The use of phosphoproteomic data to identify altered kinases and signaling pathways in cancer

By

Sara Renee Savage

Thesis

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

MASTER OF SCIENCE

in

Biomedical Informatics

August 10, 2018

Nashville, Tennessee

Approved:

Bing Zhang, Ph.D.

Carlos Lopez, Ph.D.

Qi Liu, Ph.D.

ACKNOWLEDGEMENTS

The work presented in this thesis would not have been possible without the funding provided by the NLM training grant (T15-LM007450) and the support of the Biomedical Informatics department at Vanderbilt. I am particularly indebted to Rischelle Jenkins, who helped me solve all administrative issues. Furthermore, this work is the result of a collaboration between all members of the Zhang lab and the larger CPTAC consortium. I would like to thank the other CPTAC centers for processing the data, and Chen Huang and Suhas Vasaikar in the Zhang lab for analyzing the colon cancer copy number and proteomic data, respectively. All members of the Zhang lab have been extremely helpful in answering any questions I had and offering suggestions on my work. Finally, I would like to acknowledge my mentor, Bing Zhang. I am extremely grateful for his guidance and for giving me the opportunity to work on these projects.

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ANN	artificial neural network
AUC	area under the curve
CGC	Cancer Gene Census
CPTAC	Clinical Proteomic Tumor Analysis Consortium
DEPOD	DEPhOsphorylation Database
DL	downloadable
EKPD	Eukaryotic Protein Kinase & Protein Phosphatase Database
EMT	epithelial-to-mesenchymal transition
ERα	estrogen receptor alpha
ERβ	estrogen receptor beta
FDR	false discovery rate
GO	Gene Ontology
BP	Biological Process
GDSC	Genomics of Drug Sensitivity in Cancer
GSEA	gene set enrichment analysis
HINT	High-quality INTeractomes
HMM	hidden Markov model
HPA	Human Protein Atlas
HPRD	the Human Protein Reference Database
HT	high-throughput
ICGC	International Cancer Genome Consortium
IMAC	immobilized metal affinity chromatography
iTRAQ	isobaric tags for relative and absolute quantification
KEA2	Kinase Enrichment Analysis 2
KSD	Kinase Sequence Database
LC	liquid chromatography
LKB1	liver kinase B1
lncRNA	long non-coding RNA
LT	low-throughput
MAP4	microtubule associated protein 4
MAPK	mitogen activated protein kinase
MCLP	MD Anderson Cell Lines Project
MELK	maternal embryonic leucine zipper kinase
METABRIC	Molecular Taxonomy of Breast Cancer International Consortium
MIMP	Mutations Impact on Phosphorylation
MOAC	metal oxide affinity chromatography
MS	mass spectrometry
MS/MS	tandem mass spectrometry
NES	normalized enrichment score

PCA	principal component analysis
PDK1	pyruvate dehydrogenase kinase
PDPK1	3-phosphoinositide-dependent protein kinase 1
PEG3	paternally expressed gene 3
PI3K	phosphatidylinositol-3-kinase
PPI	protein-protein interaction
PSSM	position specific scoring matrices
PTM	post-translational modifications
RB1	retinoblastoma 1
ROC	receiver operating characteristic
RPPA	reverse phase protein array
RWR	random walk with restart
SAMHD1	sterile alpha motif and HD domain containing protein 1
SNV	single nucleotide variation
ssGSEA	single sample gene set enrichment analysis
SVM	support vector machine
TCGA	The Cancer Genome Atlas
TMT	tandem mass tag
TOP2A	topoisomerase II alpha
TUBA1B	tubulin alpha-1B chain
UNSP	unspecified

CHAPTER 1

INTRODUCTION

Dysregulation of Cellular Components in Cancer

Despite more than 2000 years of research on cancer, we still do not fully understand the disease[1]. Cancer is a complicated process that can occur in nearly every cell type and is characterized by abnormal cell proliferation. It is the second leading cause of death in the United States; more than half a million people die each year due to cancer[2].

Cancer is a genetic disease driven by changes in DNA[3]. Genetic mutations include single nucleotide polymorphisms, small insertions or deletions, chromosomal translocations, and copy-number alterations of large regions. Some of these mutations might occur in the germline, while the majority are somatic. Germline mutations, which can be inherited and passed on to future generations, appear in most cells of the body. Hereditary cancer accounts for 5 to 10% of all cancer diagnoses[4]. Somatic mutations, however, cannot be inherited and usually occur in a small population of cells. Together, genetic aberrations cause changes in other cell components, leading to pathology.

The first step in cell physiology is the transcription of genes to RNA. Transcriptional regulation is itself a complicated process involving specific DNA promoter sequences, transcription factors, and regulatory proteins and RNAs. Therefore, dysregulation of transcription is one of the contributing factors to cancer pathology. Both the expression and regulation of genes can be affected[5]. For example, MYC is a transcription factor frequently mutated in cancer and these mutations often affect its function or prevent its degradation[6]. Furthermore, overexpression of MYC causes changes in the expression levels of numerous other genes and results in tumor growth[7]. In T-cell acute lymphoblastic leukemia, chromosomal translocation moves a set of genes closer to a DNA regulatory element, leading to increased expression[8]. In addition to coding RNAs, non-coding RNAs can be dysregulated in cancer. miRNAs are small non-coding RNAs responsible for finely regulating the expression of protein-coding mRNA. For example, amplification of the miR-17-92 cluster of miRNAs allows for increased cell growth in lymphoma by inhibiting tumor suppressor translation[9]. Furthermore, long non-coding RNAs (lncRNAs) regulate gene expression through various mechanisms. In breast cancer, amplification of the lncRNA HOTAIR reduces tumor suppressor expression[10].

After RNA is transcribed, it is further translated into proteins. Protein translation is another point of dysregulation in cancer. First, abnormalities from previous steps can be transferred to proteins. Changes in mRNA expression usually lead to corresponding changes in protein levels. DNA mutations can be translated into the protein sequence, leading to changes in protein structure and function. For example, single missense mutations in the TP53 transcription factor decrease its ability to bind DNA[11]. Additionally, chromosomal translocation events can create new fusion proteins. A fusion of two kinases, BCR and ABL, forms after a translocation event between chromosomes 9 and 22. The fusion affects regulation of these kinases, leading to a constitutively active enzyme[12]. Finally, the actual process of translation might contribute to cancer

development. The ribosome might have altered structure or function and altered signaling might affect the rate of translation[13].

Finally, all components of the cell can be further regulated by the addition of other molecules. DNA methylation influences gene transcription. Hypermethylation of some promoters in colon cancer inhibits tumor suppressor expression[14]. Proteins are also modified in several ways. Glycosylation, the addition of a carbohydrate to a protein, is altered in cancer. Tumors have overall changes in glycosylation compared to normal[15]. One effect is the disruption of E-cadherin junctions that contributes to metastasis[16]. Similarly, phosphorylation, the addition of a phosphate group, is dysregulated in cancer. Tumor cells have a different phosphorylation pattern than normal cells[17]. Mutations can also create or destroy phosphorylation sites, leading to altered signaling[18].

The altered cellular components are intricately connected and combine to produce cell phenotypes. While elucidating the effect of a single mutation on an individual cell is relatively easy, determining the result of the interplay between several mutations is much harder. Average solid tumors contain between 33 and 66 genes with single-base substitutions or small insertions or deletions[19]. They also have dozens of chromosomal translocations, as well as amplification of 17% of the genome and deletion of a further 16%[20,21]. Only a fraction of these mutations drive disease pathology, while the remaining are passenger mutations that do not contribute to tumor growth. Determining the driver mutations and how they contribute to cancer development and progression is essential for effective treatment. Studying these data together produce a much better overall picture of the cancer.

Multi-Omics Integration in Cancer

The development of high-throughput methods has allowed the collection of many data types, resulting in compilations of large amounts of information about specific tumors and cell lines. Each cellular component can be assayed. The whole genome can be sequenced to find mutations (genomics). RNA sequencing determines expression of coding and non-coding transcripts (transcriptomics). Copy number alterations are often identified by gene arrays. Methylation is further assayed using tiling microarrays or BeadChips. Protein levels and post-translational modifications are evaluated using mass spectrometry or protein arrays (proteomics). Finally, clinical data can be collected for each patient and phenotype data can be collected for each cell line.

There are a few large-scale studies collecting and integrating multi-omics information on cancer patients. The Cancer Genome Atlas (TCGA) was a large program that collected clinical, genomic, transcriptomic, and proteomic data for over 11,000 patients and 30 tumor types[21]. The Clinical Proteomic Tumor Analysis Consortium (CPTAC) extended this data to include mass spectrometry proteomic and phosphoproteomic data for three of the cancer types[22]. The Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) collected similar data for over 2000 breast cancer patients[23]. Finally, the International Cancer Genome Consortium (ICGC) also collected data from several different countries on 21 tumor types[24]. Much work has been done creating bioinformatics tools and resources to integrate and analyze these diverse data types.

Kinases and Phosphatases in Cancer

While many aspects of the entire cell system are involved in cancer, kinase signaling is an important component. Kinases, the second largest protein class, are enzymes that catalyze the transfer of a phosphate group from energy molecules to other substrates. The majority of kinases phosphorylate proteins, but other substrates include carbohydrates, amino acids, and lipids[25]. Phosphatases are the enzymes that reverse this process. This signaling process contributes to all of the essential cell pathways that are also hallmarks of cancer including metabolism, motility, proliferation, and the DNA damage response.

These two enzyme groups, along with their substrates, are known to be dysregulated in cancer. For example, PTEN, a phosphatase and known tumor suppressor, is frequently mutated in cancer. The mutations inactivate PTEN, leading to constitutive activation of AKT kinase signaling[26]. The phosphatidylinositol-3-kinase (PI3K)/AKT pathway regulates cell survival in response to stress and promotes cell proliferation and migration. The gene for the driving kinase of this pathway, PIK3CA, is also one of the most frequently mutated genes in cancer[27]. The activating mutations on PIK3CA increase proliferation and enhance AKT signaling[28]. Another kinase signaling pathway frequently dysregulated in cancer is the mitogen activated protein kinase (MAPK) pathway. The MAPK pathway is involved in many cell processes and can function to suppress or support tumor growth based on context[29]. In cancer, abnormal overexpression of the upstream receptor tyrosine kinases, SRC, or RAS upregulates MAPK signaling to increase proliferation and invasion[30].

Because kinases are such important targets, almost 40 kinase inhibitors have been approved by the FDA and are frequently used as anti-cancer agents[31]. Furthermore, some kinase inhibitors can be used for very specific mutations in personalized medicine. For example, vemurafenib specifically targets the kinase BRAF with the V600 mutation[32]. Imatinib targets the fusion kinase BCR-ABL, although it also inhibits the kinases c-KIT and PDGFR[33]. While kinase inhibitors have been instrumental in treating some patients, there are some downsides to using them. Because the active domains of many kinases are very similar, inhibitors frequently target many kinases[34]. Resistance to the drugs develops rapidly because of redundancy in the kinase signaling pathways[35]. Finally, toxicity to normal cells causes severe side effects for some inhibitors[36].

The Use of Phosphoproteomic Data to Study Kinase Signaling

Understanding which kinases are dysregulated in a patient and prioritizing kinases for drug development requires studying kinase signaling at a systems level. In the past, global kinase signaling dysregulation was primarily studied at the genomic level with determination of activating or inhibiting mutations. Individual kinases are studied in individual patients or cancer cell lines using molecular biology techniques. However, a method has recently been developed to study kinase signaling at the systems level for numerous cancer patients at once. Phosphorylated peptides in samples are enriched by affinity purification, optionally labeled for quantification, and identified by mass spectrometry. Because phosphorylation is the direct result of net kinase and phosphatase activity, it can be used as a read-out of active enzymes and their downstream pathways.

There are, however, numerous challenges with phosphoproteomic experiments. First, sample preparation and enrichment strategies can bias results. For example, the use of immobilized metal affinity chromatography to enrich phosphorylated peptides produces a higher proportion of multiply phosphorylated peptides than the use of metal oxide affinity chromatography[37]. Furthermore, phosphorylation is dynamic and unstable. When tumor samples are collected for phosphoproteomic analysis, cold ischemia time affects the results. Mertins et al found 24% of the phosphoproteome is regulated by cold ischemia, which complicates comparisons between patients if sample collection is not tightly controlled[38]. Finally, identification of the protein and the exact location of the phosphorylated residue is still a challenge[39]. Peptides can map to multiple proteins, which can affect site quantification. Additionally, spectra for phosphorylated sites within close proximity are hard to differentiate. In a study of 22 research groups analyzing the same phosphoproteomic data, the groups did not agree on site localization for 21% of the spectra[40].

Because mass spectrometry phosphoproteomics is a relatively new but promising technique, much work still needs to be done to understand its limitations and determine the best and most biologically relevant methods for analysis. This thesis aims to characterize the limitations of mass spectrometry in studying kinase signaling and demonstrate its utility in identifying new insights in cancer that cannot be determined using other methods.

CHAPTER 2

OVERVIEW OF RESOURCES FOR STUDYING KINASES, PHOSPHATASES, AND PHOSPHORYLATION SITES: TOOLS FOR ANALYZING PHOSPHOPROTEOMIC DATA

Introduction

Post-translational modifications (PTMs) are an essential aspect of cell signaling. The enzymatic addition or removal of a molecule or chemical to a protein allows a cell to rapidly and reversibly respond to environmental stimuli. There are many types of PTMs, but phosphorylation, the method of transferring a phosphate group to a substrate, is the best-studied and is common in mammalian cells. Cells use phosphorylation as a molecular switch and carefully regulate the process. Furthermore, phosphorylation dysregulation has been hypothesized to contribute to diseases such as cancer[17]. In mammalian cells, proteins are phosphorylated primarily on serine, threonine, and tyrosine amino acid residues. However, phosphorylation occasionally occurs on other residues including histidine, as well as on non-protein substrates[41].

Kinases are enzymes responsible for catalyzing the transfer of the phosphate group to a substrate. There are over 500 human kinases, but the actual number varies among research groups due to differences in functional prediction and annotation of pseudogenes[25,42]. Phosphatases reverse the action of kinases by catalyzing the removal of the phosphate group. There are almost 200 human phosphatase genes and, like kinases, most phosphatases dephosphorylate protein substrates[43].

Phosphoproteomic data have emerged as a mechanism to study kinase signaling. To analyze this data, resources such as knowledge bases of kinases, phosphatases, and phosphorylation sites are required. Additionally, tools for prediction, visualization, and analysis are instrumental in understanding the data.

While many tools have overlapping functions, they differ in underlying knowledge bases, algorithms, input and output format and data, accessibility, advantages, limitations, and maintenance. Additionally, a newly developed tool is usually compared to a similar, previously published tool, but comparisons often do not include real-world, biological use-cases. For example, the inference of kinase activity is a popular use for phosphoproteomic data. Methods used to infer activity include using permutation to determine non-uniform distribution of substrates, comparing mean phosphorylation of substrates using a Z-test, and assessing phosphorylation levels of regulatory sites on kinases[44,45]. There has been little validation of the methods and only one benchmarking paper study comparing a few of the methods has been published[44].

Finally, the targeted audience for many tools consists of biologists without computational backgrounds. However, biologists are rarely consulted for design input and never requested to test the final product. There is no comprehensive list of tools to aid those using phosphoproteomic data in their research. Therefore, this chapter aims to collect tools and resources that can be used to analyze phosphoproteomic data, perform some benchmarking comparisons to determine the best tool available, and assess usability of the tools from the standpoint of a user.

Methods

Collection of Tools

The OMICtools resource (https://omictools.com) is a manually curated collection of bioinformatics tools[46]. This site was searched in October 2017 for tools using the words 'kinase', 'phosphorylation', 'phospho', or 'phosphatase'. In addition, several more tools were collected from the literature. Only tools that were freely available, still accessible, and non-obsolete were included. Tools specific for organisms other than human were discarded. The year of last update was assumed to be the year of publication unless otherwise noted on the website. The method of access can be by a website (Web) or by a downloadable, locally-run tool (Tool). DL indicates database information or tool results could be downloaded. The website URLs for all resources can be found in **Appendix B**.

Testing Knowledge Bases

Each website was accessed in October 2017. A protein was submitted to the search function and the links provided in the results were tested. Data statistics were collected for human proteins from downloadable files where possible and from websites or manuscripts for online-only resources.

Identifying the Human Kinome and Phosphatome

Human protein kinases were downloaded from KinBase[25] (http://kinase.com/web/current/) and non-protein kinases annotated with the term KW-0418 were downloaded from Swiss-Prot[47] (http://www.uniprot.org) in January 2017. Human phosphatases, excluding inactive phosphatases and those with the designation 'pseudophosphatase,' were downloaded in April 2017 from the DEPhOsphorylation Database[48] (DEPOD, http://depod.bioss.uni-freiburg.de/) and Phosphatome.Net[43]. All gene identifiers (HGNC, UniProt) were updated to the June 2017 versions.

Testing Kinase-Substrate Prediction Tools

Tools predicting kinases for phosphorylation sites were accessed through local tool installation or through the tool's website. PhoScan[49] and phos_pred[50] were run locally on a Windows laptop, while NetPhorest[51], NetworKIN[52], iGPS[53], GPS[54], PhosphoPredict[55], and MusiteDeep[56] were run locally on a Mac laptop. PhosphoPICK[57], NetPhos[58], Musite[59], and pkaPS[60] were accessed via their websites. Gold standard positive and negative phosphorylation sites for five kinases (CDK1, CK2, MAPK1, PKA, and PKC) were downloaded from dbPTM[61]. Positive sites were phosphorylation sites experimentally validated to be phosphorylated by a particular kinase. Negative sites were phosphorylatable residues not known to be phosphorylated on the same proteins. Only human serine and threonine sites were used. Tools were set with the lowest threshold if they did not have an option to return scores for all sites. For each site, the maximum score was retained if the tool predicted for more than one related kinase (e.g., the maximum score of PKCalpha and PKCbeta on the same site). If a tool did not return a score for a site, the lowest possible score was given to the site. The receiver operating characteristic (ROC) curve and area under the curve (AUC) were calculated for the results from each tool using the R package ROCR[62].

Testing Kinase Activity Prediction Tools

A phosphoproteomic dataset from a cell line experiment with 20 kinase inhibitors was used to test kinase activity prediction tools[63]. The R programming environment was used to create files in the input format for each tool. Significantly downregulated sites for each inhibitor were submitted to KEA2[64] and significantly inhibited kinases were defined as those with false discovery rate (FDR) < 0.05 and at least 3 overlapping substrates. The log2 fold change for each thirteenmer phosphorylation site (±6 amino acids surrounding the phosphorylated site) was submitted to PHOXTRACK[65] (1000 permutations, minimum number of substrates = 3, weighted statistics). Significantly inhibited kinases were defined as those with FDR < 0.05 and normalized enrichment value < 0. The fold change for each site with each inhibitor was submitted to the KSEA app website[45] and significantly inhibited kinases were defined as those with FDR < 0.05, at least 3 substrates in the dataset, and a z score < 0. The substrates of kinases from PhosphoSitePlus[66] (version July 2017) and Signor[67] (version October 2017) were used for IKAP[68]. IKAP was run locally on a Mac laptop with the bounds between -11 and 11 and 50 iterations. The 5 kinases with the lowest activity scores for each experiment were chosen. The positive set were kinases known to be inhibited by each drug (supplementary table in reference 63); all other kinases were considered to be negative. The significant kinases for each tool were counted for presence in the positive and negative sets.

Creating Gold Standard Sets for Testing Kinase Activity Inference

Twenty-five known kinase targets of the 20 inhibitors used in the Wilkes' experiment with at least six substrates in the data were chosen for analysis. The gold standard positive set was the 60 known inhibitor-target pairs. The 20 inhibitors were then paired with each of the twenty-five kinases that were not known targets of that inhibitor. The gold standard negative sets were created by randomly selecting 60 inhibitor-non-target pairs 20 different times.

Method Comparison for Kinase Activity Inference

Pre-ranked gene set enrichment analysis (GSEA) was implemented using WebGestaltR[69]. The z score method from the KSEA app was also implemented in R. Briefly, the mean log2 fold change for all sites was subtracted from the mean log2 fold change for substrates in the set. This difference was multiplied by the square root of the number of sites in the substrate set and divided by the standard deviation across all sites in the dataset. Phosphorylation sites were defined as the thirteenmer peptide and the median log2 fold change was calculated for multiple peptides referring to the same site. Activity scores were calculated as the signed -log10(FDR). The sign followed the normalized enrichment score (NES) from GSEA or the z score from the z score method.

Kinase Activity Inference Using Substrate Sets with Different Experimental Evidence

Experimentally validated substrate sets were created for the 25 kinases from PhosphoSitePlus (version May 2018) annotated as *in vivo* or *in vitro* experimental evidence. *In silico* substrate sets were generated from NetworKIN using the NetworKIN score ≥ 2. Analysis was limited to the 16 kinases with at least 3 substrates in each set.

Kinase Activity Inference Using Substrate Sets from Different Databases

Substrate sets were created for the 25 kinases from 5 different databases. Thirteenmer sequences for substrates from each database were created by mapping to their respective protein sequences. Substrates for PhosphoSitePlus (version May 2018) were combined with those from Signor (version May 2018), Swiss-Prot (version May 2018), the Human Protein Reference Database[70] (HPRD, version 9), and Phospho.ELM[71] (version 9.0). Two additional sets of a combination of PhosphoSitePlus + HPRD + Swiss-Prot and a combination of all five databases were created. Analysis was limited to the 23 kinases with at least 3 substrates in every database combination.

AUC for Kinase Activity Inference

AUC was calculated using the ROCR R package for each of the 20 negative sets with the positive set. AUC values within a method were compared using ANOVA with a Tukey's post hoc analysis for differences between pairs. AUC values between methods were compared by t-test. Significance was defined as $p < 0.05$.

Results

Knowledge Bases of Kinases and Phosphatases

Knowledge bases for kinase signaling can be separated into those collecting information on the enzymes, and those collecting experimentally validated phosphorylation sites. Of the 15 different resources that collect information specifically on protein kinases and phosphatases, 12 provide data on kinases, while 4 provide data on phosphatases (**Table 1**). Only one resource, the Eukaryotic Protein Kinase & Protein Phosphatase Database (EKPD) contains information on both types of enzymes[72]. The databases contain various types of data and some include an option for downloading files, while others are only available as an online website.

The kinase knowledge bases can be further separated into two different types: those that include comprehensive data on all known protein kinases, and those that were developed for a specific purpose, such as collecting driver mutations in kinases. Notably, no kinase resource collects data on non-protein kinases. KinBase, which was developed by Gerard Manning, contains 538 protein kinases and is considered the primary source of human protein kinases and their classification[25]. Many other resources base their kinase list on KinBase.

Kinomer, Kinase Sequence Database (KSD), and KinG are general kinase sequence databases that provide very little other information and are outdated[73–75]. KinMutBase, a collection of disease-causing mutations in protein kinase domains, is also outdated, contains data on only 31 kinases, and primarily consists of broken links[76]. KinWeb and EKPD provide gene and protein identifiers, classification, description, and sequence information, but these data can also be found in other resources. However, KinWeb does have prediction of the disulfite bonding state of cysteines in the protein, as well as prediction of alpha helices, and EKPD presents data in an easy-to-read format[72,77].

Use of the remaining general resources depends on which data you want to access. KinaseNET and ProKinO contain the most comprehensive databases on protein kinases, but both are only available as online resources[78]. They include protein

sequences, links to the kinases in other databases (e.g., UniProt, Ensembl, Entrez), information on the kinase domains, expression in tissue, and disease associations. ProKinO specifically contains pathway information, mutations and their disease associations, chromosomal location of the kinase, and links to published manuscripts. KinaseNET includes PTMs, known binding partners, inhibitors, upstream kinases, downstream substrates, and information about regulation. KinaseNET provides all data on a single page, while ProKinO requires more than 10 clicks on separate tabs and pages to obtain all information on a kinase.

For studying diseases, MOKCa and Kin-Driver specifically have data on protein kinase mutations[79,80]. MOKCa has tissue specificity of mutations while Kin-Driver focuses on driver mutations and reports whether the mutation is activating or inactivating. Finally, KLIFS provides structural information for approximately half of the protein kinases bound to various ligands[81].

Because phosphatases are less well studied than kinases, there are fewer resources dedicated to their collection. EKPD provides the same information for phosphatases as it does for kinases. HuPho, however, was the first comprehensive collection of phosphatases and the database includes pathway and substrate data, as well as siRNA phenotype data and links to orthologs in other species[82]. DEPOD used data from HuPho as a starting point and therefore contains much of the same information[48]. Finally, Phosphatome.Net is the phosphatase version of KinBase[43]. The website contains basic classification and sequence information.

Name	Last Update	Method of Access	Version	Enzyme	Protein Number	Reference
KSD	2002	WebIDL		Protein Kinases	913	74
KinWeb	2005	Web		Protein Kinases	519	77
Kinomer	2008	WebIDL	1	Protein Kinases	505	73
MOKCa	2008	Web		Protein Kinases	423	79
HuPho	2012	WebIDL		Phosphatases	313	82
EKPD	2013	Web	1.1	Protein Kinases and Phosphatases	676	72
KinBase	2014	Web		Protein Kinases	538	25
Kin-Driver	2014	WebIDL		Protein Kinases	518	80
KinG	2014	Web		Protein Kinases	1813*	75
KinMutBase	2015	WebIDL	4	Protein Kinases	31	76
DEPOD	2016	WebIDL	1.1	Phosphatases	239	48
ProKinO	2016	Web	2	Protein Kinases	538	78
KinaseNET	2017	Web		Protein Kinases	>530	
Phosphatome	2017	Web	3	Phosphatases	189	43
KLIFS	2018	WebIDL	2.3	Protein Kinases	285	81,83

Table 1. Knowledge bases of human kinases and phosphatases. *Indicates the inclusion of an unknown number of non-human proteins.

Generating the List of Human Kinases and Phosphatases

After evaluation of the knowledge bases, I collected the human protein kinases from KinBase because it contained the most comprehensive list. Because no resource includes non-protein kinases, these enzymes were collected from Swiss-Prot (http://www.uniprot.org) annotated with the term KW-0418 (‘Kinase’). Notably, other ontologies do not have a term to specify proteins with phosphotransferase activity. For example, the Gene Ontology term ‘protein kinase activity’, GO:0050222, contains 1246

human genes, many of which are regulators of kinase activity and do not contain phosphotransferase activity themselves. The final collection contained 687 unique human kinases. This corresponded to 688 genes, as *CKMT1A* and *CKMT1B* produce the same protein product despite being separate genes. Of these 688 genes, 538 encode protein kinases.

Active human phosphatases were collected from DEPOD. Additional human phosphatases that were not in DEPOD and were not verified pseudophosphatases were added from Phosphatome.Net. This resulted in 255 human proteins annotated as phosphatases.

Knowledge Bases of Phosphorylation Sites

Besides information about specific kinases and phosphatases, data on phosphorylation sites are important for studying the signaling process. Phosphorylation site databases collect information on the location of phosphorylated residues in proteins from experimental data. These experiments can be low-throughput or high-throughput. High-throughput phosphorylation site identifications are assigned by probability unlike the more stringent experimental validation in low-throughput experiments, but some databases combine sites from both types of experiments without identifying the source experiment type.

In addition to phosphorylation site information, about 75% of the 23 resources collect interactions between kinases or phosphatases and their substrates (**Table 2**). These often do not include the exact phosphorylation site, but instead provide interactions between an enzyme and its substrate at the gene level.

Name	Last Update	Method of Access	Version	Sites	Proteins	Kinases	Phosphatases	Data Type	Reference
PhosphoPep	2007	WebIDL	2.0	3,980				MS	84,85
Phospho.ELM	2010	WebIDL	9.0	26,651	5,374	250		HT, LT	71,86,87
Phospho3D	2010	WebIDL	2.0	1,770		59		HT, LT	88
HPRD	2010	WebIDL	9	78,005	11,807	291	42	UNSP	70,89,90
PHOSIDA	2011	WebIDL	3.24	22,806	5,175			MS	91,92
PTMfunc	2012	Web		31,165				MS	93
HuPho	2012	WebIDL		190	121		55	UNSP	82
SubPhosDB	2012	WebIDL	1	137,153	17,297	238		UNSP	94
RegPhos	2013	WebIDL	2.0	66,301	10,849	380		UNSP	95,96
ANIA	2013	WebIDL		305	220			LT	97,98
PhosphoNetworks	2013	WebIDL			1,140	255		UNSP	99
Kinome NetworkX	2014	DL		173,460	18,610	357		UNSP	100
ProteomeScout	2014	WebIDL	2	290,007	23,387			MS	101,102
dbPTM	2015	WebIDL	v2016	76,736	10,648	>360		UNSP	61,103,104
dbPAF	2016	WebIDL	1.0	244,034	18,773			UNSP	105
PhosphoAtlas	2016	DL		2,595	1,284	501		UNSP	17
DEPOD	2016	WebIDL	1.1	253	210		88	UNSP	48
KANPHOS	2016	Web	β			73		MS	106
Phosphatome	2017	Web	3	6,008	2,000	319	106	UNSP	43
PhosphoNET	2017	Web		966,817†	22,698	488		UNSP+pred	107
PhosphoSitePlus	2018	WebIDL	May-18	235,406	20,086	370		HT, LT	66
Swiss-Prot	2018	WebIDL	May-18	40,150	7,971	350		UNSP	47,108
Signor	2018	WebIDL	2.0	4,354	1,605	415	98	UNSP	67,109

Table 2. Databases of phosphorylation sites. The number of unique kinases and phosphatases reported to phosphorylate sites in the database is included. For some databases, these numbers include enzyme groups in addition to individual enzymes. Data type indicates whether the data are from mass spectrometry

(MS) experiments, separated high-throughput (HT) and low-throughput (LT) experiments, or whether the database combines data from both HT and LT experiments without specifying (UNSP). †Indicates inclusion of predicted phosphorylation sites (pred).

The four main resources for phosphorylation sites curated data manually from the literature (**Figure 1**). HPRD and Swiss-Prot are general databases of all proteins[47,70]. The remaining two, PhosphoSitePlus and Phospho.ELM, specifically contain phosphorylation site information[66,71]. Both PhosphoSitePlus and Swiss-Prot are frequently updated, while HPRD and Phospho.ELM were last updated in 2010. All four of these databases also include kinase information for sites if known.

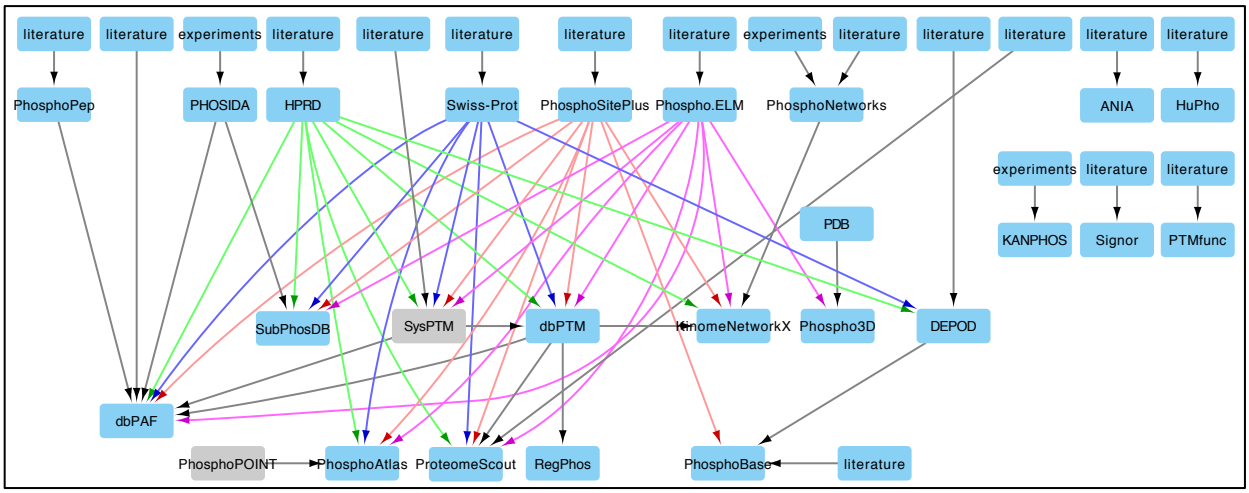


Figure 1. Network of phosphorylation site and kinase-substrate interaction databases. Gray nodes indicate databases that are no longer accessible. Arrows point from the knowledge source to the collecting database. Arrows originating from the four most highly used databases are colored by source (green=HPRD, blue=Swiss-Prot, red=PhosphoSitePlus, purple=Phospho.ELM).

Other smaller databases were generated through manual curation or publication of a laboratory's own phosphorylation site data. KANPHOS collects phosphorylation sites in neural signaling identified by high-throughput experiments[106]. PHOSIDA is another collection of data that were primarily produced in cell lines[91]. PhosphoPEP integrates mass spectrometry experiments from Cell Signaling Technology and their own laboratory[84,85]. PTMfunc collects mass spectrometry experiments and adds functional predictions from various tools for each site[93]. Signor extracts high quality signaling interactions from the literature[67]. Finally, ANIA and PhosphoNetworks curate the literature for a specific purpose. ANIA collects phosphorylation sites that serve as binding sites for 14-3-3 proteins, while PhosphoNetworks creates a kinase-substrate network curated from the literature and a protein microarray experiment[97,99].

The remaining resources integrate phosphorylation sites and kinase information from other databases (**Figure 1**). The database dbPAF collects phosphorylation sites from several databases[105]. ProteomeScout also collects phosphorylation sites from other databases along with literature-curated experiments and provides a tool for analyzing a user's data[101]. The database dbPTM collects all PTMs and the responsible enzyme from several sources[61]. KinomeNetworkX, RegPhos, and PhosphoAtlas curate and integrate

data specifically to create kinase-substrate networks[17,95,100]. PhosphoNET is an online-only tool that includes predicted phosphorylation sites in addition to those with experimental evidence[107]. SubPhosDB annotates phosphorylated proteins with subcellular localization[94]. Finally, Phospho3D specifically collects phosphorylation sites with 3D structures[88].

Five databases collect information on phosphatase-substrate interactions. As mentioned, DEPOD, HuPho, and Phosphatome.Net all curate enzyme interactions from the literature. HPRD and Signor also collect some site-specific phosphatase information.

Each database contains a different number of phosphorylation sites and enzyme-substrate relationships depending on the source and method of collection (**Table 2**). ProteomeScout, PhosphoSitePlus, and dbPAF contain the most number of experimentally validated, downloadable sites. The site numbers for these three databases include specific protein isoforms, as do several other resources. PhosphoAtlas contains substrates for the most number of individual kinases. Signor, Swiss-Prot, RegPhos, Phospho3D, dbPTM, and Phospho.ELM have substrates for individual kinases and kinase families. Finally, PhosphoSitePlus has substrates for some specific kinase isoforms.

Besides the general phosphorylation site databases, other specialized databases have formed using phosphorylated protein information. For example, PepCyber: P~PPep is an online searchable database of proteins that interact with phosphorylated proteins[110].

Errors in Substrate Databases

Based on these databases, PhosphoSitePlus is the best resource for experimentally-identified phosphorylation sites and kinases for phosphorylation sites. PhosphoSitePlus is frequently updated and well-curated. The downstream integrating databases suffer from ID mapping errors. For example, in PhosphoAtlas there is an entry for PEG (paternally expressed gene 3) phosphorylating CDC25B. PEG is not a known kinase, but pEg3 kinase (also known as maternal embryonic leucine zipper kinase, MELK) is known to phosphorylate CDC25B[111]. Many of the downstream databases also have issues with PDPK1 and PDK1. The gene *PDPK1*, 3-phosphoinositide-dependent protein kinase 1, produces a protein known to the biological community as PDK1. However, there is an additional kinase, pyruvate dehydrogenase kinase, that is produced by the gene *PDK1*. Databases that try to integrate sites frequently attribute the substrates of *PDPK1* to *PDK1*. Finally, integrating databases propagate errors from the original databases. For example, HPRD contains an entry for PTPN11 phosphorylating PTK2B although PTPN11 is a known phosphatase and not a kinase. The original manuscript connected to this entry confirmed that PTPN11 is a phosphatase and that it just binds to PTK2B at that particular site[112]. Databases that collect information from HPRD, such as RegPhos and PhosphoAtlas, include this incorrect entry for PTPN11.

Known Substrates of Kinases and Phosphatases

The four main databases together produce 493 substrate sets of individual kinases and kinase families (**Figure 2A**). PhosphoSitePlus contains the most unique sites, while other databases contribute only a few additional sites per kinase. CSNK2A1 has the most number of substrates (584), while over half of the sets contain fewer than 10 substrates.

For substrates of phosphatases, DEPOD, HPRD, and Phosphatome.Net combined produce sets for 83 phosphatases. The most unique information comes from DEPOD and Phosphatome.Net. The number of known sites for each phosphatase is far fewer than that for kinases. PPP2CA has the most substrates (167), while 70% of the phosphatases have fewer than 10 substrates (**Figure 2B**).

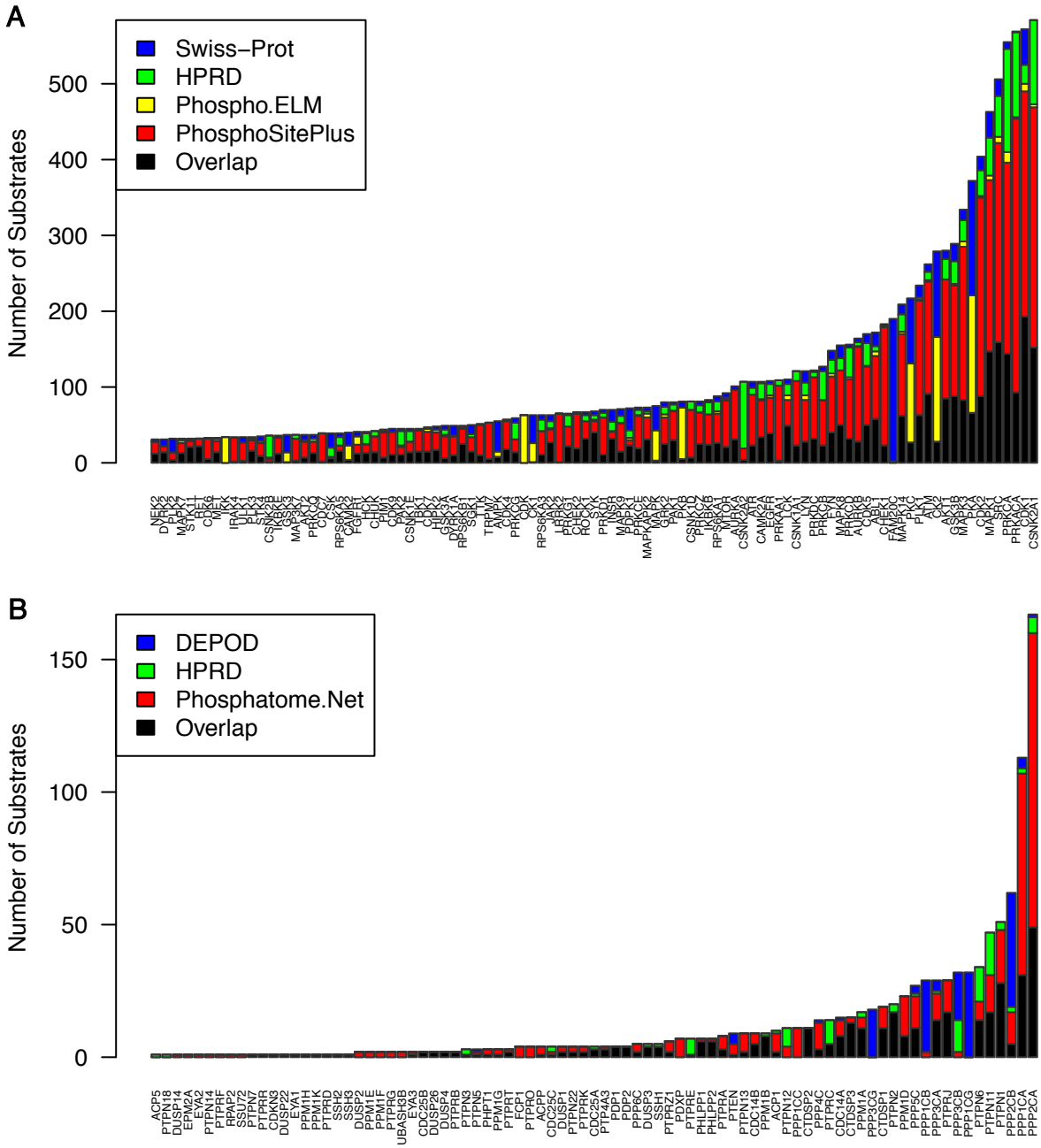


Figure 2. Number of substrates per kinase and phosphatase. A) Number of substrates for the top 100 kinases in four databases. Substrates present in more than one database are colored black while the remaining sites are unique to each database. B) Number of substrates for each phosphatase in DEPOD (blue), HPRD (green), Phosphatome.Net (red), or in more than one database (black).

Phosphorylation Site Prediction Tools

Despite decades of research, very few phosphorylation sites have known kinases or phosphatases. Of the sites in PhosphoSitePlus, only about 3% have an experimentally validated human kinase. Therefore, numerous tools have been developed to predict which sites in a protein can be phosphorylated and which kinases phosphorylate that given site.

These prediction tools were developed using a variety of features and methods and have been reviewed elsewhere[113,114]. The early versions of phosphorylation site predictors were motif-based. They generated the frequency of amino acids surrounding a site and searched for that pattern in protein sequences. Later tools used more sophisticated methods such as support vector machines (SVM), random forest, Bayesian probability, and position specific scoring matrices (PSSM)[50,115–117]. Besides amino acid sequence, tools included a vast array of features such as the 3D structure of the phosphorylation site, disorder score, cell cycle data, and co-expression of kinases and substrates[118–120]. Others, like NetworKIN and iGPS, used protein-protein interaction data to filter predictions[53,121]. **Table 3** provides an overview of all currently available tools to predict phosphorylation sites or kinases for phosphorylation sites. While a few tools have been developed to predict sites for phosphatases, only NetPhorest and NetworKIN are still accessible[121].

Tool	Last Update	Version	Prediction Type	Method	Kinases/ Phosphatases	Type	Reference
Disphos	2004	1.3	phosphorylation sites	bagged logistic regression		Web	118
PPSP	2006	1.06	phosphorylation sites of kinases	Bayesian decision theory	68	Web	116
KinasePhos2.0	2007	2.0	phosphorylation sites of kinases	SVM	58	Web	115
pkaPS	2007		phosphorylation sites of PKA	scoring function	1	WeblDL	60
PhoScan	2008		phosphorylation sites of kinases	scoring function	48	WeblTool	49
Phos3D	2009		phosphorylation sites and some kinase specificity	SVM	5	Web	119
Musite	2010	1	phosphorylation sites and some kinase specificity	SVM	13	WeblDL	59
PHOSIDA Predictor	2011	3.24	phosphorylation S and T sites	SVM		Web	91
Predikin	2011		phosphorylation sites of kinases	PSSM	any	WeblDL	117
GPS-Polo	2012	1.0	phosphorylation sites of Plk	group-based scoring function PSSM	1	WeblTool	122
iGPS	2012	1.0.1	phosphorylation sites of kinases in vivo	GPS with PPI	407	Tool	53
HMMpTM	2013		phosphorylation sites of kinases and topology	HMM	9	WeblDL	123
PKIS	2013		phosphorylation sites of kinases	SVM	56	WeblDL	124
CEASAR	2013		kinases for known phosphorylation sites	naïve Bayes	289	DL	120
GPS	2014	3.0	phosphorylation sites of kinases	group-based scoring function PSSM	464	WeblDLlTool	54
NetPhorest	2014	2.1	phosphorylation sites of kinases	ANN&PSSM	244	WeblDLlTool	51,121
NetworKIN	2014	3.0	phosphorylation sites of kinases in vivo	naïve Bayes with PPI	123	WeblDLlTool	52,121

phos_pred	2014		predicts phosphorylation sites for kinases	random forest	54	Tool†	50
PhosphoSVM	2014		phosphorylation sites	SVM		Web	125
PSEA	2014		phosphorylation sites of kinases	GSEA	33	Web	126
Scansite	2015	4	kinase motifs in proteins	PSSM	70	WeblDL	127
KSP-PUEL	2015		phosphorylation sites of kinases	SVM ensemble	2*	Tool	128
PhosphoPICK	2016		phosphorylation sites of kinases	Bayesian network	107	WeblDL	57
PhosD	2016		kinase-substrate relationships	probabilistic model	399	DL	129
PhosphoNET	2017		phosphorylation sites of kinases	PSSM	488	Web	107
NetPhos	2017	3.1	phosphorylation sites and some kinase specificity	ANN	17	WeblTool†	58,130
PhosphoPredict	2017		phosphorylation sites of kinases	random forest	12	WeblDLlTool	55
MusiteDeep	2017		phosphorylation sites and some kinase specificity	deep learning	5	Tool†	56
PhosPred-RF	2017		phosphorylation sites	random forest		Web	131

Table 3. Available phosphorylation site and kinase-substrate prediction tools. *Indicates number of trained kinases, but tool can be trained with others. †Indicates tool is not available for all three main operating systems (Linux, Mac, Windows). *SVM* – support vector machine, *PSSM* – position specific scoring matrix, *GSEA* – gene set enrichment analysis, *ANN* – artificial neural network, *HMM* – hidden Markov model, *PPI* – protein-protein interaction

Almost all phosphorylation site predictors were trained using data from Phospho.ELM (**Figure 3**). Swiss-Prot and PhosphoSitePlus were also heavily used resources. Notably, almost all tools were developed using experimentally verified substrate data as the training set. Therefore, the tools are only able to predict the responsible kinase if there is existing data for substrates of that kinase.

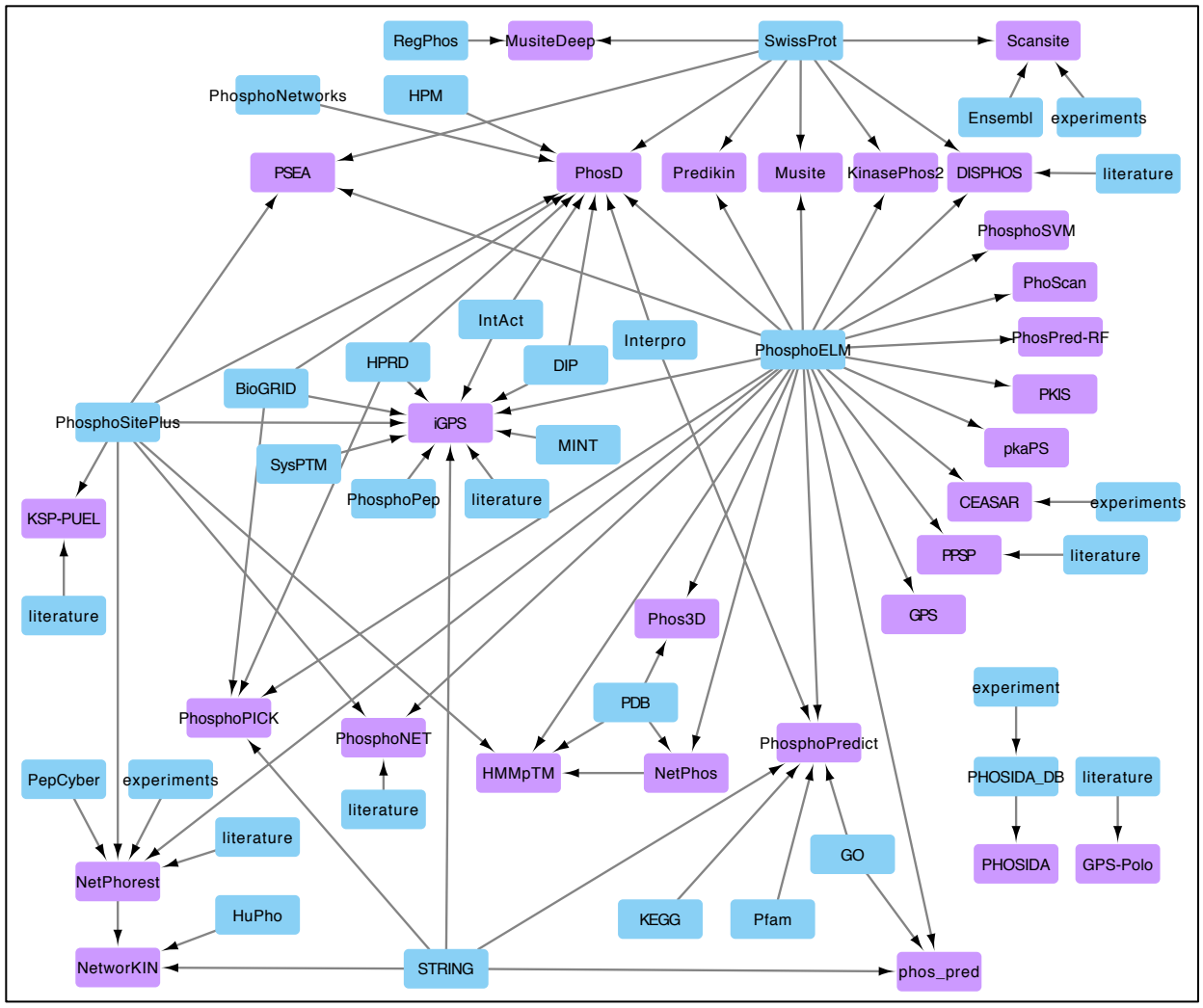


Figure 3. Network of phosphorylation site predictor tools and the resources used to make predictions. Tools are colored purple while the databases used by the tools are colored blue.

The usability of each tool differs based on purpose of use. For possible kinases phosphorylating a single substrate of interest, web-based tools would suffice. However, the limit on the number of sequences submitted for prediction and the lack of downloadable results prevent these same tools for being useful in large-scale studies. Furthermore, downloadable tools are useful for large-scale studies, but tools can be difficult to install and use. For example, phos_pred requires modifying MATLAB code to run. NetPhos is downloadable but can only be run on Linux, while PhoScan can only be run on Windows machines. Finally, tools like GPS and phos_pred provide pre-defined cutoffs for results, while others like musite and KSP-PUEL allow users to define their own thresholds or to train the models using their own data.

For large-scale kinase-substrate prediction, only 12 pre-trained tools were available that provide downloadable results. The best, unbiased way to test these tools is to use validated sites that were not used for the training of any tool. Unfortunately, most tools do not report the actual sites used for training and finding a set of sites to fit these criteria is nearly impossible. Therefore, all 12 tools were tested using gold-standard

positive and negative human sites from dbPTM for five kinases. The outcomes might be slightly biased in favor of newer tools and those that used some of these sites in their training.

ROC curves for four kinases (CDK1, CK2, MAPK1, and PKA) are shown in **Figure 4**, while the fifth curve (PKC) is shown in **Appendix A** (Figure S1). Notably, musite was unable to predict for a few random protein sequences in each submission. MusiteDeep and GPS had the highest area under the curve (AUC) for all kinases tested. The PKA-specific tool pkaPS also performed well. Performance for most tools, however, varied across kinases.

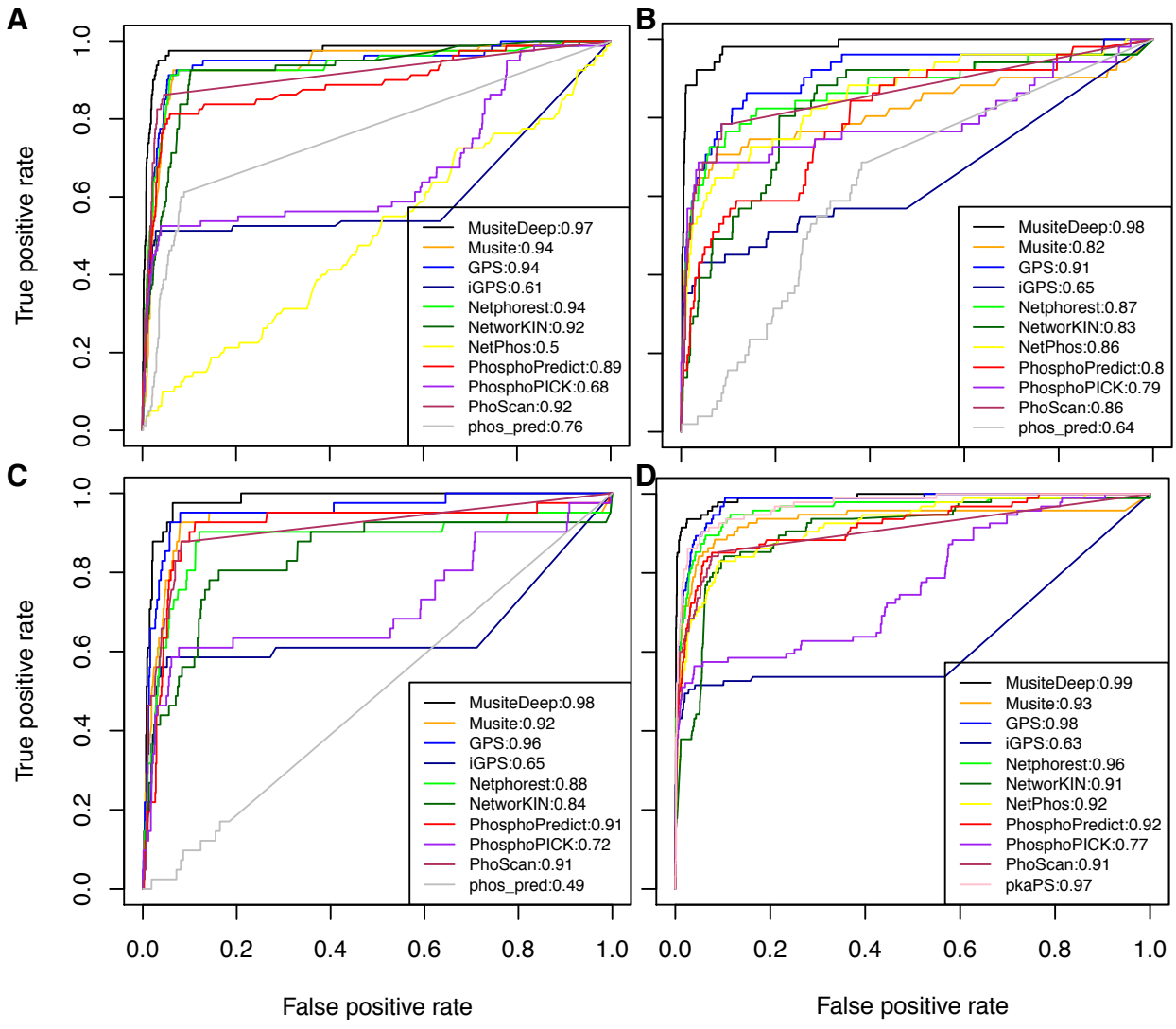


Figure 4. ROC curves for substrate prediction of four kinases. The false positive and true positive rates of substrate prediction for A) CDK1, B) CK2, C) MAPK1, and D) PKA. The AUC for each tool is listed next to the tool name. The tool pkaPS only predicts for PKA, while phos_pred does not predict for PKA and NetPhos does not predict for MAPK1.

Comparison of Kinase Activity Tools

Using known or predicted kinases for phosphorylation sites, kinase activity can be inferred from global phosphoproteomic data. Tools and methods have been developed to predict kinase activity, but there has been little effort spent towards comparing these tools or determining the most biologically-relevant set of parameters. The available tools (PHOSIDA, KEA2, KSEA App, PHOXTRACK, and IKAP) each use a different algorithm to infer activity (**Table 4**). The PHOSIDA *de novo* motif finder uses a simple method of bootstrapping to determine enrichment of sequence motifs in a set of phosphorylated peptides and then matches those to known kinase motifs[91]. Kinase Enrichment Analysis 2 (KEA2) uses over-representation analysis to determine enrichment of kinase substrates in a condition[64]. Similarly, the KSEA App uses mean phosphorylation of substrates of kinases as a proxy for activity[45]. PHOXTRACK modified pre-ranked GSEA to determine enrichment of known kinase targets[65]. IKAP extended these methods using a cost function to infer the relative contributions of multiple kinases acting on the same site[68].

Tool	Last Update	Prediction Type	Method	Input	Type	Reference
PHOSIDA Motif Finder	2011	sequence motifs	bootstrap	phosphosite 13mer	Web	91
KEA2	2012	kinase activity	Fisher's exact test	gene symbols and phosphosite	WeblDLlTool	64
CellNOpt	2013	signaling networks	logic formalisms	phosphoproteomic data	Tool	132
Sorad	2013	time-course analysis	ordinary differential equations	phosphoproteomic data	Tool	133
PHOXTRACK	2014	kinase activity	GSEA	phosphosite 13mer and log2 expression	WeblDL	65
ProteomeScout						101
CLUE	2015	time-course kinase activity	k-means clustering	phosphoproteomic data	Tool	100
PhosFox	2015	phosphorylation site comparison between groups	comparison	phosphoproteomic data	Tool	134
SELPHI	2015	phosphoproteomic data analysis	multiple functions	phosphoproteomic data	WeblDL	135
DynaPho	2016	phosphoproteomic analysis for multiple conditions	activity modules	phosphoproteomic data	WeblDL	136
IKAP	2016	kinase activity	cost function	phosphoproteomic data	Tool	68
KinasePA	2016	kinase perturbation in multiple treatments	directional hypothesis testing framework	phosphoproteomic data	WeblTool	137
KSEA	2017	kinase activity	Z score	phosphoproteomic data	WeblDLlTool	45

Table 4. Kinase activity prediction and phosphoproteomic dataset analysis tools. *GSEA* – gene set enrichment analysis

Comparison of these tools is challenging because they use different input and underlying databases. Because PHOSIDA is only available online without downloadable results, I excluded this tool from further analysis. KEA2 requires a set of sites in the format of HGNC symbol and phosphorylated amino acid residue position separated by an underscore. It contains sets for 250 different kinases. KSEA App requires a strictly

formatted comma-delimited file with the HGNC symbol, phosphorylated position, and non-log-transformed fold change. Users can choose between known sets from the July 2016 release of PhosphoSitePlus or the known + predicted site sets from PhosphoSitePlus and NetworKIN. PHOXTRACK requires a two-column file with a thirteenmer peptide and log-transformed fold change. It can use substrate sets from the four main databases or a user-supplied database. Finally, IKAP required tabular data entered into MATLAB, manual modification of MATLAB code to change parameters, and allowed a user to upload their own set of substrates. Because one thirteenmer might match multiple proteins and phosphorylated positions, the actual substrate list presented to each tool may differ slightly.

I used a phosphoproteomic experiment with 20 kinase inhibitors to compare the kinase activity predictions by KEA2, KSEA app, PHOXTRACK, and IKAP. To determine how well each tool covered the known targets of kinases, I counted the number of significantly downregulated known kinases of each inhibitor and the significantly downregulated kinases of each inhibitor that were not known targets of that inhibitor. The KSEA App made the most true positive predictions across all experiments, while IKAP made the fewest true positive predictions (**Figure 5A**). PHOXTRACK made the fewest false positive predictions (**Figure 5B**).

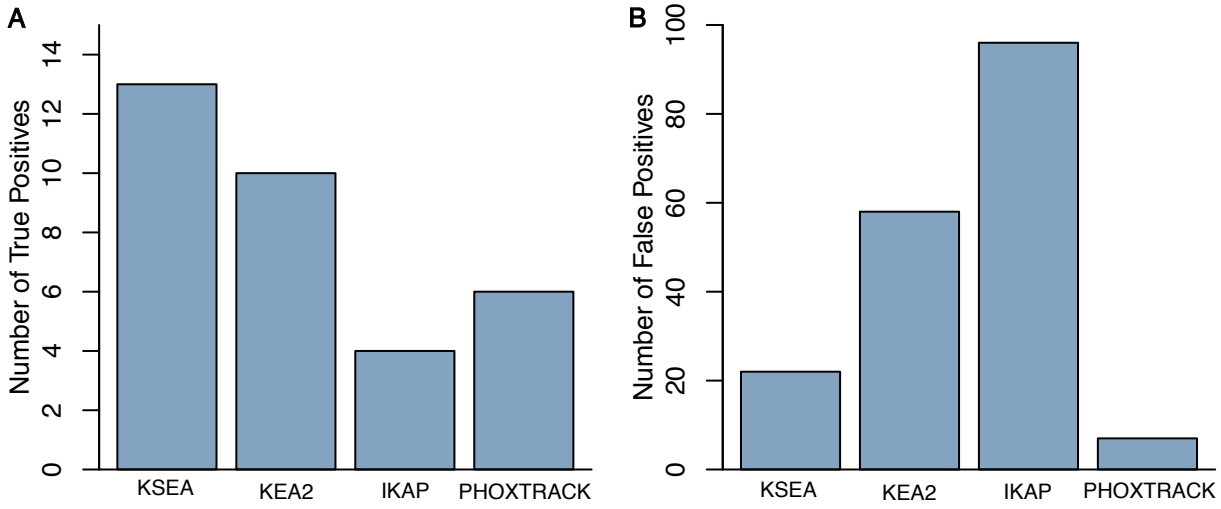


Figure 5. True and false positive predictions for kinase activity prediction tools. A) For all 20 inhibitors, the number of known targets predicted to be significantly downregulated by each tool. B) For all inhibitors, the number of all significantly downregulated kinases that do not match known inhibitor targets.

Comparison of Kinase Activity Inference Methods

Because the KSEA app and PHOXTRACK both maximized true positive results while minimizing false positive results, I performed a systematic comparison of the methods underlying these two tools following a protocol similar to Hernandez-Armenta et al in their benchmarking paper[44]. While kinase prediction tools can increase the number of substrates for each kinase, it is not known how including these substrates affects kinase activity inference. I compared the AUC for the positive set of inhibitor-target pairs and 20 different randomly permuted inhibitor-target pairs for kinase activity inference

using substrate sets from different levels of evidence. The GSEA and z score methods had similar performance, but the GSEA method performed better when the substrate sets came from *in vitro* experiments (**Figure 6A**, p = 0.002). For both methods, using substrate sets from *in silico* predictions performed significantly worse in comparison to using experimentally validated substrate sets (p = 0.0009 for GSEA and p = 0.001 for z score).

As shown above, many databases collect substrates of kinases, so I compared the performance of each database paired with PhosphoSitePlus to maximize the number of kinase targets with substrate sets. The z score method performed better than the GSEA method for all databases (**Figure 6B**, $p < 0.002$). Within each method, performance across most of the databases was similar except for Signor. Substrate sets from the combination of PhosphoSitePlus and Signor performed worse against PhosphoSitePlus alone, Phospho.ELM, and Swiss-Prot using the GSEA method ($p < 0.01$). Substrate sets from the combination of PhosphoSitePlus and Signor performed worse against Swiss-Prot and the combination of Swiss-Prot, PhosphoSitePlus, and HPRD using the z score method ($p < 0.002$). Swiss-Prot performed better than PhosphoSitePlus alone and with Phospho.ELM ($p < 0.04$) using the z score method.

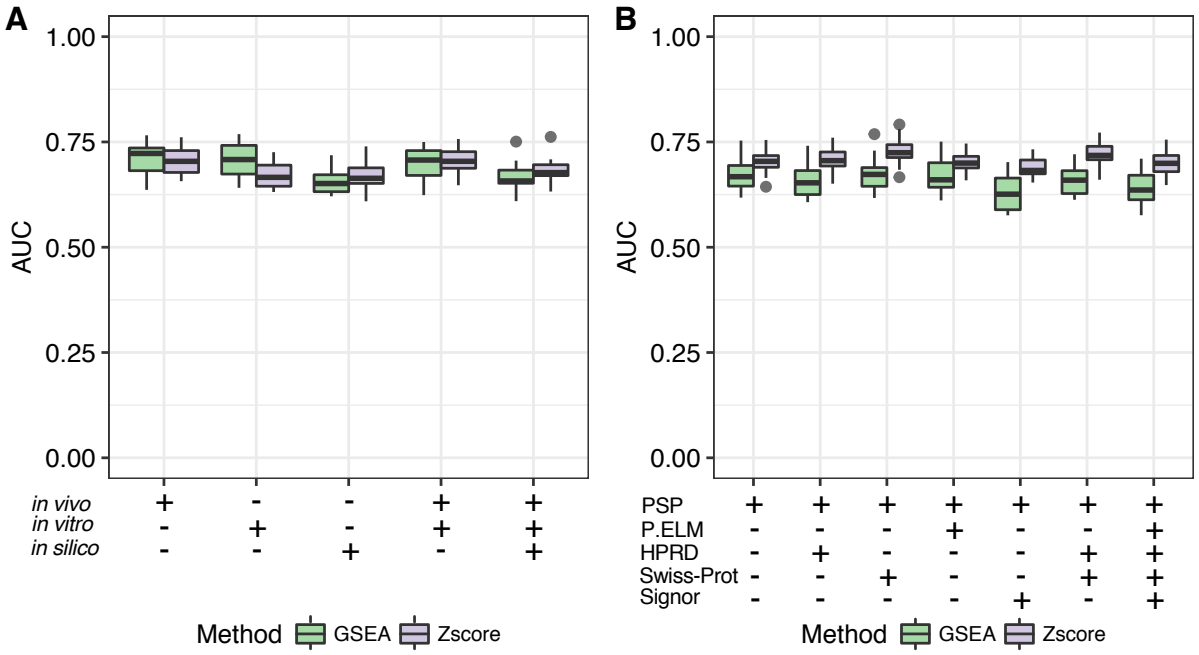


Figure 6. Kinase activity AUC for the GSEA and z score methods using various substrate sets. A) AUC values for 20 randomly permutated negative sets matched with the single positive set using two methods. Substrate sets were generated using substrates with *in vivo*, *in vitro*, and/or *in silico* evidence. B) AUC values for experiments using substrate sets from combinations of databases.

Phosphoproteomic Experiment Analysis Tools

Besides activity prediction, phosphoproteomic data can be used for other analyses. PhosFox compares phosphorylated peptides between conditions[134]. SELPHI allows biologists to quickly and easily analyze phosphoproteomic data with clustering analyses, kinase-substrate correlation, and pathway enrichment[135]. Finally, a set of tools

(CellNOpt, Sorad, CLUE, DynaPho, and KinasePA) were developed specifically for phosphoproteomic time-course or multiple condition analyses (**Table 4**)[100,132,133,136,137].

Prediction of Mutation Effect

Mutations can affect kinase function or presence of a phosphorylation site. PhosSNP is a database of known gene polymorphisms near phosphorylation sites that are categorized based on suspected effect[138]. The remaining four resources are tools to predict the effect of mutations (**Table 5**). Mutations Impact on Phosphorylation (MIMP) uses Bayesian statistics to predict whether mutations around a phosphorylation site will change which kinase binds to that site[139]. It can predict rewiring for 124 kinases using experimentally validated data, or it can be extended to predict for 322 kinases using predicted kinase-substrate relationships. ReKINect also predicts rewiring from mutations, but it further predicts the destruction or creation of phosphorylation sites and inactivation or constitutive activation of kinases[140]. PhosphoPICK-SNP is also similar to MIMP. It predicts the kinase responsible for phosphorylating a site, and whether a mutation affects its ability to phosphorylate the site[141]. Finally, wKinMut2 predicts which mutations on kinases contribute to disease[142].

Tool	Last Update	Version	Prediction Type	Method	Kinases/ Phosphatases	Method of Access	Reference
PhosSNP	2009	1.0	SNVs that might influence phosphorylation status	rules		Tool	138
MIMP	2015		missense SNV impact on kinase-substrate	Bayesian model	322	WebIDLITool	139
ReKINect	2015		effect of SNV on signaling network	PSSM		WebIDL	140
PhosphoPICK-SNP	2016		effect of SNV on phosphorylation level	Bayesian models	107	WebIDL	141
wKinMut2	2016	2	SNV effect on kinase	Random Forest	450	WebIDLITool	142

Table 5. Resources for studying the effect of mutations on kinases and phosphorylation sites. *SNV* – Single Nucleotide Variation, *PSSM* – Position Specific Scoring Matrix

Resources for Kinase Inhibitors

Which small molecules inhibit which kinases is important when considering kinases as therapeutic targets. Most resources connect known drugs to their known kinase targets (**Table 6**). KidFamMap connects kinases, their inhibitors, and associated diseases[143]. DrugKiNET shows the known inhibitors for kinases, and the kinases that a compound inhibits. It also predicts which kinases a drug can inhibit. K-Map extends these interactions to suggest the best compound to inhibit a set of kinases[144]. Finally, KinomeSelector groups kinases by sequence similarity and similarity of drug response. It then allows a user to choose a subset of kinases to target that cover the kinome[121].

Tool	Last Update	Description	Kinases	Inhibitors	Method of Access	Reference
KIDFamMap	2012	kinase-inhibitor interactions	399	35,788	Web	143
K-Map	2013	best inhibitor for a set of kinases	300 or 442	178 or 72	WebIDL	144
KinomeSelector	2014	minimal set of kinases to inhibit	>500	NA	WebIDL	121
DrugKiNET	2017	Known and prediction drug activity on kinases	>800		WebIDL	

Table 6. Kinase-inhibitor relationship resources. K-Map has two different databases – one with 178 drugs inhibiting 300 kinases and one with 72 drugs inhibiting 442 kinases.

Other Kinase Signaling Tools

The final set of bioinformatics tools related to kinase signaling, summarized in **Table 7**, cover visualization, data retrieval, and prediction tools. KinMap, PyTMs, and RegPhos2.0 are visualization tools for the kinome tree, 3D structures of phosphorylation sites, and signaling networks, respectively[95,145,146]. RegPhos2.0 also provides heatmaps for kinase and substrate mRNA expression in cancer. PhosphoLogo is used to generate sequence logos for kinases[147]. For data retrieval, RLIMS-P and eFIP are both tools that extract data on phosphorylation interactions from the literature[148,149]. CPhos identifies phosphorylation sites that are conserved across species[150]. 14-3-3-Pred predicts phosphorylation sites in protein sequences that might bind to 14-3-3 proteins[151]. KinConform takes structure files and predicts whether any kinase chains in the structure are inactive or active[152]. Kinannote predicts whether a protein sequence is a kinase[153]. Finally, CrossCheck can identify the overlap between a given list of genes and the data in a database[154].

Tool	Last Update	Version	Type	Input	Output	Method of Access	Reference
PhosphoLogo	2012		generate sequence logos	peptide sequences	logo	WeblTool	147
CPhos	2012	1.3	identifies conserved phosphorylation sites	phosphopeptides	conservation score	WeblDLlTool	150
RegPhos2.0	2013	2.0	visualization of kinase data	gene names	network visualization or cancer gene expression	WeblDL	95,96
eFIP	2014		returns publications involving phosphorylation	gene names or words	publications matching those words	Web	149
RLIMS-P	2014	2.0	returns protein phosphorylation information from literature	PMIDs or keywords	kinase, substrate, and site	WeblDL	148
PyTMs	2015	1.2	pyMOL plugin to add PTMs to protein models	protein models, PTMs	PTMs integrated in protein models	Tool	146
14-3-3-Pred	2015		predicts 14-3-3 binding phosphosite	protein sequences	predicted 14-3-3 binding sites	WeblDL	151
KinMap	2016		kinome tree visualization	kinases	tree with highlighted branches	WeblDL	145
KinConform	2017		determines which structures are kinases	structures	active or inactive kinase chains	Tool	152

Kinannote	2017	identifies and classifies kinases	protein sequences	kinase classification	Tool	153
CrossCheck	2017	identifies overlap between a database and input gene list	protein names	protein overlap	WebIDLITool	154

Table 7. Miscellaneous kinase signaling tools.

Discussion

The available databases and tools for studying kinase signaling cover diverse functions and include information on enzymes and their substrates, inhibitors, activity, and mutations. Together the tools comprise the current best standard for studying kinase signaling. Through the review of available resources, the human kinome and phosphatome were identified, substrate prediction tools were compared, and the choice of substrate sets on kinase activity inference was evaluated. Overall, these tools allow a researcher to discover vast amounts of information from their phosphoproteomic data and some tools can even perform entire sets of analyses with a single button click[135].

Despite the work that has been done, there is lots of room for advancement both in tool and method development to study kinase signaling using phosphoproteomic data. First, the majority of tools focus almost exclusively on the study of protein kinases. However, phosphatases are critical components of the kinase signaling cascade and are frequently dysregulated in cancer. Understanding the role of the interplay between kinases and phosphatases on the net phosphorylation seen in global phosphoproteomic data is essential to identifying abnormal cell signaling in disease. Furthermore, while the current tools and research are aimed at studying dysregulated protein phosphorylation, non-protein phosphorylation is also often altered in disease. For example, hexokinases, which phosphorylate glucose, drive glucose metabolism and contribute to tumor initiation in mouse models of lung and breast cancers[155]. The development of resources and tools to study non-protein kinases and phosphatases could advance research in a variety of fields.

While the current tools provide critical functions, their error rate and accuracy could be improved. Errors are frequently propagated or amplified when tools collect data from a variety of resources. However, the integration of several databases did not affect kinase activity inference compared to a single database, so the error rate in phosphorylation site databases may have a minimal effect on downstream tools.

Current tools to predict substrates of kinases perform well, but accuracy varies based on kinase. Furthermore, most tools can predict only for few kinases. MusiteDeep achieved high accuracy using their deep learning approach, but the large number of substrates required for training their method resulted in predictions for only 5 kinases. This limits its use in studying global kinase signaling. pkaPS also performed well, although it was built for a single kinase. This tool was unique because their negative set was sites experimentally validated to not be phosphorylated by PKA. Most other tools use sites not currently known to be phosphorylated by the kinase of interest, which means it is possible those sites could actually be phosphorylated by the kinase. Replicating the strategy for gold standard negative sets from pkaPS might improve accuracy for kinases in other tools.

Kinase activity inference is currently limited by the substrate number for kinases. The majority of kinases have fewer than 10 known substrates and the probability of identifying those sites in phosphoproteomic experiments is low. However, the current tools for adding substrates using prediction decreased the accuracy of kinase activity inference. The best two methods for activity inference are GSEA and z score. However, the substrate sets have a significant effect on the accuracy of the results. Because publicly available tools each use a different substrate set, this might affect the results when using different tools. In this review, I counted the number of true positives and false positives to compare the tools to understand how well each tool covered the target space of the inhibitors. The overrepresentation analysis method had a high number of false positives and the tools using the GSEA method and the z score method performed better. However, both could only predict downregulation for a small fraction of the known targets of those inhibitors. More work should be done to elucidate substrate sets and identify new ways to infer kinase activity.

For all tools, usability can be an issue, both for bioinformaticians and biologists with no computational experience. Tools are frequently platform-dependent, do not allow downloadable results, and are not well annotated. Furthermore, tools are difficult to compare or to use more than one during analysis. The input and output formats are not standardized and use a variety of protein naming conventions.

The largest challenge was deciphering input limitations and understanding results. For example, submitting a sequence with a large number of phosphorylatable residues to GPS caused the software to stall without an error message and no documentation mentioned a size limit. Musite did not provide results for a sequence or two each run without explanation. Furthermore, downloadable result files for many tools had no column headers so the column contents were unknown. For example, the downloadable file from musite has no column titles, so you have to check the table on the website to understand the results. Additionally, scores are usually presented without explanation. Only careful reading of the manuscript or the manual elucidates what value signifies a “good” response. For example, in Scansite, the score 0 is the best, with scores closest to 0 indicating the best match. But in PhosphoPICK, the score indicates probability of being phosphorylated by a kinase at that site so a score closer to 1 is better. Experts in machine learning might understand the score without explanation, but biologists likely will not.

One way to fix this challenge is to have a detailed, easy-to-find manual. The manual should include ways to run the tool, the underlying mechanism of the method, and detailed description of the results. The description of the results should also be available where results are visualized. Furthermore, sample input is helpful for a new user to test the tool and determine whether the results will be useful for their experiment before preparing their own data files.

In conclusion, there are many tools and resources that can be used to study kinase signaling and these tools will become even more essential with the continued production of phosphoproteomic data. It is essential for the biological community to research under-studied enzymes and to validate specific substrates of kinases and phosphatases. Furthermore, bioinformaticians should consider creating tools that utilize information from both sides of the enzymatic phosphorylation reaction. Finally, resources should be

carefully planned, easy to use, and well maintained. The community should work to standardize the use of enzyme IDs and phosphorylation site location.

CHAPTER 3

PROTEOMIC LANDSCAPE OF KINASE SIGNALING IN CANCER

Introduction

The massive sequencing efforts identifying the genomic and transcriptomic alterations in numerous cancer samples by groups such as TCGA, METABRIC, and ICGC provided novel insights into cancer processes and better understanding of the underlying biological mechanisms[156,157]. However, the size and complexity of the data make translating these findings into the clinic challenging. Furthermore, discoveries at the genomic level do not always translate well to the protein level. The correlation between mRNA and protein expression is low for many genes[158,159]. Additionally, many genomic alterations are in non-coding regions and the number of druggable mutations in some cancers is low[22]. Because proteins are the primary drivers of cell phenotype and are the majority of drug targets in cancer, complementing the genomic data with global proteomic changes should help to clarify the effect of genomic alterations and narrow the focus to important drug targets.

CPTAC undertook the goal of characterizing the proteome in three cancer types (breast, colorectal, and ovarian) that already had genomic data in the TCGA[22]. Additionally, they generated phosphoproteomic data for breast and ovarian cancers. To study the global proteome and phosphoproteome, they used quantitative mass spectrometry. In comparison to the antibody approach of studying proteins, mass spectrometry can discover novel peptides, elucidate the entire proteome of a sample at once, and can circumvent the non-specificity issue of antibodies.

Global protein mass spectrometry is performed by first lysing tissue and digesting proteins into short peptides. Trypsin is the most commonly used digesting enzyme, but other proteases used include chymotrypsin, LysC, LysN, AspN, GluC, and ArgC[160]. Sequence coverage depends heavily on the selection of protease and including more than one can increase coverage[160]. After digestion, peptides are either labeled with a stable isotope such as isobaric tags for relative and absolute quantification (iTRAQ) or tandem mass tag (TMT) or left unlabeled. Label-free mass spectrometry is not easily quantified due to differences in peptide ionization efficiency and lack of internal standard, although counting the number of spectra correlates well with protein abundance[161,162]. Tags such as iTRAQ or TMT provide multiplexing and easier quantification but diminish signal and can therefore discover fewer differentially expressed proteins[163,164].

Peptides are further divided into fractions, usually using liquid chromatography (LC), to allow for better peptide resolution. Ideally, the mass spectrometer should detect a single peptide at a time. Tandem mass spectrometry (MS/MS), which outputs *m/z* and intensity values, follows sample separation. Finally, peptides are identified from the mass spectrometry spectra by matching to reference protein databases using software.

Phosphoproteomics experiments are performed in much the same way, but there is an extra step after sample fractionation to enrich for phosphorylated peptides. This is usually done using immobilized metal affinity chromatography (IMAC) which attracts negatively charged phosphate groups to positive metal ions or metal oxide affinity

chromatography which attracts oxygen to metal atoms[165]. Quantification and coverage for phosphorylated peptides is lower because of lower abundance and because usually only a single peptide can be used to quantify a site compared to several peptides referring to the same protein.

While mass spectrometry can identify proteins at a global level, the technique may have some bias in the quantification of proteins. Some proteins (especially smaller proteins) do not have enough tryptic peptides, might not have unique peptides for identification, or have abundance below the limit of detection. This is a particular challenge for kinases, because kinases are usually in low abundance[166]. Therefore, we first need to understand the limits of the technology before using it to study kinase signaling.

Methods

Definition of Human Kinases, Phosphatases, and Their Substrates

Kinases and phosphatases were defined as in **Chapter 2**. Protein lengths and sequences were downloaded from UniProt in June 2017. Experimentally validated phosphorylation sites and enzyme-substrate interactions were downloaded from dbPAF[105] (http://dbpaf.biocuckoo.org/), PhosphoSitePlus (July 2017), and Signor (October 2017). dbPAF was used as a database of phosphorylation sites that did not include data from any CPTAC publications.

Cancer Proteomics Datasets

Retrospective proteomic data for breast, colorectal, and ovarian cancers and phosphoproteomic data for breast and ovarian cancers generated by CPTAC were downloaded from the respective manuscripts. The breast cancer samples consisted of 105 (plus 3 replicates and 3 normal) tissue samples from patients with invasive breast adenocarcinoma[158]. These proteomic samples were generated by LC-MS/MS performed on samples digested with LysC and trypsin. Values were log2 transformed iTRAQ ratios of the sample compared to a common reference pool.

The colorectal cancer dataset consisted of spectral counts of proteins in 95 (plus 5 replicates) tumor samples from patients with colorectal carcinoma[159]. These samples were previously digested with trypsin and quantified by LC-MS/MS using the label-free spectral counting method.

The ovarian cancer dataset consisted of 174 (plus 32 replicates) samples from patients with ovarian serous adenocarcinoma[167]. These samples were digested with trypsin, processed using LC-MS/MS, and reported as log2 transformed iTRAQ ratios between the sample peptides and a common reference pool. Three unique peptides were required to identify a protein.

Identified proteins used below are those reported in the respective manuscripts of the proteomic datasets. Clinical, genomic, and transcriptomic data corresponding to these samples were downloaded from cBioPortal (http://www.cbioportal.org/)[168,169].

Phosphorylation Site Processing

All quantified phosphorylation site peptides were mapped to UniProt protein sequences (July 2017). The canonical UniProt ID and sequence were chosen. If the peptide matched only to protein isoforms, one isoform ID and sequence was chosen as a representative of the group. Log ratios for multiple peptides corresponding to the same phosphorylation site were combined by median. Sites were defined as thirteenmer peptides.

Comparison with the Human Protein Atlas

Expression by immunohistochemistry in multiple cancer types was downloaded from the Human Protein Atlas (HPA; www.proteinatlas.org)[170] in June 2017. Categorical expression was converted into a single score for each protein in breast, colorectal, and ovarian cancers. 'Not detected' was zero points, 'low' expression was 1 point, 'medium' expression was 2 points, and 'high' expression was 3 points. Points were averaged across samples, and proteins with a score less than or equal to 0.1 were considered undetected. This corresponded to at most one sample with 'low' expression.

Gene Ontology Term Enrichment

Gene ontology (GO) biological process term enrichment was performed using the R package version of WebGestalt. The analysis was performed for undetected enzymes in all three cancers against a background reference of the human kinome or human phosphatome. FDR with a Benjamini Hochberg correction < 0.05 was considered significant. Enrichment was also performed using Fisher's exact test and kinase and phosphatase superfamilies, excluding non-protein kinases.

Quantified Data

All analyses requiring quantified data were filtered to samples passing quality control filters as described in the original publications. Proteins and sites with missing values in fewer than 50% of the samples were retained.

Kinase Correlation with Clinical Features

Kinase activity was inferred for single samples using single sample GSEA (ssGSEA) analysis in the GSVA R package for enrichment of substrate sets[171,172]. Comparisons of kinase activity scores, phosphorylation site levels on kinases, and kinase protein abundance between groups of patients were performed using the Wilcoxon ranked sum test. FDR < 0.05 was considered significant.

Prediction of Kinase Mutations

Proteogenomic variants identified in the breast cancer manuscript were submitted to ReKINect for analysis[140]. Mutations were filtered to those with predicted effect on kinases and with substrates identified in PhosphoSitePlus (version May 2018). For each mutated kinase, the mean phosphorylation level of its substrates in the mutated sample was compared to samples with no mutation.

Statistical Analysis

The R computing environment was used to perform all analyses. The package VennDiagram was used to create the Venn diagrams[173]. The package pheatmap was used to create the heatmaps[174]. Comparison of means between undetected and detected enzymes was performed using a t-test. A p value < 0.05 was considered significant.

Results

Kinases and Phosphatases Identified by Mass Spectrometry

Because of different experimental and data analysis methods, the total number of identified proteins differed between the proteomic datasets for breast, colorectal, and ovarian cancers (**Table 8**). The breast cancer dataset had the most identified proteins and consequently had the most identified kinases and phosphatases. Overall, enzymes were found across cancer samples rather than being cancer type-specific. Forty-five percent of kinases (**Figure 7A**) and 50% of phosphatases (**Figure 7B**) were identified in all three cancer types. Over 65% of the kinome and phosphatome were identified in at least two cancer types.

Cancer Type	Samples	Proteins	Kinases	Phosphatases
Breast	105	15,369	622	214
Colorectal	90	7,211	347	135
Ovarian	174	9,600	435	165

Table 8. Number of identified proteins in three cancer datasets. There are 688 total known human kinases and 255 total phosphatases.

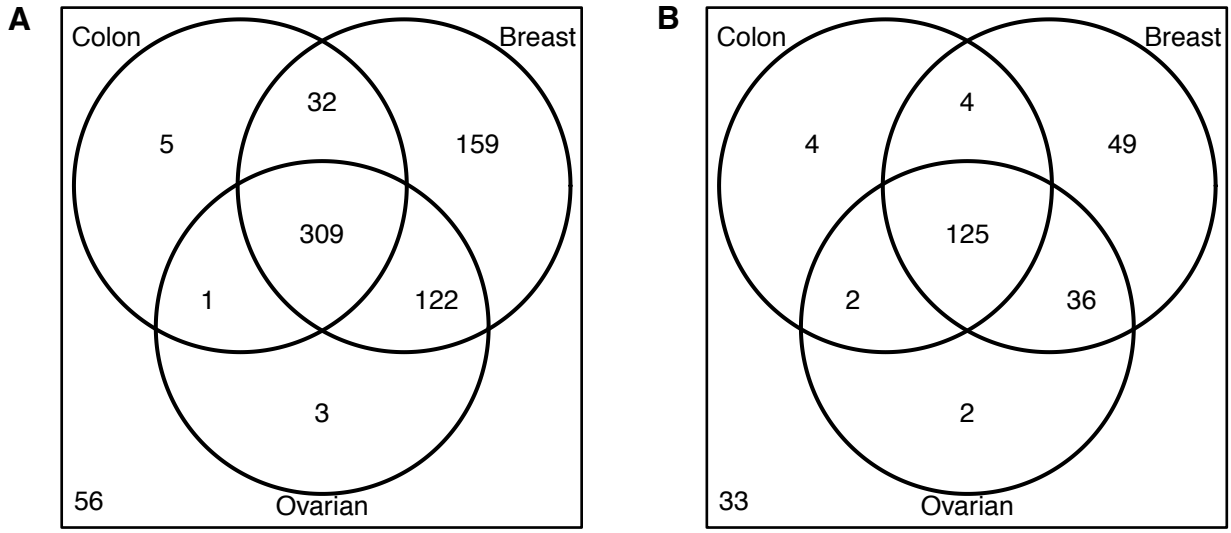


Figure 7. Intersection of detected A) kinases and B) phosphatases in three different cancer types. There were 56 kinases and 33 phosphatases undetected by mass spectrometry in any sample.

The other primary mechanism to measure protein expression is the use of antibodies. HPA evaluated the expression of proteins in a variety of cancer types using antibodies in immunohistochemistry. If proteins are undetected in a tissue by both mass

spectrometry and antibody-based methods, then the proteins are likely not expressed in that tissue.

I compared the presence of proteins in breast, colorectal, and ovarian cancers in both CPTAC and HPA data. Over 1/3 of the enzymes were identified in all three cancer types by both CPTAC and HPA. However, there was little overlap between those undetected by mass spectrometry and those undetected by antibodies (**Figure 8**). Many enzymes that were not identified by mass spectrometry in any cancer sample had high expression by antibodies. The discrepancy could not be explained by antibody reliability. Only 17 out of 49 antibodies for enzymes undetected by mass spectrometry were annotated as uncertain, while the rest were confirmed to be specific.

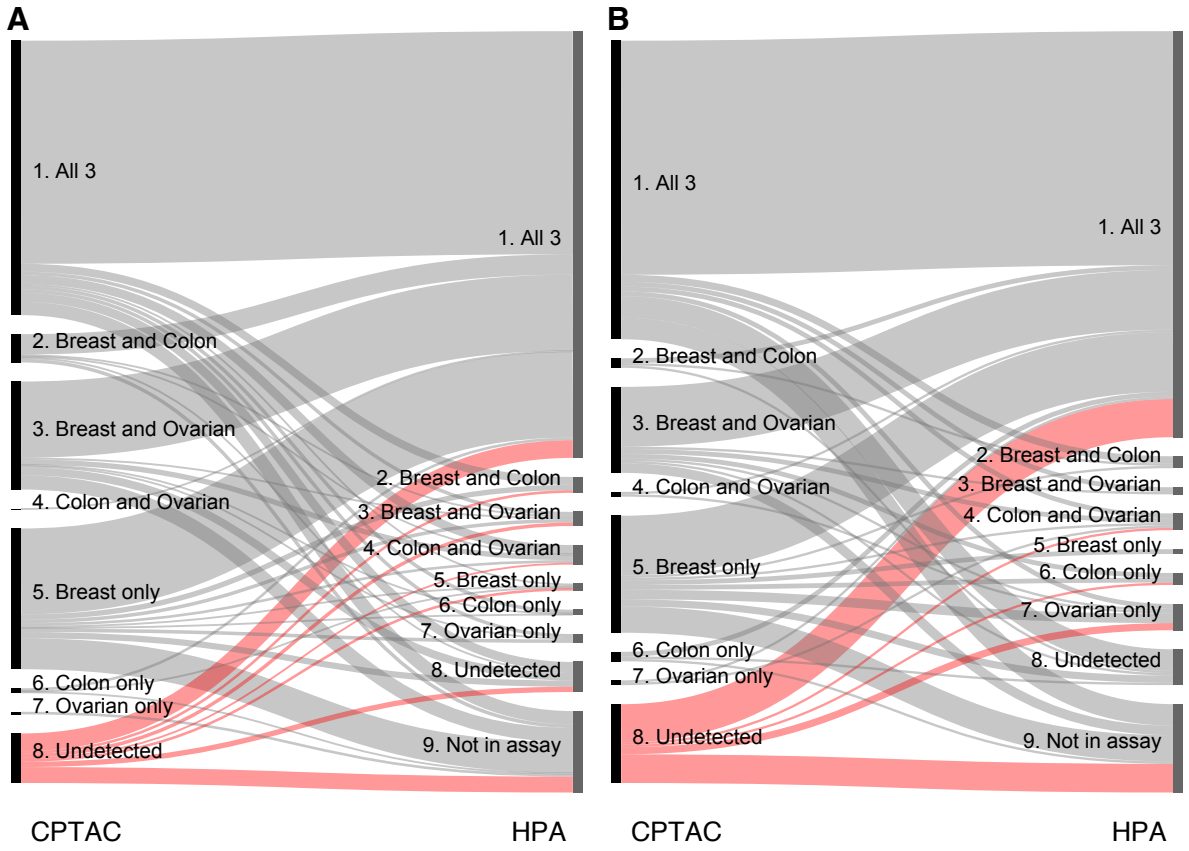


Figure 8. Enzymes in CPTAC and HPA data. A) Identification of kinases and B) phosphatases by mass spectrometry (CPTAC) and immunohistochemistry (HPA). A large proportion of enzymes were identified in all three cancer types in both CPTAC and HPA. Some proteins undetected by mass spectrometry (indicated in red) showed expression with antibodies in cancer tissue.

Protein Length

Protein length can also affect detection by mass spectrometry. Longer amino acid sequences might be digested into more peptides, which increases the chance of detection. There was a significant difference in length of undetected compared to detected proteins in breast cancer, but not colorectal or ovarian cancers (**Figure 9**).

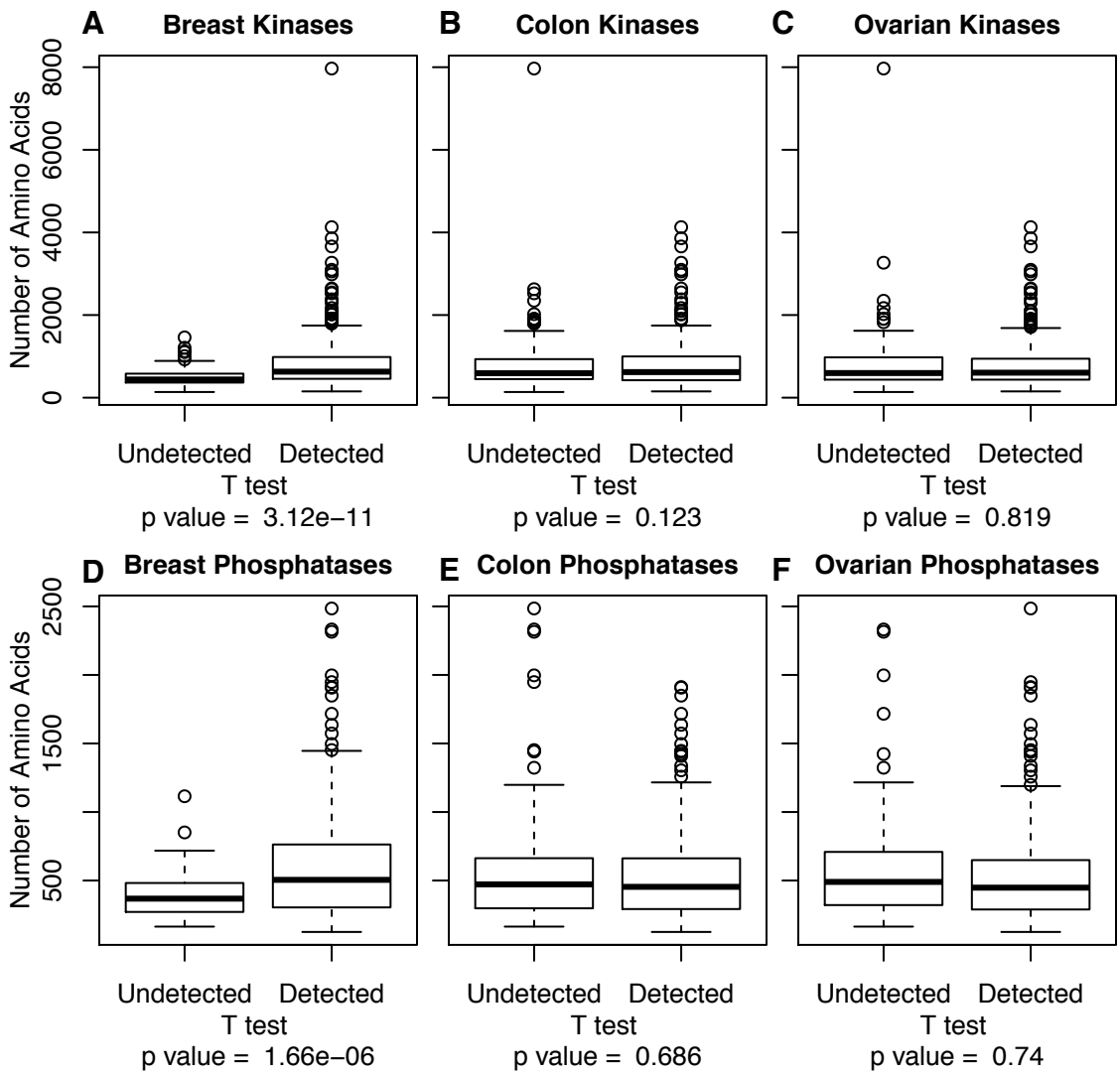


Figure 9. Number of amino acids in proteins that were and were not detected by mass spectrometry. There was a significant difference in the lengths of the A) kinases and D) phosphatases detected by mass spectrometry in breast cancer samples compared to those that were not detected. There was no difference in detected B) kinase and E) phosphatase length in colorectal cancer, nor in the C) kinases and F) phosphatases of ovarian cancer.

mRNA Expression

Proteins in low abundance may be difficult to identify in mass spectrometry. As a surrogate for protein expression, mRNA expression can be used to assess whether proteins have a low expression. In these three cancers, proteins that were not identified in individual cancers had a lower mRNA expression than those that were identified in that cancer (**Figure 10**).

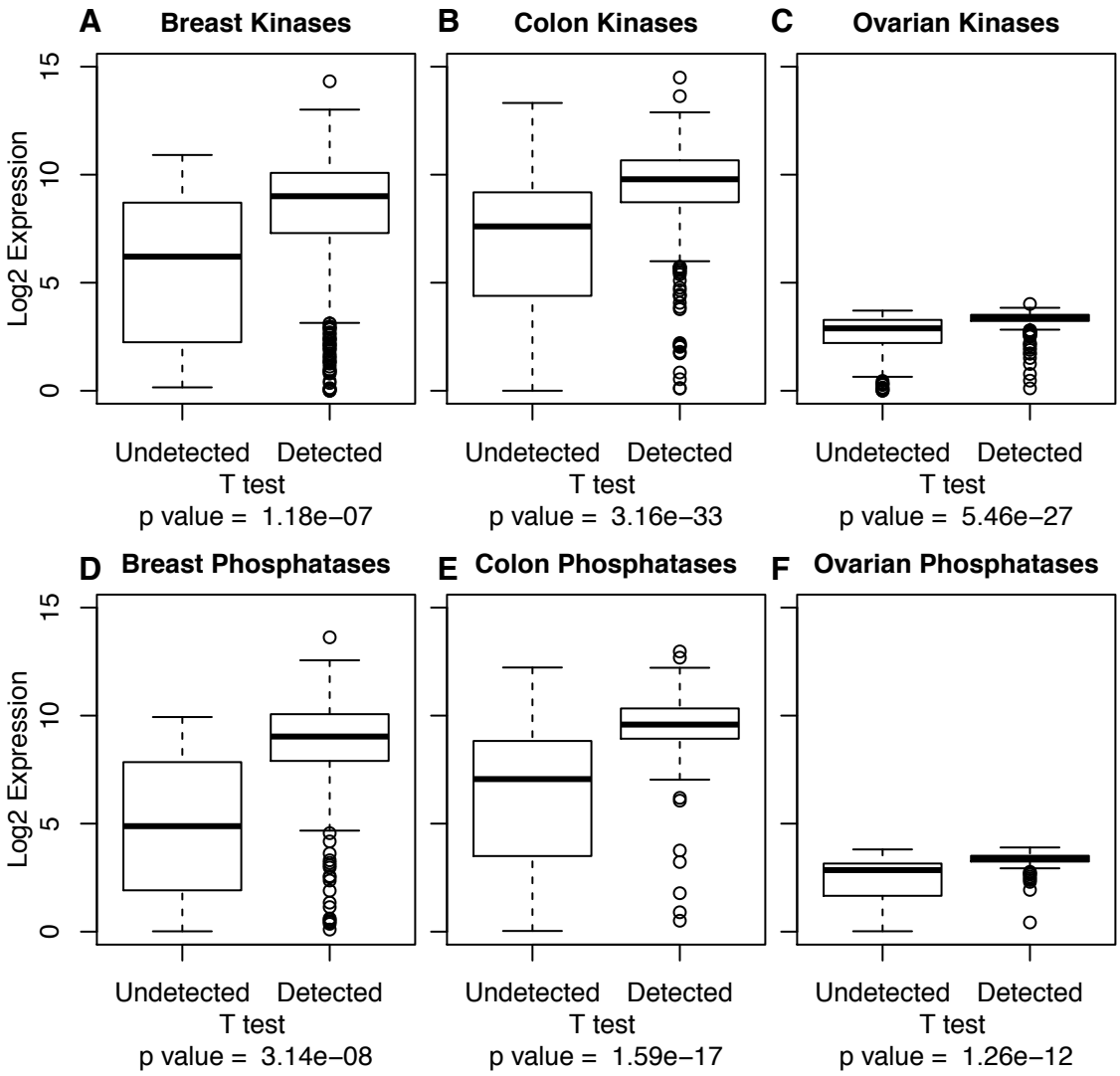


Figure 10. Log2 mRNA expression of proteins that were and were not detected by mass spectrometry. Difference in expression of A,B,C) kinases and D,E,F) phosphatases detected by mass spectrometry in breast cancer (A,D), colorectal cancer (B,E), and ovarian cancer (C,F).

Term Enrichment of Undetected Enzymes

There was no significant enrichment (FDR < 0.05) of GO biological process terms for undetected kinases or phosphatases compared to the full human kinome or phosphatome. There was also no significant enrichment of individual kinase families in the undetected kinases. However, the undetected phosphatases were enriched for the DSP, phosphatidic acid phosphatase, and the chloroperoxidase superfamilies. Fewer than half of the members of each of these superfamilies were identified by mass spectrometry in these CPTAC experiments.

Phosphoproteomic Analysis of Breast and Ovarian Cancers

The CPTAC studies for both breast and ovarian cancers contained phosphoproteomic data complementary to the proteomic data. The breast cancer data

identified almost 23,000 phosphorylation sites that had non-missing values in at least 50% of the 77 samples passing quality-control filters. This corresponded to sites on almost 6,000 unique proteins. The ovarian cancer data, however, only identified 5,000 phosphorylation sites on about 2,500 proteins in at least 50% of 66 samples. Most of these phosphorylation sites (92% in breast, 96% in ovarian) had been identified in previous low or high-throughput experiments as reported in the dbPAF database. These data captured phosphorylation sites on kinases (357 kinases in breast, 177 in ovarian) and phosphatases (96 phosphatases in breast, 41 in ovarian). Almost all the kinases and phosphatases with quantified phosphorylation sites had corresponding proteomic data.

Kinase Activity in Breast Cancer

Kinase activity can be inferred by determining the enrichment of its substrates in a sample. Out of the almost 23,000 quantified phosphorylation sites in breast cancer, only 1,486 have a known kinase or phosphatase. Using ssGSEA of these sites, activity for 158 kinases and phosphatases could be determined. The enzymes with the highest variability in activity (standard deviation ≥ 0.25) across samples were STK11, TTBK1, and CDK5. Kinase activity clustered with transcriptomic subtypes, suggesting similar kinase signaling in tumors with similar gene expression (**Figure 11**). There was a subset of primarily luminal A tumors that had higher relative activity of most kinases compared to the other samples. There was another subset of mixed luminal A and luminal B tumors that had very low relative activity of MAP2K and MAP3K kinases compared to other samples.

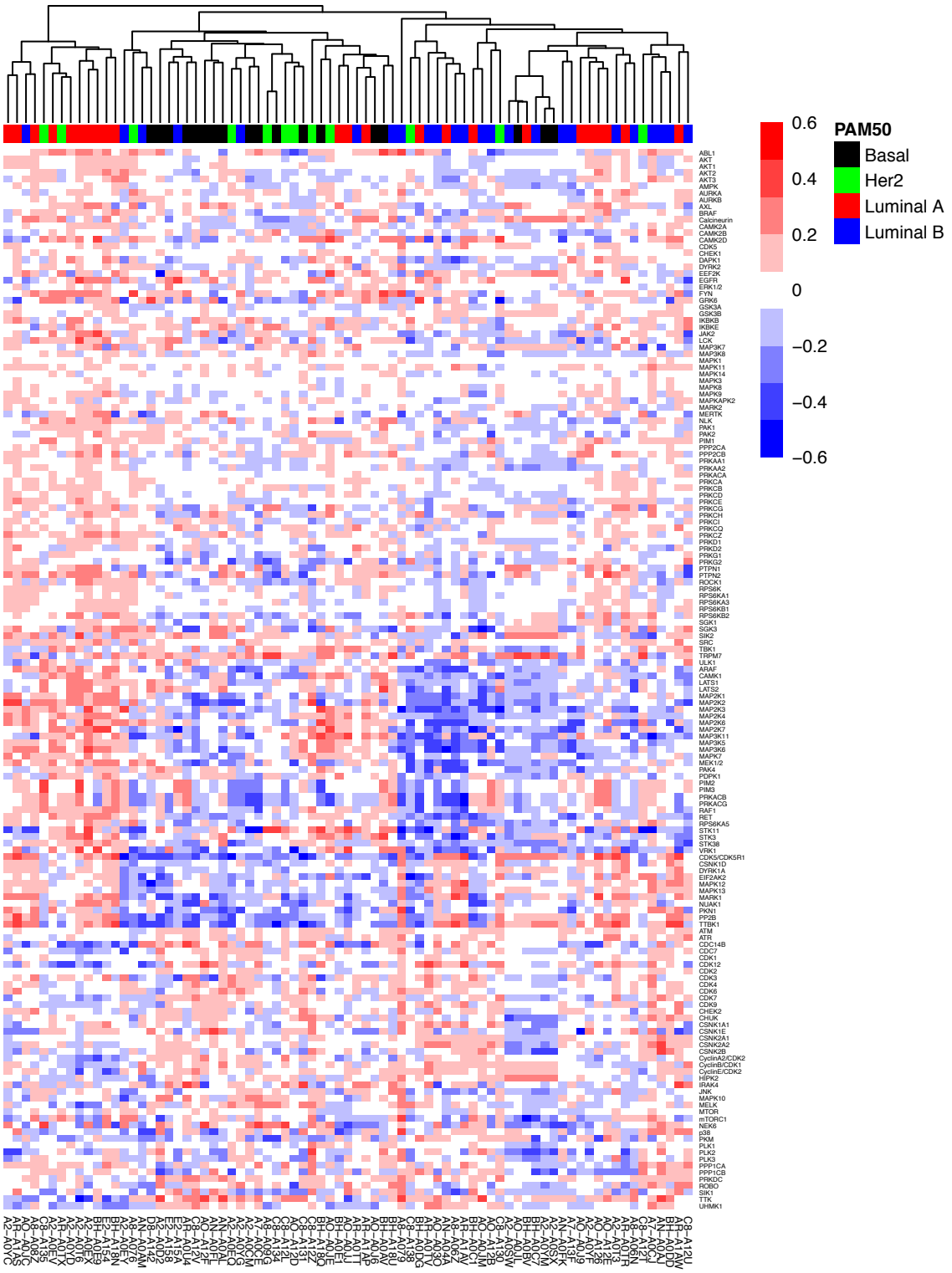


Figure 11. Heatmap of inferred kinase and phosphatase activity scores for breast cancer samples. Kinase activity was determined using ssGSEA of enzyme site-level substrates. Red indicates increased kinase activity relative to other samples, while blue indicates decreased kinase activity. The sample PAM50 subtype is indicated by black for basal, green for Her2, red for luminal A, and blue for luminal B. Rows are ordered by k means cluster (4 clusters).

To determine the effect of kinase activity on patients, I wanted to compare activity to clinical features, but unfortunately clinical variables were limited in this cohort. Therefore, I chose two interesting data features: early vs late stage tumors and tumors with and without *GATA3* mutation. The transcription factor GATA3 is mutated in about 10% of breast tumors, but the phenotypic effects are not well understood[175]. None of the kinases with highly variable activity were significantly different in any of these groups. However, the activity of MERTK was significantly lower in samples with *GATA3* mutation compared to samples without *GATA3* mutation (**Figure 12A**, FDR = 0.02).

No kinase activity scores were significantly different between early (stages 1 and 2) tumors and late (stages 3 and 4) tumors. However, the protein level of the kinase NT5C was significantly increased in stage 1&2 compared to 3&4 (**Figure 12B**, FDR = 0.01). NT5C is a nucleoside kinase and therefore does not have an activity score. No phosphorylation levels on any kinase were significantly different among these groups.

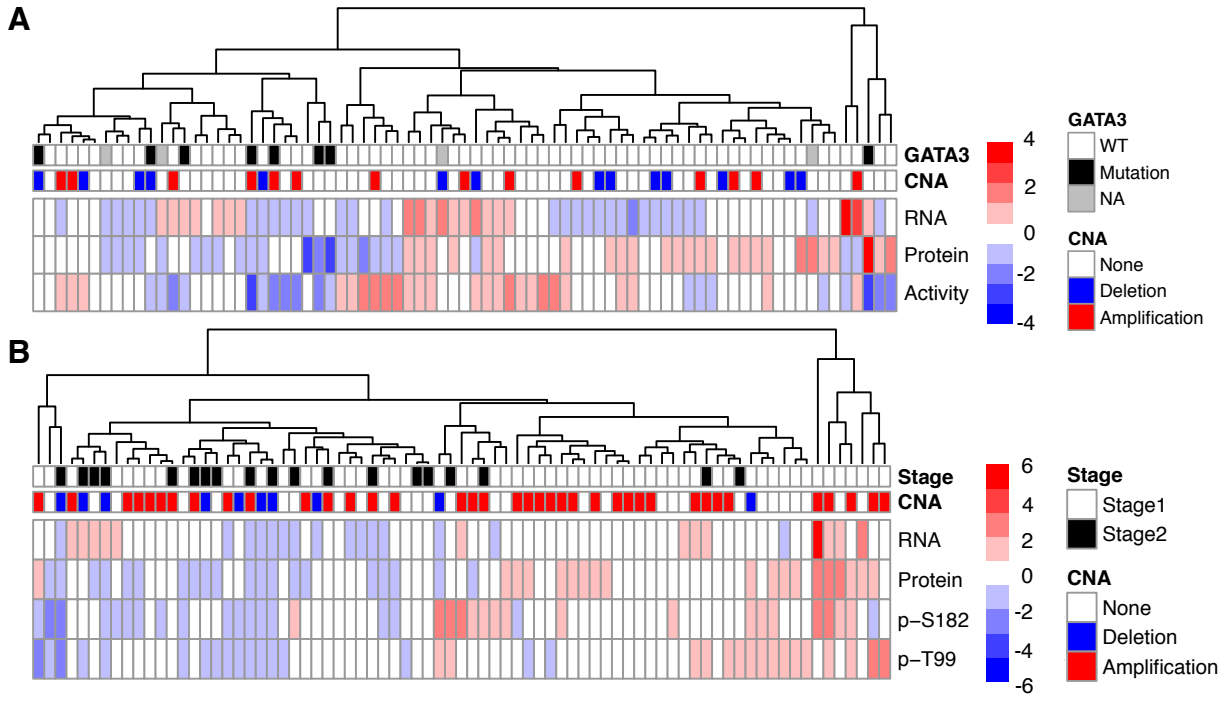


Figure 12. Multi-omics profile of MERTK and NT5C in breast cancer. A) RNA, protein, and ssGSEA activity scores for MERTK. B) RNA, protein, and phosphorylation levels at S182 and T99 for NT5C. All values were standardized using z score.

Effect of Mutations on Kinase Signaling

Finally, I wanted to determine the effect of protein mutations on kinase activity in breast cancer. To do this, I predicted the effect of the 3,658 identified protein variants using ReKINect. Of the 641 mutations that had a possible effect on kinase signaling, most were predicted to have an unknown function (**Table 9**). Phosphoproteomic data of kinase substrates could be used to evaluate the effect of these mutations. Unfortunately, mutations are unique to individual patients and proteins, which limits the statistical

analysis. However, some mutations show interesting patterns worth exploring in future work. For example, sample A2-A0YD had a D110N mutation in the kinase domain of CDK6. ReKINect predicted this mutation to be an "uninterpreted mutation on kinase domain." The mean phosphorylation levels of CDK6 substrates for that sample were 1.8 standard deviations below the mean for all samples. Therefore, this mutation might inhibit the function of CDK6. Similarly, patient A0-A126 had a T337K mutation on MASTL. Phosphorylation of its substrate was 1.5 standard deviations below the mean.

Mutation Type	Frequency
Destruction of phosphorylation site	64
Kinase downstream rewiring on kinase domain	1
Kinase downstream rewiring, Mutation around phosphorylation site on kinase domain	1
Kinase inactivation by phosphorylation destruction, Destruction of phosphorylation site on kinase domain	1
Mutation around phosphorylation site	510
Phosphomimicking mutation	9
SH2 downstream rewiring on kinase protein on SH2 domain	1
SH2 downstream rewiring on SH2 domain	1
Uninterpreted mutation on kinase	41
Uninterpreted mutation on SH2	12

Table 9. Predicted proteogenomic mutation effect from ReKINect.

Discussion

The proteomic and phosphoproteomic data generated by CPTAC provide unique opportunities to explore signaling dysregulation in cancer. Although mass spectrometry experimental protocol and analysis methods have a significant effect on the number of protein identifications, in general it can identify all aspects of kinase signaling with limited biases. Notably a minimal number of kinases and phosphatases were not detected in any of the three experiments, but there was no single reason for exclusion. Some of these enzymes are likely to be tissue specific (e.g., GRK7 is retina-specific) or exhibit low expression in these cancer tissues. Because there was no bias in identifiable kinase family, we can be confident proteomic data can be used to study global kinase signaling in cancer tissue.

Phosphoproteomic and proteomic data can identify individual kinases that are interesting in breast cancer. While these are primarily association studies and limited by sample size, they provide a starting point for future studies. In particular, STK11, CDK5, and TTBK1 appear to have differential activity across samples and may become targets of individualized treatment. *STK11* encodes liver kinase B1 (LKB1), which is a tumor suppressor involved in major cell pathways. In breast cancer, increased LKB1 activity can inhibit TGF-β1 transcription, leading to inhibition of epithelial-to-mesenchymal transition (EMT)[176]. Furthermore, lower LKB1 mRNA expression correlated with poor survival in the TCGA cohort[177], although another study found significant correlation with survival only in HER2-positive patients[178]. Few reports explore CDK5 or TTBK1 in breast cancer, but one study suggested lower CDK5 expression promotes longer metastasis-free survival[179] and CDK5 is pro-tumorigenic in several cancer types[180].

In addition to the kinases with variable activity across samples, predicted MERTK activity was significantly reduced in samples with mutations in *GATA3*. Although no associations between these two proteins are noted in the literature, it is possible that GATA3 affects MERTK expression. MERTK is predicted to be a target of GATA3 based on its binding site motif[181] and MERTK mRNA was significantly reduced in *GATA3* knock-out keratinocytes[182].

Unfortunately, clinical outcomes are limited in the breast cancer cohort, but an interesting finding was the correlation of NT5C abundance with breast cancer stage. Little is known about NT5C in breast cancer, but in leukemia patients treated with cytarabine, higher mRNA expression of NT5C correlated with worse outcomes. This finding is counter to its high expression in early stage breast cancer here[183].

Finally, the use of phosphoproteomic data to validate kinase signaling mutations holds promise. Future work to quantify both variant and wild-type peptides will help determine the effect of mutations near a phosphorylation site on the ability of that site to be phosphorylated. Furthermore, extending mutation analyses to include SNVs identified at the gene level might increase sample size and allow for statistical analysis of the effect of individual kinase mutations on the activity of that kinase.

CHAPTER 4

CHARACTERIZATION OF MOLECULAR SUBTYPE-SPECIFIC DRIVER SUBNETWORKS IN BREAST CANCER

Introduction

Breast cancer is a heterogeneous disease with different treatment modalities and outcomes. To differentiate between patients, breast cancer is split into various subtypes by histological type, stage, presence of hormone receptors, or mRNA expression. Using mRNA data, breast tumors can be classified into four different molecular subtypes based on the expression of 50 genes[184,185]. The four PAM50 subtypes, basal, Her2, luminal A, and luminal B, have different characteristics, survival rates, and treatment options[186,187]. While the expression patterns of tumors within a subtype are similar, the underlying signaling mechanisms driving these subtypes are not well understood.

Cancer is hypothesized to be driven by genomic mutations that lead to downstream signaling dysregulation and phenotype changes. However, most mutations are present in small percentages of patients. As seen in Chapter 3, mutations are frequently specific for individual patients. However, individual mutations in patients might all converge on the same pathway. For example, one patient might have an inactivating mutation on PTEN, which results in active AKT signaling. A second patient might have an activating mutation on PI3K, which also results in active AKT signaling. In this way, different individual mutations could drive the same altered expression patterns within a subtype.

One way to extract signaling networks from individual mutations is to use the random walk with restart (RWR). In this method, a random walk on a PPI network starts from mutated nodes and randomly walks to neighboring nodes at each time step. The restart value allows the random walk to teleport back to the starting node to restrict the final smoothed network to the nearest neighborhood of the starting nodes. The final probabilities of landing on individual nodes provides a smoothed network from the starting mutated nodes. Because the probability of landing on a single node immediately downstream of two different mutated genes is high, the most significant genes will be those downstream of multiple different mutations. This method has been used to identify subtypes of various cancers, extract driver subnetworks in colorectal cancer, and identify altered signaling pathways after drug treatment, among others[188–190].

The limitations to identifying driver subnetworks using RWR are validation and identifying the network direction. Mutations can be activating or inhibitory on a gene and the actual function of most individual mutations in cancer is not known. Furthermore, validating the network as a true driver of the subtype usually requires extensive molecular biology experiments. Phosphoproteomic data, however, provides a unique opportunity to both validate the extracted subnetworks and determine whether the network is under- or overactive in that subtype.

Methods

Protein-Protein Interaction Network

The PPI network was downloaded from High-quality INTeractomes[191] (HINT, http://hint.yulab.org/) in May 2017. The largest connected component was extracted using igraph[192] in R and consisted of 11,711 nodes and 110,838 edges. The nodes were normalized by degree.

Starting Probabilities

Non-silent mutation data were downloaded for 806 breast cancer samples from cBioPortal and processed as in **Chapter 3**. Mutation data for individual samples were normalized by total number of mutations within a sample, summed across all samples within a PAM50 subtype, and averaged. Nodes with higher probability indicated more samples with that mutation. Genes were randomly permutated 1000 times for statistical analysis.

Driver Subnetwork Extraction

Starting probabilities for all mutated nodes in a subtype and the permuted probabilities were submitted to the NetWalker[190] algorithm implemented in R. The restart probability was set to 0.5. Local p values were calculated as the rank of the gene in the real run across all permuted runs. Global p values were calculated as the rank of the gene across all scores. Nodes were considered significant with both a global and local $p < 0.05$. Networks were generated by filtering for edges in the HINT network between significant nodes. Additional subnetworks were extracted using edges between significant nodes that were unique to a subtype.

Driver Subnetwork Pathway Enrichment

Overrepresentation enrichment analysis of significant nodes in each network was performed using WebGestaltR and WikiPathways[193]. The minimum overlap was set to be 10 and the maximum was 500. The background was all possible nodes in the HINT network and pathways were considered significant with FDR < 0.05.

Differential Phosphorylation Site Abundance

Phosphorylation sites were processed as in **Chapter 3**. Log2 ratios were converted to z score. The z scores for each subtype were compared to the samples of all other subtypes using the Wilcoxon signed rank test. P values were adjusted using the Benjamini-Hochberg method and significance was assigned as FDR < 0.05.

Evaluation of the DNA Damage Response in the Basal Subnetwork

The number of nodes with significant differential phosphorylation were counted for each significant pathway. The pathways with the highest number of these nodes were explored. The DNA damage response pathway (WP707) in the basal subnetwork had the highest number of significant differentially phosphorylated nodes. Pathway direction was inferred from the regulatory effect of phosphorylation on those nodes.

Differential Substrate Phosphorylation

The kinases, defined as in **Chapter 2**, were extracted from each subnetwork. For every kinase in an individual subnetwork with at least 1 substrate in the phosphoproteomic data, the z score phosphorylation levels of the substrates were compared between samples in a single subtype and all other samples using a Wilcoxon rank sum test. P values were adjusted using the Benjamini-Hochberg method and were considered significant at FDR < 0.05.

Software Packages

The R computing environment was used to perform all analyses. The Venn diagram was created using Venny v2.1. Cytoscape[194] was used for network visualization and the R package pheatmap was used to create the heatmap.

Results

Generation of Driver Subnetworks

Subtype-specific driver subnetworks were created using network propagation from mutated genes in that subtype. There were 145 basal samples, 66 Her2, 416 luminal A, and 179 luminal B. The mutations in each sample were normalized by number of mutations in the sample and were combined within a subtype. For all subtypes, the network propagation started from about 4000 mutated nodes. Approximately 600 nodes in each subtype had statistically significant steady state probability after propagation. Over 50% of the nodes for each subtype were unique to that subtype (**Figure 13**). Only 29 genes were significant in all four networks and only 6.5% of the genes were present in at least 3 of the 4 networks, indicating subtypes had uniquely altered pathways.

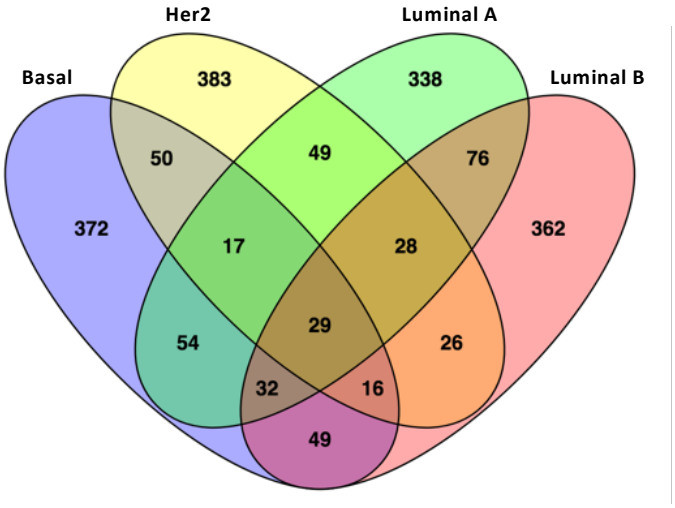


Figure 13. Venn diagram of significant nodes in subtype-specific driver subnetworks.

I filtered for edges between all significant nodes in a subtype and extracted the largest connected component for each network. The basal subnetwork had 287 nodes and 468 edges, the Her2 subnetwork had 239 nodes and 407 edges, luminal A had 321

nodes and 509 edges, and luminal B had 299 nodes and 458 edges. All networks can be found in **Appendix C**.

Using WikiPathways, I performed pathway enrichment for each of the driver subnetworks. The basal subnetwork was enriched for alpha 6 beta 4 signaling, DNA damage response, FasL pathway and stress induction of HSP, and striated muscle contraction (Table D1, FDR < 0.05). The Her2 subnetwork was enriched for signaling of hepatocyte growth factor receptor, rac1/pak1/p38/MMP-2 pathway, PDGF pathway, and focal adhesion (Table D2, FDR < 0.05). Both the luminal A and luminal B subnetworks had a large number of enriched pathways. In addition to several of the pathways enriched in basal and Her2, luminal A also had enrichment of ErbB signaling, estrogen signaling, MAPK, and PI3K-AKT-mTOR signaling pathways (Table D3, FDR < 0.05). The luminal B subnetwork uniquely had enrichment for regulation of actin cytoskeleton, miRNA targets in ECM and membrane receptors, and TGF-beta signaling (Table D4, FDR < 0.05).

Differential Abundance of Phosphorylation Sites Across Subtypes

To use phosphoproteomic data to validate driver subnetworks, phosphorylation patterns should differ between subtypes. However, few sites were significantly changed in one subtype compared to the others. There were no significant sites in Her2 samples compared to the others (**Figure 14B**). In basal, 1338 sites were significant compared to the others, 820 were in luminal A, and 20 were in luminal B (**Figure 14A,C,D**). Only 71 of the significant sites were on proteins significant in the basal subnetwork, 41 in the luminal A, and 2 in the luminal B. Despite this minimal overlap, phosphoproteomic data might still suggest a direction for the pathways.

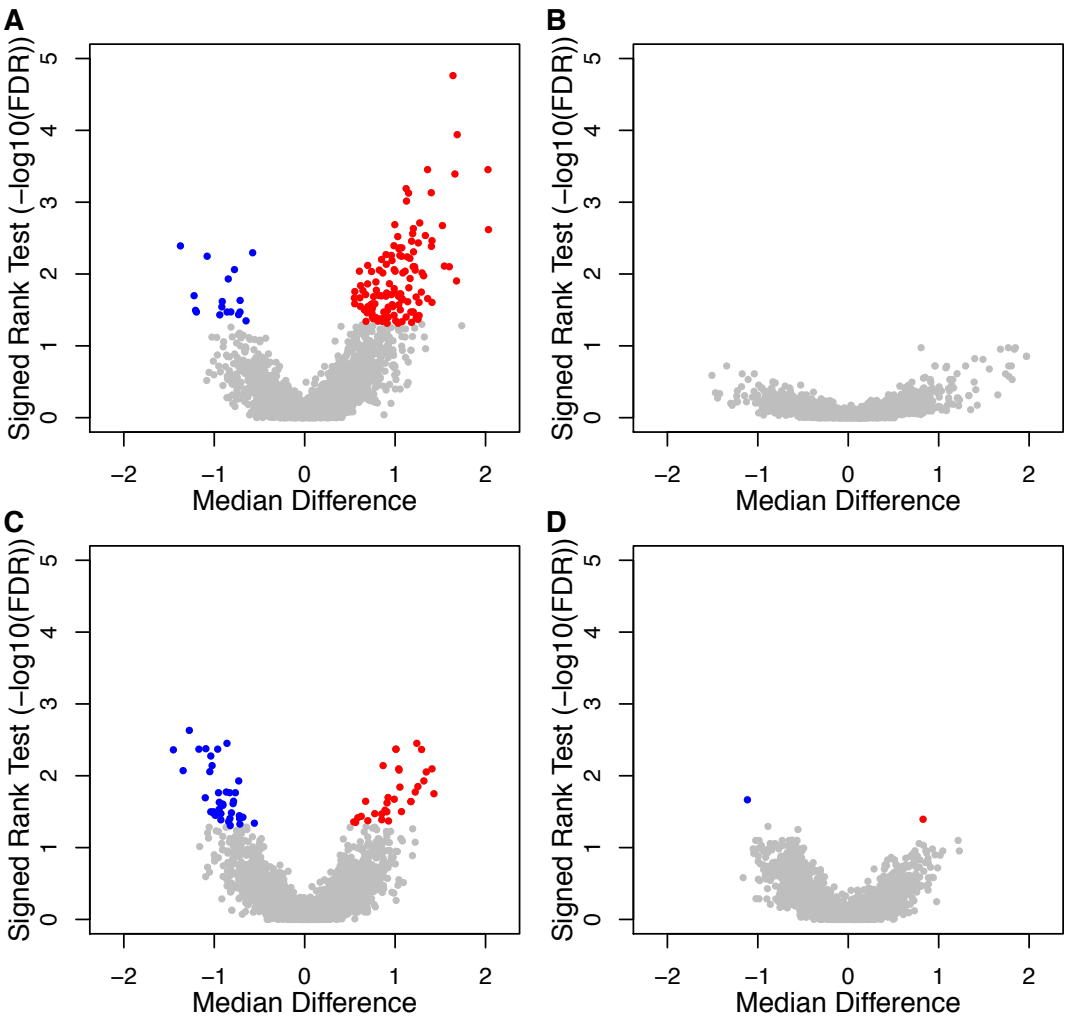


Figure 14. Relative differential phosphorylation abundance in breast cancer subtypes. Normalized phosphorylation abundance at each site on nodes in the network was compared between each subtype and the remaining samples. Red indicates significantly increased phosphorylation sites and blue indicates significantly decreased phosphorylation sites in A) basal, C) luminal A, and D) luminal B subtypes relative to all other samples. There was no significant differential phosphorylation in B) Her2 samples.

DNA Damage Response in Basal

One interesting enriched pathway in the basal subtype was the DNA damage response pathway. Eleven of the nodes in the basal subnetwork were present in the DNA damage response pathway. Four of these nodes had phosphorylation sites that were significantly increased in the basal subtype compared to all other samples (**Figure 3**). Phosphorylation on TP53 at S315 increases in response to DNA damage and promotes TP53 activity in the response to DNA damage[195]. Phosphorylation of both kinases PRKDC and CHEK2 increase their activity and also promote their response to DNA damage[196,197]. Therefore, these phosphorylation sites indicate that the DNA damage response is highly active in basal tumors.

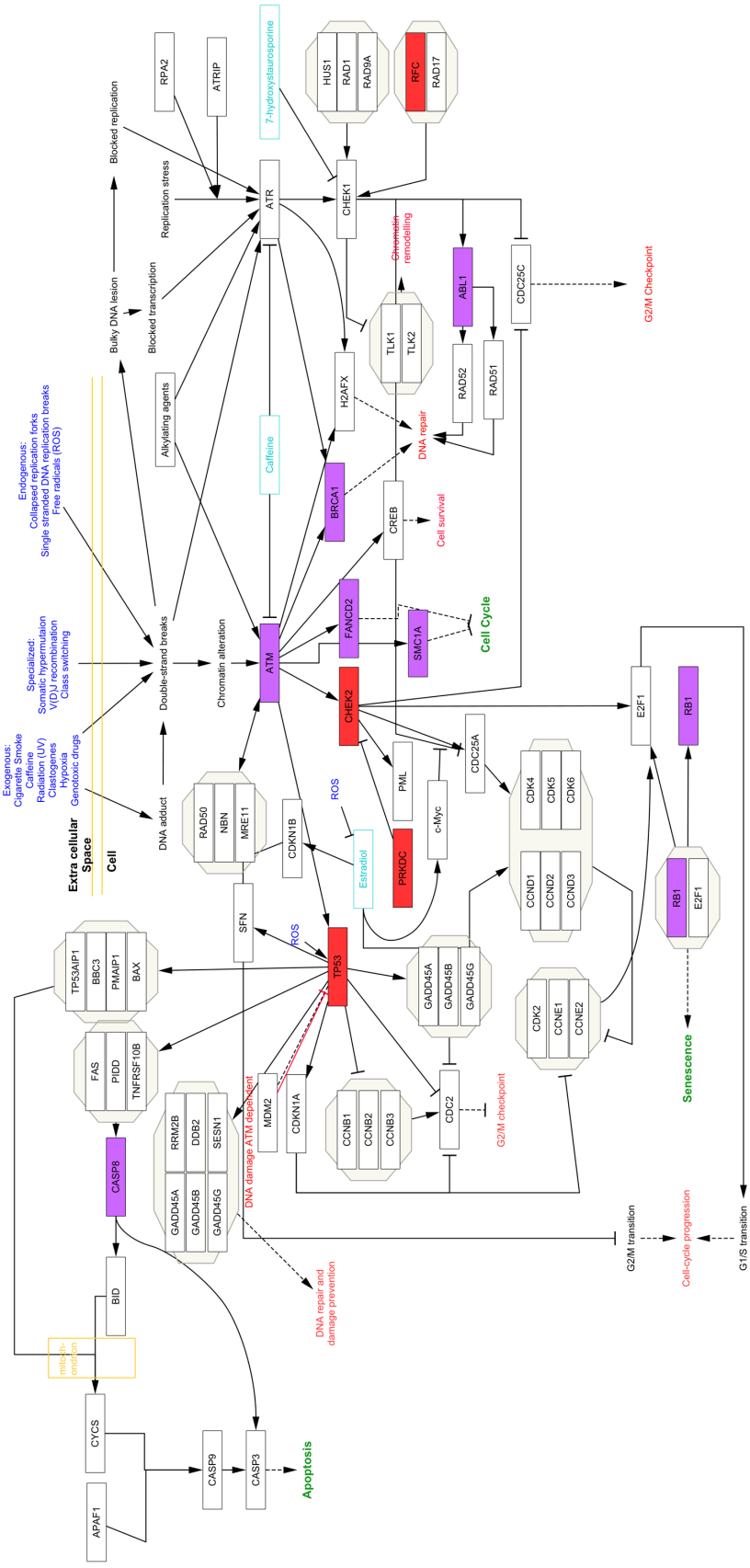


Figure 15. Overlap of the basal driver subnetwork and the DNA damage response pathway. Purple nodes indicate genes in the basal driver subnetwork and red nodes indicate genes that were in the basal driver subnetwork and also had significantly increased phosphorylation.

Subnetworks Unique for Each Subtype

To determine what makes each subtype unique, I further filtered the driver subnetworks to include only edges between nodes that were unique to an individual subtype and overlapped significant changes in phosphorylation. I kept only components consisting of more than 3 nodes.

The basal subnetwork consisted of 6 components with 63 nodes and 58 edges (**Figure 16**). Most components had at least one node with significantly altered phosphorylation in basal samples compared to all other samples. The largest component surrounded the protein estrogen receptor beta (ERβ) encoded by the ESR2 gene.

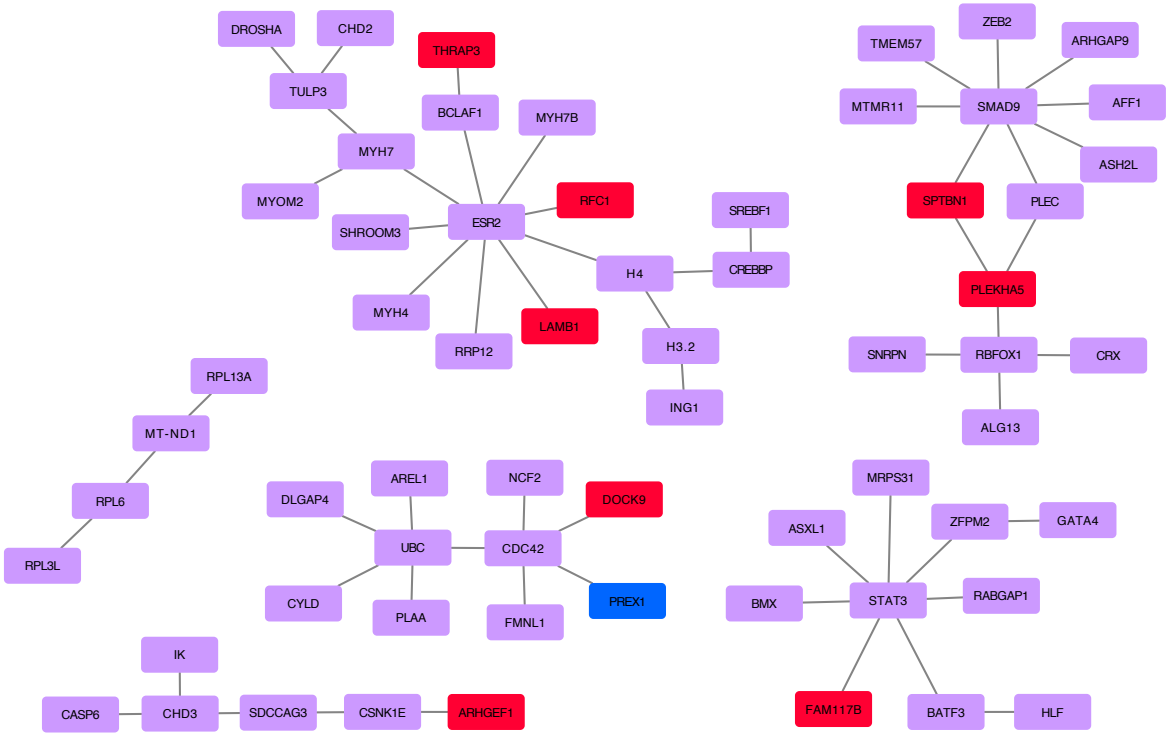


Figure 16. Nodes unique to the basal subnetwork. Nodes with significantly increased phosphorylation sites are colored red, while decreased phosphorylation is indicated by blue.

The luminal A subtype was the only other network to contain nodes with altered phosphorylation levels. It contained 6 components with 43 nodes and 39 edges (**Figure 17**). Unlike the basal subnetwork, most of the phosphorylation sites were downregulated. Furthermore, it contained a network for ESR1, which encodes ERα, instead of ESR2.

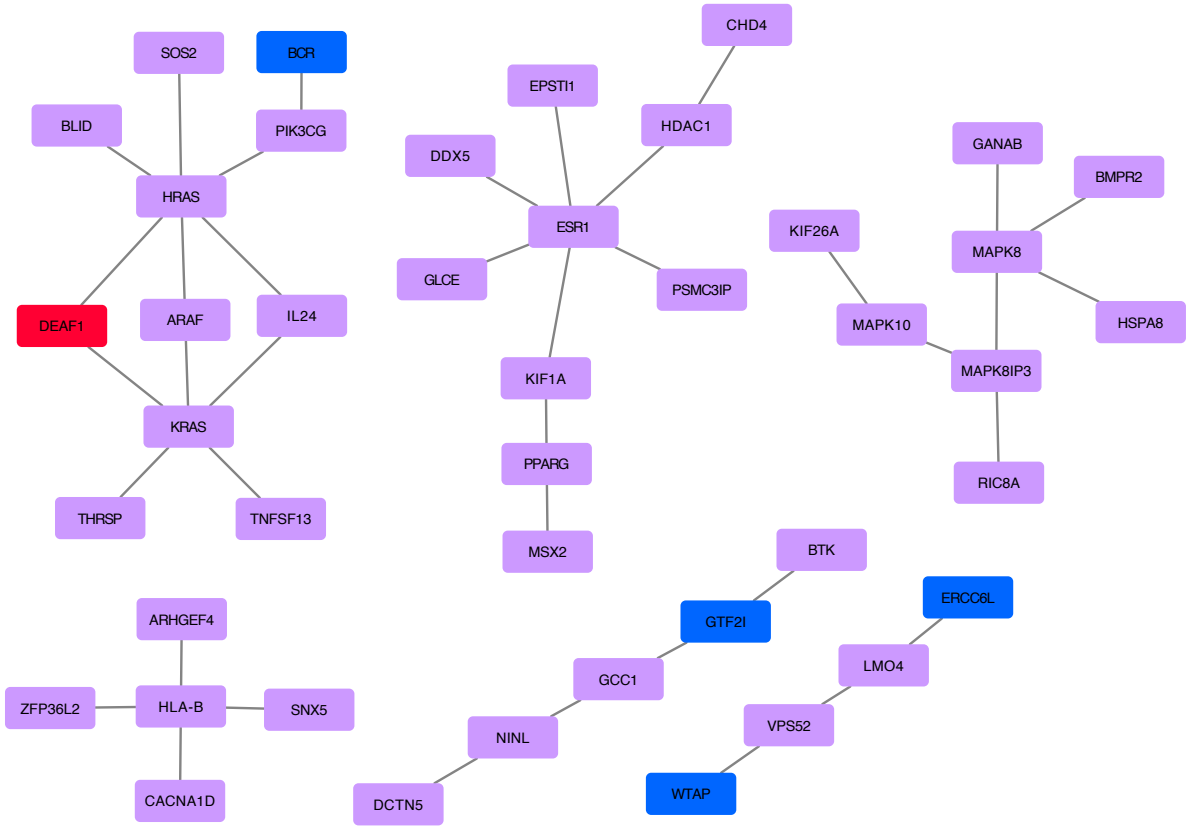


Figure 17. Nodes unique to the luminal A subnetwork. Nodes with significantly increased phosphorylation sites are colored red, while decreased phosphorylation is indicated by blue.

Altered Kinase Substrates

The networks for both basal and luminal A contained a few kinases. To assess the differences in kinase signaling in these networks, I compared the phosphorylation levels of substrates of kinases in each subtype to the other subtypes. Approximately 50 kinases were present in each subtype driver subnetwork and slightly over half had substrates in the phosphoproteomic data. For each subtype, a fraction of the kinases had altered phosphorylation of its substrates in that subtype compared to the others (**Figure 18**).

Interestingly, some of the kinases identified in the unique basal and luminal A subnetworks had substrate data. A single kinase, CSNK1E, was unique to the basal subtype. Phosphorylation of CSNK1E substrates was higher in the basal samples compared to others. The unique luminal A subnetworks had 8 kinases, 5 of which had phosphorylation data. The substrates of ARAF were significantly increased, while the substrates of MAPK10 and BCR were significantly decreased.

The best validation of the identified subnetworks would be if the kinases in the subnetworks were dysregulated only in those subtypes. This was not the case, as the substrates of about 10 kinases not identified in each driver subnetwork were also significantly dysregulated in that subtype.

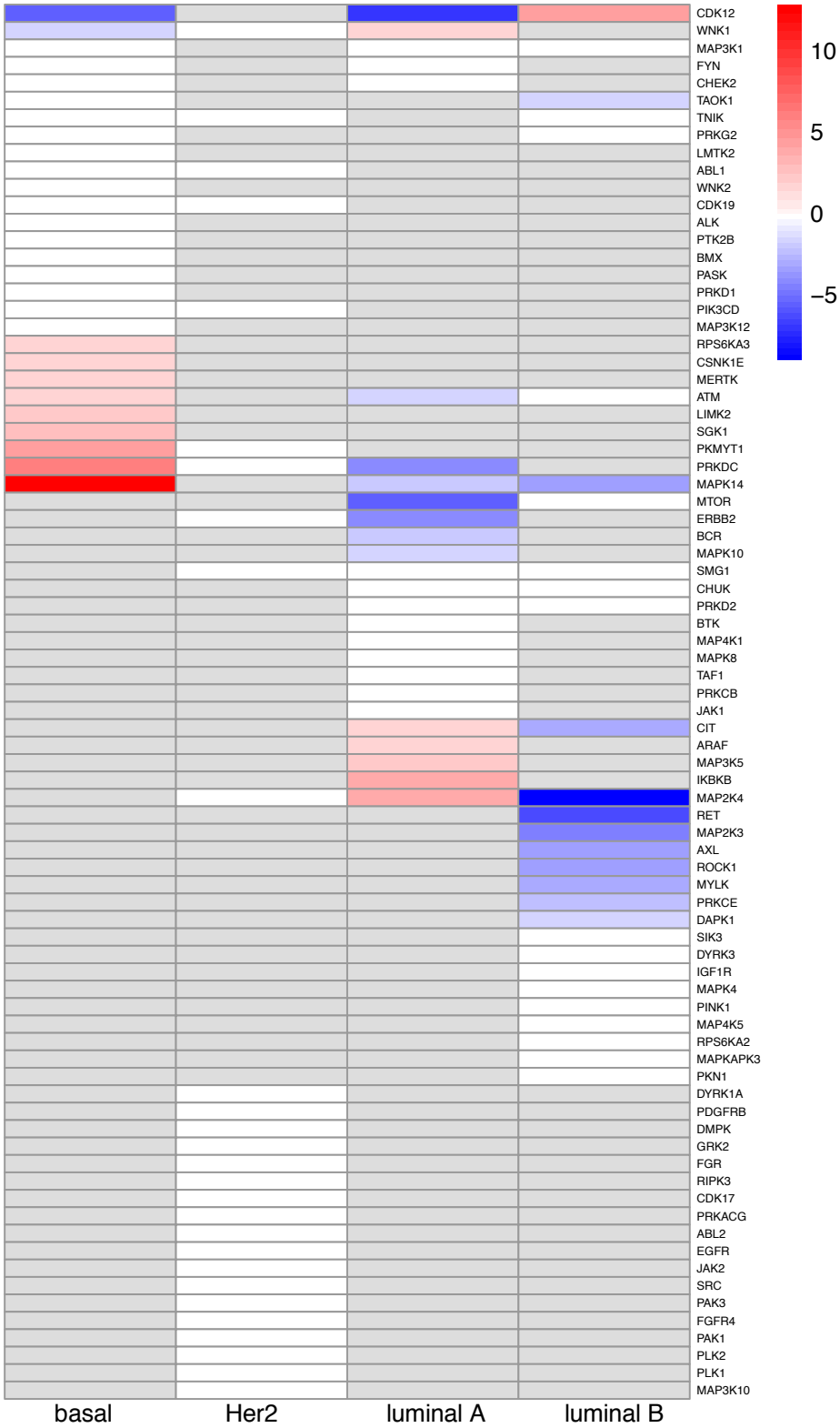


Figure 18. Relative phosphorylation of substrates of kinases. Red indicates significantly increased phosphorylation, blue indicates significantly decreased phosphorylation, and white indicates no change in phosphorylation. Color is based on -log10(FDR). Gray indicates kinases that were not significantly altered in the driver subnetworks. Grey indicates the kinase is not present in that driver subnetwork.

Discussion

Overall, genomic mutations converged on distinct subnetworks in breast cancer molecular subtypes. While phosphoproteomic data had limited utility in validation due to minimal phosphorylation differences and low signal, the data provided interesting insights into the function of these networks.

The driver subnetworks extracted for each subtype covered known dysregulated pathways. For example, the estrogen receptor signaling pathway was enriched in the luminal A subnetwork. Luminal A samples are typically ERα-positive and respond well to endocrine therapy[198]. Furthermore, PI3K is most frequently mutated in ER-positive luminal tumors and this pathway was also enriched in the luminal A subnetwork[199]. Another signaling pathway, the signaling of the hepatocyte growth factor receptor (also known as MET), was enriched in both the luminal A and Her2 subnetworks. Interestingly, MET downregulation increases ERBB2 (also known as Her2) activation[200], which is the defining characteristic of the Her2 subtype. Furthermore, low expression of MET is seen in both the luminal A and Her2 subtypes[201]. Finally, although the ErbB signaling pathway was not enriched in the Her2 subnetwork, ERBB2 was one of the most connected nodes.

The basal subtype is compelling because it has the worst prognosis and is enriched with tumors lacking expression of hormone receptors or ERBB2, which results in the fewest treatment options[202]. The basal driver subnetwork was enriched for genes related to the DNA damage response and this subtype is known to have defects in DNA repair[203]. The phosphoproteomic data indicated high phosphorylation levels of several proteins in this pathway and phosphorylation of substrates of the known DNA repair kinases ATM and PRKDC was upregulated. This might indicate a compensatory mechanism of the kinases in this pathway to regulate DNA damage.

The other pathway enriched in the basal subnetwork was the alpha 6 beta 4 signaling pathway. The alpha 6 beta 4 integrin is highly upregulated in basal breast cancer[204], so mutations in this pathway might be driving integrin expression.

Finally, an interesting area of future exploration for the basal subtype is the ESR2 pathway. ESR2 was uniquely present in the basal driver subnetwork and any of its protein-interaction partners were highly phosphorylated. ERα has known functions in breast cancer and it is rarely expressed in basal breast cancer, unlike in the other subtypes. Much less is known about the function of ERβ in breast cancer, although reports suggest it represses EMT. Since some basal breast cancers have expression of ERβ, treatment with an ERβ agonist might have a positive effect.

The main limitation of this study was validation and phosphoproteomic significance, likely due to sample size. Only 12 Her2 samples out of the total 77 had phosphoproteomic data, which likely limited the significance. Increasing the sample size, better phosphoproteomic resolution, and comparison to normal samples might produce a better signal-to-noise ratio. Furthermore, the heterogeneity within a subtype might contribute to the overall noise and reduce the signal that can be extracted. Exploring sample-specific driver subnetworks might ameliorate this problem. Finally, it is important to note that this analysis was performed comparing one subtype relative to others. While a particular pathway might be downregulated in one subtype compared to the others, it may still be overactive compared to normal and could then still be a target for therapy.

CHAPTER 5

PHOSPHOPROTEOMIC DATA REVEAL A DUAL ROLE FOR RB1 IN COLON CANCER

Introduction

Colorectal cancer is one of the leading causes of cancer and cancer-related deaths in the United States[205]. It consists primarily of adenocarcinomas arising from colon and rectal epithelial cells. Usual treatments include surgery for early stage cancer. For more advanced tumors, chemotherapy with or without targeted therapy is often used in combination with surgery. Chemotherapy consists of some combination of leucovorin, 5-fluorouracil, oxaliplatin, irinotecan, and capecitabine, which all disrupt proper cell division. Additionally, targeted anti-VEGF therapies, such as bevacizumab or aflibercept, or anti-EGFR therapies, such as cetuximab or panitumumab, are used[206].

While the survival rate for colon cancer is high for early stages (> 90% five-year survival), patients with later stages still have poor survival and few targeted therapy options[205]. Global genomic studies on colorectal cancer tried to find new targets by providing frequently mutated genes, identification of fusion genes, and alterations in transcription factor activity[207]. However, these targets have yet to become clinically relevant, partly because the effect of many mutations on protein functions are not yet known.

Proteins are the primary targets in cancer and therefore studying global changes at the protein level should help identify new therapeutic targets. Previously, CPTAC generated proteomic and phosphoproteomic datasets for three cancer types. They then extended the colorectal cancer study by generating data for a new cohort of colon cancer patients with the addition of phosphoproteomic data and matched normal samples for both the proteomic and phosphoproteomic datasets. This allows for removal of baseline protein expression and shows alterations specific to tumor samples.

This new phosphoproteomic data allows colon cancer kinase signaling to be probed. As in most cancers, kinase signaling in colon cancer is dysregulated. Ras and BRAF are both kinases that are highly mutated in colon cancer[208]. Together with upregulation of EGFR, these mutations result in activation of the MAPK pathway[209]. There are also mutations in the PI3K pathway and SRC is known to be overexpressed[210,211].

Furthermore, we used the phosphoproteomic data to explore the effect of phosphorylation on substrates and clarify a contradiction in the colon cancer literature. The tumor suppressor gene retinoblastoma 1 (RB1) produces a protein that is a master cell cycle regulator. In normal cells, RB1 is monophosphorylated and bound to the transcription factor E2F[212]. Upon phosphorylation by members of the CDK family of kinases, RB1 releases E2F to drive transcription of cell-cycle related genes.

In many cancers, RB1 is mutated or deleted, supporting its role in cancer proliferation[213–215]. However, RB1 is rarely mutated in colon cancer and it has been reported to be amplified in about 30% of colon cancer tumors with minimal deletion in other samples[216,217]. Additionally, RB1 mRNA expression and protein was increased in more than 80% of operable colorectal cancer cases[218]. Meanwhile in normal tissue, RB1

expression is confined to a small set of epithelial cells in the transition zone[219]. Furthermore, using western blot, several groups also showed that RB1 is phosphorylated in tumor tissue compared to normal[220,221]. We hypothesized that RB1 might be inactivated by phosphorylation rather than mutation or deletion in colorectal cancer. We used phosphoproteomic data to explore this idea.

Methods

Data

Data were obtained from the CPTAC consortium. The data consisted of 197 individual samples, which included 96 matched tumor-normal colon adenocarcinoma pairs, with copy number, TMT global proteomic, and TMT phosphoproteomic data. Six hundred eighty-eight human kinases and 255 phosphatases were defined as in **Chapter 2**.

Processing Phosphoproteomic Data

Phosphopeptide identification was performed using MS-GF+[222,223] to match against the RefSeq database (version April 2017). Site localization was performed using the Ascore algorithm[224]. Identified peptides were mapped to UniProt sequences (version July 2017) and named according to the canonical UniProt sequence. If the peptide matched multiple canonical UniProt sequences, the best ID was chosen based on presence of the protein in the proteomic data. If no canonical IDs had proteomic data, or if more than one protein was present in the quantified proteomics data, an ID was chosen at random. For peptides not matching a canonical protein sequence, a matching protein isoform ID was chosen. Peptides were filtered to those with an Ascore of ≥ 19 in at least one scan and a Q value < 0.01. Peptide abundances were log2 transformed and zero-centered for each gene. Data were re-centered across all samples to achieve a common median of 0. Phosphorylation site levels were determined by the median level for all peptides matching that site. Quantified sites and proteins were defined as those containing non-missing values in at least 50% of the matched samples. Data quality was assessed using principal component analysis (PCA) with the prcomp function in R.

Tumor vs Normal Analysis

Log fold change was calculated as the log2 peptide ratios for normal samples subtracted from the log2 peptide ratios for tumor samples. Differential expression was performed using the paired Wilcoxon rank sum test. P values were adjusted using the Benjamini-Hochberg correction and an FDR < 0.05 was deemed significant. Phosphorylation site tumor markers were defined as sites upregulated more than 2x fold with an FDR < 0.01.

Term and Pathway Enrichment

Overrepresentation analysis to describe proteins with upregulated and downregulated sites was performed using WebGestaltR with default parameters (minimum overlap 10, maximum 500). The background was the full list of proteins with phosphorylation data passing the same filters. The database for enrichment was Gene

Ontology Biological Process (GO BP). Results were considered significant with FDR < 0.05.

Biomarker Comparison

Proteins containing phosphorylation sites upregulated > 2x fold were compared with protein biomarkers (upregulated > 2x fold and FDR < 0.01 in tumor compared to normal) and genes identified as having mutations and activity relevant to cancer in the Cancer Gene Census (CGC)[225]. Proteome data were quantified using unshared peptides and had non-missing values in at least 50% of tumor-normal samples.

Kinase Activity Prediction Using GSEA

Unique phosphorylation sites were identified as a thirteenmer sequence. Phosphorylation sites of kinases were determined by a union of kinase-substrate interactions in PhosphoSitePlus (version May 2018), HPRD (version 9.0), and Swiss-Prot (version May 2018). The median log2 fold change of sites with at least 50% non-missing values was used to rank the phosphorylation sites and was submitted to the pre-ranked GSEA function in WebGestaltR. A minimum set size of 3 substrates was required and 1000 permutations was used to determine significance. Kinases were considered over-active with NES > 0 and FDR < 0.05. Kinases were considered under-active with an NES < 0 and FDR < 0.05.

Kinase Activity Prediction Using Regulatory Sites

Regulatory sites with known effect on kinase activity were downloaded from Signor. Differentially upregulated sites in tumor vs normal (FDR < 0.05) that were known to activate kinase activity were used to predict increased activity and differentially downregulated activating sites were used to predict decreased activity.

RB1 Characteristics

RB1 CNA and protein data were downloaded from CPTAC. The lollipop plot diagram was created using the R package trackViewer[226].

ssGSEA Activity Scores

Protein activity in individual samples was determined using ssGSEA in the GSVA package in R. Substrate sets for kinases from the GSEA method were used to determine enrichment of phosphorylation sites of kinases. Substrate sets for E2F were downloaded from ENCODE[227] and Hallmark sets were downloaded from MSigDB[228]. For enrichment of E2F targets and the Hallmark pathways, log2 fold change of proteins was used as the ranking method. Ten substrates were required for each analysis.

Correlation of RB1 Features to Cell Line Drug Sensitivity

mRNA expression and cell line drug response were downloaded from the Genomics of Drug Sensitivity in Cancer (GDSC) project[229]. Response to CDK inhibitors was compared between cell lines with the lowest mRNA expression of RB1 (expression < 3) and the highest mRNA expression (expression > 6) using the student's t-test.

Reverse phase protein array (RPPA) phosphorylation data for a subset of these cell lines were downloaded from the MD Anderson Cell Lines Project (MCLP)[230]. RB1 phosphorylation at S807/S811 was correlated with CDK inhibitor response using Pearson correlation.

Results

Phosphorylation Changes in Colon Cancer

The colon cancer dataset consisted of 197 samples (96 tumor samples with matched-normal tissue) with phosphoproteomic data. There were 71,504 unique identified peptides in the mass spectrometry data (**Table 10**). After filtering for high quality peptides, 31,339 unique phosphorylation sites could be identified. Of these sites, 7,298 sites on over 2,500 proteins were present in at least 50% of the matched tumor-normal pairs. Most of these sites (98%) have been identified in other experiments and reported in PhosphoSitePlus.

Feature	Number
Identified peptides	71,504
Filtered peptides	42,790
Phosphorylation sites	31,339
50% paired non-missing	7,295
Phosphorylated serine	6,444
Phosphorylated threonine	786
Phosphorylated tyrosine	65

Table 10. Colon phosphoproteomic data by the numbers.

The phosphorylation patterns differed between tumor and normal samples. PCA analysis showed good discrimination between the two tissue types in the first component (**Figure 19A**). Additionally, median phosphorylation levels across all of the sites were lower in tumor samples compared to normal (**Figure 19B**).

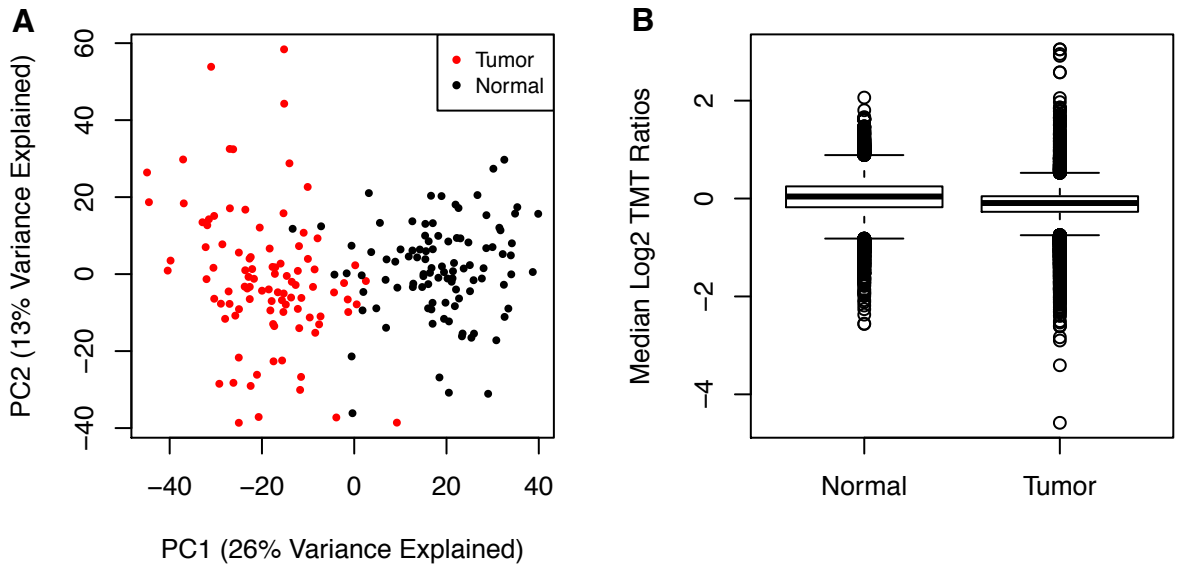


Figure 19. Difference between phosphorylation in colon tumor and normal samples. A) PCA for tumor (red) and normal (black) samples. B) Median log2 phosphorylation site levels in normal and tumor samples.

To assess changes in phosphorylation levels in individual tumors compared to normal tissue, the log2 peptide ratio for normal samples was subtracted from the log2 peptide ratio for cancer samples. Out of the 7,298 quantified sites, 5,895 of these sites also had TMT mass spectrometry protein data. The tumor log fold change in protein abundance and the tumor log fold change in phosphorylation levels were highly correlated (**Figure 20A**, Pearson correlation = 0.81, p value < 2.2×10^{-16}). Using the signed rank Wilcoxon test, differential abundance of phosphorylation sites was determined. Out of the 7,298 quantified sites, 2,363 sites were significantly upregulated in tumor compared to normal, while 3,338 were downregulated (**Figure 20B**). Proteins with phosphorylation site abundance higher in normal were enriched for stromal-related proteins. The top 10 enriched terms for downregulated sites in tumor were all related to cytoskeleton organization and cell locomotion (**Appendix E**, Table E1, FDR < 3×10^{-10} for all 10 terms). This indicated that normal samples likely contained a higher percentage of stromal and muscle cells, while tumor sample composition was primarily epithelial cells. Top enrichment terms for proteins with upregulated sites included RNA processing and splicing and chromatin organization (**Appendix E**, Table E2, FDR < 5×10^{-12}).

There was no enrichment for the 42 proteins with hypophosphorylated sites (indicated in blue in **Figure 20A**), but the 62 proteins with hyperphosphorylated sites (indicated in red in **Figure 20A**) were related to muscle structure and contraction and cytoskeleton.

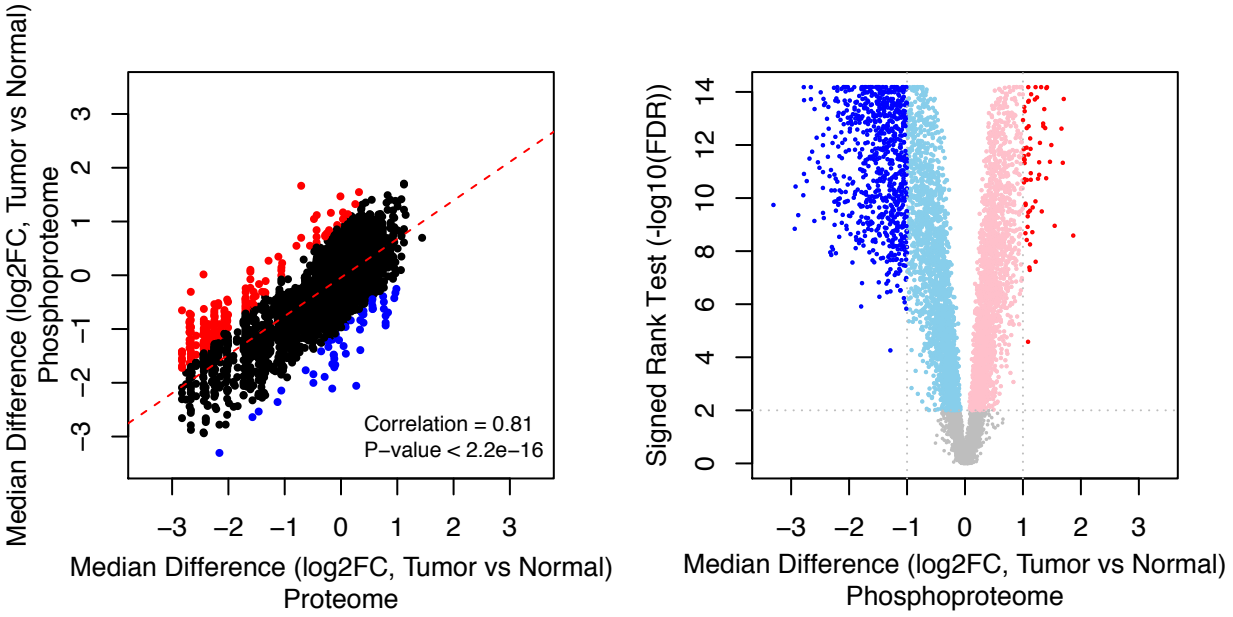


Figure 20. Characterization of phosphorylation sites in colon cancer. A) Median log2 tumor-normal fold change of proteins and sites on those proteins. Phosphorylation levels are highly correlated with protein abundance (Pearson correlation = 0.81, $p < 2.2x10^{-16}$). The dashed red line shows the linear regression between abundances. Red colored points are sites 2x fold above the 45-degree line, while blue colored points are sites 2x fold below the 45-degree line. B) Differential expression of phosphorylation sites in tumor vs normal. Sites are colored based on significance (FDR < 0.01, blue = higher in normal, pink = higher in tumor. Darker color indicates absolute fold change ≥ 2).

Comparison with Proteomic and Genomic Data

There were 63 sites on 50 proteins with > 2x fold change and FDR < 0.01 in tumor samples compared to normal (**Appendix F**). Six genes identified in the CGC as important in cancer based on mutations had highly upregulated phosphorylation sites (DEK, NPM1, PML, RB1, TFRC, CDK12). Four proteins (DDX21, RSL1D1, S100A11, TOP2A) were highly abundant in tumor compared to normal and also had high phosphorylation site levels. No protein existed in all three, suggesting that each type of data added new information about proteins involved in colon cancer (**Figure 21**). Using GO enrichment analysis, the proteins with highly upregulated sites were primarily related to cell cycle and general DNA/RNA processes (**Appendix E**, Table E3).

While the functions of most sites of these sites are unknown, a few highly upregulated sites have been studied in relation to cancer. Phosphorylation of the protein product of gene *NOLC1* at site S563 inhibits the kinase CK2, which might prevent its pro-apoptotic function[231]. Topoisomerase II alpha (TOP2A) is upregulated in proliferating cells and controls the structure of DNA during the cell cycle[232]. Phosphorylation of TOP2A at S1106 enhances its activity[233]. Furthermore, phosphorylation of microtubule-associated protein 4 (MAP4) at S787 promotes its dissociation from tubulin, which results in depolymerization of microtubules[234]. Phosphorylation of this site further associates with resistance to the chemotherapy taxol[235].

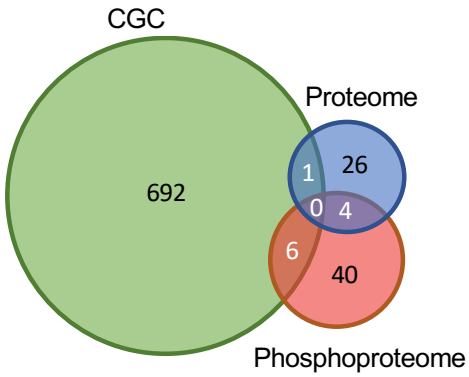


Figure 21. Overlap of cancer-associated genes and colon cancer-associated proteins and phosphoproteins. Number of proteins in the CGC, increased 2x fold in the protein data, or with sites increased 2x fold in the phosphoproteomic data.

Kinase Activity in Colon Cancer

Kinase activity was inferred using enrichment of phosphorylation sites of specific kinases. Out of 78 kinase sets with at least 3 substrates in the colon data, 7 had increased activity and 7 had relatively decreased activity in tumor compared to normal (**Figure 22A**). The kinases with increased activity were primarily cell cycle related proteins (CDK1, CDK2, CDK3, CDK4, CDK6, CDK7, and CDC7). The kinases with relatively decreased activity in tumor were GSK3A and GSK3B, CDK5, MAPK12, DYRK1A, CK1, and PDPK1.

Another way to infer kinase activity is to examine relative phosphorylation of regulatory sites on kinases and phosphatases. In the data, there were quantified sites on 185 kinases (160 of which were protein kinases) and 35 phosphatases. Sixty-one of the sites on kinases and 8 sites on phosphatases had a known effect on the enzyme. Of these sites, activating sites on 5 kinases were upregulated and activating sites on 31 kinases were downregulated in tumor (**Figure 22B**). Only 4 kinases (CDK7, GSK3A, GSK3B, and DYRK1A) were reported to be significantly over or under-active in colon cancer by both methods, while many of the remaining kinases were missing either phosphorylation or substrate information. Inhibitors for most of these kinases have undergone at least a phase I clinical trial, but only a few kinases are targeted by FDA-approved inhibitors (**Figure 22B**).

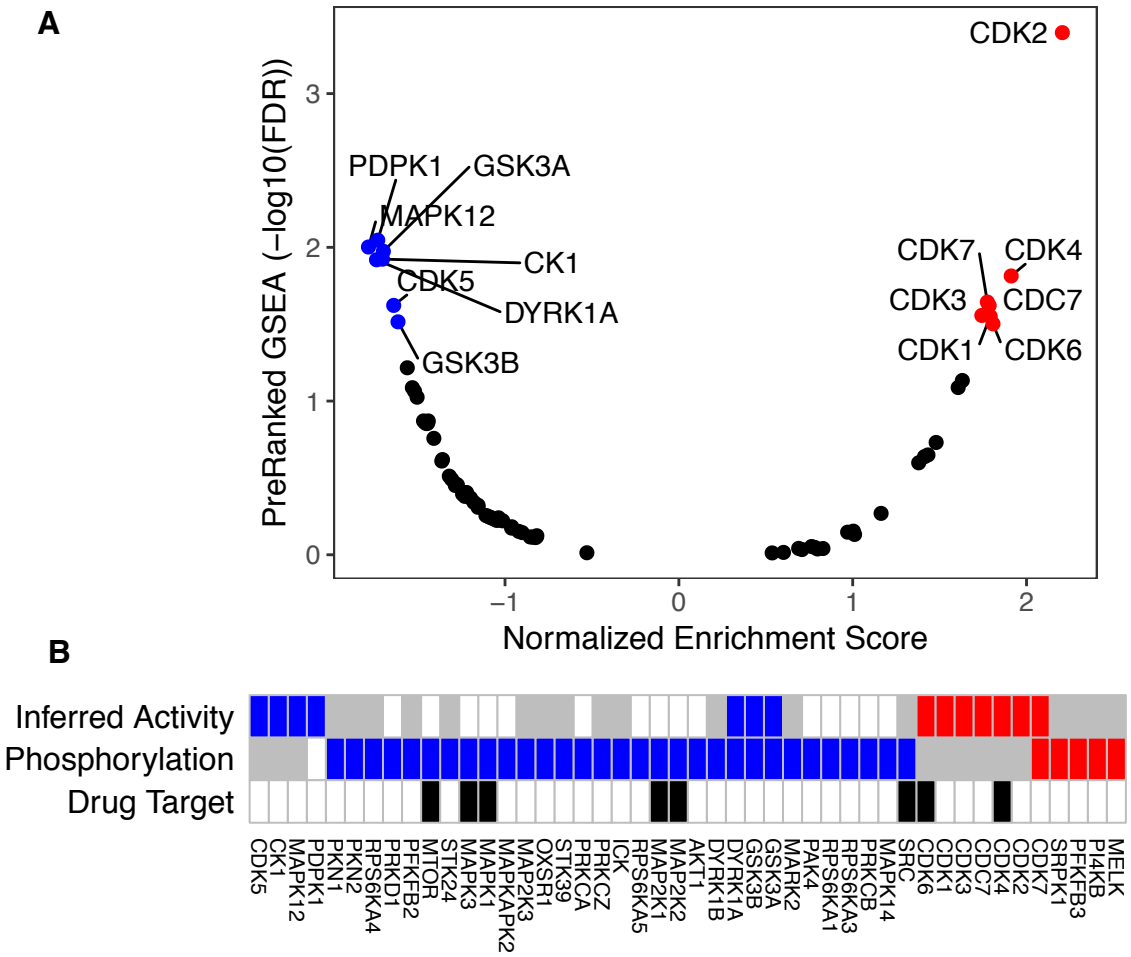


Figure 22. Kinase activity in colon tumor compared to normal. A) Kinase activity inferred using the pre-ranked GSEA method. The normalized substrate enrichment score is used as an activity score. Kinases with significantly different activity in tumor compared to normal are colored and named. Blue indicates relatively decreased activity, while red indicates increased activity. B) Summary of kinase activity using two different methods. Inferred activity are the significant kinases in (A). Phosphorylation indicates kinases with significantly increased (red) or decreased (blue) phosphorylation of their activating sites. White boxes indicate no significant difference. Gray boxes indicate kinase activity cannot be inferred for that kinase using that method. Kinases targeted by an FDA-approved inhibitor are indicated with black boxes.

RB1 Characteristics in Colon Cancer

Because cell cycle kinases were highly active in the colon cancer samples and one of the proteins with multiple highly upregulated phosphorylation sites was the master cell cycle regulator RB1 (**Appendix F**), we hypothesized RB1 was being inactivated by phosphorylation instead of mutation or deletion. The RB1 gene was amplified in a majority of the 96 colon tumor samples (**Figure 23A**). In comparison to normal, RB1 protein was also increased in almost all samples (**Figure 23B**). Furthermore, eleven sites on RB1 were identified in the phosphoproteomic data and almost all were increased in tumor compared to normal (**Figure 23D**). Only six of these sites (S37, S249, T373, S807, S811, T826) were present in > 50% of the paired tumor-normal samples and five of these were

significantly increased in tumor compared to normal (**Figure 23C**). Unlike the other 5 sites, S37 is not known to affect RB1 activity so it was excluded from further analysis.

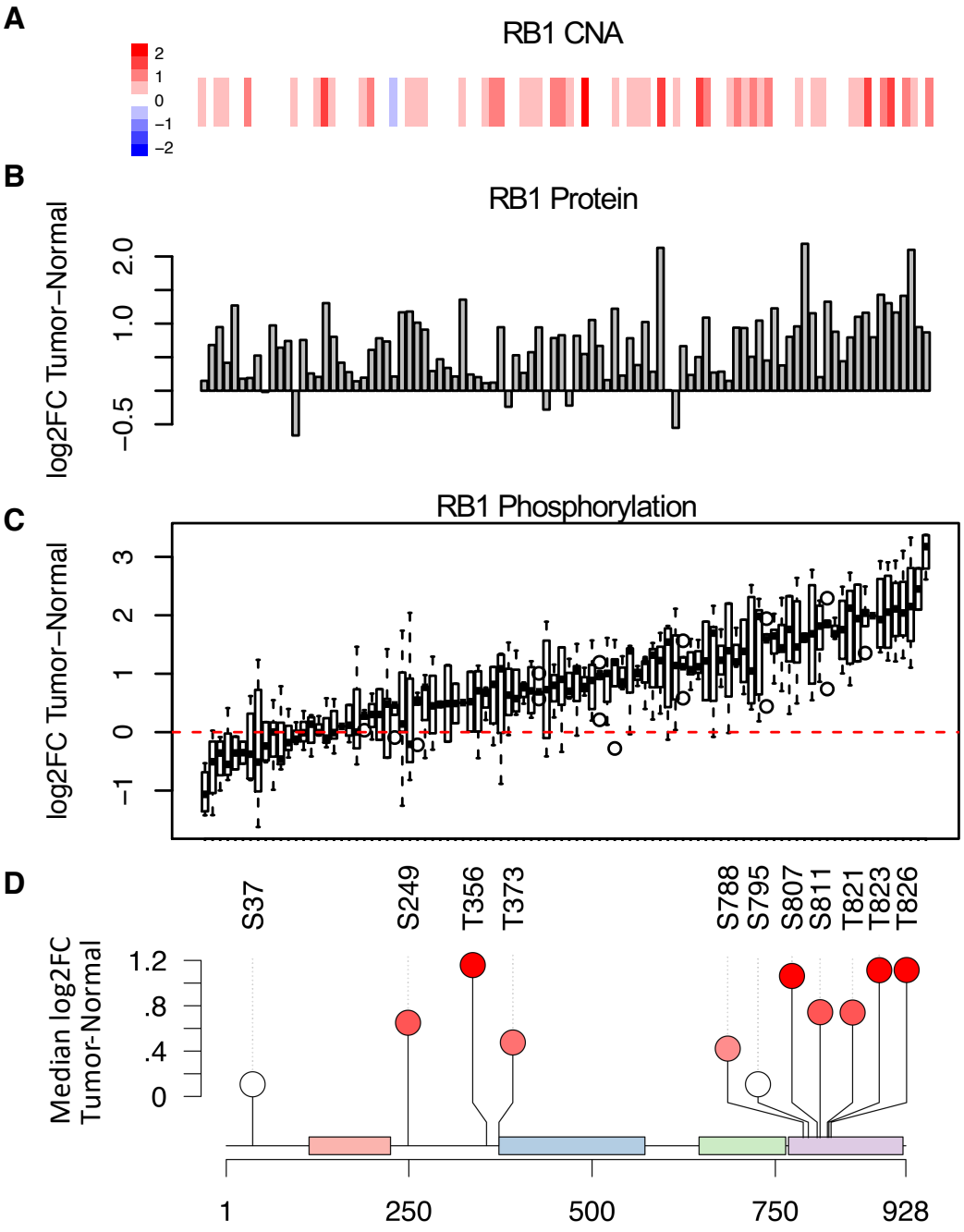


Figure 23. RB1 characteristics in colon cancer. RB1 A) CNA, B) protein fold change, and C) phosphorylation fold change for five sites (S249, T373, S807, S811, T826) across all 96 samples. Samples are ordered by increasing average phosphorylation abundance. D) All identified phosphorylation sites on RB1. Color and height indicate median log2 fold change in tumor compared to normal.

RB1 Phosphorylation Correlates with Increased Proliferation and Decreased Apoptosis

To determine the effect of RB1 phosphorylation, I correlated average RB1 phosphorylation abundance change with other data features. Average RB1 phosphorylation abundance change in tumor compared to normal was highly correlated with CDK2 activity at an individual sample level (**Figure 24A**, Pearson correlation = 0.54, p value = 1.5×10^{-8}). It was also correlated with E2F1 activity (**Figure 24B**, Pearson correlation = 0.36, p value = 3.8×10^{-4}) and tumor vs normal phosphorylation of histone H3 (**Figure 24C**, Pearson correlation = 0.45, p value = 7.2×10^{-4}), which is a marker of proliferation. Finally, average RB1 phosphorylation was negatively correlated with a set of apoptosis proteins (**Figure 24D**, Pearson correlation = -0.29, p value = 4.6×10^{-3}).

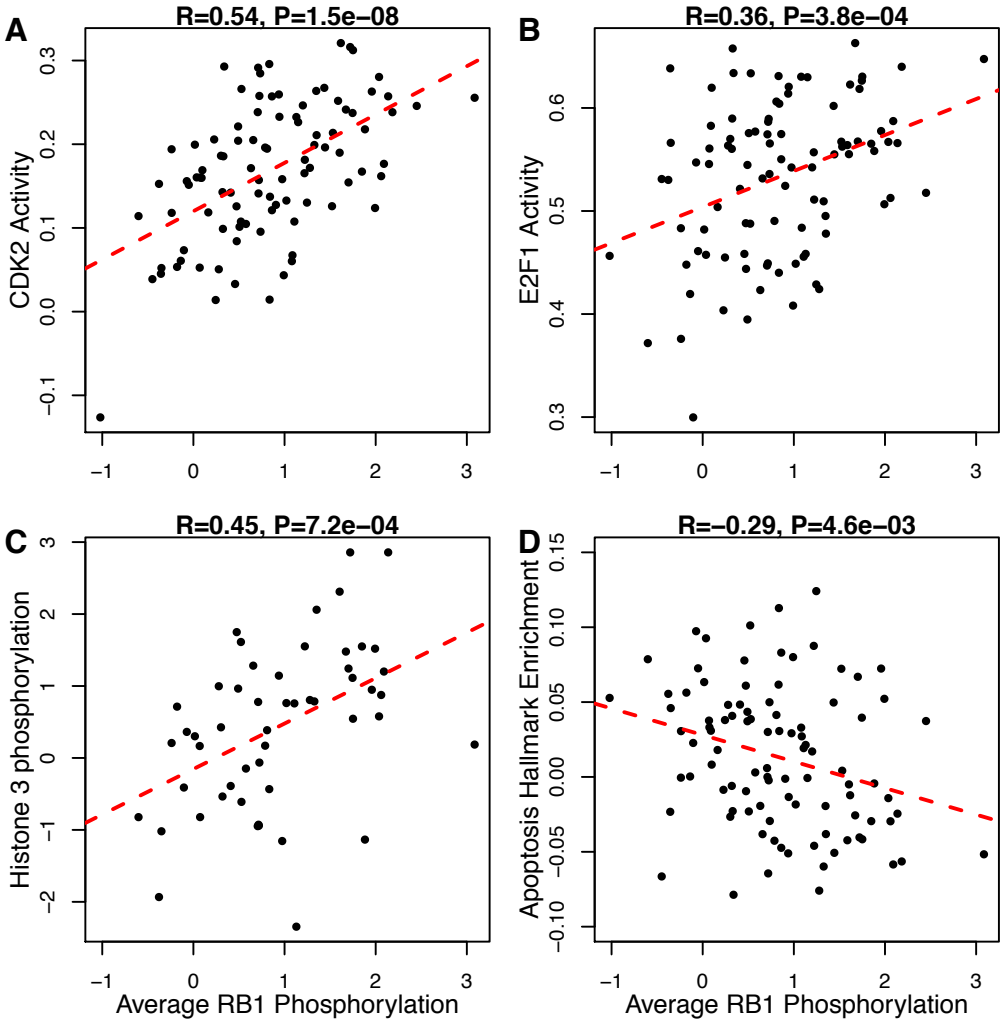


Figure 24. RB1 correlation with proliferation and apoptosis markers. Correlation of average RB1 phosphorylation change in tumor compared to normal correlation with A) ssGSEA activity scores for CDK2, B) ssGSEA enrichment of protein abundance of targets of E2F1, C) phosphorylation of histone H3.1, and D) protein abundance of apoptotic proteins.

Unlike tumors where RB1 is deleted or mutated, RB1 inactivation by phosphorylation is a possible drug target in colon cancer. RB1 could be re-activated by CDK inhibitors. Additionally, RB1 phosphorylation could be used as a marker of response to CDK drugs. Therefore, I examined RB1 characteristics in comparison to cell line response to CDK inhibitors. Using data from the GDSC, cell lines with higher RB1 mRNA expression were significantly more sensitive to two different CDK inhibitors (**Figure 25A-B**). One drug (PHA-793887) targeted CDK2/5/7, while the other (palbociclib) targeted CDK4/6. CDK2, CDK4, and CDK6 are known to phosphorylate RB1. The MCLP project assessed RB1 phosphorylation using RPPA for many of the same cell lines. RB1 phosphorylation also correlated with cell line sensitivity to these two inhibitors (**Figure 25C-D**).

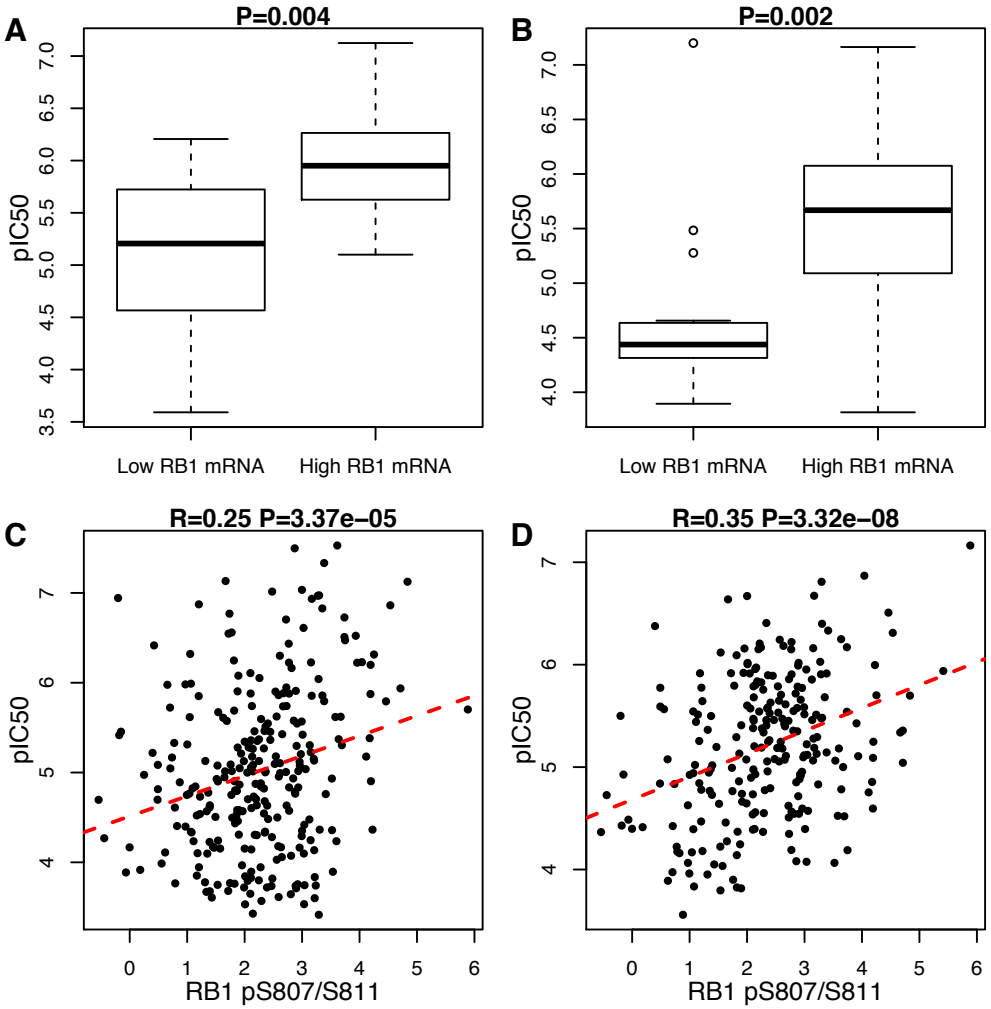


Figure 25. RB1 effect on cell line sensitivity to CDK inhibitors. Comparison of cell line sensitivity to A) PHA-793887 and B) palbociclib between cell lines with low RB1 mRNA expression (RSEM < 3) and high RB1 mRNA expression (RSEM ≥ 6). C) Correlation of RB1 phosphorylation at pS807/S811 and response to PHA-793887 or D) palbociclib.

Summary of RB1 in Colon Cancer

Together, in colon cancer RB1 was amplified at the gene level and this amplification persisted at the protein level (**Figure 26**). Furthermore, RB1 was hyperphosphorylated by active CDK2. This released E2F to drive proliferation but also contributed to inhibition of apoptosis, which gave RB1 two different ways to promote tumor progression and survival.

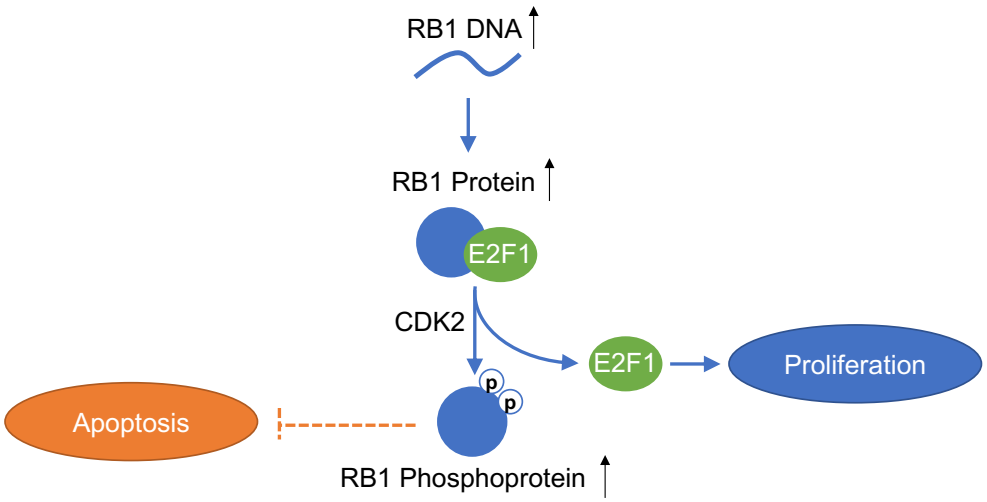


Figure 26. Summary of RB1 activity in colon cancer. RB1 is amplified at the DNA level and also the proteomic level. High levels of CDK2 activity results in increased RB1 phosphorylation, releasing E2F1 to drive proliferation. Furthermore, RB1 contributes to apoptosis suppression.

Discussion

Overall, phosphoproteomic data provided both a complementary and unique perspective of colon cancer compared to other data types. Phosphoproteomic data confirmed some of the colon cancer targets known by genomic or proteomic data. High phosphorylation was found on DDX21, RSL1D1, S100A11, and TOP2A, which were proteins also highly over-expressed in tumor samples compared to normal. TOP2A is involved in the cell cycle and is frequently overexpressed in cancers, including colon cancer[236,237]. DDX21 is an RNA helicase that is also upregulated in cancer, correlates with proliferation, and drives rRNA processing[238]. RSL1D1 is involved in cellular senescence and apoptosis and has also been associated with prostate cancer severity[239]. Finally, S100A11 is a calcium binding protein that helps repair damaged cell membranes[240]. While all of these proteins are upregulated in cancer, they are not frequently mutated and are therefore not known as cancer drivers.

Interestingly, although many of the proteins with highly upregulated phosphorylation sites in cancer were also upregulated at the protein level, some highly abundant sites were on proteins with low abundance in cancer. Tubulin alpha-1B chain (TUBA1B) and its associated protein MAP4 were both highly phosphorylated but downregulated at the protein level in cancer. These phosphorylation sites likely result in destabilized microtubules and promote motility in colon cancer[241]. Sterile α motif and HD domain containing protein 1 (SAMHD1), a nucleotide phosphatase, regulates dNTP homeostasis, has a role in genome stability, and has tumor suppressor functions[242]. In

our study, SAMHD1 was slightly downregulated 1.13x fold, but phosphorylated at T592 2.24x fold. Phosphorylation at this site might further reduce its activity by reducing protein stability[242]. Finally, the kinase MAP3K20, also known as ZAK, had a 3.2x fold increase of phosphorylation at S637, although the protein was downregulated 1.6x fold. ZAK can act as a cell fate switch and its downregulation in lung cancer leads to increased tumor cell survival[243]. However, nothing is known about the phosphorylation at S637, so it could be an interesting topic for future exploration.

Phosphorylation was also high on six known cancer drivers: CDK12, DEK, PML, NPM1, RB1, and TFRC. RB1 in particular was interesting because it is a known tumor suppressor and is unusually amplified in colon cancer. The phosphorylation levels of RB1 explain the mechanism of RB1 inactivation in colon cancer. Furthermore, average RB1 phosphorylation was negatively correlated with apoptosis. RB1 has been linked to apoptosis in the literature[244]. Although its role is not clearly defined, hyperphosphorylated RB1 possibly binds specifically to E2F on apoptotic gene promoters while releasing E2F on cell proliferation gene promoters[245]. It might also control the subcellular localization of the cell survival protein Bag-1[246].

Finally, phosphoproteomic data showed dysregulation of several kinases in colon cancer. As expected, cell cycle regulating kinases are highly active in cancer compared to normal. However, increased phosphorylation of kinase activating sites also suggested dysregulation of a few other kinases. Some of these kinases have been linked to cancer. For example, SRPK1 expression is increased in several cancers and this promotes cancer stemness, angiogenesis, and metastasis[247]. PFKFB3 is also high in proliferating cells and is linked to proliferation and glycolysis[248]. However, these findings should be more carefully validated. MELK appeared to be a good cancer target as it was highly correlated with proliferation and small molecule inhibitors were developed as anti-cancer agents. However, complete knock-down of MELK had no effect on cell survival or proliferation rates and therefore is unlikely to have an effect on cancer patients[249].

CHAPTER 6

CONCLUSIONS AND FUTURE DIRECTIONS

Summary

Throughout this thesis, I have described the ways phosphoproteomic data can be used to identify single interesting kinases, altered signaling pathways, and ultimately pathway dysregulation at the disease level. Phosphoproteomic data provides complementary information, extends that information, and occasionally contradicts the information generated by genomic, transcriptomic, and proteomic data. Furthermore, phosphoproteomic data validates existing knowledge, generates new areas of exploration, and highlights possible therapeutic targets. This chapter summarizes the findings from this work and provides future research directions.

Evaluation of Technology

While this work was clearly limited by sample size and signal resolution above noise, mass spectrometry technology continues to rapidly improve. Ten years ago, most experiments generated around 1000 phosphorylation sites[250–252]. Now, improvements in machine sensitivity, algorithms, and enrichment methods have expanded this 10x fold to over 10,000 quantifiable sites per experiment. Most kinases and phosphatases can already be identified using mass spectrometry and with the continued interest in this technology, the methods and analysis tools will only improve. One current extension is the enrichment and identification of kinases using kinase inhibitors attached to beads[253]. This allows for focused analysis on kinase signaling using small amounts of sample.

Improvements in bioinformatics tools for phosphoproteomic analysis will also promote future discoveries. Tools have thus far focused on very specific areas of kinase signaling. Future work should integrate phosphatases and combine various methods. For example, tools could be developed that assess phosphorylation abundance after mutations or that allow users to visualize changes over entire signaling pathways rather than at the individual kinase level.

Additionally, while many databases have collected information on kinases and phosphatases, programmatic use of this information is challenging. There are no standard terms for functions or regulation and most databases do not have downloadable data. Future work to identify phosphorylation sites on kinases that affect enzymatic activity and characterization of the regulatory mechanisms of each kinase's activity will help improve kinase activity inference. Furthermore, new tools could integrate the various mechanisms for kinase activity inference, such as prediction from substrate phosphorylation and altered phosphorylation of regulatory sites. In my study, these two methods were complementary, and many kinases could only be identified using one of the methods. The combination might improve inference for kinases with few substrates or for kinases lacking information on their regulatory sites.

Standardization is also a challenge for phosphorylation site identification. Phosphorylation sites are usually identified by their position in the protein sequence. However, protein sequences vary across databases and are frequently updated.

Additionally, position varies based on the splice isoform and the species. Identifying phosphorylation sites is a challenge when using multiple databases or analyzing different phosphoproteomic datasets generated using different protein sequence databases. Using the sequence surrounding a site helps, but the sequence can be identical for related proteins and slight variations can occur due to polymorphisms.

Finally, the methods used here can easily be extended to study other PTMs or signaling pathways. Other PTMs are just as important as phosphorylation in both understanding cell signaling status and protein function. Additionally, many can affect the phosphorylation levels of cells. For example, a bacteria acetyltransferase can acetylate MAP2K and prevent its activating phosphorylation[254]. Finally, this work focused specifically on kinase signaling. However, other types, such as response to calcium or GPCR signaling are both important and frequently integrate with kinases. To understand the overall picture of cell signaling, future work could better integrate these pathways.

Kinase Signaling in Cancer

One overall result from this work was the acknowledgement that kinase signaling clearly differs among subgroups of patients. These data could be used to stratify patients and understand their response to treatment. For example, patients with high levels of phosphorylated RB1 will likely respond well to CDK inhibitors, while patients with low levels of RB1 phosphorylation or protein may have little effect. Additionally, patients with kinases predicted to be highly active based on substrate phosphorylation might respond well to inhibitors of those kinases.

Moreover, genomic mutations drive alterations in signaling pathways and these might be used to predict alterations to kinase signaling. Conversely, phosphorylation of kinase substrates could be used to predict the effect of understudied protein mutations. These can help predict how a patient will respond to therapy. Ideally, future work will connect kinase activity predictions from phosphoproteomic data with response to inhibitor treatment to help validate the activity predictions.

Kinase Signaling Comparison Among Tissues

The addition of normal tissue greatly improves the signal-to-noise ratio. Normal tissue helps remove the general background level for an individual patient. Furthermore, stoichiometric analyses of phosphorylation are an exciting future direction. In this work, increased phosphorylation could result from increased protein abundance and/or hyper-phosphorylation of that individual site. Increasing phosphorylation regardless of protein abundance more clearly shows the actual activity of that protein. While all of these studies did have proteomic data, correcting for protein abundance is a challenge. Correction for protein abundance in the colon cancer dataset showed mostly altered abundance on cytoskeletal and stromal proteins. This might be due to tissue composition affecting normalization. Mechanisms to improve standardization, quantification, and normalization might ameliorate these effects.

APPENDIX A

ROC CURVE FOR SUBSTRATE PREDICTION OF PKC

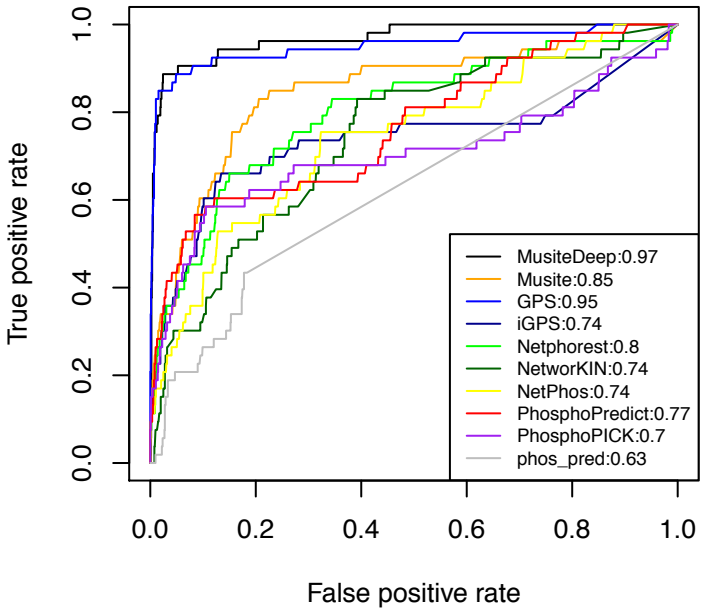


Figure S1. ROC curve for substrate prediction of PKC. The false positive and true positive rates for PKC substrate prediction. The AUC for each tool is listed next to the tool name.

APPENDIX B

WEBSITE URLS FOR BIOINFORMATICS TOOLS

Name	Website URL
14-3-3-Pred	http://www.compbio.dundee.ac.uk/1433pred
ANIA	https://ania-1433.lifesci.dundee.ac.uk/prediction/webserver/index.py
CEASAR	http://www.phosphonetworks.org
CellNOpt	http://www.cellnopt.org
CLUE	https://cran.r-project.org/web/packages/ClueR/index.html
CPhos	https://hpcwebapps.cit.nih.gov/CPhos/
CrossCheck	http://www.proteinguru.com/toolbox/crosscheck/
dbPAF	http://dbpaf.biocuckoo.org
dbPTM	http://dbptm.mbc.nctu.edu.tw
DEPOD	http://depod.bioss.uni-freiburg.de
DrugKiNET	http://www.drugkinet.ca
DynaPho	http://140.112.52.89/dynapho/?p=100&tid=none
eFIP	http://research.bioinformatics.udel.edu/eFIPonline/index.php
EKPD	http://ekpd.biocuckoo.org
GPS	http://gps.biocuckoo.org
GPS-Polo	http://polo.biocuckoo.org
HMMpTM	http://aias.biol.uoa.gr/HMMpTM/
HPRD	http://hprd.org
HuPho	http://hupho.uniroma2.it
iGPS	http://igps.biocuckoo.org
IKAP	https://github.com/marcel-mischnik/IKAP
K-Map	http://tanlab.ucdenver.edu/kMap
KANPHOS	https://kanphos.neuroinf.jp
KEA2	http://www.maayanlab.net/KEA2
KIDFamMap	http://gemdock.life.nctu.edu.tw/KIDFamMap/
Kin-Driver	http://kin-driver.leloir.org.ar/index.php
Kinannote	https://sourceforge.net/projects/kinannote/
KinaseNET	http://www.kinasenet.ca
KinasePA	http://www.maths.usyd.edu.au/u/pengyi/software/KinasePA.html
KinasePhos2.0	http://kinasephos2.mbc.nctu.edu.tw
KinBase	http://kinase.com/web/current/
KinConform	https://github.com/esbg/kinconform
KinG	http://king.mbu.iisc.ernet.in
KinMap	http://www.kinhub.org/kinmap/
KinMutBase	http://structure.bmc.lu.se/idbase/KinMutBase/

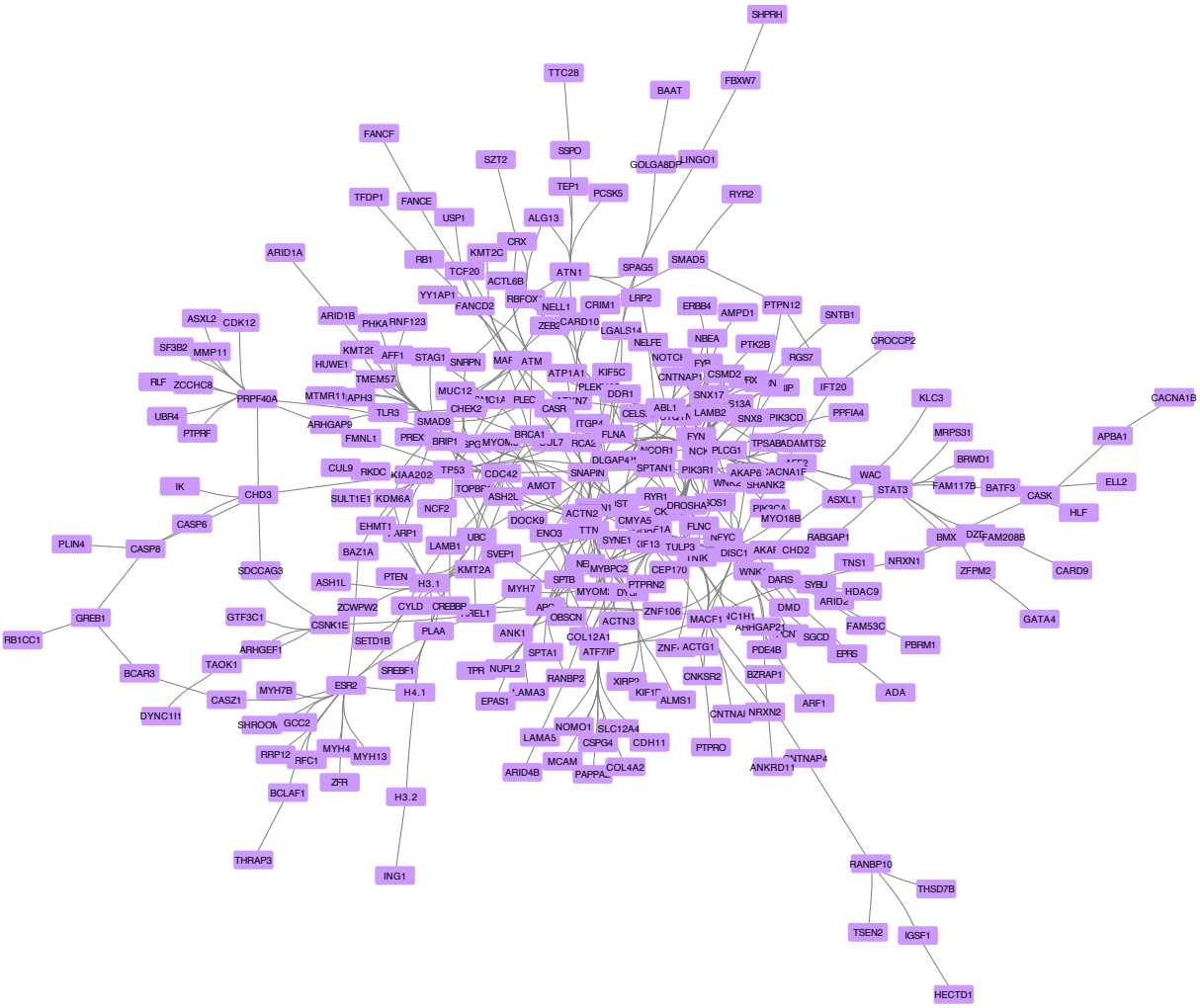
Kinome NetworkX	https://bioinfo.uth.edu/kinomenetworkX/
Kinomer	http://www.compbio.dundee.ac.uk/kinomer/
KinomeSelector	http://kinomeselector.jensenlab.org/index.html
KinWeb	http://www.itb.cnr.it/kinweb/index.htm
KLIFS	http://klifs.vu-compmedchem.nl
KSD	http://sequoia.ucsf.edu/ksd/
KSEA	https://casecpb.shinyapps.io/ksea/
KSP-PUEL	https://github.com/PengyiYang/KSP-PUEL
MIMP	http://mimp.baderlab.org
MOKCa	http://strubiol.icr.ac.uk/extra/mokca/
Musite	http://musite.net
MusiteDeep	https://github.com/duolinwang/MusiteDeep
NetPhorest	http://netphorest.info
NetPhos	http://www.cbs.dtu.dk/services/NetPhos/
NetworKIN	http://networkin.info/index.shtml
phos_pred	http://bioinformatics.ustc.edu.cn/phos_pred/
Phos3D	http://phos3d.mpimp-golm.mpg.de
PhoScan	http://bioinfo.au.tsinghua.edu.cn/phoscan/
PhosD	http://comp-sysbio.org/phosd/
PhosFox	https://bitbucket.org/phintsan/phosfox
PHOSIDA	http://141.61.102.18/phosida/index.aspx
Phosphatome	http://phosphatome.net/3.0/
Phospho.ELM	http://phospho.elm.eu.org/index.html
Phospho3D	http://www.phospho3d.org
PhosphoAtlas	http://cancer.ucsf.edu/phosphoatlas
PhosphoLogo	https://hpcwebapps.cit.nih.gov/PhosphoLogo/
PhosphoNET	http://www.phosphonet.ca
PhosphoNetworks	http://www.phosphonetworks.org
PhosphoPep	http://www.phosphopep.org
PhosphoPICK	http://bioinf.scmb.uq.edu.au/phosphopick/phosphopick
PhosphoPICK-SNP	http://bioinf.scmb.uq.edu.au/phosphopick/snpanalysis
PhosphoPredict	http://phosphopredict.erc.monash.edu/webserver.html
PhosphoSitePlus	http://www.phosphosite.org
PhosphoSVM	http://sysbio.unl.edu/PhosphoSVM/
PhosPred-RF	http://server.malab.cn/PhosPred-RF/index.jsp
PhosSNP	http://phossnp.biocuckoo.org
PHOXTRACK	http://phoxtrack.molgen.mpg.de
pkaPS	http://mendel.imp.ac.at/pkaPS/
PKIS	http://bioinformatics.ustc.edu.cn/pkis/
PPSP	http://ppsp.biocuckoo.org

Predikin	http://predikin.biosci.uq.edu.au
ProKinO	http://vulcan.cs.uga.edu/prokino/about/browser
ProteomeScout	https://proteomescout.wustl.edu/
PSEA	http://bioinfo.ncu.edu.cn/PKPred_Home.aspx
PTMfunc	http://ptmfunc.com
PyTMs	https://pymolwiki.org/index.php/Pytms
RegPhos	http://regphos.mbc.nctu.edu.tw/
RegPhos2.0	http://csb.cse.yzu.edu.tw/RegPhos2/
ReKINect	http://rekinect.info/home
RLIMS-P	http://research.bioinformatics.udel.edu/rlimsp/
Scansite	http://scansite4.mit.edu/#home
SELPHI	http://rothwebprod.mshri.on.ca:8081
Signor	http://signor.uniroma2.it
Sorad	http://research.cs.aalto.fi/csb/software/sorad/
SubPhosDB	http://bioinfo.ncu.edu.cn/SubPhosDB.aspx
Swiss-Prot	http://www.uniprot.org
wKinMut2	http://kinmut2.bioinfo.cnio.es/KinMut2

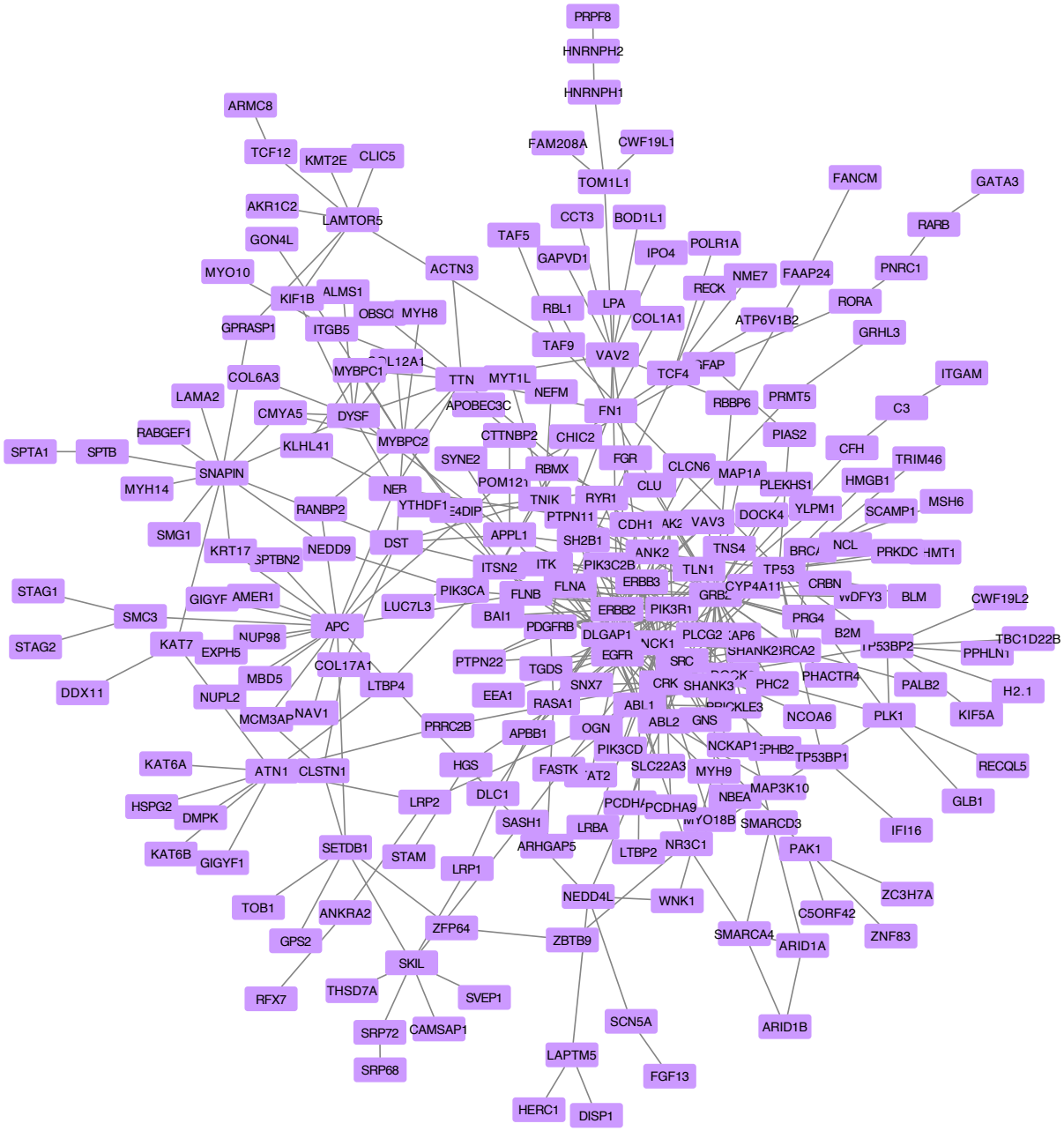
APPENDIX C

BREAST CANCER SUBTYPE DRIVER SUBNETWORKS

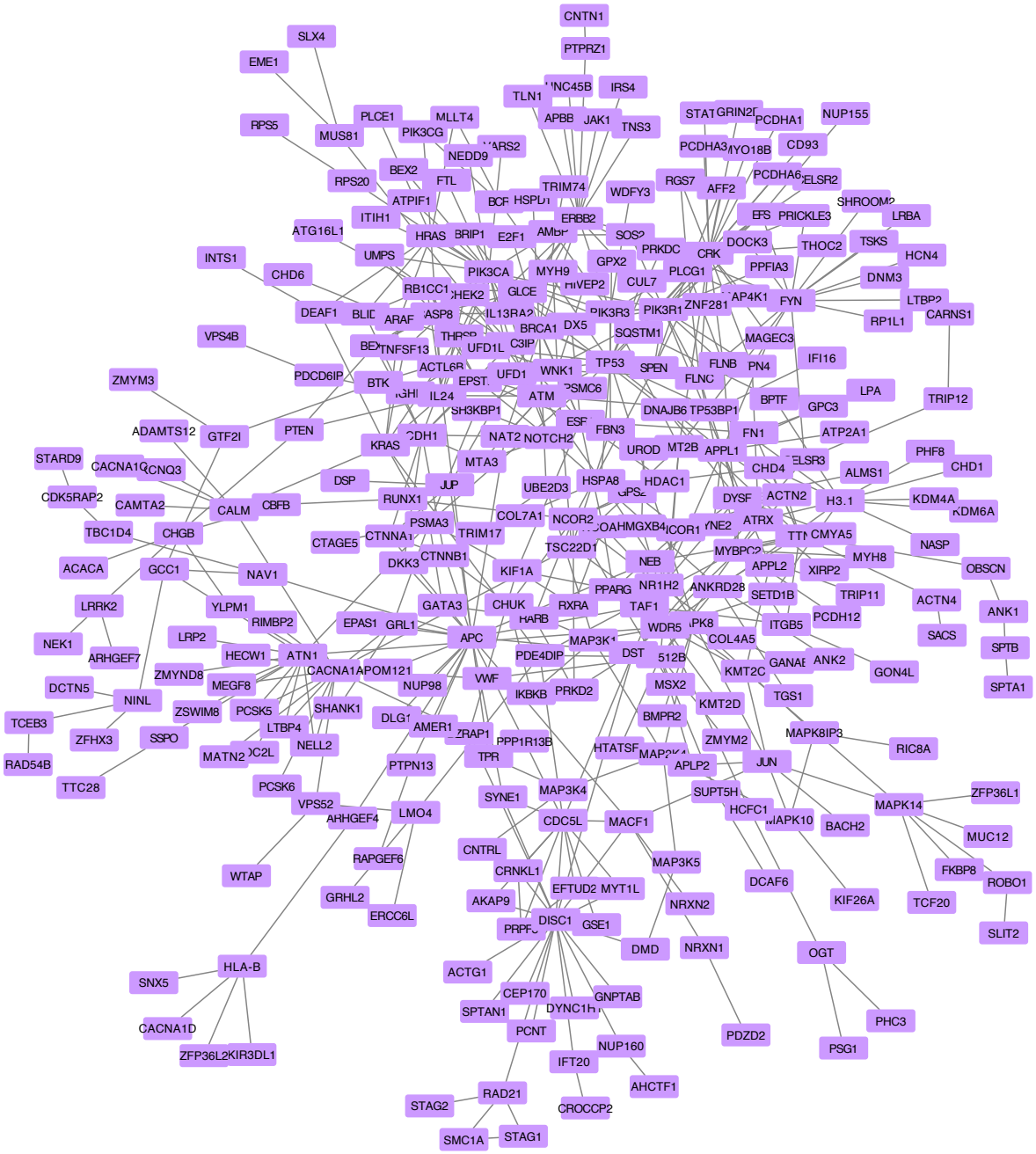
Basal Driver Subnetwork



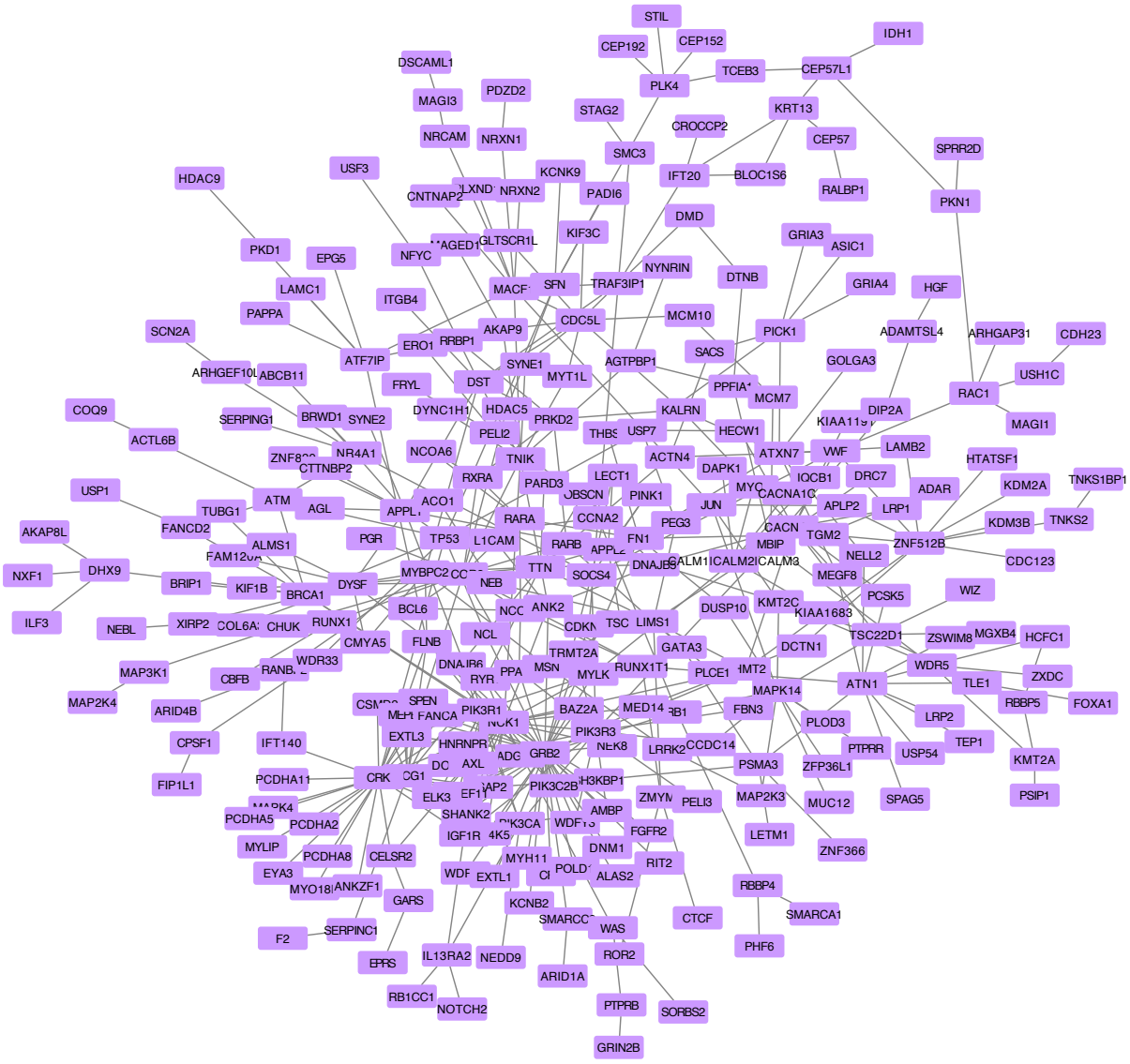
Her2 Driver Subnetwork



Luminal A Driver Subnetwork



Luminal B Driver Subnetwork



APPENDIX D

WIKIPATHWAYS ENRICHMENT RESULTS FOR SUBTYPE DRIVER SUBNETWORKS

Table D1. WikiPathways enrichment in the basal driver subnetwork.

Gene Set	Description	Overlap	Enrichment Score	Normalized Enrichment Score	P Value	FDR
WP3651	Pathways Affected in Adenoid Cystic Carcinoma	15	3.2	4.7	2.7E-07	7.2E-05
WP244	Alpha 6 Beta 4 signaling pathway	8	1.7	4.7	1.9E-04	2.6E-02
WP314	Fas Ligand (FasL) pathway and Stress induction of Heat Shock Proteins (HSP) regulation	9	2.4	3.7	5.1E-04	3.5E-02
WP2377	Integrated Pancreatic Cancer Pathway	22	10.5	2.1	6.2E-04	3.5E-02
WP2118	Arrhythmogenic Right Ventricular Cardiomyopathy	10	3.1	3.2	7.6E-04	3.5E-02
WP383	Striated Muscle Contraction	8	2.1	3.9	7.8E-04	3.5E-02
WP707	DNA Damage Response	11	3.8	2.9	1.1E-03	4.1E-02
WP710	DNA Damage Response (only ATM dependent)	14	5.6	2.5	1.2E-03	4.1E-02

Table D2. WikiPathways enrichment in the Her2 driver subnetwork.

Gene Set	Description	Overlap	Enrichment Score	Normalized Enrichment Score	P Value	FDR
WP2261	Signaling Pathways in Glioblastoma	17	4.8	3.5	3.2E-06	0.001
WP3303	Rac1/Pak1/p38/MMP-2 pathway	14	3.8	3.7	1.4E-05	0.002
WP306	Focal Adhesion	24	9.6	2.5	2.0E-05	0.002
WP3651	Pathways Affected in Adenoid Cystic Carcinoma	12	3.2	3.7	5.9E-05	0.004
WP2377	Integrated Pancreatic Cancer Pathway	23	10.6	2.2	2.9E-04	0.013
WP2526	PDGF Pathway	9	2.3	3.9	3.1E-04	0.013
WP3680	Association Between Physico-Chemical Features and Toxicity Associated Pathways	11	3.3	3.3	3.3E-04	0.013
WP313	Signaling of Hepatocyte Growth Factor Receptor	8	2.0	3.9	7.0E-04	0.024

Table D3. WikiPathways enrichment in the luminal A driver subnetwork.

Gene Set	Description	Overlap	Enrichment Score	Normalized Enrichment Score	P Value	FDR
WP3651	Pathways Affected in Adenoid Cystic Carcinoma	21	3.6	5.8	6.1E-12	1.7E-09
WP710	DNA Damage Response (only ATM dependent)	21	6.3	3.3	6.0E-07	8.1E-05
WP3844	PI3K-AKT-mTOR signaling pathway and therapeutic opportunities	11	1.9	5.7	1.1E-06	9.8E-05
WP2118	Arrhythmogenic Right Ventricular Cardiomyopathy	14	3.5	4.0	4.2E-06	2.8E-04
WP2377	Integrated Pancreatic Cancer Pathway	28	11.8	2.4	1.0E-05	5.5E-04
WP3915	Angiopoietin Like Protein 8 Regulatory Pathway	21	8.2	2.6	4.4E-05	1.6E-03
WP481	Insulin Signaling	24	10.1	2.4	4.7E-05	1.6E-03
WP422	MAPK Cascade	9	1.9	4.8	5.0E-05	1.6E-03
WP2261	Signaling Pathways in Glioblastoma	16	5.3	3.0	5.2E-05	1.6E-03
WP3303	Rac1/Pak1/p38/MMP-2 pathway	13	4.2	3.1	1.9E-04	5.2E-03
WP314	Fas Ligand (FasL) pathway and Stress induction of Heat Shock Proteins (HSP) regulation	10	2.7	3.7	2.5E-04	6.3E-03
WP306	Focal Adhesion	23	10.7	2.2	3.0E-04	6.4E-03

WP1971	Integrated Cancer Pathway	10	2.8	3.6	3.1E-04	6.4E-03
WP2034	Leptin signaling pathway	14	5.0	2.8	3.3E-04	6.4E-03
WP23	B Cell Receptor Signaling Pathway	16	6.3	2.6	3.8E-04	6.8E-03
WP673	ErbB Signaling Pathway	11	3.4	3.2	4.0E-04	6.8E-03
WP437	EGF/EGFR Signaling Pathway	22	10.4	2.1	5.3E-04	8.4E-03
WP1984	Integrated Breast Cancer Pathway	20	9.1	2.2	6.0E-04	9.1E-03
WP2526	PDGF Pathway	9	2.5	3.6	6.6E-04	9.4E-03
WP2380	Brain-Derived Neurotrophic Factor (BDNF) signaling pathway	20	9.3	2.2	7.3E-04	9.9E-03
WP383	Striated Muscle Contraction	8	2.3	3.4	1.7E-03	2.2E-02
WP615	Senescence and Autophagy in Cancer	15	6.5	2.3	1.8E-03	2.2E-02
WP1545	miRNAs involved in DNA damage response	5	1.0	5.0	2.2E-03	2.5E-02
WP195	IL-1 signaling pathway	10	3.5	2.8	2.2E-03	2.5E-02
WP2203	Thymic Stromal LymphoPoietin (TSLP) Signaling Pathway	9	3.0	3.0	2.4E-03	2.5E-02
WP712	Estrogen signaling pathway	6	1.5	4.1	2.5E-03	2.5E-02
WP2018	RANKL/RANK (Receptor activator of NFKB (ligand)) Signaling Pathway	10	3.6	2.8	2.5E-03	2.5E-02
WP3680	Association Between Physico-Chemical Features and Toxicity Associated Pathways	10	3.7	2.7	2.9E-03	2.8E-02
WP1544	MicroRNAs in cardiomyocyte hypertrophy	12	5.0	2.4	3.5E-03	3.3E-02
WP1403	AMP-activated Protein Kinase (AMPK) Signaling	10	3.8	2.6	3.8E-03	3.4E-02
WP2857	Mesodermal Commitment Pathway	15	7.1	2.1	3.9E-03	3.4E-02
WP3668	Hypothesized Pathways in Pathogenesis of Cardiovascular Disease	6	1.6	3.8	4.0E-03	3.4E-02
WP382	MAPK Signaling Pathway	20	10.7	1.9	4.4E-03	3.7E-02
WP2036	TNF related weak inducer of apoptosis (TWEAK) Signaling Pathway	8	2.7	2.9	4.8E-03	3.7E-02
WP2032	Human Thyroid Stimulating Hormone (TSH) signaling pathway	10	3.9	2.5	4.9E-03	3.7E-02
WP2870	Extracellular vesicle-mediated signaling in recipient cells	6	1.7	3.6	5.0E-03	3.7E-02
WP75	Toll-like Receptor Signaling Pathway	13	6.0	2.2	6.0E-03	4.2E-02
WP1433	Nucleotide-binding Oligomerization Domain (NOD) pathway	7	2.3	3.1	6.1E-03	4.2E-02
WP313	Signaling of Hepatocyte Growth Factor Receptor	7	2.3	3.1	6.1E-03	4.2E-02
WP2643	Nanoparticle-mediated activation of receptor signaling	6	1.8	3.3	7.4E-03	4.9E-02
WP585	Interferon type I signaling pathways	9	3.5	2.5	7.4E-03	4.9E-02

Table D4. WikiPathways enrichment in the luminal B driver subnetwork.

Gene Set	Description	Overlap	Enrichment Score	Normalized Enrichment Score	P Value	FDR
WP1544	MicroRNAs in cardiomyocyte hypertrophy	18	4.8	3.8	5.2E-07	0.0001
WP306	Focal Adhesion	25	10.1	2.5	1.6E-05	0.0022
WP2261	Signaling Pathways in Glioblastoma	16	5.1	3.2	2.8E-05	0.0025
WP437	EGF/EGFR Signaling Pathway	23	9.9	2.3	9.3E-05	0.0063
WP2911	miRNA targets in ECM and membrane receptors	6	0.9	6.8	1.2E-04	0.0064
WP2795	Cardiac Hypertrophic Response	11	3.2	3.4	2.6E-04	0.0117
WP244	Alpha 6 Beta 4 signaling pathway	8	1.8	4.4	3.1E-04	0.0119
WP2380	Brain-Derived Neurotrophic Factor (BDNF) signaling pathway	20	8.8	2.3	3.8E-04	0.0129
WP3651	Pathways Affected in Adenoid Cystic Carcinoma	11	3.4	3.2	4.4E-04	0.0132
WP2377	Integrated Pancreatic Cancer Pathway	23	11.2	2.1	6.3E-04	0.0167
WP2572	Primary Focal Segmental Glomerulosclerosis FSGS	11	3.6	3.0	7.1E-04	0.0167

WP51	Regulation of Actin Cytoskeleton	18	7.9	2.3	7.4E-04	0.0167
WP2032	Human Thyroid Stimulating Hormone (TSH) signaling pathway	11	3.7	2.9	9.7E-04	0.0201
WP3844	PI3K-AKT-mTOR signaling pathway and therapeutic opportunities	7	1.8	3.8	1.8E-03	0.0339
WP2526	PDGF Pathway	8	2.4	3.3	2.1E-03	0.0367
WP2828	Bladder Cancer	7	1.9	3.7	2.2E-03	0.0367
WP1528	Physiological and Pathological Hypertrophy of the Heart	6	1.5	4.1	2.5E-03	0.0391
WP366	TGF-beta Signaling Pathway	17	8.2	2.1	2.8E-03	0.0419

APPENDIX E

GO BP ENRICHMENT FOR PROTEINS WITH ALTERED PHOSPHORYLATION IN COLON CANCER

Table E1. GO BP Enrichment for Proteins with Downregulated Phosphorylation Sites in Tumor.

Gene Set	Description	Overlap	Enrichment Score	Normalized Enrichment Score	P Value	FDR
GO:0006928	movement of cell or subcellular component	253	181.395349	1.39474359	0	0
GO:0007010	cytoskeleton organization	261	176.22739	1.48104106	0	0
GO:0030029	actin filament-based process	179	110.594315	1.61852804	0	0
GO:0030036	actin cytoskeleton organization	157	97.1576227	1.61593085	0	0
GO:0032970	regulation of actin filament-based process	87	51.1627907	1.70045455	6.44E-15	3.28E-12
GO:1902589	single-organism organelle organization	298	226.356589	1.31650685	1.18E-14	4.99E-12
GO:0032956	regulation of actin cytoskeleton organization	74	42.8940568	1.72518072	1.40E-13	5.10E-11
GO:0007015	actin filament organization	95	58.9147287	1.6125	4.60E-13	1.34E-10
GO:0003008	system process	161	111.627907	1.44229167	4.73E-13	1.34E-10
GO:0040011	locomotion	213	156.072351	1.36475166	7.95E-13	2.02E-10

Table E2. GO BP Enrichment for Proteins with Upregulated Phosphorylation Sites in Tumor.

Gene Set	Description	Overlap	Enrichment Score	Normalized Enrichment Score	P Value	FDR
GO:0006396	RNA processing	216	109.10422	1.97975843	0	0
GO:0006397	mRNA processing	135	71.2265289	1.89536121	0	0
GO:0008380	RNA splicing	115	60.5219638	1.90013662	0	0
GO:0016071	mRNA metabolic process	157	87.6950904	1.79029407	0	0
GO:0022613	ribonucleoprotein complex biogenesis	89	46.5236865	1.91300404	0	0
GO:0034470	ncRNA processing	71	31.7019811	2.23960767	0	0
GO:0034660	ncRNA metabolic process	92	43.6416882	2.1080761	0	0
GO:0042254	ribosome biogenesis	59	27.1731266	2.17126284	2.22E-16	7.07E-14
GO:0006364	rRNA processing	49	21.4091301	2.28874316	3.33E-16	8.48E-14
GO:0016072	rRNA metabolic process	49	21.4091301	2.28874316	3.33E-16	8.48E-14

Table E3. GO BP Enrichment for Proteins with Highly Upregulated Phosphorylation Sites in Tumor.

Gene Set	Description	Overlap	Enrichment Score	Normalized Enrichment Score	P Value	FDR
GO:0042254	ribosome biogenesis	10	1.36434109	7.32954545	4.79E-07	0.00122032
GO:0051052	regulation of DNA metabolic process	11	1.88113695	5.84752747	1.25E-06	0.00158957
GO:0022613	ribonucleoprotein complex biogenesis	11	2.33591731	4.7090708	1.10E-05	0.00932361
GO:0006364	rRNA processing	7	1.0749354	6.51201923	6.80E-05	0.03464428
GO:0016072	rRNA metabolic process	7	1.0749354	6.51201923	6.80E-05	0.03464428
GO:0006259	DNA metabolic process	13	4.13436693	3.144375	0.00012511	0.04759933
GO:0034470	ncRNA processing	8	1.59173127	5.02597403	0.00013087	0.04759933

APPENDIX F

COLON CANCER-ASSOCIATED PHOSPHORYLATION SITES

HGNC Symbol	UniProt ID	Site	Median Log2(Tumor)-Log2(Normal)	FDR	-Log10(FDR)
CLDN2	P57739	S208	1.868	2.59E-09	8.586238901
RSL1D1	O76021	S427	1.704	1.85E-14	13.73308206
RSL1D1	O76021	S443	1.6875	4.69E-12	11.32846341
MAP3K20	Q9NYL2	S637	1.6645	2.42E-13	12.6155395
PARP1	P09874	S257	1.547	1.11E-09	8.953587331
NOP2	P46087	S67	1.4895	1.00E-12	11.99799569
UBE2M	P61081	S28	1.466	4.30E-12	11.36608691
DDX21	Q9NR30	S121	1.415	6.59E-15	14.18100695
NCL	P19338	S67	1.4035	7.17E-15	14.14464472
FOXK1	P85037	S223	1.401	1.79E-11	10.74712642
NPM1	P06748	S260	1.384	2.31E-13	12.63704785
TOP2A	P11388	S1106	1.36775	4.44E-14	13.35281059
TCOF1	Q13428	S156	1.351	1.97E-13	12.70597241
TFRC	P02786	S34	1.349	1.55E-13	12.80929415
SRRM2	Q9UQ35	T2599	1.325	3.18E-10	9.496918431
RPL4	P36578	S295	1.3135	6.59E-15	14.18100695
TNS4	Q8IZW8	S350	1.293	8.32E-13	12.07968007
PNPLA2	Q96AD5	S428	1.28	1.33E-11	10.87632875
TOP2A	P11388	S1247	1.2795	4.48E-12	11.34870137
FOSL2	P15408	S17	1.2595	1.86E-11	10.72961408
LIG3	P49916	S913	1.243	1.61E-14	13.79312313
RSL1D1	O76021	S392	1.219	2.51E-08	7.600505014
NUP153	P49790	S192	1.2175	6.59E-15	14.18100695
NOP2	P46087	S732	1.216	1.33E-10	9.875157307
NOLC1	Q14978	S538	1.20525	1.92E-14	13.71671823
CDK13	Q14004	S340	1.16825	2.25E-10	9.647615714
SAMHD1	Q9Y3Z3	T592	1.162	1.93E-09	8.714899114
NOC2L	Q9Y3T9	S49	1.141	1.22E-14	13.91287263
ILF3	Q12906	S382	1.1355	2.04E-11	10.68932899
SUB1	P53999	S19	1.135	1.15E-11	10.9375928
RB1	P06400	T826	1.131	1.40E-12	11.853356
RPP30	P78346	S251	1.125	5.66E-13	12.24683109
PML	P29590	S512	1.122	5.57E-09	8.253812672
S100A11	P31949	S6	1.122	2.99E-10	9.52434254
MAP4	P27816	S787	1.118	5.12E-08	7.291009472
CDK12	Q9NYV4	S303	1.111	4.13E-08	7.384294752
CDK2	P24941	Y15	1.101	2.17E-14	13.66453617
TMPO	P42166	S184	1.098	1.96E-13	12.70724838
NUP93	Q8N1F7	S767	1.095	2.30E-12	11.6390753
NOLC1	Q14978	S563	1.0935	3.45E-10	9.462403982
MYBBP1A	Q9BQG0	S1267	1.09175	6.59E-15	14.18100695
RIF1	Q5UIP0	S782	1.091	2.35E-14	13.62859382
SPP1	P10451	S234	1.088	2.61E-05	4.582709325
RB1	P06400	S807	1.082	4.87E-10	9.312207505
OSBPL3	Q9H4L5-2	S251	1.08	1.22E-09	8.913524704
MKI67	P46013	S2505	1.068	2.08E-11	10.68172081
SSFA2	P28290	S759	1.0655	3.80E-09	8.420183162
SMC3	Q9UQE7	S1065	1.06525	2.31E-13	12.63704785
DEK	P35659	S306	1.06375	1.26E-12	11.89895398
TUBA1B	P68363	S48	1.0445	1.83E-10	9.738036853

AMPD2	Q01433	S190	1.0445	1.63E-10	9.788566187
CDK13	Q14004	S342	1.043	2.12E-09	8.67438279
CDK2	P24941	T14	1.03075	1.95E-11	10.70913932
DEK	P35659	S307	1.03	5.36E-14	13.27111673
NOM1	Q5C9Z4	S321	1.029	3.59E-12	11.44443685
SRSF9	Q13242	S204	1.0285	4.79E-12	11.31930057
WDR4	P57081	S391	1.02525	1.34E-12	11.8724523
NOC2L	Q9Y3T9	S56	1.02225	3.70E-14	13.43162797
NOM1	Q5C9Z4	S320	1.022	2.62E-12	11.58204193
NOM1	Q5C9Z4	S317	1.0185	3.47E-12	11.45970211
NCAPG	Q9BPX3	S674	1.018	2.56E-11	10.59101817
BMS1	Q14692	S552	1.014	1.44E-12	11.8427087
DEK	P35659	S303	1.001	1.15E-13	12.93857546

REFERENCES

1. Weinstein, I. B. & Case, K. The History of Cancer Research: Introducing an AACR Centennial Series. *Cancer Res* **68,** 6861–6862 (2008).
2. Siegel, R. L., Miller, K. D. & Jemal, A. Cancer statistics, 2017. *CA: A Cancer Journal for Clinicians* **67,** 7–30 (2017).
3. Knudson, A. G. Two genetic hits (more or less) to cancer. *Nature Reviews Cancer* **1,** 157–162 (2001).
4. Garber, J. E. & Offit, K. Hereditary Cancer Predisposition Syndromes. *JCO* **23,** 276–292 (2005).
5. Bradner, J. E., Hnisz, D. & Young, R. A. Transcriptional Addiction in Cancer. *Cell* **168,** 629–643 (2017).
6. Bahram, F., von der Lehr, N., Cetinkaya, C. & Larsson, L. G. c-Myc hot spot mutations in lymphomas result in inefficient ubiquitination and decreased proteasome-mediated turnover. *Blood* **95,** 2104–2110 (2000).
7. Leder, A., Pattengale, P. K., Kuo, A., Stewart, T. A. & Leder, P. Consequences of widespread deregulation of the c-myc gene in transgenic mice: Multiple neoplasms and normal development. *Cell* **45,** 485–495 (1986).
8. Sur, I. & Taipale, J. The role of enhancers in cancer. *Nat. Rev. Cancer* **16,** 483–493 (2016).
9. Fuziwara, C. S. & Kimura, E. T. Insights into Regulation of the miR-17-92 Cluster of miRNAs in Cancer. *Front Med (Lausanne)* **2,** (2015).
10. Bhan, A., Soleimani, M. & Mandal, S. S. Long Noncoding RNA and Cancer: A New Paradigm. *Cancer Res.* **77,** 3965–3981 (2017).
11. Rivlin, N., Brosh, R., Oren, M. & Rotter, V. Mutations in the p53 Tumor Suppressor Gene. *Genes Cancer* **2,** 466–474 (2011).
12. Salesse, S. & Verfaillie, C. M. BCR/ABL: from molecular mechanisms of leukemia induction to treatment of chronic myelogenous leukemia. *Oncogene* **21,** 8547–8559 (2002).
13. Pandolfi, P. P. Aberrant mRNA translation in cancer pathogenesis: an old concept revisited comes finally of age. *Oncogene* **23,** 3134–3137 (2004).
14. Ashktorab, H. & Brim, H. DNA Methylation and Colorectal Cancer. *Curr Colorectal Cancer Rep* **10,** 425–430 (2014).
15. Pinho, S. S. & Reis, C. A. Glycosylation in cancer: mechanisms and clinical implications. *Nature Reviews Cancer* **15,** nrc3982 (2015).
16. Carvalho, S. *et al.* Preventing E-cadherin aberrant N-glycosylation at Asn-554 improves its critical function in gastric cancer. *Oncogene* **35,** 1619–1631 (2016).
17. Olow, A. *et al.* An Atlas of the Human Kinome Reveals the Mutational Landscape Underlying Dysregulated Phosphorylation Cascades in Cancer. *Cancer Res.* **76,** 1733–1745 (2016).
18. Radivojac, P. *et al.* Gain and loss of phosphorylation sites in human cancer. *Bioinformatics* **24,** i241–i247 (2008).
19. Vogelstein, B. *et al.* Cancer Genome Landscapes. *Science* **339,** 1546–1558 (2013).

20. Beroukhim, R. *et al.* The landscape of somatic copy-number alteration across human cancers. *Nature* **463,** 899–905 (2010).
21. Tomczak, K., Czerwińska, P. & Wiznerowicz, M. The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. *Contemp Oncol (Pozn)* **19,** A68–A77 (2015).
22. Ellis, M. J. *et al.* Connecting genomic alterations to cancer biology with proteomics: the NCI Clinical Proteomic Tumor Analysis Consortium. *Cancer Discov* **3,** 1108–1112 (2013).
23. Pereira, B. *et al.* The somatic mutation profiles of 2,433 breast cancers refines their genomic and transcriptomic landscapes. *Nat Commun* **7,** 11479 (2016).
24. Consortium, T. I. C. G. International network of cancer genome projects. *Nature* **464,** nature08987 (2010).
25. Manning, G., Whyte, D. B., Martinez, R., Hunter, T. & Sudarsanam, S. The Protein Kinase Complement of the Human Genome. *Science* **298,** 1912–1934 (2002).
26. Chalhoub, N. & Baker, S. J. PTEN and the PI3-Kinase Pathway in Cancer. *Annu Rev Pathol* **4,** 127–150 (2009).
27. Samuels, Y. & Waldman, T. Oncogenic Mutations of PIK3CA in Human Cancers. *Curr Top Microbiol Immunol* **347,** 21–41 (2010).
28. Karakas, B., Bachman, K. E. & Park, B. H. Mutation of the PIK3CA oncogene in human cancers. *Br J Cancer* **94,** 455–459 (2006).
29. Burotto, M., Chiou, V. L., Lee, J.-M. & Kohn, E. C. The MAPK pathway across different malignancies: A new perspective. *Cancer* **120,** 3446–3456 (2014).
30. Roberts, P. J. & Der, C. J. Targeting the Raf-MEK-ERK mitogen-activated protein kinase cascade for the treatment of cancer. *Oncogene* **26,** 3291–3310 (2007).
31. Knapp, S. New opportunities for kinase drug repurposing and target discovery. *British Journal of Cancer* **118,** 936–937 (2018).
32. Kim, A. & Cohen, M. S. The discovery of vemurafenib for the treatment of BRAF-mutated metastatic melanoma. *Expert Opin Drug Discov* **11,** 907–916 (2016).
33. Iqbal, N. & Iqbal, N. Imatinib: A Breakthrough of Targeted Therapy in Cancer. *Chemotherapy Research and Practice* (2014). doi:10.1155/2014/357027
34. Bain, J. *et al.* The selectivity of protein kinase inhibitors: a further update. *Biochem J* **408,** 297–315 (2007).
35. Barouch-Bentov, R. & Sauer, K. Mechanisms of Drug-Resistance in Kinases. *Expert Opin Investig Drugs* **20,** 153–208 (2011).
36. Chrisoulidou, A. *et al.* Treatment compliance and severe adverse events limit the use of tyrosine kinase inhibitors in refractory thyroid cancer. *Onco Targets Ther* **8,** 2435–2442 (2015).
37. Jensen, S. S. & Larsen, M. R. Evaluation of the impact of some experimental procedures on different phosphopeptide enrichment techniques. *Rapid Commun. Mass Spectrom.* **21,** 3635–3645 (2007).
38. Mertins, P. *et al.* Ischemia in Tumors Induces Early and Sustained Phosphorylation Changes in Stress Kinase Pathways but Does Not Affect Global Protein Levels. *Mol Cell Proteomics* **13,** 1690–1704 (2014).

39. Lee, D. C. H., Jones, A. R. & Hubbard, S. J. Computational phosphoproteomics: From identification to localization. *Proteomics* **15,** 950–963 (2015).
40. Chalkley, R. J. & Clauser, K. R. Modification Site Localization Scoring: Strategies and Performance. *Mol Cell Proteomics* **11,** 3–14 (2012).
41. Adam, K. & Hunter, T. Histidine kinases and the missing phosphoproteome from prokaryotes to eukaryotes. *Lab. Invest.* (2017). doi:10.1038/labinvest.2017.118
42. Park, J. *et al.* Building a human kinase gene repository: Bioinformatics, molecular cloning, and functional validation. *Proc Natl Acad Sci U S A* **102,** 8114–8119 (2005).
43. Chen, M. J., Dixon, J. E. & Manning, G. Genomics and evolution of protein phosphatases. *Sci Signal* **10,** (2017).
44. Hernandez-Armenta, C., Ochoa, D., Gonçalves, E., Saez-Rodriguez, J. & Beltrao, P. Benchmarking substrate-based kinase activity inference using phosphoproteomic data. *Bioinformatics* **33,** 1845–1851 (2017).
45. Wiredja, D. D., Koyutürk, M. & Chance, M. R. The KSEA App: a web-based tool for kinase activity inference from quantitative phosphoproteomics. *Bioinformatics* (2017). doi:10.1093/bioinformatics/btx415
46. Henry, V. J., Bandrowski, A. E., Pepin, A.-S., Gonzalez, B. J. & Desfeux, A. OMICtools: an informative directory for multi-omic data analysis. *Database (Oxford)* **2014,** (2014).
47. The UniProt Consortium. UniProt: the universal protein knowledgebase. *Nucleic Acids Res.* **45,** D158–D169 (2017).
48. Duan, G., Li, X. & Köhn, M. The human DEPhOsphorylation database DEPOD: a 2015 update. *Nucleic Acids Res.* **43,** D531-535 (2015).
49. Li, T., Li, F. & Zhang, X. Prediction of kinase-specific phosphorylation sites with sequence features by a log-odds ratio approach. *Proteins* **70,** 404–414 (2008).
50. Fan, W. *et al.* Prediction of protein kinase-specific phosphorylation sites in hierarchical structure using functional information and random forest. *Amino Acids* **46,** 1069–1078 (2014).
51. Miller, M. L. *et al.* Linear motif atlas for phosphorylation-dependent signaling. *Sci Signal* **1,** ra2 (2008).
52. Linding, R. *et al.* Systematic discovery of in vivo phosphorylation networks. *Cell* **129,** 1415–1426 (2007).
53. Song, C. *et al.* Systematic analysis of protein phosphorylation networks from phosphoproteomic data. *Mol. Cell Proteomics* **11,** 1070–1083 (2012).
54. Xue, Y. *et al.* GPS 2.1: enhanced prediction of kinase-specific phosphorylation sites with an algorithm of motif length selection. *Protein Eng. Des. Sel.* **24,** 255–260 (2011).
55. Song, J. *et al.* PhosphoPredict: A bioinformatics tool for prediction of human kinase-specific phosphorylation substrates and sites by integrating heterogeneous feature selection. *Scientific Reports* **7,** 6862 (2017).
56. Wang, D. *et al.* MusiteDeep: a deep-learning framework for general and kinase-specific phosphorylation site prediction. *Bioinformatics* **33,** 3909–3916 (2017).

57. Patrick, R., Lê Cao, K.-A., Kobe, B. & Bodén, M. PhosphoPICK: modelling cellular context to map kinase-substrate phosphorylation events. *Bioinformatics* **31,** 382–389 (2015).
58. Blom, N., Gammeltoft, S. & Brunak, S. Sequence and structure-based prediction of eukaryotic protein phosphorylation sites. *J. Mol. Biol.* **294,** 1351–1362 (1999).
59. Gao, J., Thelen, J. J., Dunker, A. K. & Xu, D. Musite, a Tool for Global Prediction of General and Kinase-specific Phosphorylation Sites. *Molecular & Cellular Proteomics : MCP* **9,** 2586 (2010).
60. Neuberger, G., Schneider, G. & Eisenhaber, F. pkaPS: prediction of protein kinase A phosphorylation sites with the simplified kinase-substrate binding model. *Biol. Direct* **2,** 1 (2007).
61. Huang, K.-Y. *et al.* dbPTM 2016: 10-year anniversary of a resource for post-translational modification of proteins. *Nucleic Acids Res.* **44,** D435-446 (2016).
62. Sing, T., Sander, O., Beerenwinkel, N. & Lengauer, T. ROCR: visualizing classifier performance in R. *Bioinformatics* **21,** 3940–3941 (2005).
63. Wilkes, E. H., Casado, P., Rajeeve, V. & Cutillas, P. R. Kinase activity ranking using phosphoproteomics data (KARP) quantifies the contribution of protein kinases to the regulation of cell viability. *Mol. Cell Proteomics* **16,** 1694–1704 (2017).
64. Lachmann, A. & Ma'ayan, A. KEA: kinase enrichment analysis. *Bioinformatics* **25,** 684–686 (2009).
65. Weidner, C., Fischer, C. & Sauer, S. PHOXTRACK-a tool for interpreting comprehensive datasets of post-translational modifications of proteins. *Bioinformatics* **30,** 3410–3411 (2014).
66. Hornbeck, P. V. *et al.* PhosphoSitePlus, 2014: mutations, PTMs and recalibrations. *Nucleic Acids Res.* **43,** D512-520 (2015).
67. Perfetto, L. *et al.* SIGNOR: a database of causal relationships between biological entities. *Nucleic Acids Res* **44,** D548–D554 (2016).
68. Mischnik, M. *et al.* IKAP: A heuristic framework for inference of kinase activities from Phosphoproteomics data. *Bioinformatics* **32,** 424–431 (2016).
69. Wang, J. *WebGestaltR: The R Version of WebGestalt*. (2017).
70. Keshava Prasad, T. S. *et al.* Human Protein Reference Database--2009 update. *Nucleic Acids Res.* **37,** D767-772 (2009).
71. Dinkel, H. *et al.* Phospho.ELM: a database of phosphorylation sites--update 2011. *Nucleic Acids Res.* **39,** D261-267 (2011).
72. Wang, Y. *et al.* EKPD: a hierarchical database of eukaryotic protein kinases and protein phosphatases. *Nucleic Acids Res.* **42,** D496-502 (2014).
73. Martin, D. M. A., Miranda-Saavedra, D. & Barton, G. J. Kinomer v. 1.0: a database of systematically classified eukaryotic protein kinases. *Nucleic Acids Res.* **37,** D244-250 (2009).
74. Buzko, O. & Shokat, K. M. A kinase sequence database: sequence alignments and family assignment. *Bioinformatics* **18,** 1274–1275 (2002).
75. Krupa, A., Abhinandan, K. R. & Srinivasan, N. KinG: a database of protein kinases in genomes. *Nucleic Acids Res.* **32,** D153-155 (2004).

76. Ortutay, C., Väliaho, J., Stenberg, K. & Vihinen, M. KinMutBase: a registry of disease-causing mutations in protein kinase domains. *Hum. Mutat.* **25,** 435–442 (2005).
77. Milanesi, L. *et al.* Systematic analysis of human kinase genes: a large number of genes and alternative splicing events result in functional and structural diversity. *BMC Bioinformatics* **6,** S20 (2005).
78. McSkimming, D. I. *et al.* ProKinO: A Unified Resource for Mining the Cancer Kinome. *Human Mutation* **36,** 175 (2015).
79. Richardson, C. J. *et al.* MoKCa database--mutations of kinases in cancer. *Nucleic Acids Res.* **37,** D824-831 (2009).
80. Simonetti, F. L., Tornador, C., Nabau-Moretó, N., Molina-Vila, M. A. & Marino-Buslje, C. Kin-Driver: a database of driver mutations in protein kinases. *Database (Oxford)* **2014,** (2014).
81. van Linden, O. P. J., Kooistra, A. J., Leurs, R., de Esch, I. J. P. & de Graaf, C. KLIFS: a knowledge-based structural database to navigate kinase-ligand interaction space. *J. Med. Chem.* **57,** 249–277 (2014).
82. Liberti, S. *et al.* HuPho: the human phosphatase portal. *FEBS J.* **280,** 379–387 (2013).
83. Kooistra, A. J. *et al.* KLIFS: a structural kinase-ligand interaction database. *Nucleic Acids Res.* **44,** D365-371 (2016).
84. Bodenmiller, B. *et al.* PhosphoPep--a phosphoproteome resource for systems biology research in Drosophila Kc167 cells. *Mol. Syst. Biol.* **3,** 139 (2007).
85. Bodenmiller, B. *et al.* PhosphoPep--a database of protein phosphorylation sites in model organisms. *Nat. Biotechnol.* **26,** 1339–1340 (2008).
86. Diella, F. *et al.* Phospho.ELM: a database of experimentally verified phosphorylation sites in eukaryotic proteins. *BMC Bioinformatics* **5,** 79 (2004).
87. Diella, F., Gould, C. M., Chica, C., Via, A. & Gibson, T. J. Phospho.ELM: a database of phosphorylation sites--update 2008. *Nucleic Acids Res.* **36,** D240-244 (2008).
88. Zanzoni, A. *et al.* Phospho3D 2.0: an enhanced database of three-dimensional structures of phosphorylation sites. *Nucleic Acids Res.* **39,** D268-271 (2011).
89. Peri, S. *et al.* Development of human protein reference database as an initial platform for approaching systems biology in humans. *Genome Res.* **13,** 2363–2371 (2003).
90. Mishra, G. R. *et al.* Human protein reference database--2006 update. *Nucleic Acids Res.* **34,** D411-414 (2006).
91. Gnad, F., Gunawardena, J. & Mann, M. PHOSIDA 2011: the posttranslational modification database. *Nucleic Acids Res.* **39,** D253-260 (2011).
92. Gnad, F. *et al.* PHOSIDA (phosphorylation site database): management, structural and evolutionary investigation, and prediction of phosphosites. *Genome Biol.* **8,** R250 (2007).
93. Beltrao, P. *et al.* Systematic functional prioritization of protein posttranslational modifications. *Cell* **150,** 413–425 (2012).

94. Chen, X., Shi, S.-P., Suo, S.-B., Xu, H.-D. & Qiu, J.-D. Proteomic analysis and prediction of human phosphorylation sites in subcellular level reveal subcellular specificity. *Bioinformatics* **31,** 194–200 (2015).
95. Huang, K.-Y. *et al.* RegPhos 2.0: an updated resource to explore protein kinase-substrate phosphorylation networks in mammals. *Database (Oxford)* **2014,** bau034 (2014).
96. Lee, T.-Y., Bo-Kai Hsu, J., Chang, W.-C. & Huang, H.-D. RegPhos: a system to explore the protein kinase-substrate phosphorylation network in humans. *Nucleic Acids Res.* **39,** D777-787 (2011).
97. Tinti, M. *et al.* ANIA: ANnotation and Integrated Analysis of the 14-3-3 interactome. *Database (Oxford)* **2014,** bat085 (2014).
98. Tinti, M., Johnson, C., Toth, R., Ferrier, D. E. K. & Mackintosh, C. Evolution of signal multiplexing by 14-3-3-binding 2R-ohnologue protein families in the vertebrates. *Open Biol* **2,** 120103 (2012).
99. Hu, J. *et al.* PhosphoNetworks: a database for human phosphorylation networks. *Bioinformatics* **30,** 141–142 (2014).
100. Cheng, F., Jia, P., Wang, Q. & Zhao, Z. Quantitative network mapping of the human kinome interactome reveals new clues for rational kinase inhibitor discovery and individualized cancer therapy. *Oncotarget* **5,** 3697–3710 (2014).
101. Matlock, M. K., Holehouse, A. S. & Naegle, K. M. ProteomeScout: a repository and analysis resource for post-translational modifications and proteins. *Nucleic Acids Res.* **43,** D521-530 (2015).
102. Naegle, K. M., Welsch, R. E., Yaffe, M. B., White, F. M. & Lauffenburger, D. A. MCAM: multiple clustering analysis methodology for deriving hypotheses and insights from high-throughput proteomic datasets. *PLoS Comput. Biol.* **7,** e1002119 (2011).
103. Lee, T.-Y. *et al.* dbPTM: an information repository of protein post-translational modification. *Nucleic Acids Res.* **34,** D622-627 (2006).
104. Lu, C.-T. *et al.* DbPTM 3.0: an informative resource for investigating substrate site specificity and functional association of protein post-translational modifications. *Nucleic Acids Res.* **41,** D295-305 (2013).
105. Ullah, S. *et al.* dbPAF: an integrative database of protein phosphorylation in animals and fungi. *Scientific Reports* **6,** srep23534 (2016).
106. Nagai, T., Yoshimoto, J., Kannon, T., Kuroda, K. & Kaibuchi, K. Phosphorylation Signals in Striatal Medium Spiny Neurons. *Trends Pharmacol. Sci.* **37,** 858–871 (2016).
107. Safaei, J., Maňuch, J., Gupta, A., Stacho, L. & Pelech, S. Prediction of 492 human protein kinase substrate specificities. *Proteome Sci* **9 Suppl 1,** S6 (2011).
108. Quintaje, S. B. & Orchard, S. The Annotation of Both Human and Mouse Kinomes in UniProtKB/Swiss-Prot: One Small Step in Manual Annotation, One Giant Leap for Full Comprehension of Genomes*. *Molecular & Cellular Proteomics : MCP* **7,** 1409 (2008).

109. Lo Surdo, P., Calderone, A., Cesareni, G. & Perfetto, L. SIGNOR: A Database of Causal Relationships Between Biological Entities-A Short Guide to Searching and Browsing. *Curr Protoc Bioinformatics* **58,** 8.23.1-8.23.16 (2017).
110. Gong, W. *et al.* PepCyber:P~PEP: a database of human protein protein interactions mediated by phosphoprotein-binding domains. *Nucleic Acids Res.* **36,** D679-683 (2008).
111. Davezac, N., Baldin, V., Blot, J., Ducommun, B. & Tassan, J.-P. Human pEg3 kinase associates with and phosphorylates CDC25B phosphatase: a potential role for pEg3 in cell cycle regulation. *Oncogene* **21,** 7630–7641 (2002).
112. Chauhan, D. *et al.* SHP2 mediates the protective effect of interleukin-6 against dexamethasone-induced apoptosis in multiple myeloma cells. *J. Biol. Chem.* **275,** 27845–27850 (2000).
113. Trost, B. & Kusalik, A. Computational prediction of eukaryotic phosphorylation sites. *Bioinformatics* **27,** 2927–2935 (2011).
114. Miller, M. L. & Blom, N. Kinase-specific prediction of protein phosphorylation sites. *Methods Mol. Biol.* **527,** 299–310, x (2009).
115. Wong, Y.-H. *et al.* KinasePhos 2.0: a web server for identifying protein kinase-specific phosphorylation sites based on sequences and coupling patterns. *Nucleic Acids Res.* **35,** W588-594 (2007).
116. Xue, Y., Li, A., Wang, L., Feng, H. & Yao, X. PPSP: prediction of PK-specific phosphorylation site with Bayesian decision theory. *BMC Bioinformatics* **7,** 163 (2006).
117. Saunders, N. F. W., Brinkworth, R. I., Huber, T., Kemp, B. E. & Kobe, B. Predikin and PredikinDB: a computational framework for the prediction of protein kinase peptide specificity and an associated database of phosphorylation sites. *BMC Bioinformatics* **9,** 245 (2008).
118. Iakoucheva, L. M. *et al.* The importance of intrinsic disorder for protein phosphorylation. *Nucleic Acids Res.* **32,** 1037–1049 (2004).
119. Durek, P., Schudoma, C., Weckwerth, W., Selbig, J. & Walther, D. Detection and characterization of 3D-signature phosphorylation site motifs and their contribution towards improved phosphorylation site prediction in proteins. *BMC Bioinformatics* **10,** 117 (2009).
120. Newman, R. H. *et al.* Construction of human activity-based phosphorylation networks. *Molecular Systems Biology* **9,** 655 (2013).
121. Horn, H. *et al.* KinomeXplorer: an integrated platform for kinome biology studies. *Nat. Methods* **11,** 603–604 (2014).
122. Liu, Z. *et al.* Systematic analysis of the Plk-mediated phosphoregulation in eukaryotes. *Brief Bioinform* **14,** 344–360 (2013).
123. Tsaousis, G. N., Bagos, P. G. & Hamodrakas, S. J. HMMpTM: improving transmembrane protein topology prediction using phosphorylation and glycosylation site prediction. *Biochim. Biophys. Acta* **1844,** 316–322 (2014).
124. Zou, L. *et al.* PKIS: computational identification of protein kinases for experimentally discovered protein phosphorylation sites. *BMC Bioinformatics* **14,** 247 (2013).

125. Dou, Y., Yao, B. & Zhang, C. PhosphoSVM: prediction of phosphorylation sites by integrating various protein sequence attributes with a support vector machine. *Amino Acids* **46,** 1459–1469 (2014).
126. Suo, S.-B., Qiu, J.-D., Shi, S.-P., Chen, X. & Liang, R.-P. PSEA: Kinase-specific prediction and analysis of human phosphorylation substrates. *Sci Rep* **4,** 4524 (2014).
127. Obenauer, J. C., Cantley, L. C. & Yaffe, M. B. Scansite 2.0: Proteome-wide prediction of cell signaling interactions using short sequence motifs. *Nucleic Acids Res.* **31,** 3635–3641 (2003).
128. Yang, P., Humphrey, S. J., James, D. E., Yang, Y. H. & Jothi, R. Positive-unlabeled ensemble learning for kinase substrate prediction from dynamic phosphoproteomics data. *Bioinformatics* **32,** 252–259 (2016).
129. Qin, G.-M., Li, R.-Y. & Zhao, X.-M. PhosD: inferring kinase-substrate interactions based on protein domains. *Bioinformatics* **33,** 1197–1204 (2017).
130. Blom, N., Sicheritz-Pontén, T., Gupta, R., Gammeltoft, S. & Brunak, S. Prediction of post-translational glycosylation and phosphorylation of proteins from the amino acid sequence. *Proteomics* **4,** 1633–1649 (2004).
131. Wei, L., Xing, P., Tang, J. & Zou, Q. PhosPred-RF: A Novel Sequence-Based Predictor for Phosphorylation Sites Using Sequential Information Only. *IEEE Trans Nanobioscience* **16,** 240–247 (2017).
132. Terfve, C. *et al.* CellNOptR: a flexible toolkit to train protein signaling networks to data using multiple logic formalisms. *BMC Syst Biol* **6,** 133 (2012).
133. Äijö, T., Granberg, K. & Lähdesmäki, H. Sorad: a systems biology approach to predict and modulate dynamic signaling pathway response from phosphoproteome time-course measurements. *Bioinformatics* **29,** 1283–1291 (2013).
134. Söderholm, S., Hintsanen, P., Öhman, T., Aittokallio, T. & Nyman, T. A. PhosFox: a bioinformatics tool for peptide-level processing of LC-MS/MS-based phosphoproteomic data. *Proteome Sci* **12,** 36 (2014).
135. Petsalaki, E. *et al.* SELPHI: correlation-based identification of kinase-associated networks from global phospho-proteomics data sets. *Nucleic Acids Res.* **43,** W276-282 (2015).
136. Hsu, C.-L., Wang, J.-K., Lu, P.-C., Huang, H.-C. & Juan, H.-F. DynaPho: a web platform for inferring the dynamics of time-series phosphoproteomics. *Bioinformatics* (2017). doi:10.1093/bioinformatics/btx443
137. Yang, P. *et al.* KinasePA: Phosphoproteomics data annotation using hypothesis driven kinase perturbation analysis. *Proteomics* **16,** 1868–1871 (2016).
138. Ren, J. *et al.* PhosSNP for Systematic Analysis of Genetic Polymorphisms That Influence Protein Phosphorylation. *Molecular & Cellular Proteomics : MCP* **9,** 623 (2010).
139. Wagih, O., Reimand, J. & Bader, G. D. MIMP: predicting the impact of mutations on kinase-substrate phosphorylation. *Nat. Methods* **12,** 531–533 (2015).
140. Creixell, P. *et al.* Kinome-wide Decoding of Network-Attacking Mutations Rewiring Cancer Signaling. *Cell* **163,** 202–217 (2015).

141. Patrick, R., Kobe, B., Lê Cao, K.-A. & Bodén, M. PhosphoPICK-SNP: quantifying the effect of amino acid variants on protein phosphorylation. *Bioinformatics* **33,** 1773–1781 (2017).
142. Vazquez, M., Pons, T., Brunak, S., Valencia, A. & Izarzugaza, J. M. G. wKinMut-2: Identification and Interpretation of Pathogenic Variants in Human Protein Kinases. *Hum. Mutat.* **37,** 36–42 (2016).
143. Chiu, Y.-Y. *et al.* KIDFamMap: a database of kinase-inhibitor-disease family maps for kinase inhibitor selectivity and binding mechanisms. *Nucleic Acids Res.* **41,** D430-440 (2013).
144. Kim, J., Yoo, M., Kang, J. & Tan, A. C. K-Map: connecting kinases with therapeutics for drug repurposing and development. *Hum. Genomics* **7,** 20 (2013).
145. Eid, S., Turk, S., Volkamer, A., Rippmann, F. & Fulle, S. KinMap: a web-based tool for interactive navigation through human kinome data. *BMC Bioinformatics* **18,** 16 (2017).
146. Warnecke, A., Sandalova, T., Achour, A. & Harris, R. A. PyTMs: a useful PyMOL plugin for modeling common post-translational modifications. *BMC Bioinformatics* **15,** 370 (2014).
147. Douglass, J. *et al.* Identifying protein kinase target preferences using mass spectrometry. *Am. J. Physiol., Cell Physiol.* **303,** C715-727 (2012).
148. Torii, M. *et al.* RLIMS-P: an online text-mining tool for literature-based extraction of protein phosphorylation information. *Database (Oxford)* **2014,** (2014).
149. Arighi, C. N. *et al.* eFIP: a tool for mining functional impact of phosphorylation from literature. *Methods Mol. Biol.* **694,** 63–75 (2011).
150. Zhao, B., Pisitkun, T., Hoffert, J. D., Knepper, M. A. & Saeed, F. CPhos: a program to calculate and visualize evolutionarily conserved functional phosphorylation sites. *Proteomics* **12,** 3299–3303 (2012).
151. Madeira, F. *et al.* 14-3-3-Pred: improved methods to predict 14-3-3-binding phosphopeptides. *Bioinformatics* **31,** 2276–2283 (2015).
152. McSkimming, D. I., Rasheed, K. & Kannan, N. Classifying kinase conformations using a machine learning approach. *BMC Bioinformatics* **18,** 86 (2017).
153. Goldberg, J. M. *et al.* Kinannote, a computer program to identify and classify members of the eukaryotic protein kinase superfamily. *Bioinformatics* **29,** 2387–2394 (2013).
154. Najafov, J. & Najafov, A. CrossCheck: an open-source web tool for high-throughput screen data analysis. *Sci Rep* **7,** 5855 (2017).
155. Patra, K. C. *et al.* Hexokinase 2 is required for tumor initiation and maintenance and its systemic deletion is therapeutic in mouse models of cancer. *Cancer Cell* **24,** 213–228 (2013).
156. Gao, Q. *et al.* Driver Fusions and Their Implications in the Development and Treatment of Human Cancers. *Cell Rep* **23,** 227-238.e3 (2018).
157. Ding, L. *et al.* Perspective on Oncogenic Processes at the End of the Beginning of Cancer Genomics. *Cell* **173,** 305-320.e10 (2018).
158. Mertins, P. *et al.* Proteogenomics connects somatic mutations to signalling in breast cancer. *Nature* **534,** 55–62 (2016).

159. Zhang, B. *et al.* Proteogenomic characterization of human colon and rectal cancer. *Nature* **513,** 382–387 (2014).
160. Giansanti, P., Tsiatsiani, L., Low, T. Y. & Heck, A. J. R. Six alternative proteases for mass spectrometry–based proteomics beyond trypsin. *Nature Protocols* **11,** 993–1006 (2016).
161. Fiehn, O. & Weckwerth, W. Mass Spectrometry: Quantitation. in *Encyclopedic Reference of Genomics and Proteomics in Molecular Medicine* 1030–1034 (Springer, Berlin, Heidelberg, 2006). doi:10.1007/3-540-29623-9_3550
162. Cooper, B., Feng, J. & Garrett, W. M. Relative, label-free protein quantitation: spectral counting error statistics from nine replicate MudPIT samples. *J. Am. Soc. Mass Spectrom.* **21,** 1534–1546 (2010).
163. Dowle, A. A., Wilson, J. & Thomas, J. R. Comparing the Diagnostic Classification Accuracy of iTRAQ, Peak-Area, Spectral-Counting, and emPAI Methods for Relative Quantification in Expression Proteomics. *J. Proteome Res.* **15,** 3550–3562 (2016).
164. Latosinska, A. *et al.* Comparative Analysis of Label-Free and 8-Plex iTRAQ Approach for Quantitative Tissue Proteomic Analysis. *PLOS ONE* **10,** e0137048 (2015).
165. Yang, D.-S. *et al.* Design and synthesis of an immobilized metal affinity chromatography and metal oxide affinity chromatography hybrid material for improved phosphopeptide enrichment. *Journal of Chromatography A* **1505,** 56–62 (2017).
166. Daub, H. *et al.* Kinase-Selective Enrichment Enables Quantitative Phosphoproteomics of the Kinome across the Cell Cycle. *Molecular Cell* **31,** 438–448 (2008).
167. Zhang, H. *et al.* Integrated Proteogenomic Characterization of Human High-Grade Serous Ovarian Cancer. *Cell* **166,** 755–765 (2016).
168. Cerami, E. *et al.* The cBio Cancer Genomics Portal: An Open Platform for Exploring Multidimensional Cancer Genomics Data. *Cancer Discov* **2,** 401–404 (2012).
169. Gao, J. *et al.* Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci Signal* **6,** pl1 (2013).
170. Uhlén, M. *et al.* Proteomics. Tissue-based map of the human proteome. *Science* **347,** 1260419 (2015).
171. Barbie, D. A. *et al.* Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. *Nature* **462,** 108–112 (2009).
172. Hänzelmann, S., Castelo, R. & Guinney, J. GSVA: gene set variation analysis for microarray and RNA-seq data. *BMC Bioinformatics* **14,** 7 (2013).
173. Chen, H. & Boutros, P. C. VennDiagram: a package for the generation of highly-customizable Venn and Euler diagrams in R. *BMC Bioinformatics* **12,** 35 (2011).
174. Kolde, R. pheatmap: Pretty Heatmaps. *R package* **1.0.10,** (2018).
175. Takaku, M., Grimm, S. A. & Wade, P. A. GATA3 in breast cancer: tumor suppressor or oncogene? *Gene Expr* **16,** 163–168 (2015).

176. Li, N.-S. *et al.* LKB1/AMPK inhibits TGF-β1 production and the TGF-β signaling pathway in breast cancer cells. *Tumour Biol.* **37,** 8249–8258 (2016).
177. Sengupta, S. *et al.* Activation of tumor suppressor LKB1 by honokiol abrogates cancer stem-like phenotype in breast cancer via inhibition of oncogenic Stat3. *Oncogene* **36,** 5709–5721 (2017).
178. Chen, I.-C. *et al.* Clinical Relevance of Liver Kinase B1(LKB1) Protein and Gene Expression in Breast Cancer. *Sci Rep* **6,** 21374 (2016).
179. Chiker, S. *et al.* Cdk5 promotes DNA replication stress checkpoint activation through RPA-32 phosphorylation, and impacts on metastasis free survival in breast cancer patients. *Cell Cycle* **14,** 3066–3078 (2015).
180. Pozo, K. & Bibb, J. A. The Emerging Role of Cdk5 in Cancer. *Trends Cancer* **2,** 606–618 (2016).
181. Rouillard, A. D. *et al.* The harmonizome: a collection of processed datasets gathered to serve and mine knowledge about genes and proteins. *Database (Oxford)* **2016,** (2016).
182. Kurek, D., Garinis, G. A., van Doorninck, J. H., van der Wees, J. & Grosveld, F. G. Transcriptome and phenotypic analysis reveals Gata3-dependent signalling pathways in murine hair follicles. *Development* **134,** 261–272 (2007).
183. Galmarini, C. M. *et al.* 5'-(3')-nucleotidase mRNA levels in blast cells are a prognostic factor in acute myeloid leukemia patients treated with cytarabine. *Haematologica* **89,** 617–619 (2004).
184. Parker, J. S. *et al.* Supervised risk predictor of breast cancer based on intrinsic subtypes. *J. Clin. Oncol.* **27,** 1160–1167 (2009).
185. Sørlie, T. *et al.* Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc Natl Acad Sci U S A* **98,** 10869–10874 (2001).
186. Fallahpour, S., Navaneelan, T., De, P. & Borgo, A. Breast cancer survival by molecular subtype: a population-based analysis of cancer registry data. *CMAJ Open* **5,** E734–E739 (2017).
187. Mao, J.-H., Diest, P. J. van, Perez-Losada, J. & Snijders, A. M. Revisiting the impact of age and molecular subtype on overall survival after radiotherapy in breast cancer patients. *Scientific Reports* **7,** 12587 (2017).
188. Hofree, M., Shen, J. P., Carter, H., Gross, A. & Ideker, T. Network-based stratification of tumor mutations. *Nat. Methods* **10,** 1108–1115 (2013).
189. Zhu, J. *et al.* Deciphering genomic alterations in colorectal cancer through transcriptional subtype-based network analysis. *PLoS ONE* **8,** e79282 (2013).
190. Zhang, B. *et al.* Relating protein adduction to gene expression changes: a systems approach. *Mol Biosyst* **7,** 2118–2127 (2011).
191. Das, J. & Yu, H. HINT: High-quality protein interactomes and their applications in understanding human disease. *BMC Syst Biol* **6,** 92 (2012).
192. Csardi, G. & Nepusz, T. The igraph software package for complex network research. *InterJournal* **Complex Systems,** 1695 (2006).
193. Slenter, D. N. *et al.* WikiPathways: a multifaceted pathway database bridging metabolomics to other omics research. *Nucleic Acids Res* **46,** D661–D667 (2018).

194. Shannon, P. *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13,** 2498–2504 (2003).
195. Blaydes, J. P. *et al.* Stoichiometric Phosphorylation of Human p53 at Ser315Stimulates p53-dependent Transcription. *J. Biol. Chem.* **276,** 4699–4708 (2001).
196. Gabant, G. *et al.* Autophosphorylated residues involved in the regulation of human chk2 in vitro. *J. Mol. Biol.* **380,** 489–503 (2008).
197. Chen, B. P. C. *et al.* Ataxia Telangiectasia Mutated (ATM) Is Essential for DNA-PKcs Phosphorylations at the Thr-2609 Cluster upon DNA Double Strand Break. *Journal of Biological Chemistry* **282,** 6582–6587 (2007).
198. Zhang, M. H., Man, H. T., Zhao, X. D., Dong, N. & Ma, S. L. Estrogen receptor-positive breast cancer molecular signatures and therapeutic potentials (Review). *Biomed Rep* **2,** 41–52 (2014).
199. Paplomata, E. & O'Regan, R. The PI3K/AKT/mTOR pathway in breast cancer: targets, trials and biomarkers. *Ther Adv Med Oncol* **6,** 154–166 (2014).
200. Paulson, A. K. *et al.* MET and ERBB2 are coexpressed in ERBB2+ breast cancer and contribute to innate resistance. *Mol. Cancer Res.* **11,** 1112–1121 (2013).
201. Zagouri, F. *et al.* Low Protein Expression of MET in ER-positive and HER2-positive Breast Cancer. *Anticancer Res* **34,** 1227–1231 (2014).
202. Bertucci, F., Finetti, P. & Birnbaum, D. Basal Breast Cancer: A Complex and Deadly Molecular Subtype. *Curr Mol Med* **12,** 96–110 (2012).
203. Ribeiro, E. *et al.* Triple Negative Breast Cancers Have a Reduced Expression of DNA Repair Genes. *PLOS ONE* **8,** e66243 (2013).
204. Perou, C. M. *et al.* Molecular portraits of human breast tumours. *Nature* **406,** 747–752 (2000).
205. American Cancer Society. *Cancer Facts & Figures 2018*. (American Cancer Society, 2018).
206. Dienstmann, R., Salazar, R. & Tabernero, J. Personalizing colon cancer adjuvant therapy: selecting optimal treatments for individual patients. *J. Clin. Oncol.* **33,** 1787–1796 (2015).
207. Comprehensive Molecular Characterization of Human Colon and Rectal Cancer. *Nature* **487,** 330–337 (2012).
208. Gonsalves, W. I. *et al.* Patient and tumor characteristics and BRAF and KRAS mutations in colon cancer, NCCTG/Alliance N0147. *J. Natl. Cancer Inst.* **106,** (2014).
209. Oikonomou, E., Koustas, E., Goulielmaki, M. & Pintzas, A. BRAF vs RAS oncogenes: are mutations of the same pathway equal? differential signalling and therapeutic implications. *Oncotarget* **5,** 11752–11777 (2014).
210. Chen, J., Elfiky, A., Han, M., Chen, C. & Saif, M. W. The Role of Src in Colon Cancer and Its Therapeutic Implications. *Clinical Colorectal Cancer* **13,** 5–13 (2014).
211. Fruman, D. A. & Rommel, C. PI3K and Cancer: Lessons, Challenges and Opportunities. *Nat Rev Drug Discov* **13,** 140–156 (2014).

212. Dyson, N. J. RB1: a prototype tumor suppressor and an enigma. *Genes Dev* **30,** 1492–1502 (2016).
213. Lohmann, D. R. RB1 gene mutations in retinoblastoma. *Hum. Mutat.* **14,** 283–288 (1999).
214. Herschkowitz, J. I., He, X., Fan, C. & Perou, C. M. The functional loss of the retinoblastoma tumour suppressor is a common event in basal-like and luminal B breast carcinomas. *Breast Cancer Res* **10,** R75 (2008).
215. Cancer Genome Atlas Research Network. Comprehensive genomic characterization of squamous cell lung cancers. *Nature* **489,** 519–525 (2012).
216. Gope, R. *et al.* Increased expression of the retinoblastoma gene in human colorectal carcinomas relative to normal colonic mucosa. *J. Natl. Cancer Inst.* **82,** 310–314 (1990).
217. Meling, G. I. *et al.* Genetic alterations within the retinoblastoma locus in colorectal carcinomas. Relation to DNA ploidy pattern studied by flow cytometric analysis. *British Journal of Cancer* **64,** 475 (1991).
218. Poller, D. N., Baxter, K. J. & Shepherd, N. A. p53 and Rb1 protein expression: are they prognostically useful in colorectal cancer? *Br. J. Cancer* **75,** 87–93 (1997).
219. Ali, A. A. *et al.* RB1 protein in normal and malignant human colorectal tissue and colon cancer cell lines. *FASEB J.* **7,** 931–937 (1993).
220. Gope, R. & Gope, M. L. Abundance and state of phosphorylation of the retinoblastoma susceptibility gene product in human colon cancer. *Mol. Cell. Biochem.* **110,** 123–133 (1992).
221. Pandey, S., Gordon, P. H. & Wang, E. Expression of proliferation-specific genes in the mucosa adjacent to colon carcinoma. *Dis. Colon Rectum* **38,** 462–467 (1995).
222. Kim, S., Gupta, N. & Pevzner, P. A. Spectral probabilities and generating functions of tandem mass spectra: a strike against decoy databases. *J. Proteome Res.* **7,** 3354–3363 (2008).
223. Kim, S. & Pevzner, P. A. MS-GF+ makes progress towards a universal database search tool for proteomics. *Nat Commun* **5,** 5277 (2014).
224. Beausoleil, S. A., Villén, J., Gerber, S. A., Rush, J. & Gygi, S. P. A probability-based approach for high-throughput protein phosphorylation analysis and site localization. *Nat. Biotechnol.* **24,** 1285–1292 (2006).
225. Forbes, S. A. *et al.* COSMIC: somatic cancer genetics at high-resolution. *Nucleic Acids Res* **45,** D777–D783 (2017).
226. Jianhong Ou, Yong-Xu Wang & Lihua Julie Zhu. trackViewer: A R/Bioconductor package for drawing elegant interactive tracks or lollipop plot to facilitate integrated analysis of multi-omics data. *Bioconductor* **R package version 1.16.0,** (2018).
227. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489,** 57–74 (2012).
228. Liberzon, A. *et al.* The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst* **1,** 417–425 (2015).
229. Yang, W. *et al.* Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. *Nucleic Acids Res.* **41,** D955-961 (2013).

230. Li, J. *et al.* Characterization of Human Cancer Cell Lines by Reverse-phase Protein Arrays. *Cancer Cell* **31,** 225–239 (2017).
231. Lee, W.-K. *et al.* Structural and functional insights into the regulation mechanism of CK2 by IP6 and the intrinsically disordered protein Nopp140. *Proc. Natl. Acad. Sci. U.S.A.* **110,** 19360–19365 (2013).
232. Delgado, J. L., Hsieh, C.-M., Chan, N.-L. & Hiasa, H. Topoisomerases as anticancer targets. *Biochem. J.* **475,** 373–398 (2018).
233. Chikamori, K. *et al.* Phosphorylation of Serine 1106 in the Catalytic Domain of Topoisomerase IIα Regulates Enzymatic Activity and Drug Sensitivity. *J. Biol. Chem.* **278,** 12696–12702 (2003).
234. Kitazawa, H. *et al.* Ser787 in the proline-rich region of human MAP4 is a critical phosphorylation site that reduces its activity to promote tubulin polymerization. *Cell Struct. Funct.* **25,** 33–39 (2000).
235. Poruchynsky, M. S. *et al.* Accompanying protein alterations in malignant cells with a microtubule-polymerizing drug-resistance phenotype and a primary resistance mechanism11Abbreviations: MTs, microtubules; MAPs, microtubule-associated proteins; MAP4, microtubule-associated protein-4; PTX, paclitaxel; EPOA, epothilone A; EPOB, epothilone B; EPOA-R, epothilone A-resistant; COL, colchicine; VCR, vincristine; and VBL, vinblastine. *Biochemical Pharmacology* **62,** 1469–1480 (2001).
236. Coss, A. *et al.* Increased topoisomerase IIalpha expression in colorectal cancer is associated with advanced disease and chemotherapeutic resistance via inhibition of apoptosis. *Cancer Lett.* **276,** 228–238 (2009).
237. de Resende, M. F. *et al.* Prognostication of prostate cancer based on TOP2A protein and gene assessment: TOP2A in prostate cancer. *J Transl Med* **11,** 36 (2013).
238. Zhang, Y., Baysac, K. C., Yee, L.-F., Saporita, A. J. & Weber, J. D. Elevated DDX21 regulates c-Jun activity and rRNA processing in human breast cancers. *Breast Cancer Res.* **16,** 449 (2014).
239. LI, X.-P. *et al.* Overexpression of ribosomal L1 domain containing 1 is associated with an aggressive phenotype and a poor prognosis in patients with prostate cancer. *Oncol Lett* **11,** 2839–2844 (2016).
240. Jaiswal, J. K. *et al.* S100A11 is required for efficient plasma membrane repair and survival of invasive cancer cells. *Nature Communications* **5,** 3795 (2014).
241. De, S., Tsimounis, A., Chen, X. & Rotenberg, S. A. Phosphorylation of α-Tubulin by Protein Kinase C Stimulates Microtubule Dynamics in Human Breast Cells. *Cytoskeleton (Hoboken)* **71,** 257–272 (2014).
242. Gelais, C. S. *et al.* A Putative Cyclin-binding Motif in Human SAMHD1 Contributes to Protein Phosphorylation, Localization, and Stability. *J. Biol. Chem.* **291,** 26332–26342 (2016).
243. Yang, J.-J. *et al.* ZAK inhibits human lung cancer cell growth via ERK and JNK activation in an AP-1-dependent manner. *Cancer Sci.* **101,** 1374–1381 (2010).

244. Yamamoto, H. *et al.* Paradoxical increase in retinoblastoma protein in colorectal carcinomas may protect cells from apoptosis. *Clin. Cancer Res.* **5,** 1805–1815 (1999).
245. Indovina, P., Pentimalli, F., Casini, N., Vocca, I. & Giordano, A. RB1 dual role in proliferation and apoptosis: Cell fate control and implications for cancer therapy. *Oncotarget* **6,** 17873–17890 (2015).
246. Arhel, N. J. *et al.* The retinoblastoma protein interacts with Bag-1 in human colonic adenoma and carcinoma derived cell lines. *Int. J. Cancer* **106,** 364–371 (2003).
247. Bullock, N. & Oltean, S. The many faces of SRPK1. *J Pathol* **241,** 437–440 (2017).
248. Shi, L., Pan, H., Liu, Z., Xie, J. & Han, W. Roles of PFKFB3 in cancer. *Signal Transduction and Targeted Therapy* **2,** 17044 (2017).
249. Giuliano, C. J., Lin, A., Smith, J. C., Palladino, A. C. & Sheltzer, J. M. MELK expression correlates with tumor mitotic activity but is not required for cancer growth. *eLife* **7,**
250. Xie, X. *et al.* A comparative phosphoproteomic analysis of a human tumor metastasis model using a label-free quantitative approach. *Electrophoresis* **31,** 1842–1852 (2010).
251. Gannon, J., Staunton, L., O'Connell, K., Doran, P. & Ohlendieck, K. Phosphoproteomic analysis of aged skeletal muscle. *Int. J. Mol. Med.* **22,** 33–42 (2008).
252. Zimman, A. *et al.* Activation of aortic endothelial cells by oxidized phospholipids: a phosphoproteomic analysis. *J Proteome Res* **9,** 2812–2824 (2010).
253. Klaeger, S. *et al.* The target landscape of clinical kinase drugs. *Science* **358,** eaan4368 (2017).
254. Mukherjee, S. *et al.* Yersinia YopJ Acetylates and Inhibits Kinase Activation by Blocking Phosphorylation. *Science* **312,** 1211–1214 (2006).