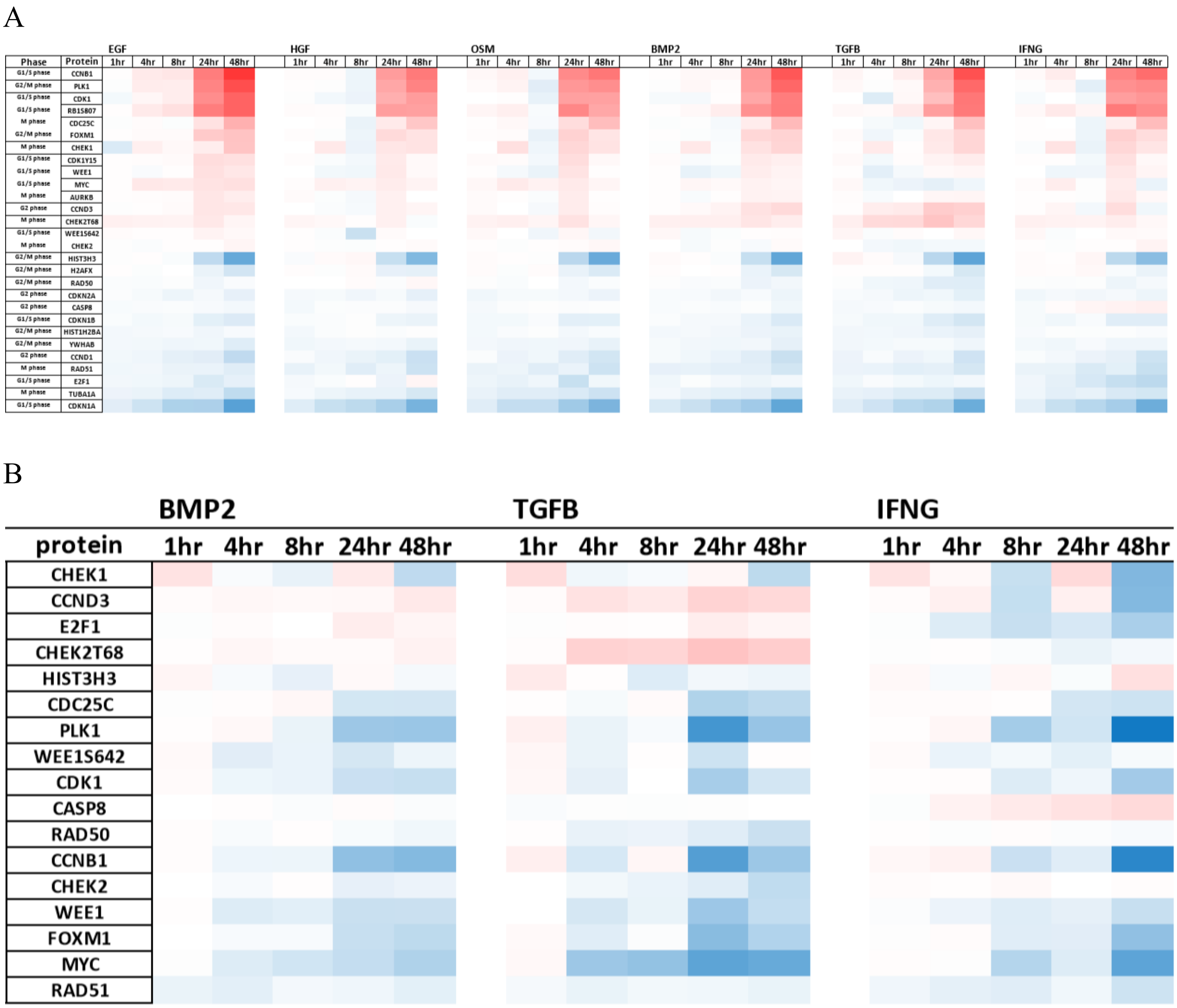


**Figure 4.9. Canonical cell cycle genes.**
(A) Showing canonical cell cycle genes with PBS as a control in all 6 ligands. (B) Showing canonical cell cycle genes with EGF as a control in BMP2, TGFB, and IFNG.

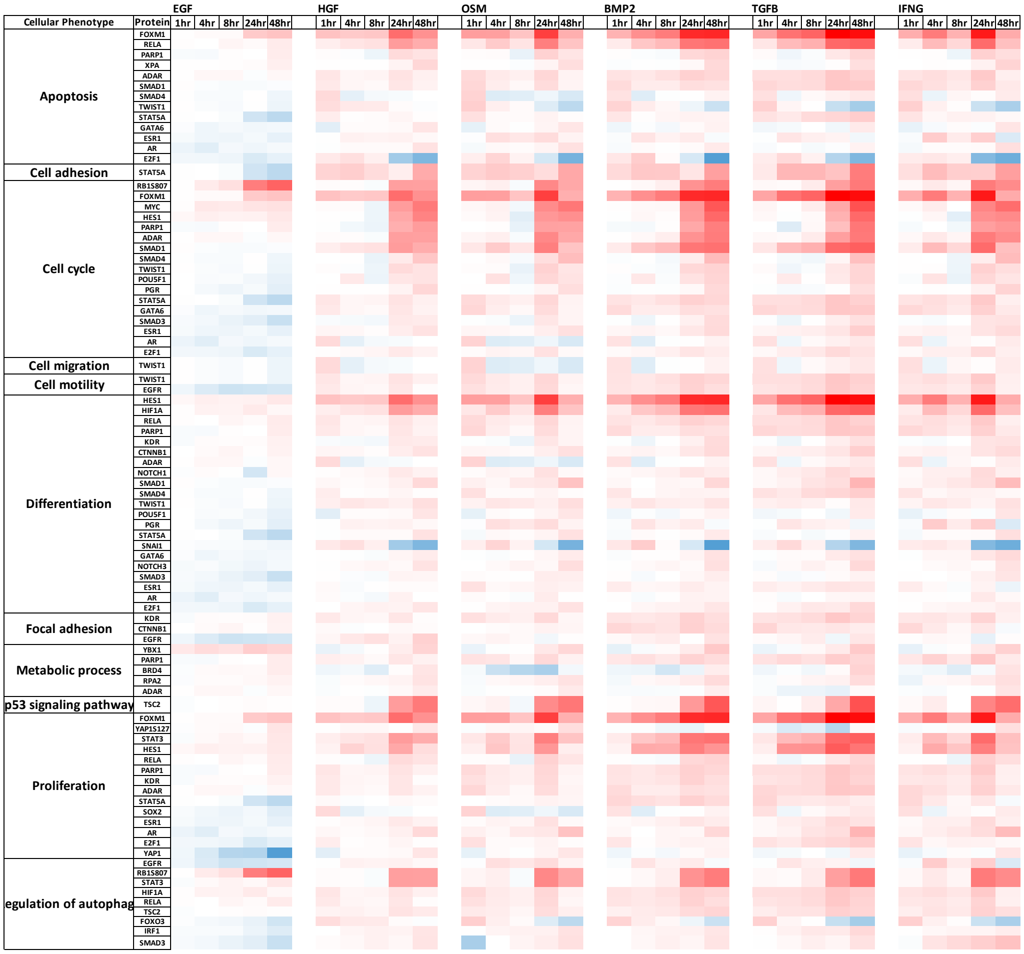
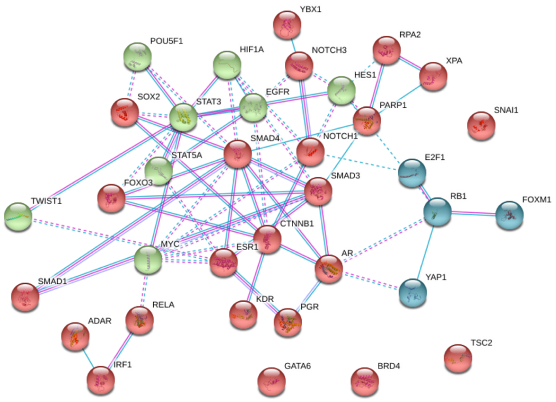


**Figure 4.10. 35 RPPA TFs.**
Showing signaling profile of 35 TFs with respect to PBS control.



| cluster | proteins | functional enrichment |
|---|---|---|
| cluster1 | AR, RELA, SMAD4, SMAD3, CTNNB1 | hsa05200:Pathways in cancer |
|  | NOTCH3, NOTCH1, TSC2, ESR1, CTNNB1 | hsa04919:Thyroid hormone signaling pathway |
|  | SOX2, SMAD4, SMAD3, SMAD1, CTNNB1 | hsa04550:Signaling pathways regulating pluripotency of stem cells |
| cluster2 | EGFR, HIF1A, MYC, STAT3, TWIST1 | hsa05205:Proteoglycans in cancer |
|  | EGFR, HIF1A, STAT5A, MYC, STAT3 | hsa05200:Pathways in cancer |
|  | EGFR, STAT5A, MYC | hsa04012:ErbB signaling pathway |
| cluster3 | E2F1, RB1 | hsa04110:Cell cycle |
|  | E2F1, RB1 | hsa05219:Bladder cancer |
|  | E2F1, RB1 | hsa05223:Non-small cell lung cancer |

**Figure 4.11. 35 Differential TF signaling proteins and their functional enrichments.** Depicting  three modules in 35 differential TF signaling proteins. The table shows their functional enrichments.

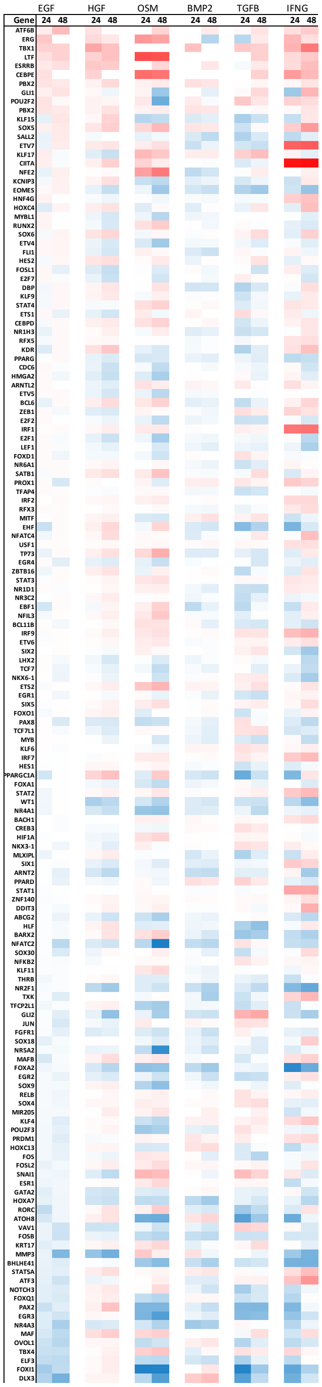


**Figure 4.12. 158 TFs in RNASeq.**
Showing gene expression profile of 158 TFs with respect to PBS control.

A

| Ligand | # of unique TF | Unique TF |
|---|---|---|
| EGF | 1 | KRT17 |
| HGF | 3 | HES2 FOSL1 TBX1 |
| OSM | 7 | LTF TP73 ERG HMGA1 GFI1 FOXA1 ETS2 |
| BMP2 | 1 | FLI1 |
|  |  | NR3C2 GLI2 HLF VAV1 EGR1 MLXIPL RELB |
| TGFB | 13 | CARF TFAP4 E2F2 MYB KLF9 NR6A1 |
|  |  | TXK NKX6-1 NKX2-5 STAT5A MIR205 SIX2 |
| IFNG | 9 | MAFB ESR1 |

B

| Ligand | # of unique T | Unique TF |
|---|---|---|
| EGF | 1 | ELF3<br>OVOL1 |
| HGF | 3 | PAX2 |
| OSM | 14 | OVOL1 ERG NR5A2 KLF11 TBX4 NR4A3 ETV1 ETV5 LTF CEBPE ETV4 SATB1 ESR1 PDX1 |
| BMP2 | 4 | RXRG HOXA7 SALL4 NR5A1 |
| TGFB | 2 | TFAP4<br>NR6A1 |
| IFNG | 11 | RFX3 WT1 ESRRB SOX30 RNF112 GLI1 TXK SOX18 STAT1 TBX1 HN |

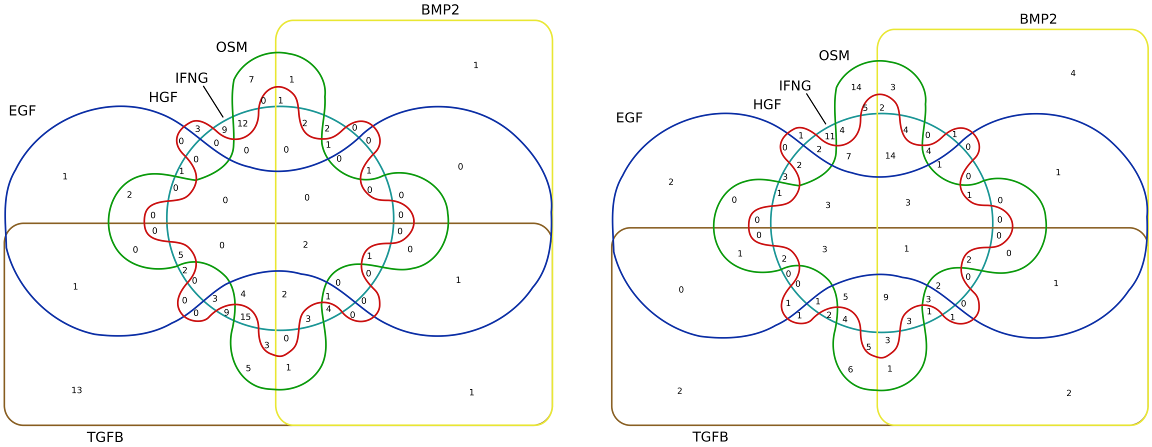
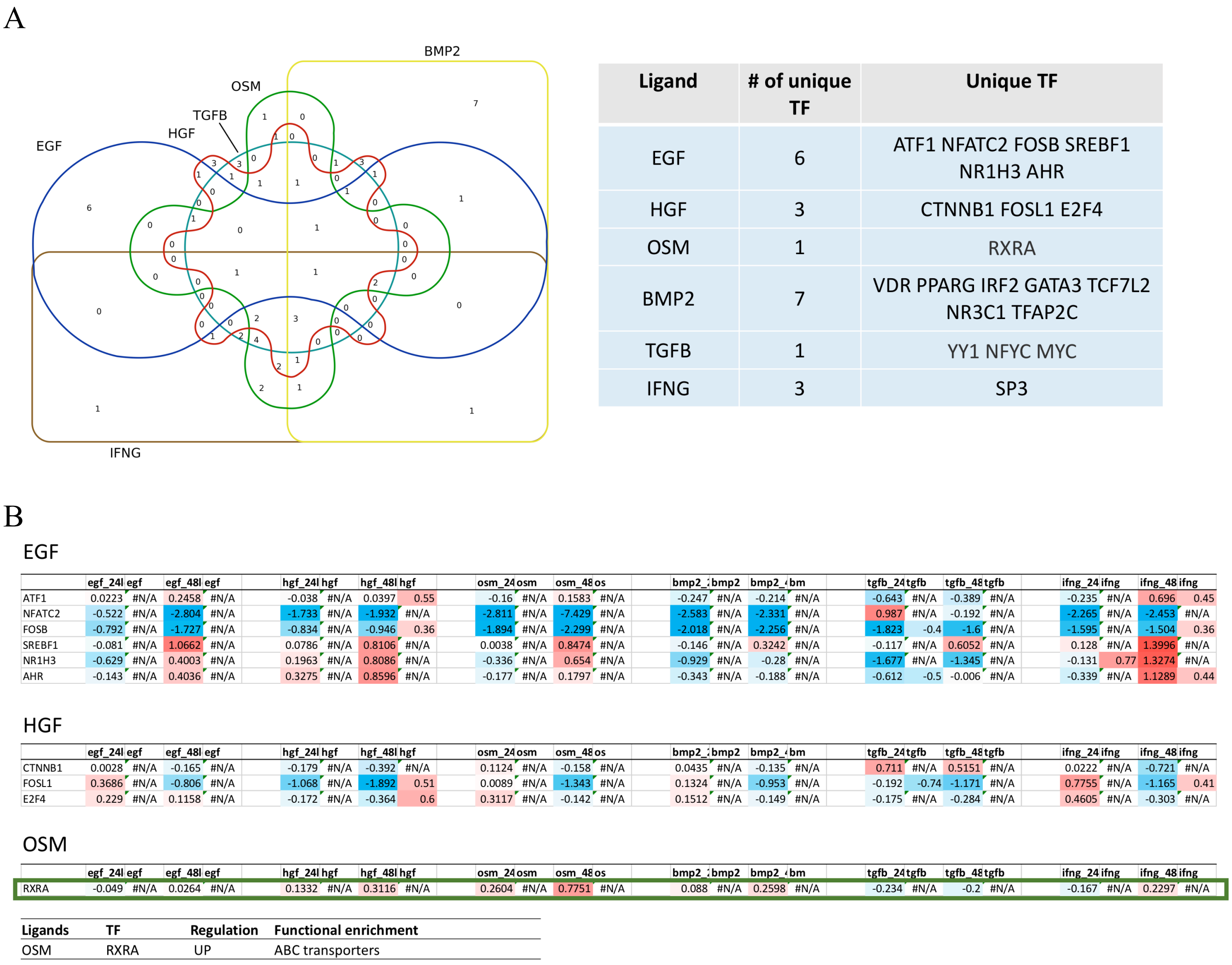


**Figure 4.13. Distribution of TFs in transcriptomic profiles.**
(A)Showing distribution of TFs in gene expression data for each ligand with respect to PBS at 24 hour. (B) at 48 hour.



| Ligand | # of unique TF | Unique TF |
|---|---|---|
| EGF | 6 | ATF1 NFATC2 FOSB SREBF1 NR1H3 AHR |
| HGF | 3 | CTNNB1 FOSL1 E2F4 |
| OSM | 1 | RXRA |
| BMP2 | 7 | VDR PPARG IRF2 GATA3 TCF7L2 NR3C1 TFAP2C |
| TGFB | 1 | YY1 NFYC MYC |
| IFNG | 3 | SP3 |

EGF

| | egf_24l | egf | egf_48l | egf | hgf_24l | hgf | hgf_48l | hgf | osm_24 | osm | osm_48 | os | bmp2_2 | bmp2 | bmp2_4 | bm | tgfb_24 | tgfb | tgfb_48 | tgfb | ifng_24 | ifng | ifng_48 | ifng |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ATF1 | 0.0223 | #N/A | 0.2458 | #N/A | -0.038 | #N/A | 0.0397 | 0.55 | -0.16 | #N/A | 0.1583 | #N/A | -0.247 | #N/A | -0.214 | #N/A | -0.643 | #N/A | -0.389 | #N/A | -0.235 | #N/A | 0.696 | 0.45 |
| NFATC2 | -0.522 | #N/A | -2.804 | #N/A | -1.733 | #N/A | -1.932 | #N/A | -2.811 | #N/A | -7.429 | #N/A | -2.583 | #N/A | -2.331 | #N/A | 0.987 | #N/A | -0.192 | #N/A | -2.265 | #N/A | -2.453 | #N/A |
| FOSB | -0.792 | #N/A | -1.727 | #N/A | -0.834 | #N/A | -0.946 | 0.36 | -1.894 | #N/A | -2.299 | #N/A | -2.018 | #N/A | -2.256 | #N/A | -1.823 | -0.4 | -1.6 | #N/A | -1.595 | #N/A | -1.504 | 0.36 |
| SREBF1 | -0.081 | #N/A | 1.0662 | #N/A | 0.0786 | #N/A | 0.8106 | #N/A | 0.0038 | #N/A | 0.8474 | #N/A | -0.146 | #N/A | 0.3242 | #N/A | -0.117 | #N/A | 0.6052 | #N/A | 0.128 | #N/A | 1.3996 | #N/A |
| NR1H3 | -0.629 | #N/A | 0.4003 | #N/A | 0.1963 | #N/A | 0.8086 | #N/A | -0.336 | #N/A | 0.654 | #N/A | -0.929 | #N/A | -0.28 | #N/A | -1.677 | #N/A | -1.345 | #N/A | -0.131 | 0.77 | 1.3274 | #N/A |
| AHR | -0.143 | #N/A | 0.4036 | #N/A | 0.3275 | #N/A | 0.8596 | #N/A | -0.177 | #N/A | 0.1797 | #N/A | -0.343 | #N/A | -0.188 | #N/A | -0.612 | -0.5 | -0.006 | #N/A | -0.339 | #N/A | 1.1289 | 0.44 |

HGF

| | egf_24l | egf | egf_48l | egf | hgf_24l | hgf | hgf_48l | hgf | osm_24 | osm | osm_48 | os | bmp2_2 | bmp2 | bmp2_4 | bm | tgfb_24 | tgfb | tgfb_48 | tgfb | ifng_24 | ifng | ifng_48 | ifng |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| CTNNB1 | 0.0028 | #N/A | -0.165 | #N/A | -0.179 | #N/A | -0.392 | #N/A | 0.1124 | #N/A | -0.158 | #N/A | 0.0435 | #N/A | -0.135 | #N/A | 0.711 | #N/A | 0.5151 | #N/A | 0.0222 | #N/A | -0.721 | #N/A |
| FOSL1 | 0.3686 | #N/A | -0.806 | #N/A | -1.068 | #N/A | -1.892 | 0.51 | 0.0089 | #N/A | -1.343 | #N/A | 0.1324 | #N/A | -0.953 | #N/A | -0.192 | -0.74 | -1.171 | #N/A | 0.7755 | #N/A | -1.165 | 0.41 |
| E2F4 | 0.229 | #N/A | 0.1158 | #N/A | -0.172 | #N/A | -0.364 | 0.6 | 0.3117 | #N/A | -0.142 | #N/A | 0.1512 | #N/A | -0.149 | #N/A | -0.175 | #N/A | -0.284 | #N/A | 0.4605 | #N/A | -0.303 | #N/A |

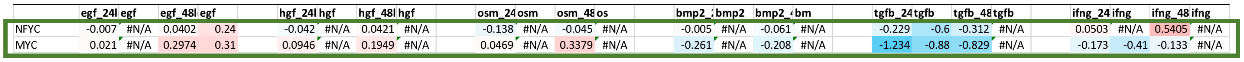
OSM

| | egf_24l | egf | egf_48l | egf | hgf_24l | hgf | hgf_48l | hgf | osm_24 | osm | osm_48 | os | bmp2_2 | bmp2 | bmp2_4 | bm | tgfb_24 | tgfb | tgfb_48 | tgfb | ifng_24 | ifng | ifng_48 | ifng |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| RXRA | -0.049 | #N/A | 0.0264 | #N/A | 0.1332 | #N/A | 0.3116 | #N/A | 0.2604 | #N/A | 0.7751 | #N/A | 0.088 | #N/A | 0.2598 | #N/A | -0.234 | #N/A | -0.2 | #N/A | -0.167 | #N/A | 0.2297 | #N/A |

| Ligands | TF | Regulation | Functional enrichment |
|---|---|---|---|
| OSM | RXRA | UP | ABC transporters |

**Figure 4.14. Functional enrichments of each ligand using RPPA, RNAseq, ATACSeq data.** (A) Showing TF distribution of unique ligands with respect to PBS control. (B) Showing functional enrichments of each unique ligand by mapping TFs from RPPA, RNASeq, and ATACSeq data.

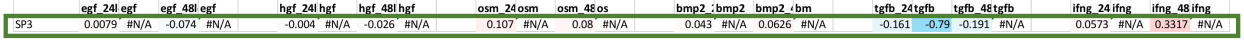
TGFB



| | egf_24l | egf | egf_48l | egf | | hgf_24l | hgf | hgf_48l | hgf | | osm_24 | osm | osm_48 | os | | bmp2_2 | bmp2 | bmp2_4 | bm | | tgfb_24 | tgfb | tgfb_48 | tgfb | | ifng_24 | ifng | ifng_48 | ifng |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| NFYC | -0.007 | #N/A | 0.0402 | 0.24 | | -0.042 | #N/A | 0.0421 | #N/A | | -0.138 | #N/A | -0.045 | #N/A | | -0.005 | #N/A | -0.061 | #N/A | | -0.229 | -0.6 | -0.312 | #N/A | | 0.0503 | #N/A | 0.5405 | #N/A |
| MYC | 0.021 | #N/A | 0.2974 | 0.31 | | 0.0946 | #N/A | 0.1949 | #N/A | | 0.0469 | #N/A | 0.3379 | #N/A | | -0.261 | #N/A | -0.208 | #N/A | | -1.234 | -0.88 | -0.829 | #N/A | | -0.173 | -0.41 | -0.133 | #N/A |

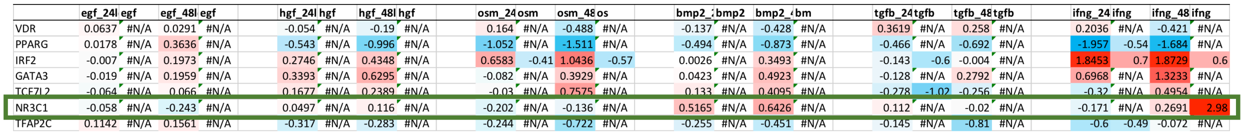
| Ligands | TF | Regulation | Functional enrichment |
|---|---|---|---|
| TGFB | NFYC | Mixed | Breast cancer |
| | MYC | Down | Cell cycle |

IFNG



| | egf_24l | egf | egf_48l | egf | | hgf_24l | hgf | hgf_48l | hgf | | osm_24 | osm | osm_48 | os | | bmp2_2 | bmp2 | bmp2_4 | bm | | tgfb_24 | tgfb | tgfb_48 | tgfb | | ifng_24 | ifng | ifng_48 | ifng |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| SP3 | 0.0079 | #N/A | -0.074 | #N/A | | -0.004 | #N/A | -0.026 | #N/A | | 0.107 | #N/A | 0.08 | #N/A | | 0.043 | #N/A | 0.0626 | #N/A | | -0.161 | -0.79 | -0.191 | #N/A | | 0.0573 | #N/A | 0.3317 | #N/A |

| Ligands | TF | Regulation | Functional enrichment |
|---|---|---|---|
| IFNG | SP3 | Down | Pathway in cancers |
| | SP3 | Down | Breast cancers |
| | SP3 | Down | PI3K-AKT signaling |
| | SP3 | Down | Cell cycle |
| | SP3 | Mixed | Focal adhesion |



| | egf_24l | egf | egf_48l | egf | | hgf_24l | hgf | hgf_48l | hgf | | osm_24 | osm | osm_48 | os | | bmp2_2 | bmp2 | bmp2_4 | bm | | tgfb_24 | tgfb | tgfb_48 | tgfb | | ifng_24 | ifng | ifng_48 | ifng |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| VDR | 0.0637 | #N/A | 0.0291 | #N/A | | -0.054 | #N/A | -0.19 | #N/A | | 0.164 | #N/A | -0.488 | #N/A | | -0.137 | #N/A | -0.428 | #N/A | | 0.3619 | #N/A | 0.258 | #N/A | | 0.2036 | #N/A | -0.421 | #N/A |
| PPARG | 0.0178 | #N/A | 0.3636 | #N/A | | -0.543 | #N/A | -0.996 | #N/A | | -1.052 | #N/A | -1.511 | #N/A | | -0.494 | #N/A | -0.873 | #N/A | | -0.466 | #N/A | -0.692 | #N/A | | -1.957 | -0.54 | -1.684 | #N/A |
| IRF2 | -0.007 | #N/A | 0.1973 | #N/A | | 0.2746 | #N/A | 0.4348 | #N/A | | 0.6583 | -0.41 | 1.0436 | -0.57 | | 0.0026 | #N/A | 0.3493 | #N/A | | -0.143 | -0.6 | -0.004 | #N/A | | 1.8453 | 0.7 | 1.8729 | 0.6 |
| GATA3 | -0.019 | #N/A | 0.1959 | #N/A | | 0.3393 | #N/A | 0.6295 | #N/A | | -0.082 | #N/A | 0.3929 | #N/A | | 0.0423 | #N/A | 0.4923 | #N/A | | -0.128 | #N/A | 0.2792 | #N/A | | 0.6968 | #N/A | 1.3233 | #N/A |
| TCF7L2 | -0.064 | #N/A | 0.066 | #N/A | | 0.1677 | #N/A | 0.2389 | #N/A | | -0.03 | #N/A | 0.7575 | #N/A | | 0.133 | #N/A | 0.4095 | #N/A | | -0.278 | -1.02 | -0.256 | #N/A | | -0.32 | #N/A | 0.4954 | #N/A |
| NR3C1 | -0.058 | #N/A | -0.243 | #N/A | | 0.0497 | #N/A | 0.116 | #N/A | | -0.202 | #N/A | -0.136 | #N/A | | 0.5165 | #N/A | 0.6426 | #N/A | | 0.112 | #N/A | -0.02 | #N/A | | -0.171 | #N/A | 0.2691 | 2.98 |
| TFAP2C | 0.1142 | #N/A | 0.1561 | #N/A | | -0.317 | #N/A | -0.283 | #N/A | | -0.244 | #N/A | -0.722 | #N/A | | -0.255 | #N/A | -0.451 | #N/A | | -0.145 | #N/A | -0.81 | #N/A | | -0.6 | -0.49 | -0.072 | #N/A |

| Ligands | TF | TF_target | Regulation | Functional enrichment |
|---|---|---|---|---|
| BMP2 | NR3C1 | NCOA2 | UP | Serotonergic synapse |
| | | KLF15 | UP | Arginine and proline metabolism |
| | | ALOX15B | UP | NOD-like receptor signaling pathway |
| | | SGK1 | UP | |
| | | NOS3 | UP | |
| | | CADM3 | UP | |
| | | TSC22D3 | UP | |
| | | CACNA1C | UP | |
| | | NR3C1 | UP | |
| | | KLF9 | UP | |
| | | NFKBIA | UP | |
| | | DUSP1 | UP | |
| | | MAOA | UP | |
| | | TNFAIP3 | UP | |
| | | FKBP5 | UP | |

**Figure 4.14. Functional enrichments of each ligand using RPPA, RNAseq, ATACSeq data.** (A) Showing TF distribution of unique ligands with respect to PBS control. (B) Showing functional enrichments of each unique ligand by mapping TFs from RPPA, RNASeq, and ATACSeq data (continued).

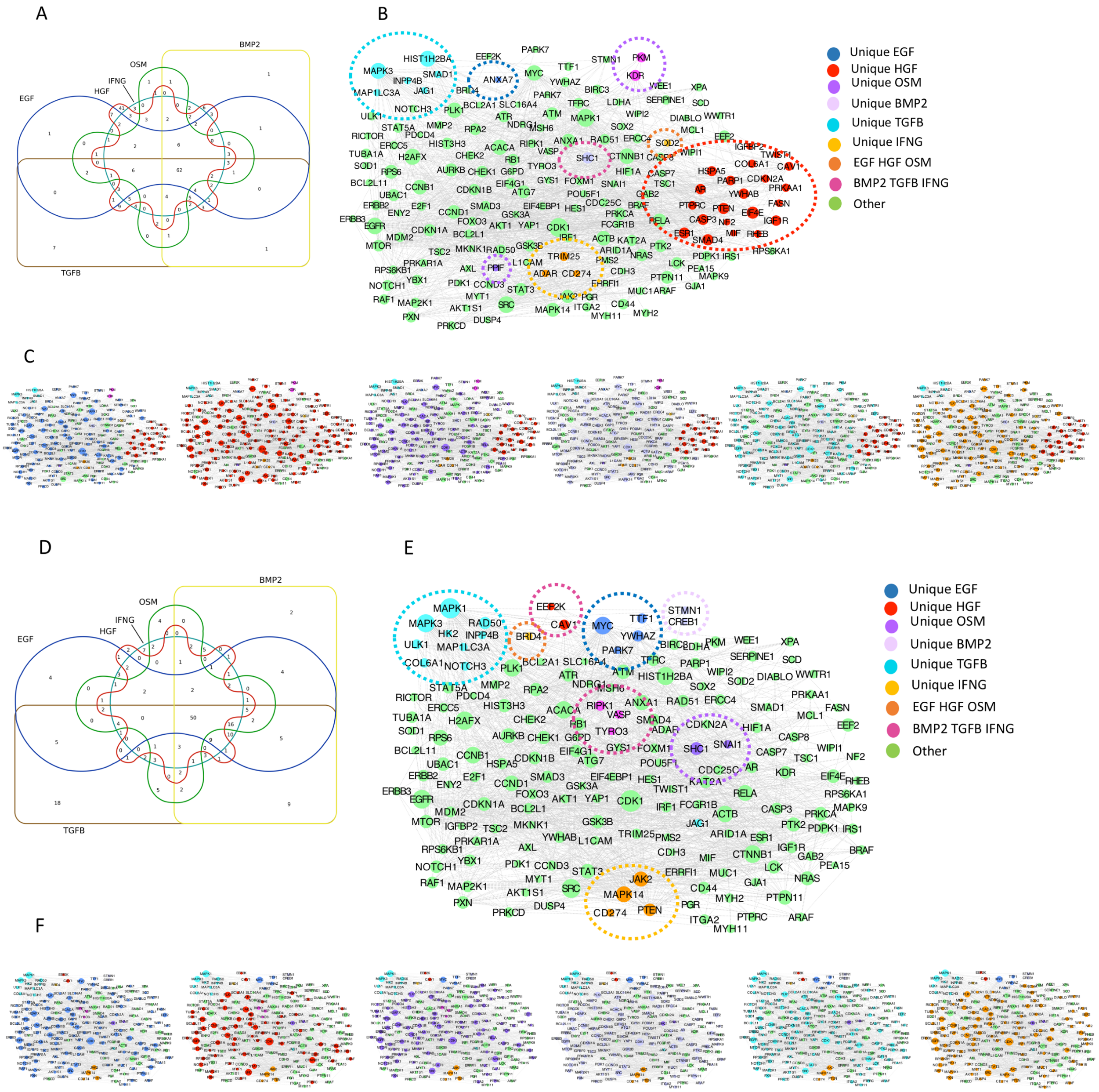


**Figure 4.15. PPI network cluster of signaling DE proteins.**

**(**A - F**)** Venn diagram showing distribution of unique and overlapped DE proteins in all ligands at 24 hour with respect to PBS control. (B) Showing a subset of a network topology associated with unique functional modules denoted by dashed ovals at 24 hour. Some nodes are shown due to the low node degree. (C) Thumbnail showing distribution of DE proteins for each ligand in the network at 24 hour with respect to PBS control. (D) showing distribution of unique and overlapped DE proteins in all ligands at 48 hour. (E) Subset of a functional modules at 48 hour with respect to PBS control. (F) Thumbnail showing distribution of DE proteins for each ligand in the network at 48 hour with respect to PBS control.

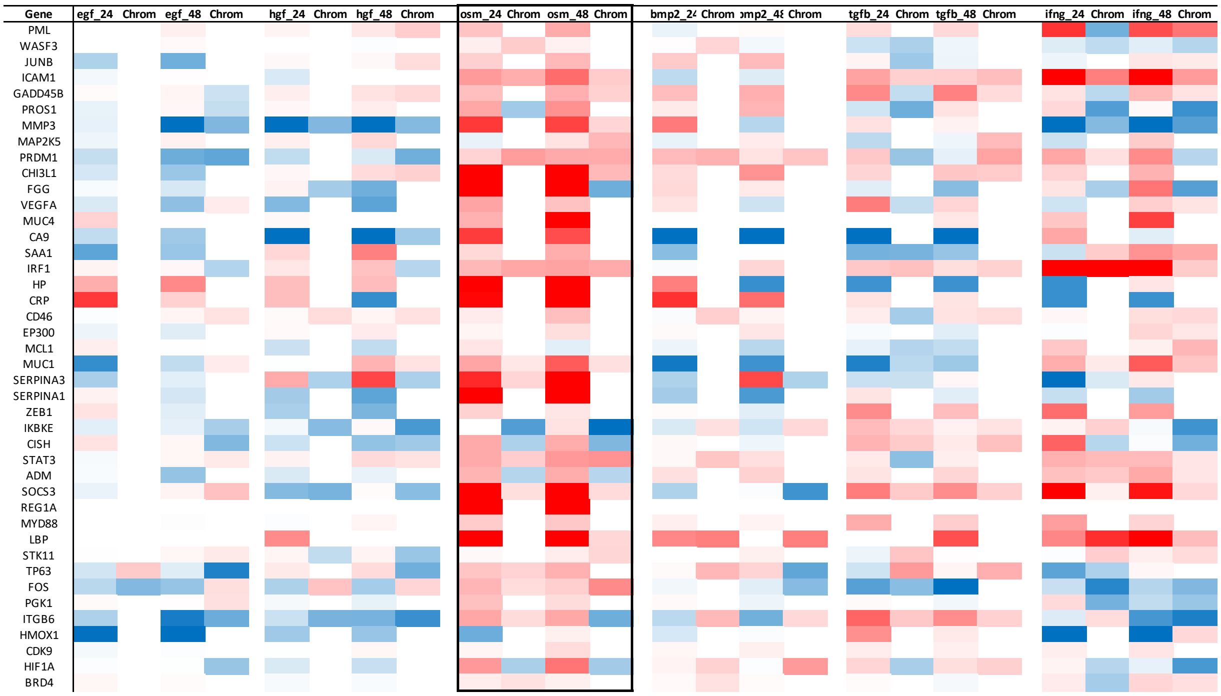


**Figure 4.16. Increased gene expressions of STAT3 target DE genes on OSM.**
Depicting  upregulations of STAT3 target DE genes in OSM (red). BRD4 is upregulated in OSM and its target HMOX1 is downregulated in OSM (blue).

## 4.7 Tables

**Table 4.1: Number of differential proteins and genes with respect to PBS for each time point.**

(A) Number of DE proteins in proteomic data and (B) number of DE genes in transcriptomic data.

(A)

RPPA (P-value <0.05, FC>=1.2)

| Condition | 1 hr | 4 hr | 8 hr | 24 hr | 48 hr |
|---|---|---|---|---|---|
| EGF | 61 | 88 | 94 | 128 | 148 |
| HGF | 51 | 54 | 38 | 212 | 95 |
| OSM | 78 | 80 | 80 | 76 | 121 |
| BMP2 | 60 | 87 | 95 | 112 | 149 |
| TGFB | 71 | 76 | 98 | 143 | 175 |
| IFNG | 62 | 89 | 89 | 137 | 140 |

(B)

RNASeq (P-value <0.01, FC >=log2 of 1.5)

| Condition | 24 hr | 48 hr |
|---|---|---|
| EGF | 505 | 834 |
| HGF | 341 | 1596 |
| OSM | 1628 | 2344 |
| BMP2 | 1017 | 1520 |
| TGFB | 2076 | 1821 |
| IFNG | 2378 | 2842 |

## 4.8 Acknowledgements

Chapter 4, is currently being prepared for submission for publication of the material. Shamim Mollah and Shankar Subramaniam. The dissertation author was a primary investigator and author of this paper.

## 4.9 References

Dale, Jason, Gordon Cain, and Brad Zell. 2009. "Accelerating Image Processing Algorithms Using Graphics Processors." *Proceedings of SPIE*. https://doi.org/10.1117/12.819156.

"DiffBind." n.d. Bioconductor. Accessed August 1, 2019. http://bioconductor.org/packages/DiffBind/.

"edgeR." n.d. Bioconductor. Accessed August 1, 2019. http://bioconductor.org/packages/edgeR/.

"Homer Software and Data Download." n.d. Accessed August 1, 2019. http://homer.ucsd.edu/homer/.

Hou, Tieying, Sutapa Ray, and Allan R. Brasier. 2007. "The Functional Role of an Interleukin 6-Inducible CDK9.STAT3 Complex in Human Gamma-Fibrinogen Gene Expression." *The Journal of Biological Chemistry* 282 (51): 37091–102.

Huang, Da Wei, Brad T. Sherman, and Richard A. Lempicki. 2009. "Systematic and Integrative Analysis of Large Gene Lists Using DAVID Bioinformatics Resources." *Nature Protocols*. https://doi.org/10.1038/nprot.2008.211.

Hu, Jun, Huanying Ge, Matt Newman, and Kejun Liu. 2012. "OSA: A Fast and Accurate Alignment Tool for RNA-Seq." *Bioinformatics*. https://doi.org/10.1093/bioinformatics/bts294.

Robinson, Mark D., and Alicia Oshlack. 2010. "A Scaling Normalization Method for Differential Expression Analysis of RNA-Seq Data." *Genome Biology* 11 (3): R25.

"TxDb.Hsapiens.UCSC.hg38.knownGene." n.d. Bioconductor. Accessed August 1, 2019. http://bioconductor.org/packages/TxDb.Hsapiens.UCSC.hg38.knownGene/.

# Conclusions

In this dissertation, we have looked into the effects of various drugs and growth factors on chromatin alterations that influence cellular phenotypes of cancer and healthy cells. We explored two major biological questions that guided our specific projects:  1) What are the mechanisms of action of drugs on chromatin alterations that can be exploited to develop chromatin-based targeted cancer therapy, 2) How do various growth factors exert their regulatory influences on healthy cells to change their cellular phenotypes.

First and second projects address the first question, in the first project, we applied latent space models to develop an integrative framework for reconstructing a 3D phosphoprotein-histone-drug network (iPhDnet) from large-scale multi-experiment multivariate high-throughput data sets.  We used an unsupervised clustering technique to generate histone signatures. We then used a  multivariate linear-model and statistical hypothesis testing (t-test) of the coefficients of the model to find significant or potentially causal connections between the histone signatures and phosphoprotein signaling networks. The iPhDnet represents our global chromatin fingerprints that can serve as a signature for assessing the efficacy of a given drug in treating cancer.

In the second project, we used protein-protein interaction database and statistical hypothesis testing (one-way ANOVA) of the iPhDnet to further reconstruct a mechanistic causal network connection among the histone signatures and the phosphoprotein signaling networks at various time points. Our mechanistic causal network implicates, CDK inhibitors flavopiridol and dinaciclib as potential therapeutics mediated by BRD4, NSD3, EZH2 and MYC targeting H3K27me3K36me3 status change in breast cancer.

To address the second biological question, in the third project, we characterized the effects of six growth factors on normal breast cells to various cellular phenotypes. We used an integrative framework and statistical hypothesis testing (t-test) of measure genes, proteins, and chromatin accessibility readouts from large-scale multi-experiment multivariate high-throughput data sets to find canonical footprints of each ligand. Our results show possible STAT3 induction promoting BRD4 to regulate cell proliferation when treated with OSM.

## Limitations and Future Directions

This dissertation systematically assesses and generates plausible hypotheses for mechanisms of actions of specific therapeutic and growth agents on chromatin remodeling influencing various cellular phenotypes. While this effort represents a significant step towards the understanding of chromatin remodeling of breast cancer therapy, the approaches described in this dissertation also have significant limitations and require experimental validations. In the future, we can introduce a machine learning approach to generate scores for characterizing growth factors responses to specific phenotypes. We believe iPhDnet can serve as a valuable method for integrative analysis of multi-omics datasets. The framework can also be extended to accommodate other types of omics data, for instance, metabolomic, cycIF, scRNAseq, and Hi-C, data. In the future, iPhDnet can be extended to incorporate nonlinear models in the framework to capture the dynamics of cellular mechanisms more accurately. The observations made by applying iPhDnet to analyze omics data have implications for the discovery of cancer therapeutics, suggesting mechanistically-associated targets.