Identification of dephosphorylated sites in the proximity of recurrent mutations in PP2A targets

MD Fakhrul Islam Faruque
Master's Thesis
Master's degree programme in Bioinformatics
Department of Future Technologies
University of Turku
December 2018

UNIVERSITY OF TURKU
Department of Future Technologies
FARUQUE, MD FAKHRUL ISLAM: Identification of dephosphorylated sites in the proximity of recurrent mutations in PP2A targets. Master's Thesis, 46 p., Appendix, 11 p.
Master's Degree Programme in Bioinformatics
December 2018

Protein Phosphatase 2A (PP2A), a major serine/threonine phosphatase, is known to be involved in the wide range of cellular functions in many cell types. Notably, PP2A's tumor suppressor function has a great potential for therapeutic use in cancer patients. However, understanding basic functions as well as translational potential of PP2A is complex and is at its infancy. Identification of PP2A targets, and especially target sites nearby the recurrent mutations can potentially provide insights on PP2A's function in the context of cancer. Thus, the focus of this thesis was to identify target sites which are directly or indirectly dephosphorylated by PP2A and thereby map the sites nearby the significant recurrent mutations in cancer samples.

Thesis presented here in made use of in-house as well as published phosphosproteomics datasets to identify the potential targets of PP2A. A protein was considered as a target of PP2A if its phosphopeptide was significantly regulated as assessed either from student's t-test or alternatively defined in respective publications. In order to identify the significantly mutated residues as compared to background mutation rate, **ActiveDriver methodology** was employed in this study. ActiveDriver tests the null hypotheses given mutational, intrinsic disorder and phosphosite information. The null hypothesis assumes that mutations in protein sequences follows Poisson distribution.

As an example of PP2A target identification process, an in-house generated B56 dataset of PP2A phosphpoproteomics dataset was used. Two sided students t-test was performed to find differentially regulated peptides and the analysis revealed 1249 out of 6739 peptides were statistically significant (unadjusted p value < 0.05). Volcano plot and heatmap for the analysis of B56 dataset were used to visualise most significant peptides. A comprehensive dataset of non-redundant phosphosites from various PP2A phosphoproteomics datasets (three groups of PP2A families) was built to reflect the broader coverage of PP2A targets. ActiveDriver analysis on cBioportal pancancer study revealed that there were 19 genes with 248 active regions (p-value < 0.05). Similar analysis on COSMIC mutational dataset revealed 57 genes with 2,723 active regions (p-value < 0.05). Network analysis was carried out on proteins having at least one significant active region. The resulting protein-protein interaction network from STRING database for the target list of proteins after ActiveDriver analysis

is significantly enriched as compared to any random network and it was also significant (5e-15). Functional enrichment analysis also provided strong evidence among those analysis and PPI enrichment p- value also significant in both cases. Based on false discovery rate, biological and molecular function among the selected genes also showed significant.

This mutational study provides better understand to identify target sites which are directly or indirectly dephosphorylated by PP2A and thereby likely provide potential clues for mechanisms of action for PP2A function.

Keywords: Protein Phosphatase 2A, Mutation, Dephosphorylation, cBioportal, COSMIC, ActiveDriver

Table of Contents

Abbreviations

CCDS- Consensus Coding Sequence

COSMIC- Catalogue Of Somatic Mutations In Cancer

DEPOD- the human DEPhOsphorylation Database

DisProt- Database of Protein Disorder

GLM- Generalized Linear Model

HPRD- Human Protein References Database

PP- Protein Phosphatase

PP2A- Protein Phosphatase 2A

PPI- Protein Protein Interaction

PPP_s- Phosphoprotein Phosphatase

PTM- Post Translational modification

PTP- Protein Tyrosine Phosphatase

PSP- Protein Serine/Tyrosine Phosphatase

List of figures

List of Tables

1. Introduction

1.1 Phosphorylation and dephosphorylation

The most functions of proteins in human are regulated by Post-Translational Modifications (PTM) which play a major role in many biological functions including cell apoptosis, cell division, proliferation, survival and development. Phosphorylation is one of the predominant post-translational modification. The process of addition and subtraction of phosphate (PO_4^{3-}) by and from protein is called phosphorylation and dephosphorylation. Phosphate group binds with hydroxyl group of serine, threonine or tyrosine amino acid side chain to form complex phosphate monoesters (Lad, Williams, & Wolfenden, 2003). Depending on the condition in cell, protein shifts from phosphorylation to dephosphorylation and vice versa. The phosphorylation and dephosphorylation process are controlled dynamically by counteracting protein kinase and phosphatase respectively. Protein kinase and phosphatase act like as control switches and regulators (Figure: 1) (Mumby & Walter, 1993).

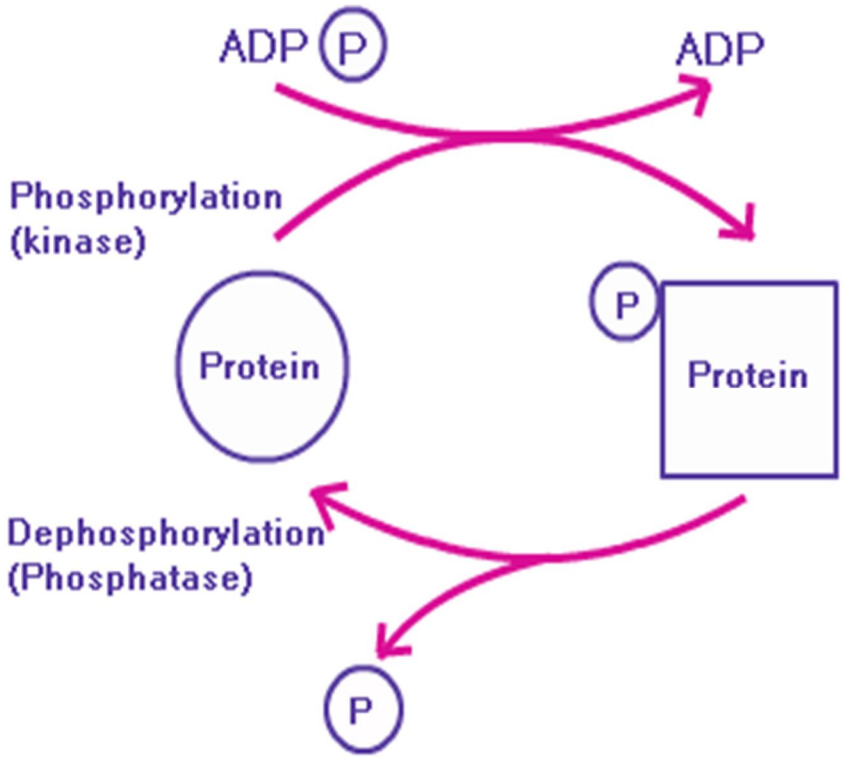


Figure 1: Overview of phosphorylation and dephosphorylation processes. Addition of phosphate group to a protein is called phosphorylation while removal phosphate group from protein is called dephosphorylation.

The opposite process of phosphorylation, dephosphorylation, is also important for cellular function (INGEBRITSEN & COHEN, 1983).

Phosphorylation normally occurs on serine, threonine and tyrosine residues in eukaryotes and acts as a major mediator of intracellular signaling transduction (Jin & Pawson, 2012). The rate of phosphorylation is most common on phosphoserine among the three amino acid residues. The rate on phospshoserine is about 86.4% followed by phosphothreonine at 11.8% and phosphotyrosine is 1.8% (Krijgsveld, 2012). In the presence of water molecule, PP_s attack the phosphate group for catalysis and dephosphorylate these phosphorylated residues.

There are two major classes of phosphatases, namely, protein tyrosine phosphatases (PTP_s) and protein serine/threonine phosphatases (PSP_s). PSPs have three subfamilies: metal dependent protein phosphatases (PPM_s), aspartate based phosphatase and phosphoprotein phosphatases (PPP_s). PPP_s also divided into subfamilies known as PP1, PP2A, PP2B (calcium activated), PP4, PP5, PP6 and PP7. The PPM_s family subdivided into PP2C and pyruvate dehydrogenase phosphatases. PTP_s remove phosphate group from post transnationally modified tyrosine residues (Seshacharyulu, Pandey, Datta, & Batra, 2013). In human, there are 119 protein phosphatases of which 98 are protein tyrosine specific phosphatases and 21 are protein serine/ threonine phosphatases (Seshacharyulu et al., 2013).

1.2 Protein Phosphatase 2A

Protein phosphatase 2A (PP2A) is a widely expressed serine threonine phosphatase, which plays a crucial role in cellular processes such as cell proliferation, signal transduction and apoptosis (Sablina & Hahn, 2007). PP2A controls the activity of serine and threonine residues as an enzyme by removing phosphate modifications from them. The presence of PP2A found almost 1% content of cellular protein (Kremmer, Ohst, Kiefer, & Brewis, 1997). Moreover, PP2A phosphatase has a well-established tumor suppressor function although understanding of the mechanisms by which PP2A achieves this relevant function is well known (Sangodkar et al., 2016). PP2A is known to exhibit both positive and negative regulation in signalling networks as consequence of its complex roles in cellular functions (Thompson & Williams, 2018).

1.2.1 Functional and structural complexity of PP2A

PP2A is a heterotrimeric holoenzyme, which has three subunits in mammals namely a structural/ scaffolding A subunit, a catalytic C subunit and a regulatory B subunit. In PP2A, A and C subunits recruit with B subunit to control substrate binding and form heterotrimeric protein complex (Figure: 2).

The regulatory subunit B has four subfamilies [B(PR53), B′(PR55 or PR61), B′′(PR72), B′′′(PR93 or PR110)], with minimum 16 classes (Xu et al., 2006). The both scaffolding and catalytic subunits have two isoforms (α,β) on each and both share high sequence similarity. However, it is opposite to regulatory subunit which share low sequence similarity among subfamilies (Janssens & Goris, 2001).

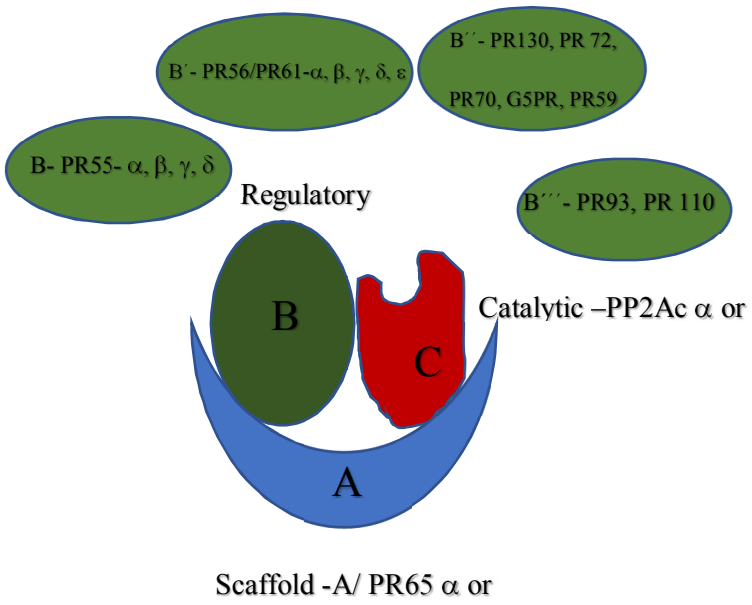


Figure 2: Structural overview of heterotrimeric Protein Phosphatase 2A. PP2A consist of scaffolding A subunit (blue), regulatory B subunit (green) and catalytic C subunit (red). Subunits A and C, each consist of two different isoforms α and β. Regulatory subunit B is subdivided into four subfamilies with different isoforms.

The combinational assembly of various subunits permit to form many distinct complexes which have been implicated in the control of cellular processes such as cell survival, proliferation and adhesion. Many cancer cell lines appear to lack of B56 γ protein expression and overexpression of PP2A B56 γ in such cancer cell lines partially reverse the tumorigenic phenotype of the cells (Westermarck & Hahn, 2008).

1.3 Mutations and cancer

Tumor genome sequencing has revealed complex landscape of somatic DNA mutations in multiple types of tissues and cancer including pancreas, blood, bone, breast, colon, lung, liver, and brain (Wood et al., 2007) and one of the main goals in cancer research has been to characterize and identify driver mutations. These ranges from small piece of DNA mutations to genomic copy number changes, alteration in gene expression and epigenetic regulation. Although, wide-range of tumor genome sequencing approaches have discovered thousands of gene mutation, it is challenging to identify tumor driver mutations from passenger mutations (Reimand & Bader, 2013).

1.3.1 Cancer

Cancer does not refer to a collection of malignancies with diverse characteristics. Different alterations requires for the progression of cancer, which may occur on epigenetic and genetic level. The proliferation and uncontrolled growth of cell lead to fatal condition if it can continue and spread. Based on characteristics of cancer, there are mainly two types of growth patterns named tumor and metastasis.

a) Uncontrolled cell division lead to overgrowth of cells called tumor.

b)The process of spreading of tumor to build-up new tumor in the body called metastasis (Griffiths, 2005).

The study of the cancer genomics is based on oncogene expression and DNA sequence which differ between tumor and the normal cells. According to Bert Volgelstein et. al study (Vogelstein et al., 2013), has shown till now that around 140 genes can promote or drive tumorigenesis and a typical tumor contains 2 to 8 driver genes and the remaining are passenger genes (Vogelstein et al., 2013).

1.3.2 Mutation and Cancer

A mutation is the change of gene pattern information that occur in DNA sequence, either due to mistake when DNA information is copied or as the result of environmental factors. Mutation can disrupt normal cell function and causes disease such as cancer. Cancer is most common human genetic disease and it is caused by mutations occurring in growth controlling genes. In common tumors, which is derived from the breast, colon, pancreas or brain, an average of 33 to 66 genes display somatic mutations that are expected to change protein functions. Among the mutations, 95% of them are single base substitution and the remainder are insertions or deletions of one or few bases (Vogelstein et al., 2013). The driver mutation is a mutation within a gene that confers a selective growth advantage and push cells lead to form cancer. On the other side, the cells which also functionally change but do not provide a growth advantage called passenger mutation. Each driver mutation provides only a few selective growth advantages to the cell and about 0.4% increases in the cell apoptosis process (Yachida et al., 2010). It is really difficult to identify driver and passenger mutation in somatic cell. However, it is important to point out between driver gene and driver mutation. Although, driver gene contains driver gene mutation, but driver gene also contain passenger mutations. Several statistical methods are available for identification of driver genes. Cancer can be driven by mutation in protein involved phosphorylation signalling and gene centric method called Active driver helps to detect such mutations comprehensively (Reimand, Wagih, & Bader, 2013).

1.3.3 Mutational data sources

Different databases are available for searching and analysing mutational data. One can easily obtain mutational information and analyse them as required. cBioportal (http://www.cbioportal.org/) and COSMIC (https://cancer.sanger.ac.uk/cosmic) databases are examples of most commonly used mutational databases for cancer genome research. In this study data was collected from these two databases.

Large-scale cancer genomics data from different platforms pose a great challenge to perform data integration, analysis and exploration, especially for biologists without a computational skills. The cBioportal (http://www.cbioportal.org/) server is specially designed for biological researchers to facilitate easy access to the complex dataset. The cBioportal provides a web tool for visualizing, exploring and analysing multidimensional oncogenes data. To date, the portal contains almost 220

different cancer studies for which data is available. This web source provides graphical representation of gene level data from multiple platforms and mutational status of the specific gene. The portal also provides information about the network visualization and analysis, survival analysis and software programmatic access (Gao et al., 2013). The summary of graphical representation of specific gene mutation status is given in Figure 3. cBioportal always needs HUGO gene symbols or gene aliases information for data input.

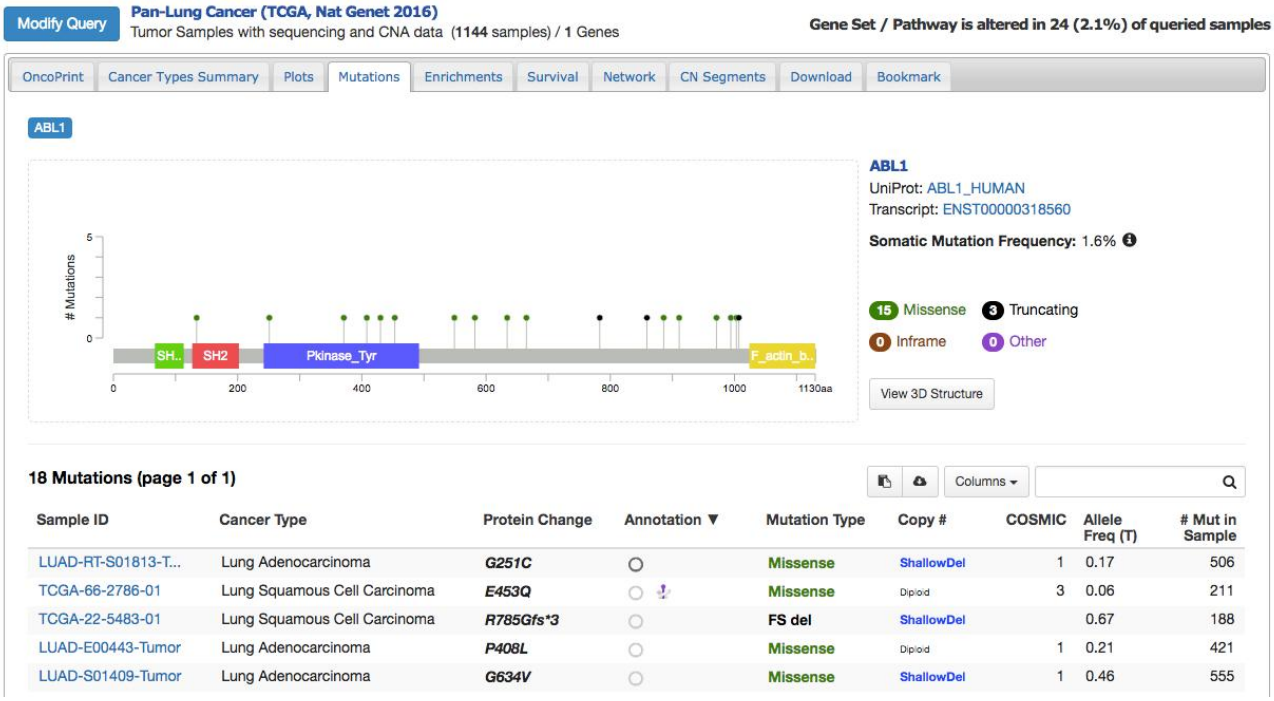


Figure 3: The query of mutation status of ABL1 in the pan lung cancer study. Five of the 18 ABL1 mutations in pan lung cancer occur in a hotspot in the kinase domain. The graphical view has been shown in the Pfam protein domains and the position of the specific mutations. As also shown in the Figure, somatic mutation frequency of ABL1 is 1.6% and this specific gene had 15 missense and 3 truncating mutations. Additionally, the tabular view is provided for more information about all mutations in specific query gene.

COSMIC (Catalogue Of Somatic Mutation In Cancer) is the largest and the most comprehensive resource in the world for exploring of somatic mutations in human cancer. COSMIC is divided into several projects and each section presented separate dataset. COSMIC has two main types of data,: high precision data, and genome wide screen data. High precision data section is manually curated by analysts and work with different targeted gene screening panels, metadata and so on. Genome wide screening section not only provides peer reviewed large scale genome screening data, but also provides unbiased, genome level profiling. It can be used to find novel driver genes. These two-section compilations of data provide extensive coverage of the cancer genomics landscape from a somatic perspective (Forbes et al., 2016).

1.4 Statistical analysis of mutation data

The degree of specific mutations selection may depend on the type of amino acid change information in the protein sequences. In specific, splice site and nonsense mutations can lead to reduced or truncated mutation, respectively (Greenman, Wooster, Futreal, Stratton, & Easton, 2006). Many different models of mutation processes have been identified and explored to model the mutational analysis of cancer genome. Recently introduced ActiveDriver tool, which is based on generalized linear regression model, was employed to help us find phosphosite whose mutations are unexpected given the backbone of mutation rate (Reimand & Bader, 2013).

1.4.1 Regression model and Generalized linear models

Regression analysis is a fundamental tool for analysing and modelling data. The regression analysis can be used to establish the relationship between independent variables and dependent variables. This technique is used for different mutation analysis in cancer genome to find the causal relationship between the variables (Yusuff, Mohamad, Ngah, & Yahaya, 2012).

Logistic regression is one of the most popular and commonly used multivariate tools used in biomedical informatics. In logistics regression, the predicted odd ratio is expressed as positive outcome of variables. Variable is formed by multiplying the values of its coefficient and its independent variables (Yusuff et al., 2012). Since the detection of cancer and prediction of mutation changes information is important, many types of research has been conducted in this area.

Linear regression model build-up a relationship between dependent variable (Y) and one or more independent variables (X) using a regression line and it represent by an equation

$$y = \beta_1 + \beta_2 x + \in$$

where, β_1 is intercept, β_2 is the slop line and $\in$ stands for error. There is a different between simple linear regression and multiple linear regression, whereas simple linear model has only one variable but multiple regression model has more than one independent variables.

Logistic regression is a type of regression model which is used for the finding the probability event which has two conditions as success or failure. Logistic regression model uses the dependent variable is binary (0/1, True/False, Yes/No) in nature. Along with logistic regression model there are

several categories of regression model under the generalized linear models based on the link function (Table 1).

Table 1: The overview of different generalised liner models (Agresti, 2003).

Model	**Random**	**Link**	**Systematic**
Linear Regression	Normal	Identity	Continuous
ANOVA	Normal	Identity	Categorical
ANCOVA	Normal	Identity	Mixed
Logistic Regression	Binomial	Logit	Mixed
Loglinear	Poisson	Log	Categorical
Poisson Regression	Poisson	Log	Mixed
Multinomial Response	Multinomial	Generalized Logit	Mixed

Protein intrinsic disorder sequences are in binary format and it is used in mutational analysis by logistic regression method.

Generalized Linear Models (GLMs) are commonly used to predict the relationship between one response and one or more covariates. A GLM has three parts. The first part is called the linear predictor,

$\eta = \beta_1 + \beta_2 x$ and the second part is the link function assuming,

$$\mu = E(y)$$

$$g(\mu) = \eta$$

Where g is a smooth, monotonic function. The linear predictor builds the relationship between η and the covariate x. Here, the assumption is that, there exists a linear relationship between η and x where β_1 is the intercept and β_2 is the slope. The link function is a function that links the expected value

μ of the response variable to the linear predictor η. The third component is the random or stochastic component. The stochastic component specifies the distribution of the response variable y. The

observations $y_1,\ldots\ldots y_n$ are assumed to be independent and it is assumed that the density of y_i is from the exponential family.

1.4.2 Poisson Distribution

A variable is considered to follow Poisson distribution when the values are count. In Poisson regression, the mean μ_i is explained in terms of explanatory variables with an appropriate link function. So, we can write the Poisson regression model as,

$$y_i \sim P(\mu_i); g(\mu_i) = x_i'\beta$$

Popular choices for $g(\mu_i)$ are the identity link $\mu_i = x_i'\beta$ and log link $\log \mu_i = x_i'\beta$ If we use log link $\mu_i = e^{x_i'\beta}$ is positive, but with the identity link, positivity is not always true (De Jong & Heller, 2008).

2. Aim of the study

Protein Phosphatase 2A (PP2A) is an important and ubiquitously expressed serine threonine phosphatase and plays a critical role in various cellular processes. PP2A constitutes ~ 1% of total cellular proteins. However, understanding basic functions as well as translational potential of PP2A is complex and is at its infancy. The common theme in this study has been the use of mutational approaches of PP2A functions in cancer biology.

The specific aims of the thesis study are listed below-

I. Identify target sites which are directly or indirectly dephosphorylated by PP2A.
II. Tried to understand whether mutations in oncogenes are in the proximity of PP2A-regulated phosphorylation sites.

3. Methods

PP2A is involved in tumor suppressor functions in addition to playing important role in signal transduction in the human cell. This thesis tried to perform systematic investigation of mutational landscape nearby the amino acids dephosphorylated by PP2A.

3.1 Phosphorylation data

Phosphosites and kinases associated with these sites were retrieved from three different publicly available databases named PhosphoELM (Dinkel et al., 2010), PhosphoSitePlus (Hornbeck et al., 2011) and Human Protein References Database(HPRD) (Keshava Prasad et al., 2008). Consensus Coding Sequence (CCDS) database was used for mapping of the phosphosites to high confidence protein sequences. In this study, phosphopeptides were mapped to CCDS sequences using exact sequence matching to avoid discrepancies between the protein isoforms and to discard the unwanted or non-matching peptides. Phosphosites with overlapping protein isoforms sequence were merged together into new continuous regions. In this study, Hugo Gene Nomenclature Committee (HGNC) symbol is used for collecting all gene information from different databases.

PhosphoELM (version 9.0) (Dinkel et al., 2011) dataset contains around 43,000 non redundant instances of phosphorylated residues in over 11,000 different protein sequences. All those validated sites, over 37,000 belongs to Homo sapiens. Among those phosphorylation sites, 90% of all phosphorylation occurs in serine or threonine residues and the number of phosphorylation sites among serine, threonine and tyrosine sites are 27421, 6256 and 3467 respectively.

3.2 Dephosphorylation data

Dephosphorylation is an independent mechanism in allosteric control of protein function. For further understanding of serine or threonine behavior in protein function, dephosphorylation mechanisms hold a huge potential for therapeutic modulation of cell signaling. Dephosphorylation data is needed for understanding of protein phosphatases and their roles in cancer cell signaling.

The human DEPhOsphorylation Database (DEPOD) (version 1.1) (Duan, Li, & Köhn, 2014) is manually curated database which collects information on human phophatases and their substrates along with information on dephosphorylation site. In this study, DEPOD used for studying human phosphatases and to understand their molecular mechanisms. It also connecting phosphatases with

kinases through their common substrate and compiling the human dephosphorylation network junction. DEPOD focuses only human phosphatases with enzymatic activities.

PP2A data collected from different research groups and their collaboration work platforms. For example, from Westermarck's lab (Kauko, Imanishi, et al., 2018), used the cell line of Hela and A549 and data was statistically significant, whereas Narla´s (Wiredja et al., 2017) data was statistically insignificant.

3.3 Mutational data

Somatic mutation from different cancer project are downloaded from cBioportal (Gao et al., 2013) and COSMIC (Forbes et al., 2016) data sources. Mutation data was formatted to be compatible with ActiveDriver software. ActiveDriver needs wild type residue, position and mutated residue. Mutation information from cBioportal and COSMIC database was split into three different categories called wild type residue, position and mutated residue.

cBioportal offers 216 different cancer datasets (as on 21.06.2018) from different projects. Additionally, webserver helps mining of underlying mutational of data. For example, the mutational tab provides both graphical and customized table about the nonsynonymous mutations identified in specific gene. The position and frequency of all mutations in the context of Pfam protein domain find in the graphical summary section. In this study, we had chosen the large data sets with different cancer types for example pan cancer (MSKCC, Nat Med 2017; 10945 samples) (Zehir et al., 2017), breast cancer (METABRIC, Nature 2012 & Nat Commun 2016; 2509 samples) (Pereira et al., 2016). The graphical summary of all nonsynonymous mutations is presented in table format. This table can be filtered and sorted, provides the information about case ID, amino acid change, type of mutation (nonsense, missense, splice site, frameshift insertion or deletion, in-frame insertion or deletion, nonstop, non-start), predicted functional impact of missense mutations, mutation status, validation status and exact genomic position. To avoid potential errors while using ActiveDriver, data were further filtered to remove the data for which amino acid information is missing and these missing data were mainly coming from non-coding mutation, frameshift, splice and truncated mutation. Isoform of specific gene was chosen based on their matching with phosphosites isoform information and selection of the right isoform was hard in this study.

Mutation data downloaded from COSMIC (version85) for targeting specific gene with its amino acid change information. All genes followed with entrez ID and HGNC ID.

3.4 Intrinsic disorder score

Intrinsic disorder scores play crucial role pathologies associated with aggregation and mis-folding of protein. The lacking of tertiary structure of a protein called intrinsic disorder of protein. Manycomputational methods are developed to predict whether a protein is disordered, given its amino acid residues. For this study intrinsic disorder scores of all proteins were computed using DisProt (Vucetic et al., 2005) as available from DP2 database (http://d2p2.pro/). If intrinsically disordered protein in the absence of mutation information for these amino acid sequences, thenthe scores are then binerised as required by ActiveDriver software. This dataset of intrinsic disorder scores are given as input to ActiveDriver tool along with phosphosite dataset.

3.5 Statistics analysis of phosphosite and mutation data

ActiveDriver package was used in R (RStudio version 1.0.1336 with R v 3.0.1) for statistical analysis. It uses the GLM (generalized linear regression model) approach which helps find out the driver genes for cancer with frequent mutations in protein signaling sites such as phosphosites. It uses the Poisson regression model which finds out the genes where the mutations in signaling sites are more frequent by assuming that missense mutations follow Poisson probability distribution in cancer gene sequences. The idea behind Poisson regression and estimating of the parameters has been shown below step by step.

The Poisson distribution has the following distribution function

$$P(y;\mu) = \frac{\mu^y \exp(-\mu)}{y!} \text{ with } E(Y) = \mu \; and \; Var(Y) = \mu$$

Where $y \geq 0, y \in N$ the observed number of amino acid sequences and $\mu > 0, \mu \in R$ the average rate of mutations of the protein sequence in a gene. Since it is considered as a regression model an independent variable x must be considered and a simple linear model can be written as,

$$\mu_i = x_i^{'}\beta \text{----------(1)} \quad \text{with an identity link function}$$

or

$$\log \mu_i = x_i^{'}\beta \text{ ----------(2)} \quad \text{with a log-link function}$$

Equation (2) can be expressed as

$$\mu_i = e^{x_i'\beta} \text{----------(3)}$$

Where $\beta = \beta_1 \dots \dots \beta_k$ a vector of coefficients. One common drawback of the identity link is that the independent variable x can have any real value whereas the mean μ on the left-hand side must be non-negative since it represents the expected value of a count variable. However, this problem can be solved by considering a log-link function which is equation (2). As mentioned in the introduction a Poisson linear regression model does not directly models the dependent variable y with the independent variable x but it considers the function of the mean of y which is commonly known as linear predictor.

To find out the parameters of a distribution function the maximum likelihood approach is used. In summary to find out the maximum estimate for the parameter μ, derivative of the log likelihood function for the probability distribution function (for Poisson distribution) would be set equal to zero and thus the value of μ will be estimated and this value will be the maximum value. Here, considering the link function in equation (2) the likelihood function for Poisson distribution will be,

$$\text{L}= \sum_{i=1}^{n} \frac{e^{-\mu_i} \mu_i^{y_i}}{y_i}$$

and the log-likelihood function will be,

$$\log L(\beta) = \sum_{i=1}^{n} \{y_i \log(\mu_i) - \mu_i - \log(y_i!)\}$$

considering the value of μ_i from equation (3)

$$\log L(\beta) = \sum_{i=1}^{n} \left\{y_i \log\left(e^{x_i'\beta}\right) - e^{x_i'\beta} - \log(y_i!)\right\}$$

$$= \sum_{i=1}^{n} \left\{y_i \left(x_i'\beta\right) - e^{x_i'\beta} - \log(y_i!)\right\} \quad \text{as we know } \log e = 0 \text{-------- (4)}$$

If the model considers only one independent variable then it can also be expressed as,

$$\log \mu_i = x_i'\beta = \beta_1 + \beta_2 x_i + \in$$

Now going back to equation (4)

$$= \sum_{i=1}^{n} \{y_i (\beta_1 + \beta_2 x_i) - e^{\beta_1 + \beta_2 x_i}\} \text{--------(5)}$$

The part of ∈ and $\log(y_i!)$ has been ignored since they are constant. The next step will be to find out the differentiation of (5) with respect to β_1 that is why the constant parts in the equation can be ignored since they will be equal to zero.

Now differentiating of (5) with respect to β_1,

$$\frac{d}{d\beta_1}\left[\sum_{i=1}^{n}\{y_i\,(\beta_1+\beta_2 x_i)-\ e^{\beta_1+\beta_2 x_i}\}\right]=0$$

$$=>\sum_{i=1}^{n}\left(y_i-\ e^{\beta_1+\beta_2 x_i}\right)\ =0$$

$$=>\sum_{i=1}^{n}y_i=\sum_{i=1}^{n}e^{\beta_1+\beta_2 x_i}$$

There is no closed form solution for β_1 in the above equation. To find the optimal value an iteratively reweighted least squares (IRLS) is used (Fox & Monette, 2002). Similar equation can be shown for β_2 as well. For this thesis work the Poisson GLM has been used to see if a phosphosite region has significant mutation rate than other parts of the gene. Like every other regression model this Poisson model has considered null $(h_0)$and alternative (h_1)hypothesis.

$$\boldsymbol{h_0}\colon \mu = e^{\beta_1+\beta_2 x_{2i}}=e^{\beta_1}e^{\beta_2 x_{2i}}=e^{\beta_1}e^{\beta_2}e^{x_{2i}}=\ e^{(\beta_1+\beta_2+x_{2i})}$$

The null hypothesis assumes that mutations in protein sequences follows Poisson distribution with intercept parameter β_1 linearly combined with a predictor or independent variable which represent disordered or non-ordered protein sequence and corresponding coefficient β_2.

$$\boldsymbol{h_1}\colon \mu = e^{\beta_1+\beta_2 x_{2i}+\beta_3 x_{3i}}=e^{\beta_1}e^{\beta_2 x_{2i}}e^{\beta_3 x_{3i}}=e^{\beta_1}e^{\beta_2}e^{\beta_3}e^{x_{2i}}e^{x_{3i}}=\ e^{(\beta_1+\beta_2+x_{2i}+\beta_3+x_{3i})}$$

In the alternative hypothesis, the mutations in the phosphosite region three might have effect on the observed number of protein sequences. The effect of other independent variables on mutation are set to zero in amino acid sequences outside the considered phosphosite region and encode relative phosphosites position within the region. A flanking region (+/- 7) of residues around the sequence position *i* has been used in ActiveDriver methodology.

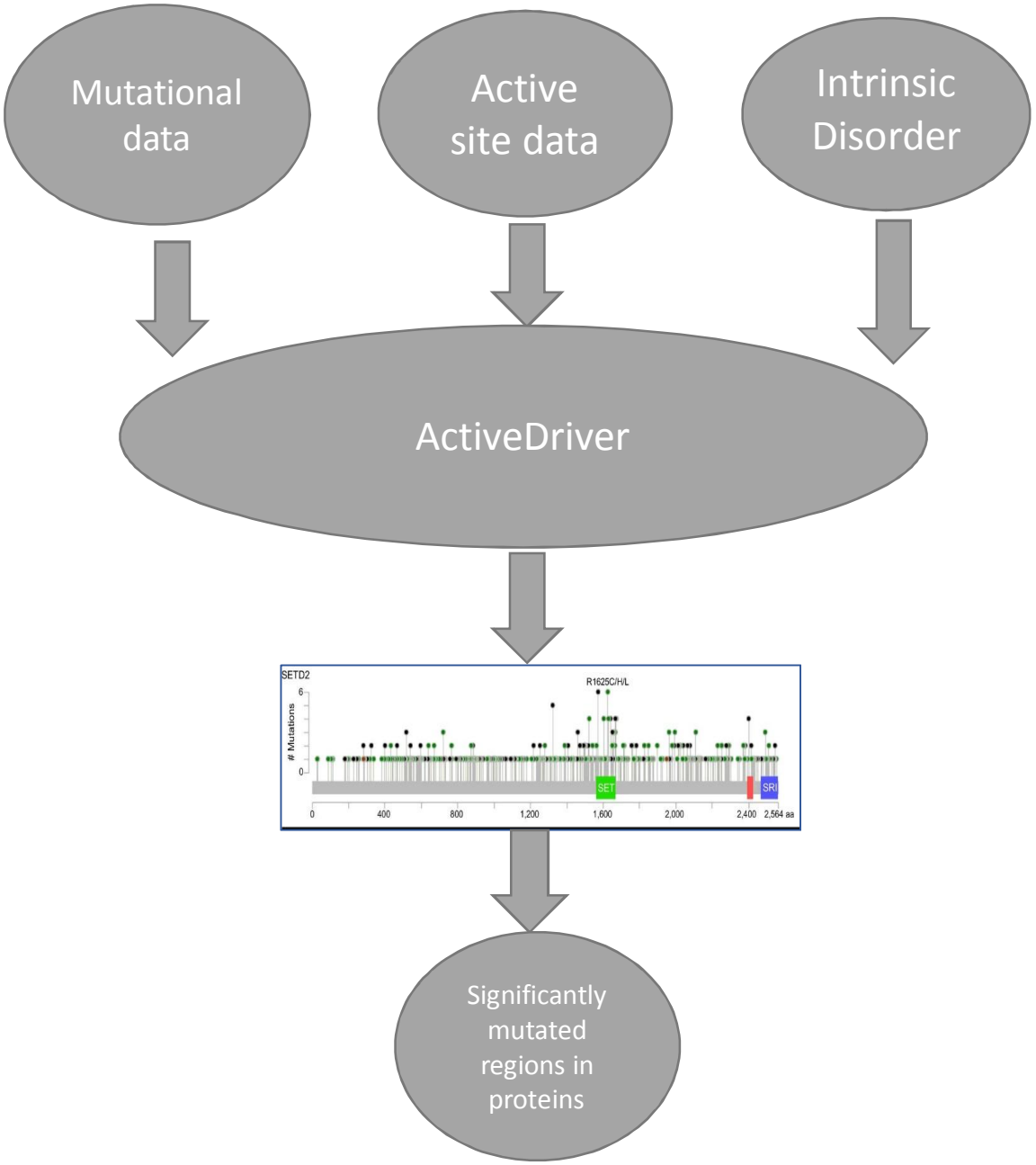
Benjamini and Hochberg (1995) method for controlling FDR can be used. The formula for calculating adjusted p-value according to Benjamini and Hochberg (1995) is,

$$q_i = p_i \frac{N}{i}$$

Where p_i is the i^{th} smallest P-value out of N total P-values for the experiment.

The general methodology overview in this study is given bellow.

Scheme to identify active sites that are
nearby significant mutations in cancer genomes



4. Results and Discussion

The focus of this thesis was to identify the dephosphorylation sites of PP2A target proteins in the neighbourhood of significant recurrent mutations in cancer samples. To achieve this goal, large scale mutational and phosphorylational datasets were integrated. Before performing the final analysis, the data have been cleaned to obtain accurate results. This results and discussion section has been divided into the following sub-sections in order to achieve the main aim of this thesis.

4.1 Identification of PP2A targets

Most cancer therapy resistance pathways are controlled by PP2A which regulates large number of cellular processes (Kauko, O'Connor, et al., 2018). The mechanisms by which PP2A performs its functions is very unclear. The identification of PP2A targets and functions behind those target proteins may provide valuable insights about PP2A biology. In this section, target identification process for PP2A was described using an in-house PP2A phosphoproteomics dataset (hereafter referred to as PP2A-B56 dataset), which was generated by knocking-down of a subunit of B56 (i.e., PPP2R5A). Student's t-test for each peptide followed by multiple hypotheses correction (Greenwood et al., 2016) (FDR analysis) was performed to identify the targets of PP2A.

Two sample t-test on log2-transformed data was performed to assess whether the means of two groups (i.e., control *vs.* siRNA knock-down) from B56 samples were significantly different for each peptide. As the statistical test was performed for 6739 peptides in this dataset simultaneously, multiple hypothesis correction was performed to reduce the chances of type 1 errors which happens when a null hypothesis has incorrectly been rejected. Top hits coming from student's t-test analysis of B56 subset is presented in Table 2.

Table 2: Top hits coming from student's t-test analysis in PP2A-B56 dataset.

Gene	Peptides	p-sites	Log fold Change	*P*-value	*P*-adjusted
NBN	O60934_IPNYQLSPTKLPSINK,[7] Phospho (S)	S432	1.05	0	0
SSRP1	Q08945_GLKEGMNPSYDEYADSDEDQHDAYLER,[6] Oxidation; (M)|[16] Phospho (S)	S444	1.01	0	0
CPD	O75976_SLLSHEFQDETDTEEETLYSSKH,[11] Phospho (T)	T1368	1.04	0	0
SPEN	Q96T58_SNSPRGEAQKLLELK,[3] Phospho (S)	S1857	1.06	0	0
DDX41	Q9UJV9_TDEVPAGGSRSEAEDEDDEDYVPYVPLR,[11] Phospho (S)	S23	1.09	0	0
LMNB1	P20700_LLEGEEERLKLSPSPSSR,[12] Phospho (S)	S391	1.17	0	0
CYBRD1	Q53TN4_NLALDEAGQRSTM,[12] Phospho; (T)|[13] Oxidation (M)	T285	0.88	0	0
NBN	O60934_IPNYQLSPTKLPSINK,[7] Phospho (S)	S432	1.06	0	0
ZC3HAV1	Q7Z2W4_FLENGSQEDLLHGNPGSTYLASNSTSAPNWK,[6] Phospho (S)	S335	1.05	0	0
TICRR	Q7Z2Z1_NLFNQELLSPSKR,[9] Phospho (S)	S923	1.18	0	0
FAM53C	Q9NYF3_FSLSPSLGPQASR,[4] Phospho (S)	S234	1.07	0	0
NCL	P19338_AIRLELQGPRGSPNAR,[12] Phospho (S)	S563	1.04	0.001	0.14
THRAP3	Q9Y2W1_RIDISPSTFR,[5] Phospho (S)	S682	1.05	0.001	0.14
TOP2A	P11388_KPIKYLEESDEDDLF,[9] Phospho (S)	S1525	1.04	0.001	0.14
SIPA1L1	O43166_TLSDESIYNSQREHFFTSR,[3] Phospho (S)	S1585	1.05	0.001	0.14
KRT7	P08729_LSSARPGGLGSSSLYGLGASRPR,[12] Phospho (S)	S37	1.13	0.001	0.14
NUFIP2	Q7Z417_GLERNDSWGSFDLR,[7] Phospho (S)	S652	1.05	0.001	0.14
CDC23	Q9UJX2_RVSPLNLSSVTP,[3] Phospho (S)	S588	1.05	0.001	0.14
BCLAF1	Q9NYF8_LKDLFDYSPPLHKNLDAR,[8] Phospho (S)	S512	1.06	0.001	0.14
PIEZO1	Q92508_TASELLLDRR,[3] Phospho (S)	S1646	1.09	0.001	0.14

Resulting data revealed that 1249 out of 6739 peptides were statistically significant (p value < 0.05). However, adjustment of multiple hypothesis correction resulted in 11 significant peptides (adjusted p-value < 0.05).

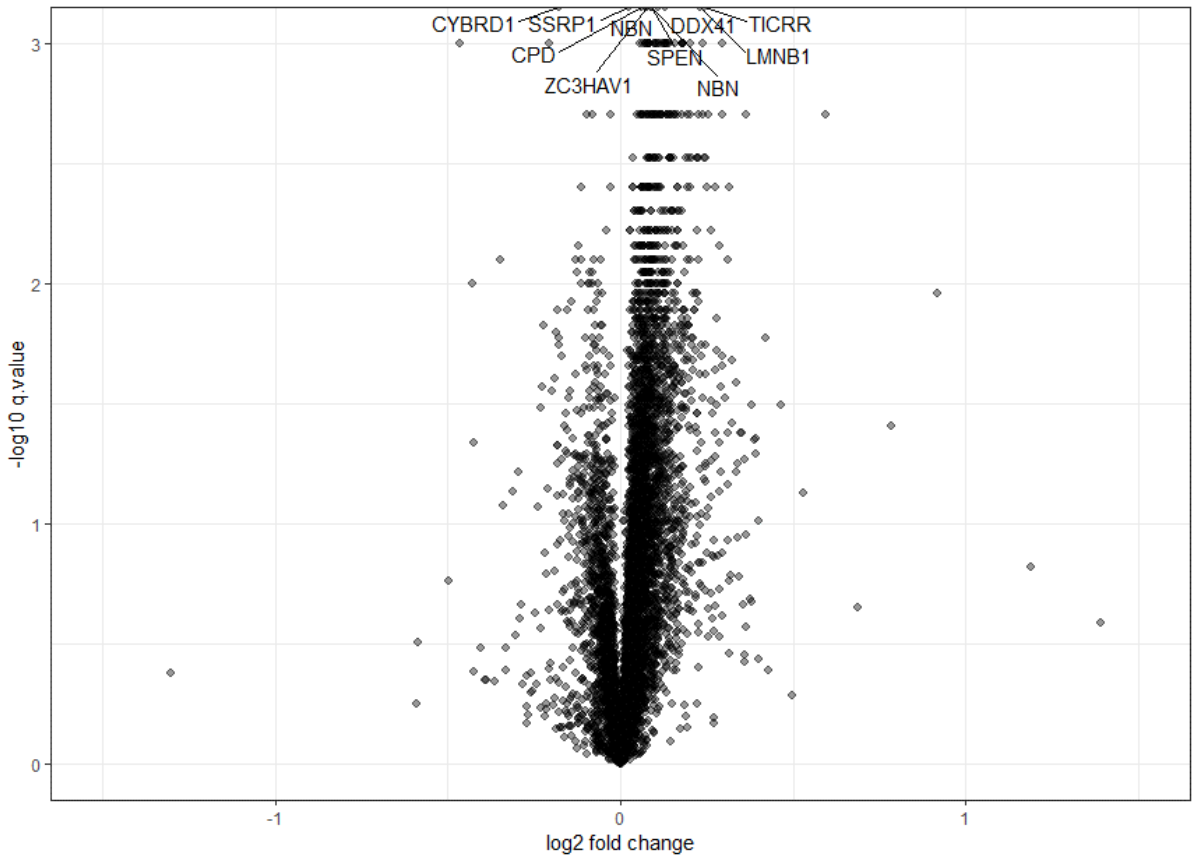


Figure 4: Volcano plot for the analysis of PP2A-B56 dataset, comparing knock-down vs. control group. Adjusted p-values (q values) on y-axis and log2 fold change values on x-axis are shown. Most significant peptides are marked.

In order to facilitate easy interpretation of significant hits from PP2A dephoshorylation experiment, results from t-test are visualized using volcano plot (Figure 4) where most significant top ten proteins with peptides are shown. Additionally, scaled expression values for the top ten peptides are shown in heatmap (Figure 5).

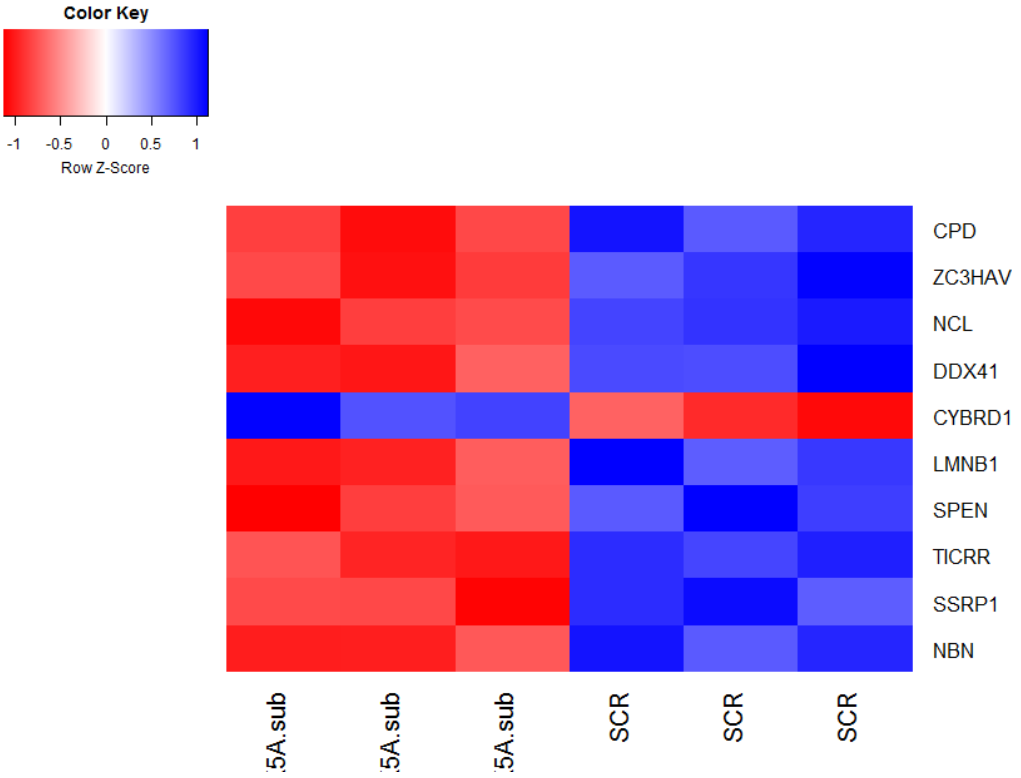


Figure 5: Heatmap shows top ten significant genes in PP2A-B56 dataset, comparing control and siRNA knock-down groups in Hela cell lines. Scale bar on the upper left shows scaled expression levels of peptides and the degree of redness, blueness colors represent negative and positive values of z-scores.

We retained all significant peptides (adjusted $p < 0.05$) to construct a PP2A dephosphorylome dataset.

4.2 Building a comprehensive PP2A dephosphorylome

PP2A target identification was explained using a PP2A-B56 dataset as described in earlier section. Comprehensive compilation of PP2A targets can facilitate our understanding about the role of PP2A in a broader context. We therefore combined the targets from PP2A-B56 dataset with other PP2A manipulations (Table 3). The target identification process for each dataset was described in respective publication. Three groups of PP2A families was combined to build a comprehensive dataset.

Table 3: Collection of datasets for building comprehensive PP2A dephosphorylome.

Conditions	**Source /Publication**
PP2R1A, CIP2A, SET, and PME1	(Kauko, Imanishi, et al., 2018)
SMAPs	(Wiredja et al., 2017)
B56	(Hertz et al., 2016)
B55	(Cundell et al., 2016)

Only unique entries were retained in final PP2A dephosphorylation database which comprise 3,398 peptides. This database was later used as input data in ActiveDriver software.

4.3 Collection of mutational data

Recently published data from Pancancer study (MSK-IMPACT clinical sequencing cohort) from cBioportal was used as a dataset for recurrent mutations (Zehir et al., 2017). Mutational information of the dataset was collected from 10,336 patients. The study revealed 530 mutated genes of which, "TP53" gene is most frequently mutated gene (frequency rate 41.67 %). The information from mutational data include amino acid change, its wild type and mutation information and position. This information is necessary for the input of ActiveDriver software. As dephosphorylation sites of PP2A were identified using canonical isoform of proteins in uniprot database, the mutations data that are mapped to canonical isoform of uniprot proteins are retained. Based on different mutational type, only thirty-five genes with their mutation type, amino acid change and chromosome number information have been shown in Table 4.

Table 4: Mutational data for Top thirty-five genes from cBioportal cancer genomic web source. Wt= Wild type; Mt= Mutation; Chr= Chromosome.

Gene	Mutation type	Amino acid change	Wt residue	Position	Mt residue	Chr
HLA-B	Frame_Shift_Del	L154Rfs*18	L	154	Rfs*18	6
HLA-C	Frame_Shift_Del	R7Efs*13	R	7	Efs*13	6
ACVR1	Frame_Shift_Del	T507Lfs*14	T	507	Lfs*14	2
ARAF	Frame_Shift_Del	R255Gfs*37	R	255	Gfs*37	23
APC	Frame_Shift_Del	K581Gfs*20	K	581	Gfs*20	5
PMAIP1	Frame_Shift_Del	L43*	L	43	*	18
ASXL1	Frame_Shift_Del	Y700*	Y	700	*	20
ASXL1	Frame_Shift_Del	W796Gfs*3	W	796	Gfs*3	20
ARID1B	Frame_Shift_Del	K1130Sfs*68	K	113	0Sfs*68	6
ARID5B	Frame_Shift_Del	I497*	I	497	*	10
ARID5B	Frame_Shift_Del	Q941Hfs*16	Q	941	Hfs*16	10
ARID2	Frame_Shift_Del	R80Efs*10	R	80	fs*10	12
ARID2	Frame_Shift_Del	Q961Hfs*14	Q	961	Hfs*14	12
ARID2	Frame_Shift_Del	G735Efs*23	G	735	Efs*23	12
CDKN2A	Frame_Shift_Del	V95Afs*22	V	95	fs*22	9
ARID1A	Frame_Shift_Del	L2016Cfs*14	L	201	6Cfs*14	1
ARID1A	Frame_Shift_Del	P224Rfs*8	P	224	Rfs*8	1
TP53	Missense_Mutation	C238W	C	238	W	17
TP53	Frame_Shift_Del	H297Tfs*48	H	297	Tfs*48	17
BRAF	Missense_Mutation	P422A	P	422	A	7

BRAF	Missense_Mutation	I572F	I	572	F	7
CIC	Nonsense_Mutation	S1105*	S	110	5*	19
CEBPD	Nonsense_Mutation	Y194*	Y	194	*	8
CENPA	Nonsense_Mutation	R52*	R	52		2
CHEK1	Nonsense_Mutation	E183*	E	183	*	11
KIT	In_Frame_Del	W557_K558del	W	557	_K558del	4
KIT	In_Frame_Del	M552_Y570del	M	552	_Y570del	4
MDC1	In_Frame_Del	G207_F214del	G	207	_F214del	6
MDC1	In_Frame_Del	G207_F214del	G	207	_F214del	6
MN1	In_Frame_Del	Q550del	Q	550	del	22
MN1	In_Frame_Del	Q549_Q550del	Q	549	_Q550del	22
MN1	In_Frame_Del	Q549_Q550del	Q	549	_Q550del	22
MAP2K1	In_Frame_Del	E102_I103del	E	102	_I103del	15
TP53	Nonsense_Mutation	Q144*	Q	144	*	17
TP53	Nonsense_Mutation	Q144*	Q	144	*	17

4.4 Integration of DEPOD data with significant recurrent mutations

In order to find the extend of overlapping sites between documented dephosphorylation sites of any phosphatase and known phosphosites nearby significantly mutated residues, ActiveDriver was used. The known phosphosite information of all proteins were collected from PhosphoSitePlus (https://www.phosphosite.org/homeAction.action) and HPRD (http://www.hprd.org/). Resulting data obtained from ActiveDriver were integrated with DEPOD data which has valuable information regarding dephosphorylational sites catalogued against respective phosphatases. The data integration was accomplished by matching with PTM position in the output of ActiveDriver results with dephoshorylation sites of DEPOD. The merged results for top fifteen genes with significant active region *p*-values are shown in Table 5. Strikingly and as expected, we found only fewer known dephosphorylated sites were overlapped with the results from ActiveDriver and the poor overlapping can in part be attributed to limited characterisation target sites of human phosphatases. For example, these analyses demonstrated cases where significantly mutated amino acid is nearby the dephosphorylation sites by any phosphatase. For example, TP53 gene had phosphorylated serine at position at 37 and 315 and were found to be nearby significantly mutated amino acids (p-value < 0.05) and DEPOD data also showed those phosphosites were dephosphorylated by CDC14A and CDC14B respectively. Other genes, MET and JAK2, which has residues named “Tyrosine”. DEPOD also gave same position for MET and JAK2 and its phosphatase name were PTP1B CBL gene had PTM position at 731 which has residue name “Tyrosine” and its showing significant (p-value < 0.05), DEPOD also gave at same position and its phosphatase name was RPTPeta.

Overall, this small exploratory study done by integrating DEPOD data with results from ActiveDriver revealed that very little is known on human phosphatases to link to those functions that mutated residues play in cancer studies. Dephosphorylated sites nearby recurrent mutations may play similar signalling role as mutated amino acids. However, PP2A related target sites were very poorly elucidated. Thus, there is greater need for identifying the targets of important phosphatases such as PP2A which is known to play tumour suppression functions in cancer.

Table 5: Merged results from DEPOD and ActiveDriver results from MSK-IMPACT clinical sequencing cohort (Zehir et al., 2017) mutational data.

Gene	PTM position	Residue	Active region p value	Phosphatase	Position
TP53	37	S	1.31E-196	CDC14A	Ser-37
TP53	315	S	4.41E-36	CDC14B	Ser-315
RET	952	Y	NA	RPTPeta	Tyr-905
RET	905	Y	0.023199	RPTPeta	Tyr-905
PTEN	398	T	4.42E-10	PTEN	Ser-380
PDGFRB	857	Y	NA	LMW-PTP	Tyr-857
MET	1234	Y	0.010267	PTP1B	Tyr-1234
MET	1003	Y	2.22E-12	PTP1B	Tyr-1234
KIT	553	Y	3.54E-09	SHP1	N/A
JAK2	1007	Y	0.11465	PTP1B	Tyr-1007
EGFR	511	S	0.000288	CDC25A	N/A
EGFR	290	T	1.26E-21	CDC25A	N/A
CDK4	172	T	0.147357	CDC25A	N/A
CBL	731	Y	0.002554	RPTPeta	Tyr-731
BRCA1	967	T	0.056432	PP1alpha	Ser-988

Location of dephosphorylated sites can be important for the functional insights. Therefore, lollipop plot (Jay & Brouwer, 2016) was used to display phosphosites. As an example, Figure 6 shows the lollipop plot for TP53 gene which shows S37 and S315 positions in active region 1 and 9. Both of these dephosphorylated positions are located in disordered regions of TP53.

Figure 6: A lollipop plot of TP53 gene showing dephosphorylated sites at S37 and S315.

4.5 Identification of PP2A target sites nearby recurrent mutations

Once PP2A target sites are known, the information can then be integrated with mutation information in cancer samples using ActiveDriver method. Datasets downloaded from cBioportal and COSMIC were used as example mutation datasets in this thesis. Specifically, data generated from a recent pancancer study on 10,945 samples (MSK-IMPACT clinical sequencing cohort, Nat Med 2017) from cBioportal databases was used. The mutational, phosphorylational and intrinsic disorder data were used as input datasets for ActiveDriver software. The summary of resulting analysis as obtained as merge report from ActiveDriver is shown in Table 6. The merge report provides information such as gene name, active region, mutational position, post-translational modification, kinase and its active region *p*-value. Table is sorted based on active region p-value and the only significant values are shown here in table format. We found 19 genes with 248 active regions have significant *p*-value < 0.05. Most significant result found in "TP53" gene and its S315 and S392 positions, which identify the influence of mutation in binding of PP2A target (Figure 9 a). There are 75 genes are non-significant. The active region *p*- values are depicted with histogram in Figure 7.

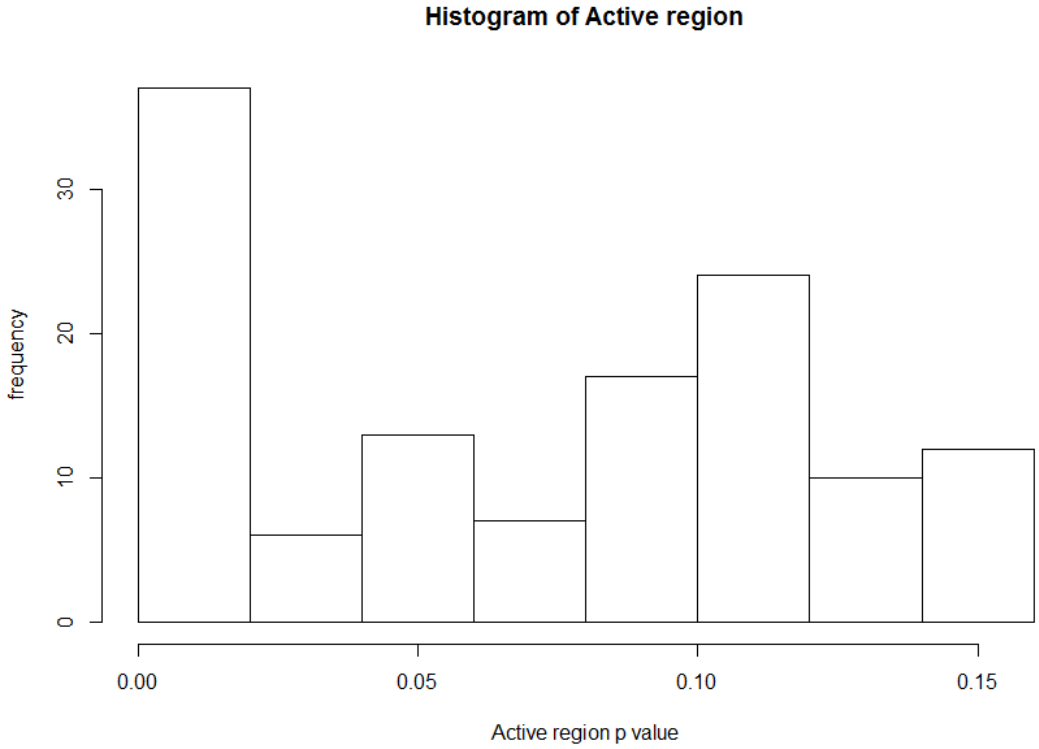


Figure 7: Histogram of active region p-values from ActiveDriver as analysed from C-bioportal dataset. X-axis represents active region p-values and Y-axis represents for frequency of significant level.

However, there are number of genes which did not give any *p*-values. It might have been caused with little or no data for modelling in ActiveDriver because analysis was restricted to flank region (+/- 7) around phosphosite position. Similar data analysis has been performed based on the data from "Breast cancer (METABRIC, Nature 2012 and common 2016)" from cBioportal (Appendix).

Table 6: Merge report of genes with significant active regions in the analysis with pan-cancer dataset (Zehir et al., 2017) from cBioportal.

Gene	Mutation position	Active region	Frequency	PTM position	Residue	Dataset source	Active region *p* Value
TP53	392	2	385	392	S	PIPs	6.27E-17
TP53	392	2	393	392	S	PIPs	6.27E-17
TP53	317	1	308	315	S	A549_SMAP	3.53E-10
RB1	251	1	242	249	S	A549_SMAP	1.07E-06
RB1	251	1	256	249	S	A549_SMAP	1.07E-06
RB1	255	1	242	249	S	A549_SMAP	1.07E-06
RB1	255	1	256	249	S	A549_SMAP	1.07E-06
KMT2A	1855	2	1865	1858	S	A549_SMAP	0.000121
KMT2A	1855	2	1851	1858	S	A549_SMAP	0.000121
KMT2A	1862	2	1865	1858	S	A549_SMAP	0.000121
KMT2A	1862	2	1851	1858	S	A549_SMAP	0.000121
KMT2A	1864	2	1865	1858	S	A549_SMAP	0.000121
KMT2A	1864	2	1851	1858	S	A549_SMAP	0.000121
KMT2A	1852	2	1865	1858	S	A549_SMAP	0.000121
KMT2A	1852	2	1851	1858	S	A549_SMAP	0.000121
NCOR1	2185	6	2191	2184	S	PIPs	0.001952
NCOR1	2185	6	2177	2184	S	PIPs	0.001952
NCOR1	2183	6	2191	2184	S	PIPs	0.001952
NCOR1	2183	6	2177	2184	S	PIPs	0.001952
FOXK1	420	2	452	420	S	PIPs	0.00396
FOXK1	420	2	452	416	S	PIPs	0.00396
FOXK1	420	2	452	413	S	PIPs	0.00396
FOXK1	420	2	400	420	S	PIPs	0.00396
FOXK1	420	2	400	416	S	PIPs	0.00396
FOXK1	420	2	400	413	S	PIPs	0.00396
KMT2D	4849	5	4856	4849	S	A549_SMAP	0.003975
KMT2D	4849	5	4842	4849	S	A549_SMAP	0.003975
SETD2	831	2	824	831	S	H358_SMAP	0.004247
SETD2	831	2	838	831	S	H358_SMAP	0.004247
TOE1	7	1	1	5	S	PIPs	0.005099
TOE1	7	1	1	5	S	A549_SMAP	0.005099
TOE1	7	1	12	5	S	PIPs	0.005099
TOE1	7	1	12	5	S	A549_SMAP	0.005099
CTNNB1	67	1	67	60	S	H358_SMAP	0.00745
CTNNB1	67	1	53	60	S	H358_SMAP	0.00745

CTNNB1	63	1	67	60	S	H358_SMAP	0.00745
CTNNB1	63	1	53	60	S	H358_SMAP	0.00745
GSK3A	284	2	272	279	Y	A549_SMAP	0.008155
GSK3A	284	2	286	279	Y	A549_SMAP	0.008155
TSC1	500	1	498	505	S	A549_SMAP	0.009679
TSC1	500	1	512	505	S	A549_SMAP	0.009679
TSC1	509	1	498	505	S	A549_SMAP	0.009679
TSC1	509	1	512	505	S	A549_SMAP	0.009679
TSC1	502	1	498	505	S	A549_SMAP	0.009679
TSC1	502	1	512	505	S	A549_SMAP	0.009679
ATRX	895	5	882	889	S	A549_SMAP	0.010994
ATRX	895	5	896	889	S	A549_SMAP	0.010994
ATRX	886	5	882	889	S	A549_SMAP	0.010994
ATRX	886	5	896	889	S	A549_SMAP	0.010994
ATRX	892	5	882	889	S	A549_SMAP	0.010994
ATRX	892	5	896	889	S	A549_SMAP	0.010994
ATRX	885	5	882	889	S	A549_SMAP	0.010994
ATRX	885	5	896	889	S	A549_SMAP	0.010994
ATRX	896	5	882	889	S	A549_SMAP	0.010994
ATRX	896	5	896	889	S	A549_SMAP	0.010994
SRSF2	25	1	33	25	T	A549_SMAP	0.015415
SRSF2	25	1	33	26	S	PIPs	0.015415
SRSF2	25	1	18	25	T	A549_SMAP	0.015415
SRSF2	25	1	18	26	S	PIPs	0.015415
CDK12	319	1	316	325	S	PIPs	0.015929
CDK12	319	1	316	323	S	PIPs	0.015929
CDK12	319	1	332	325	S	PIPs	0.015929
CDK12	319	1	332	323	S	PIPs	0.015929
CDK12	327	1	316	325	S	PIPs	0.015929
CDK12	327	1	316	323	S	PIPs	0.015929
CDK12	327	1	332	325	S	PIPs	0.015929
CDK12	327	1	332	323	S	PIPs	0.015929
CDK12	329	1	316	325	S	PIPs	0.015929
CDK12	329	1	316	323	S	PIPs	0.015929
CDK12	329	1	332	325	S	PIPs	0.015929
CDK12	329	1	332	323	S	PIPs	0.015929
CDK12	322	1	316	325	S	PIPs	0.015929
CDK12	322	1	316	323	S	PIPs	0.015929
CDK12	322	1	332	325	S	PIPs	0.015929
CDK12	322	1	332	323	S	PIPs	0.015929
NCOA3	218	1	221	214	S	A549_SMAP	0.021552
NCOA3	218	1	207	214	S	A549_SMAP	0.021552
NCOA3	217	1	221	214	S	A549_SMAP	0.021552

NCOA3	217	1	207	214	S	A549_SMAP	0.021552
NCOA3	213	1	221	214	S	A549_SMAP	0.021552
NCOA3	213	1	207	214	S	A549_SMAP	0.021552
TGFBR2	349	1	345	352	S	A549_SMAP	0.041689
TGFBR2	349	1	359	352	S	A549_SMAP	0.041689
TGFBR2	356	1	345	352	S	A549_SMAP	0.041689
TGFBR2	356	1	359	352	S	A549_SMAP	0.041689
TGFBR2	355	1	345	352	S	A549_SMAP	0.041689
TGFBR2	355	1	359	352	S	A549_SMAP	0.041689
TGFBR2	357	1	345	352	S	A549_SMAP	0.041689
TGFBR2	357	1	359	352	S	A549_SMAP	0.041689
MAX	10	1	18	11	S	PIPs	0.046041
MAX	10	1	1	11	S	PIPs	0.046041
MAX	8	1	18	11	S	PIPs	0.046041
MAX	8	1	18	2	S	PIPs	0.046041
MAX	8	1	1	11	S	PIPs	0.046041
MAX	8	1	1	2	S	PIPs	0.046041
MAX	15	1	18	11	S	PIPs	0.046041
MAX	15	1	1	11	S	PIPs	0.046041
NPM1	125	2	144	125	S	PIPs	0.054797
NPM1	125	2	144	125	S	A549_SMAP	0.054797
NPM1	125	2	118	125	S	PIPs	0.054797
NPM1	125	2	118	125	S	A549_SMAP	0.054797

In order to find the influential substrates of PP2A, the network analysis was also performed on PP2A target substrates for which there exist at least one significant active region. The analysis was performed using STRING database (https://string-db.org). TP53 gene was a densely connected node (degree 14) with other 14 genes. Other important genes include, NCOR1, KMT2A, NPM1 which have eight, seven and six as their degree respectively. However, some genes such as FOXK1 has zero degree, showing no interaction with other genes. The degree distribution of all genes are shown in Table 7.

Table 7: Most significant genes and their degree. Here, degree means number of undirected edges.

Gene	Degree
TP53	14
NCOR1	8
KMT2A	7
NPM1	6
CTNNB1	6
RB1	5
SETD2	5
MAX	5
KMT2D	4
CDK12	4
NCOA3	3
ATRX	3
TGFBR2	3
GSK3A	2
SRSF2	2
SPEN	1
TOE1	1
TSC1	1
FOXK1	0

Functional enrichment analysis was performed on PP2A targets from above results to explore their association with any cancer –specific functions. The analysis was done by comparing the input above gene set (Table 7) to each of the bins in the gene ontology. In this analysis, minimum required interaction score was 0.40. For functional enrichment analysis and network statistics, these significant genes also have been performed on Among these genes, there are 24 number of edges are connected and average node degree is 5.33 (Figure: 8). An important results from the analysis is that PPI network is enriched significantly (p-value < 2.68e-1) which means above 19 genes have more interactions among themselves than at random, indicating co-ordinated biological role of these genes. The false discovery rate also found significant in this analysis.

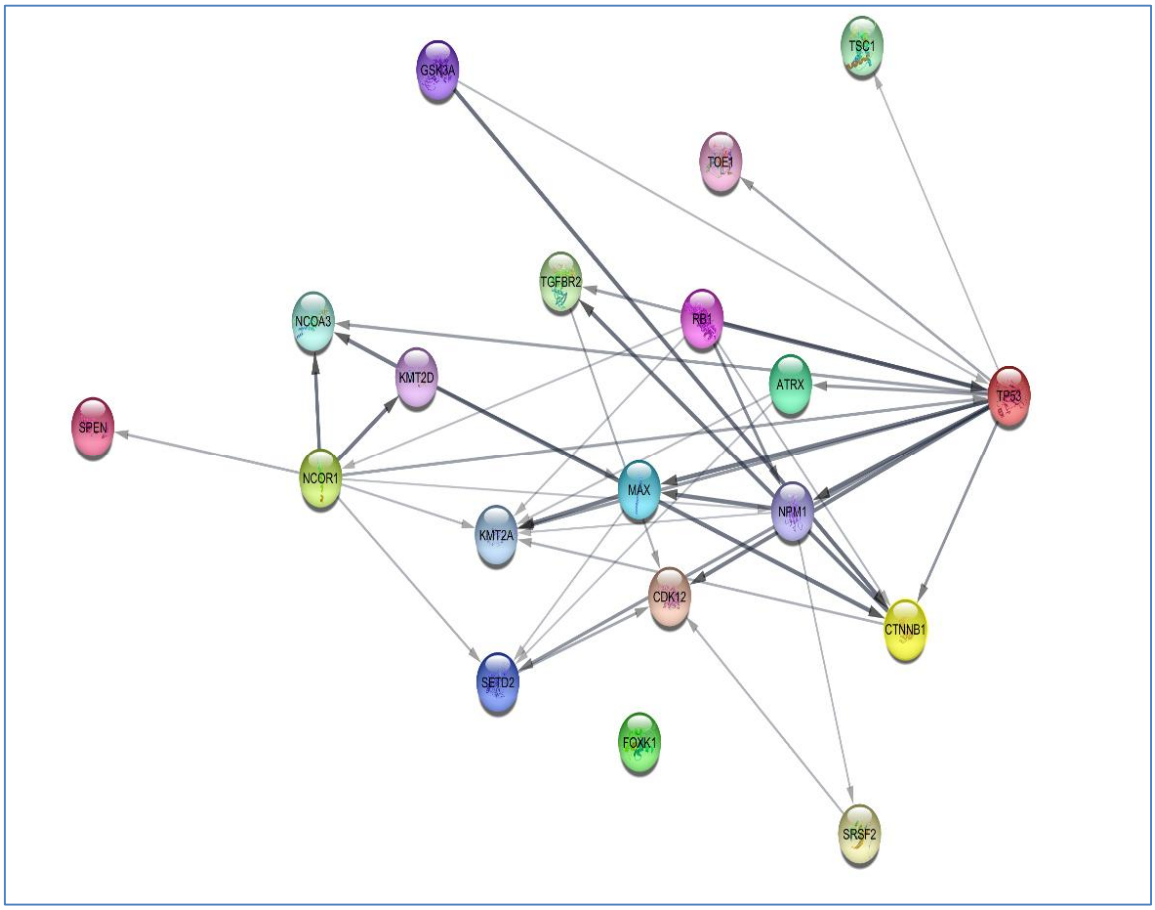


Figure 8: Dephosphylated regulatory network of genes nearby the significant mutations (MSK-IMPACT clinical sequencing cohort, Nat Med 2017). Confidence (score) cut off value was 0.4. Hierarchical layout of the string network is displayed here. The different colors indicates different genes. The arrow of the nodes indicates the degree of connectivity of the nodes. Among 19 genes TP53 has been shown highly influence directly with 14 other genes. On the other hand, FOXK1 has been shown least influence to others. The figure is generated from STRING database (version 10.5).

Functional enrichment analysis was performed based on neighborhood algorithm. Biological process and molecular function process also showed significant output. For example, gene expression counted 16 genes among 17 and it's false discovery rate was 5.81 e-07.

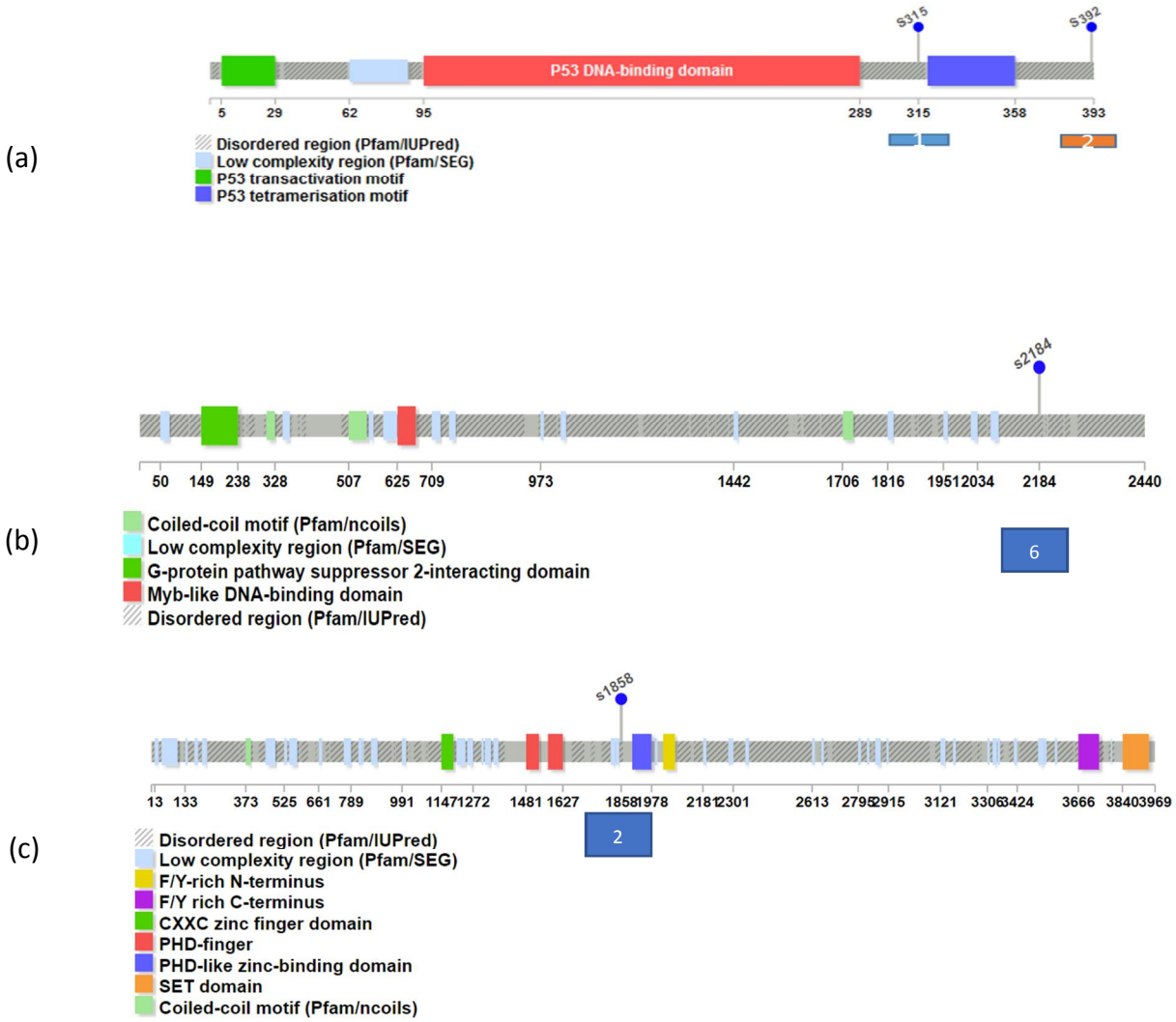


Figure 9: Lollipop plot of three significant genes. Outside boxes (blue color) indicate the active region position (a) PP2A dephosphorylation effect on TP53 shows it has S315 and S392 position with active region 1 and 2 and it has mutation position at 317 and 392 (b) NCOR1 gene has also PP2A dephosphorylation at S2184 position, which has highly significant and influence to bind with mutational position at 2183 (c) The p-value of KMT2A is near zero and it has also dephosphorylation effect of PP2A at S1858 position.

4.6 Dephosphorylation of PP2A output from COSMIC data

Similar analysis as performed in section 4.5 was repeated with mutational data from COSMIC database (https://cancer.sanger.ac.uk/cosmic/download). The mutational, phosphorylational and intrinsic disorder data were used as input datasets for ActiveDriver software. The summary of resulting analysis from is shown in Table 8. The merge report provides information such as gene name, active region, mutational position, post-translational modification, kinase and its active region *p*-value. Table is sorted based on active region p-value and the only significant values are shown here in table format. We found 57 genes with 2,723 active region which gave significant p value (< 0.05) with different position

Table 8: Merge report of top 50 significant genes with their dephosphorylation effect on PP2A target (COSMIC coding point mutation data) from COSMIC.

Gene	Mut position	Active region	PTM position	Residue	Dataset source	Active region *p* value
BRAF	449	1	447	S	A549_SMAP	4.56E-248
BRAF	454	1	447	S	A549_SMAP	4.56E-248
BRAF	451	1	447	S	A549_SMAP	4.56E-248
BRAF	442	1	447	S	A549_SMAP	4.56E-248
BRAF	441	1	447	S	A549_SMAP	4.56E-248
BRAF	453	1	447	S	A549_SMAP	4.56E-248
BRAF	444	1	447	S	A549_SMAP	4.56E-248
BRAF	443	1	447	S	A549_SMAP	4.56E-248
BRAF	446	1	447	S	A549_SMAP	4.56E-248
BRAF	447	1	447	S	A549_SMAP	4.56E-248
BRAF	440	1	447	S	A549_SMAP	4.56E-248
BRAF	450	1	447	S	A549_SMAP	4.56E-248
USP8	725	1	718	S	A549_SMAP	1.59E-185
USP8	721	1	718	S	A549_SMAP	1.59E-185
USP8	722	1	718	S	A549_SMAP	1.59E-185
USP8	713	1	718	S	A549_SMAP	1.59E-185
USP8	720	1	718	S	A549_SMAP	1.59E-185
USP8	715	1	718	S	A549_SMAP	1.59E-185
USP8	718	1	718	S	A549_SMAP	1.59E-185
USP8	716	1	718	S	A549_SMAP	1.59E-185
USP8	719	1	718	S	A549_SMAP	1.59E-185
USP8	717	1	718	S	A549_SMAP	1.59E-185
TP53	391	1	392	S	PIPs	2.44E-116
TP53	389	1	392	S	PIPs	2.44E-116

TP53	386	1	392	S	PIPs	2.44E-116
TP53	390	1	392	S	PIPs	2.44E-116
TP53	392	1	392	S	PIPs	2.44E-116
CTNNB1	546	2	551	T	A549_SMAP	1.02E-92
CTNNB1	550	2	551	T	A549_SMAP	1.02E-92
CTNNB1	547	2	551	T	A549_SMAP	1.02E-92
CTNNB1	549	2	551	T	A549_SMAP	1.02E-92
CTNNB1	555	2	551	T	A549_SMAP	1.02E-92
CTNNB1	553	2	551	T	A549_SMAP	1.02E-92
CTNNB1	545	2	551	T	A549_SMAP	1.02E-92
CTNNB1	56	1	60	S	H358_SMAP	3.26E-49
CTNNB1	58	1	60	S	H358_SMAP	3.26E-49
CTNNB1	55	1	60	S	H358_SMAP	3.26E-49
CTNNB1	54	1	60	S	H358_SMAP	3.26E-49
CTNNB1	53	1	60	S	H358_SMAP	3.26E-49
CTNNB1	65	1	60	S	H358_SMAP	3.26E-49
CTNNB1	67	1	60	S	H358_SMAP	3.26E-49
CTNNB1	61	1	60	S	H358_SMAP	3.26E-49
CTNNB1	60	1	60	S	H358_SMAP	3.26E-49
CTNNB1	59	1	60	S	H358_SMAP	3.26E-49
CTNNB1	57	1	60	S	H358_SMAP	3.26E-49
CTNNB1	66	1	60	S	H358_SMAP	3.26E-49
RB1	248	1	249	S	A549_SMAP	1.48E-08
RB1	251	1	249	S	A549_SMAP	1.48E-08
RB1	255	1	249	S	A549_SMAP	1.48E-08
RB1	249	1	249	S	A549_SMAP	1.48E-08

Functional and networks analysis on 57 significant genes from COSMIC data have been analysed using STRING database (https://string-db.org). Among 57 number of nodes, there are 166 number of edges are connected and average node degree is 5.82 (Figure: 10). An important results from the analysis is that PPI network is enriched significantly *p*-value < 5e-15) which means above 57 genes have more interactions among themselves than at random, indicating co-ordinated biological role of these genes. The false discovery rate also found in this analysis and it is less than 0.05, which means the influence level of mutation in binding of PP2A to these genes are high in COSMIC data.

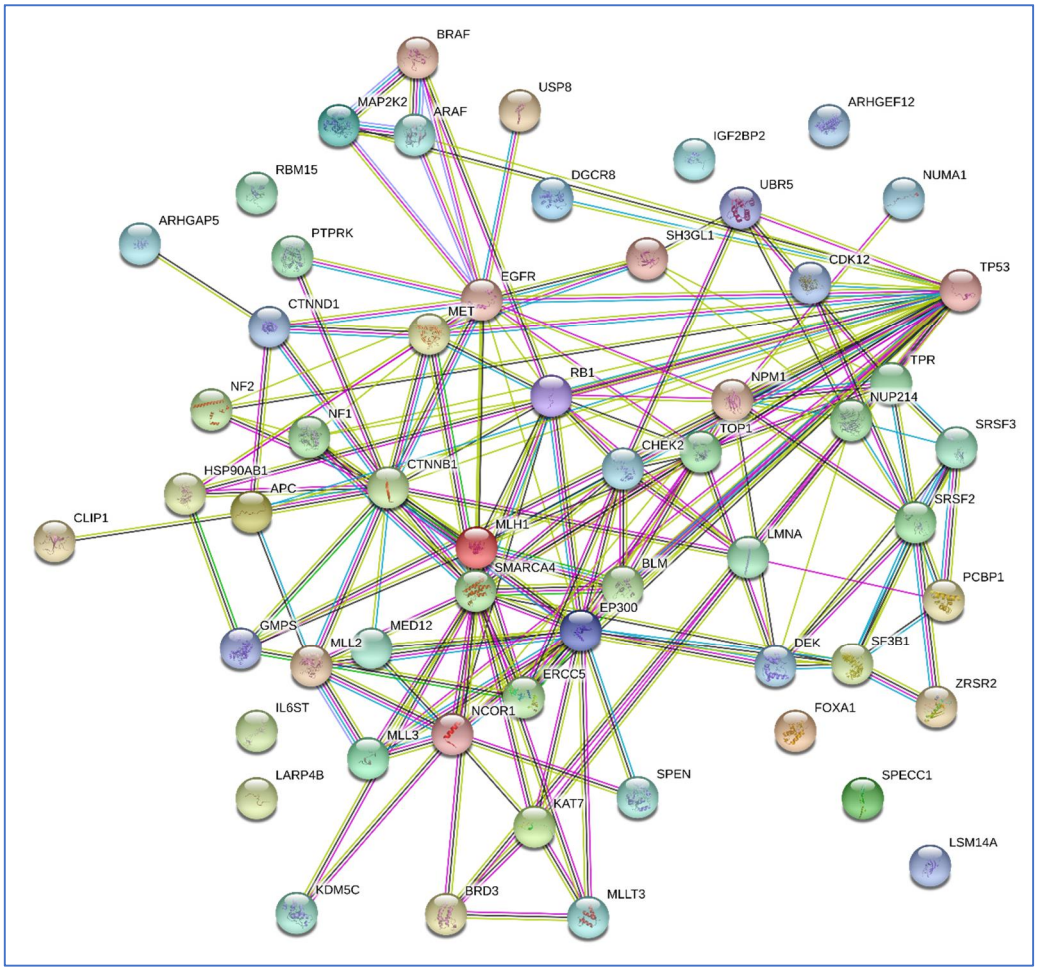


Figure 10: Dephosphorylated regulatory network of significant genes from COSMIC. Confidence (score) cut off value was 0.7. Hierarchical layout of the string network is displayed here. The arrow of the nodes indicates the degree of connectivity of the nodes. Among 57 genes TP53 has been connected densely with 17 other genes. On the other hand, some genes are loosely connected, indicating less interaction with others. This data have been analysed from STRING tool version 10.5.

5. Conclusions

In this thesis, systematic investigation of mutational landscape nearby PP2A-driven dephosphorylated sites in PP2A targets was performed. To achieve this goal, large scale mutational and phosphorylational datasets were integrated. Before performing the final analysis, the data was cleaned to obtain accurate results. In order to accomplish the proposed aim in my thesis, we utilised in-house as well as published phosphosproteomics datasets as starting point to identify the potential targets of PP2A. Student t-test and false discovery rate (FDR) analysis were performed all phosphosproteomics dataset to identify the significant peptides. The most significant genes also showed in the volcano plot and their p value and false discovery rate are near to 0. We collected all significant peptides to form comprehensive dataset for targeting PP2A dephosphorylation database.

Publicly available large-scale cancer genomics data resources such as cBioportal (http://www.cbioportal.org/index.do) and COSMIC (http://cancer.sanger.ac.uk/cosmic) were utilised in our analysis. From cBioportal, pancancer study data from MSK-IMPACT clinical sequencing cohort study and Breast cancer study (METABRIC, Nature 2012 and Nat commun 2016) for mutational analysis.

A study named "MSK-IMPACT clinical sequencing cohort, Nat Med 2017" dataset has been chosen from cBioportal. From 10,945 samples, found 19 genes 248 active regions gave significant p value (<0.05) with different position, which have dephosphorylation effect on PP2A target. There are 75 genes are non-significant. Another mutational data from COSMIC database with PP2A dephosphorylation dataset also performed in ActiveDriver. From 508,561 samples from COSMIC database, found statistically significant 57 genes with 2,723 active regions, which have dephosphorylation effect on PP2A target.

The network analysis was performed among 19 significant genes from cBioportal data. TP53 gene highly connected with 14 other genes among 19. However, there is no connection between FOXK1 and other genes, which means FOXK1 does not have any interaction with other genes. NCOR1, KMT2A, NPM1 have number of connections between other genes which have eight, seven and six respectively. Among these genes, there are 24 number of edges are connected and average node degree is 5.33 and PPI enrichment value also significant and it is 2.68e-12. The false discovery rate also found in this analysis and it is less than 0.05, which means the influence level of mutation

in binding of PP2A to these genes are high. The network analysis also performed among 57 significant genes from COSMIC database. Among 57 number of nodes, there are 166 number of edges are connected and average node degree is 5.82 (Figure: 10). PPI enrichment p- value has significant and it is 5e-15.

Last but not least, finally we can conclude that mutational study help us understand whether existing mutations correlate with dephosphorylation sites of PP2A targets and thereby likely provide potential clues for mechanisms of action for PP2A function.

6. References

Agresti, A. (2003). *Categorical data analysis* (Vol. 482). John Wiley & Sons.

Cundell, M. J., Hutter, L. H., Bastos, R. N., Poser, E., Holder, J., Mohammed, S., … Barr, F. A. (2016). A PP2A-B55 recognition signal controls substrate dephosphorylation kinetics during mitotic exit. *Journal of Cell Biology*, *214*(5), 539–554.

De Jong, P., & Heller, G. Z. (2008). *Generalized linear models for insurance data* (Vol. 10). Cambridge University Press Cambridge.

Dinkel, H., Chica, C., Via, A., Gould, C. M., Jensen, L. J., Gibson, T. J., & Diella, F. (2010). Phospho. ELM: a database of phosphorylation sites—update 2011. *Nucleic Acids Research*, *39*(suppl_1), D261–D267.

Dinkel, H., Chica, C., Via, A., Gould, C. M., Jensen, L. J., Gibson, T. J., & Diella, F. (2011). Phospho.ELM: a database of phosphorylation sites—update 2011. *Nucleic Acids Research*, *39*(suppl_1), D261–D267.

Duan, G., Li, X., & Köhn, M. (2014). The human DEPhOsphorylation database DEPOD: a 2015 update. *Nucleic Acids Research*, *43*(D1), D531–D535.

Forbes, S. A., Beare, D., Boutselakis, H., Bamford, S., Bindal, N., Tate, J., … Ponting, L. (2016). COSMIC: somatic cancer genetics at high-resolution. *Nucleic Acids Research*, *45*(D1), D777–D783.

Fox, J., & Monette, G. (2002). *An R and S-Plus companion to applied regression*. Sage.

Gao, J., Aksoy, B. A., Dogrusoz, U., Dresdner, G., Gross, B., Sumer, S. O., … Larsson, E. (2013). Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci. Signal.*, *6*(269), pl1-pl1.

Greenman, C., Wooster, R., Futreal, P. A., Stratton, M. R., & Easton, D. F. (2006). Statistical analysis of pathogenicity of somatic mutations in cancer. *Genetics*, *173*(4), 2187–2198.

Greenwood, E. J. D., Matheson, N. J., Wals, K., van den Boomen, D. J. H., Antrobus, R., Williamson, J. C., & Lehner, P. J. (2016). Temporal proteomic analysis of HIV infection reveals remodelling of the host phosphoproteome by lentiviral Vif variants. *Elife*, *5*, e18296.

Griffiths, A. J. F. (2005). *An introduction to genetic analysis*. Macmillan.

Hertz, E. P. T., Kruse, T., Davey, N. E., López-Méndez, B., Sigurðsson, J. O., Montoya, G., … Nilsson, J. (2016). A Conserved Motif Provides Binding Specificity to the PP2A-B56 Phosphatase. *Molecular Cell*, *63*(4), 686–695.

Hornbeck, P. V, Kornhauser, J. M., Tkachev, S., Zhang, B., Skrzypek, E., Murray, B., … Sullivan, M. (2011). PhosphoSitePlus: a comprehensive resource for investigating the structure and function of experimentally determined post-translational modifications in man and mouse. *Nucleic Acids Research*, *40*(D1), D261–D270.

INGEBRITSEN, T. S., & COHEN, P. (1983). The protein phosphatases involved in cellular regulation. *The FEBS Journal*, *132*(2), 255–261.

Janssens, V., & Goris, J. (2001). Protein phosphatase 2A: a highly regulated family of serine/threonine phosphatases implicated in cell growth and signalling. *Biochemical Journal*, *353*(Pt 3), 417.

Jay, J. J., & Brouwer, C. (2016). Lollipops in the clinic: information dense mutation plots for precision medicine. *PloS One*, *11*(8), e0160519.

Jin, J., & Pawson, T. (2012). Modular evolution of phosphorylation-based signalling systems. *Phil. Trans. R. Soc. B*, *367*(1602), 2540–2555.

Kauko, O., Imanishi, S. Y., Kulesskiy, E., Laajala, T. D., Yetukuri, L., Laine, A., … Yadav, B. (2018). Rules for PP2A-controlled phosphosignalling and drug responses. *BioRxiv*, 271841.

Kauko, O., O'Connor, C. M., Kulesskiy, E., Sangodkar, J., Aakula, A., Izadmehr, S., … Westermarck, J. (2018). PP2A inhibition is a druggable MEK inhibitor resistance mechanism in KRAS-mutant lung cancer cells. *Science Translational Medicine*, *10*(450).

Keshava Prasad, T. S., Goel, R., Kandasamy, K., Keerthikumar, S., Kumar, S., Mathivanan, S., … Venugopal, A. (2008). Human protein reference database—2009 update. *Nucleic Acids Research*, *37*(suppl_1), D767–D772.

Kremmer, E., Ohst, K. I. M., Kiefer, J., & Brewis, N. (1997). Separation of PP2A Core Enzyme and Holoenzyme with Monoclonal Antibodies against the Regulatory A Subunit : Abundant Expression of Both Forms in Cells. *Molecular and Cellular Biology*, *17*(3), 1692–1701.

Krijgsveld, J. (2012). Proteiomics of Biological Systems: Protein Phosphorylation Using Mass Spectrometry Techniques. By Bryan M. Ham. *ChemBioChem*, *13*(15), 2301–2302.

Lad, C., Williams, N. H., & Wolfenden, R. (2003). The rate of hydrolysis of phosphomonoester dianions and the exceptional catalytic proficiencies of protein and inositol phosphatases. *Proceedings of the National Academy of Sciences*, *100*(10), 5607–5610.

Mumby, M. C., & Walter, G. (1993). Protein serine/threonine phosphatases: structure, regulation, and functions in cell growth. *Physiological Reviews*, *73*(4), 673–699.

Pereira, B., Chin, S.-F., Rueda, O. M., Vollan, H.-K. M., Provenzano, E., Bardwell, H. A., … Sammut, S.-J. (2016). The somatic mutation profiles of 2,433 breast cancers refine their genomic and transcriptomic landscapes. *Nature Communications*, *7*, 11479.

Reimand, J., & Bader, G. D. (2013). Systematic analysis of somatic mutations in phosphorylation signaling predicts novel cancer drivers. *Molecular Systems Biology*, *9*(1), 637.

Reimand, J., Wagih, O., & Bader, G. D. (2013). The mutational landscape of phosphorylation signaling in cancer. *Scientific Reports*, *3*, 2651.

Sablina, A. A., & Hahn, W. C. (2007). The Role of PP2A A Subunits in Tumor Suppression. *Cell Adhesion & Migration*, *1*(3), 140–141. https://doi.org/10.4161/cam.1.3.4986

Sangodkar, J., Farrington, C. C., McClinch, K., Galsky, M. D., Kastrinsky, D. B., & Narla, G. (2016). All roads lead to PP2A: exploiting the therapeutic potential of this phosphatase. *The FEBS Journal*, *283*(6), 1004–1024.

Seshacharyulu, P., Pandey, P., Datta, K., & Batra, S. K. (2013). Phosphatase: PP2A structural importance, regulation and its aberrant expression in cancer. *Cancer Letters*, *335*(1), 9–18.

Thompson, J. J., & Williams, C. S. (2018). Protein Phosphatase 2A in the Regulation of Wnt Signaling, Stem Cells, and Cancer. *Genes*, *9*(3), 121.

Vogelstein, B., Papadopoulos, N., Velculescu, V. E., Zhou, S., Diaz, L. A., & Kinzler, K. W. (2013). Cancer genome landscapes. *Science*, *339*(6127), 1546–1558.

Vucetic, S., Obradovic, Z., Vacic, V., Radivojac, P., Peng, K., Iakoucheva, L. M., … Dunker, A. K. (2005). DisProt: a database of protein disorder. *Bioinformatics*, *21*(1), 137–140.

Westermarck, J., & Hahn, W. C. (2008). Multiple pathways regulated by the tumor suppressor PP2A in transformation. *Trends in Molecular Medicine*, *14*(4), 152–160.

Wiredja, D. D., Ayati, M., Mazhar, S., Sangodkar, J., Maxwell, S., Schlatzer, D., … Chance, M. R. (2017). Phosphoproteomics Profiling of Nonsmall Cell Lung Cancer Cells Treated with a Novel Phosphatase Activator. *Proteomics*, *17*(22).

Wood, L. D., Parsons, D. W., Jones, S., Lin, J., Sjöblom, T., Leary, R. J., … Ptak, J. (2007). The genomic landscapes of human breast and colorectal cancers. *Science*, *318*(5853), 1108–1113.

Xu, Y., Xing, Y., Chen, Y., Chao, Y., Lin, Z., Fan, E., … Shi, Y. (2006). Structure of the protein phosphatase 2A holoenzyme. *Cell*, *127*(6), 1239–1251.

Yachida, S., Jones, S., Bozic, I., Antal, T., Leary, R., Fu, B., … Nowak, M. A. (2010). Distant metastasis occurs late during the genetic evolution of pancreatic cancer. *Nature*, *467*(7319), 1114.

Yusuff, H., Mohamad, N., Ngah, U., & Yahaya, A. (2012). Breast cancer analysis using logistic regression. *International Journal of Research And Applied Studies*, *11*.

Zehir, A., Benayed, R., Shah, R. H., Syed, A., Middha, S., Kim, H. R., … Devlin, S. M. (2017). Mutational landscape of metastatic cancer revealed from prospective clinical sequencing of 10,000 patients. *Nature Medicine*, *23*(6), 703.

Acknowledgements

This thesis was written as a part of my master's degree in Bioinformatics program at the Faculty of Future Technologies under University of Turku, in the period between October 2017 and December 2018.

Frist and foremost, I would like to express my gratitude to my supervisor Professor Jukka Westermarck for giving me opportunity to work under his supervision. Without his direct and clear guidelines, it would not be possible to complete this thesis.

Beside my supervisor, I want to thanks to Laxman Yetukuri for his continuous support, patience, motivation, and enthusiasm. His guidance helped me in all the time for writing of this thesis. A special thanks to Laxman for always keeping your door open and tolerating my "Silly" questions.

I want to give special thanks to my departmental supervisor Martti Tolvanen for his technical supporting for this thesis. He is such a nice person to be very supportive from the very beginning.

I really relished working on this thesis and have learned lots of new things which I did not imagine when I started.

Most importantly, I would like to convey my gratitude to my parents, younger sister and lovely wife. They always keep inspiring me from the long way home so that I can complete master's degree from University of Turku.

At last but not at least, I am very grateful to the all members of Cancer Cell Signaling group and other stuffs at Turku centre for Biotechnology. Their support greatly helped me to carry on this thesis in a calm manner.

Finally, I would like to thank all my classmates and teachers who provided good studying environment and motivation.

MD Fakhrul Islam Faruque

December, 2018.

Appendix

1. The Onco Query Language (OQL) data types and functionality (Gao et al., 2013).

Date type	Keyword	Code	Description	Example
Mutation	MUT	MUT= x	Show case with specific mutation or mutation types.	TP53: MUT=MISSENSE;TP53: MUT=NONSENSE; TP53: MUT=NON-START; TP53: MUT=NONSTOP; TP53: MUT=FRAMESHIFT; TP53: MUT=INFRAME; TP53: MUT=INFRAME; TP53: MUT=SPLICE; TP53: MUT=TRUNCATED

2. Merge report of "Breast cancer (METABRIC, Nature 2012 and common 2016)" study from cBioportal and genes with their dephosphorylation effect on PP2A target.

Gene	Mut_position	Active_region	PTM_position	Residue	Dataset source	Active_region_p
SMARCC2	303	2	304	S	A549_MEK	0.002179
SMARCC2	303	2	302	S	A549_MEK	0.002179
SMARCC2	303	2	304	S	A549_MEK	0.002179
SMARCC2	303	2	302	S	A549_MEK	0.002179
SMARCC2	303	2	304	S	A549_MEK	0.002179
SMARCC2	303	2	302	S	A549_MEK	0.002179
SMARCC2	303	2	304	S	A549_MEK	0.002179
SMARCC2	303	2	302	S	A549_MEK	0.002179
AHNAK2	1260	8	1253	S	A549_MEK	0.003844
AHNAK2	1260	8	1253	S	A549_MEK	0.003844
AHNAK2	1251	8	1253	S	A549_MEK	0.003844
AHNAK2	1251	8	1253	S	A549_MEK	0.003844
AHNAK2	1248	8	1253	S	A549_MEK	0.003844
AHNAK2	1248	8	1253	S	A549_MEK	0.003844
AHNAK2	1252	8	1253	S	A549_MEK	0.003844
AHNAK2	1252	8	1253	S	A549_MEK	0.003844
AHNAK2	1254	8	1253	S	A549_MEK	0.003844
AHNAK2	1254	8	1253	S	A549_MEK	0.003844
AHNAK2	1252	8	1253	S	A549_MEK	0.003844
AHNAK2	1252	8	1253	S	A549_MEK	0.003844

AHNAK	210	1	210	S	A549_MEK	0.005277
AHNAK	210	1	212	S	A549_MEK	0.005277
AHNAK	210	1	216	S	A549_SMAP	0.005277
AHNAK	210	1	210	S	A549_MEK	0.005277
AHNAK	210	1	212	S	A549_MEK	0.005277
AHNAK	210	1	216	S	A549_SMAP	0.005277
AHNAK	218	1	212	S	A549_MEK	0.005277
AHNAK	218	1	220	S	A549_SMAP	0.005277
AHNAK	218	1	216	S	A549_SMAP	0.005277
AHNAK	218	1	212	S	A549_MEK	0.005277
AHNAK	218	1	220	S	A549_SMAP	0.005277
AHNAK	218	1	216	S	A549_SMAP	0.005277
AHNAK	210	1	210	S	A549_MEK	0.005277
AHNAK	210	1	212	S	A549_MEK	0.005277
AHNAK	210	1	216	S	A549_SMAP	0.005277
AHNAK	210	1	210	S	A549_MEK	0.005277
AHNAK	210	1	212	S	A549_MEK	0.005277
AHNAK	210	1	216	S	A549_SMAP	0.005277
AHNAK	5795	26	5794	T	A549_MEK	0.008118
AHNAK	5795	26	5790	S	A549_MEK	0.008118
AHNAK	5795	26	5794	T	A549_MEK	0.008118
AHNAK	5795	26	5790	S	A549_MEK	0.008118
AHNAK	5799	26	5794	T	A549_MEK	0.008118
AHNAK	5799	26	5794	T	A549_MEK	0.008118
AHNAK	5779	26	5780	S	A549_MEK	0.008118
AHNAK	5779	26	5784	S	PIPs	0.008118
AHNAK	5779	26	5780	S	A549_MEK	0.008118
AHNAK	5779	26	5784	S	PIPs	0.008118
AHNAK	5779	26	5780	S	A549_MEK	0.008118
AHNAK	5779	26	5784	S	PIPs	0.008118
AHNAK	5779	26	5780	S	A549_MEK	0.008118
AHNAK	5779	26	5784	S	PIPs	0.008118
AHNAK	5779	26	5780	S	A549_MEK	0.008118
AHNAK	5779	26	5784	S	PIPs	0.008118
AHNAK	5779	26	5780	S	A549_MEK	0.008118
AHNAK	5779	26	5784	S	PIPs	0.008118
AHNAK	5779	26	5780	S	A549_MEK	0.008118
AHNAK	5779	26	5784	S	PIPs	0.008118
AHNAK	5779	26	5780	S	A549_MEK	0.008118
AHNAK	5779	26	5784	S	PIPs	0.008118
ATR	433	1	435	S	A549_MEK	0.008968
ATR	433	1	435	S	A549_MEK	0.008968
ATR	436	1	435	S	A549_MEK	0.008968
ATR	436	1	435	S	A549_MEK	0.008968
CHD1	96	1	90	S	Francisbar	0.011364

CHD1	96	1	90	S	Francisbar	0.011364
CHD1	96	1	90	S	Francisbar	0.011364
CHD1	96	1	90	S	Francisbar	0.011364
CHD1	96	1	90	S	Francisbar	0.011364
CHD1	96	1	90	S	Francisbar	0.011364
AHNAK	5578	22	5582	S	H358_SMAP	0.017355
AHNAK	5578	22	5582	S	H358_SMAP	0.017355
AHNAK	5586	22	5589	S	H358_SMAP	0.017355
AHNAK	5586	22	5582	S	H358_SMAP	0.017355
AHNAK	5586	22	5589	S	H358_SMAP	0.017355
AHNAK	5586	22	5582	S	H358_SMAP	0.017355
AHNAK	5592	22	5589	S	H358_SMAP	0.017355
AHNAK	5592	22	5589	S	H358_SMAP	0.017355
AHNAK	5592	22	5589	S	H358_SMAP	0.017355
AHNAK	5592	22	5589	S	H358_SMAP	0.017355
SMARCC1	322	1	328	S	H358_SMAP	0.033298
SMARCC1	322	1	328	S	H358_SMAP	0.033298
SMARCC1	335	1	328	S	H358_SMAP	0.033298
SMARCC1	335	1	330	S	H358_SMAP	0.033298
SMARCC1	335	1	328	S	H358_SMAP	0.033298
SMARCC1	335	1	330	S	H358_SMAP	0.033298
SMARCC1	337	1	330	S	H358_SMAP	0.033298
SMARCC1	337	1	330	S	H358_SMAP	0.033298
AHNAK	4908	16	4908	S	H358_SMAP	0.048773
AHNAK	4908	16	4908	S	H358_SMAP	0.048773
AHNAK	5326	19	5332	S	H358_SMAP	0.048773
AHNAK	5326	19	5332	S	H358_SMAP	0.048773
AHNAK	5332	19	5332	S	H358_SMAP	0.048773
AHNAK	5332	19	5332	S	H358_SMAP	0.048773
AHNAK	516	5	511	S	PIPs	0.048773
AHNAK	516	5	511	S	PIPs	0.048773
AHNAK	510	5	511	S	PIPs	0.048773
AHNAK	510	5	511	S	PIPs	0.048773
AHNAK	511	5	511	S	PIPs	0.048773
AHNAK	511	5	511	S	PIPs	0.048773
ARID1A	1184	1	1184	S	A549_MEK	0.057132
ARID1A	1184	1	1184	S	A549_MEK	0.057132
CHD1	1097	2	1100	S	PIPs	0.065391
CHD1	1097	2	1098	S	PIPs	0.065391
CHD1	1097	2	1096	S	PIPs	0.065391
CHD1	1097	2	1100	S	PIPs	0.065391
CHD1	1097	2	1098	S	PIPs	0.065391
CHD1	1097	2	1096	S	PIPs	0.065391
CHD1	1094	2	1100	S	PIPs	0.065391
CHD1	1094	2	1098	S	PIPs	0.065391

CHD1	1094	2	1096	S	PIPs	0.065391
CHD1	1094	2	1100	S	PIPs	0.065391
CHD1	1094	2	1098	S	PIPs	0.065391
CHD1	1094	2	1096	S	PIPs	0.065391
SETD2	2080	1	2082	S	PIPs	0.066734
SETD2	2080	1	2080	S	PIPs	0.066734
SETD2	2080	1	2082	S	PIPs	0.066734
SETD2	2080	1	2080	S	PIPs	0.066734
SETD2	2079	1	2082	S	PIPs	0.066734
SETD2	2079	1	2080	S	PIPs	0.066734
SETD2	2079	1	2082	S	PIPs	0.066734
SETD2	2079	1	2080	S	PIPs	0.066734
AHNAK2	305	1	298	T	A549_MEK	0.12026
AHNAK2	305	1	298	T	A549_MEK	0.12026
AHNAK2	593	4	598	T	A549_MEK	0.12026
AHNAK2	593	4	593	S	H358_SMAP	0.12026
AHNAK2	593	4	598	T	A549_MEK	0.12026
AHNAK2	593	4	593	S	H358_SMAP	0.12026
AHNAK2	5720	15	5715	T	A549_SMAP	0.13349
AHNAK2	5720	15	5715	T	A549_SMAP	0.13349
AHNAK	5854	27	5857	S	H358_SMAP	0.134511
AHNAK	5854	27	5851	S	H358_SMAP	0.134511
AHNAK	5854	27	5857	S	H358_SMAP	0.134511
AHNAK	5854	27	5851	S	H358_SMAP	0.134511
AHNAK	4711	15	4715	S	A549_MEK	0.135898
AHNAK	4711	15	4715	S	A549_MEK	0.135898
AHNAK	4710	15	4715	S	A549_MEK	0.135898
AHNAK	4710	15	4715	S	A549_MEK	0.135898
AHNAK2	920	7	923	S	A549_MEK	0.148609
AHNAK2	920	7	923	S	A549_MEK	0.148609
AHNAK2	928	7	923	S	A549_MEK	0.148609
AHNAK2	928	7	923	S	A549_MEK	0.148609
AHNAK2	927	7	923	S	A549_MEK	0.148609
AHNAK2	927	7	923	S	A549_MEK	0.148609
AHNAK2	923	7	923	S	A549_MEK	0.148609
AHNAK2	923	7	923	S	A549_MEK	0.148609

3. Programming coding on R platform (RStudio version 1.0.1336 with R v 3.0.1).

```
# install pakage
install.packages("ActiveDriver")
library(ActiveDriver)

data(ActiveDriver_data)

phos_results = ActiveDriver(sequences, sequence_disorder, mutations, phosphosites)
phos_results
#Overian cancer mutation
ovarian_mutations = mutations[grep("ovarian", mutations$sample_id),]
ovarian_mutations
gene_name_ovarian_mutations= ovarian_mutations[,1]
#Breast cancer
breast_cancer_mutations = mutations[grep("breast_cancer", mutations$sample_id),]
breast_cancer_mutations
gene_name_breast_cancer= breast_cancer_mutations[,1]
gene_name_breast_cancer
length(gene_name_breast_cancer)
#Pancancer_mutation
pancancer_mutations = mutations[grep("pancreatic_cance", mutations$sample_id),]
pancancer_mutations
gene_name_pancancer_mutations= pancancer_mutations[,1]
gene_name_pancancer_mutations
#GBM_muts
GBM_muts = mutations[grep("glioblastoma", mutations$sample_id),]
GBM_muts
gene_name_GBM_muts= GBM_muts[,1]
gene_name_GBM_muts

kin_rslt_GBM = ActiveDriver(sequences, sequence_disorder, GBM_muts, kinase_domains,
simplified=TRUE)

kin_results = ActiveDriver(sequences, sequence_disorder, mutations, kinase_domains,
simplified=TRUE)
ls()

data(ActiveDriver_data)

phos_results = ActiveDriver(sequences, sequence_disorder, mutations, phosphosites)
phos_results
```

```
############

library(seqinr)
all_seq_dis<-
read.table("D:/fif_data/Desktop/prdos_scores_human_final_seq_binary.txt",header=TRUE)

unique(all_seq_dis[,3])
x=as.vector(all_seq_dis[,3])
x[1:3]
y=as.vector(all_seq_dis[,4])
aafile<- read.fasta("D:/fif_data/Desktop/sequences_for_TCGA_pancancer.fa", seqtype =
"AA")
Seqs=unlist(getSequence(aafile, as.string=T))
names(Seqs)=names(aafile)
names(Seqs[1:3])


disfile<- read.fasta("D:/fif_data/Desktop/sequence_disorder_for_TCGA_pancancer.fa",
seqtype = "AA")
dis=unlist(getSequence(disfile, as.string=T))
names(dis)=names(disfile)
names(dis[1:3])

m=read.table("D:/fif_data/Desktop/all_mutations_for_TCGA_pancancer.tab",
header=TRUE)



p=read.table("D:/fif_data/Desktop/all_phosphosites_for_TCGA_pancancer.tab",
header=TRUE)


phosresults.pan = ActiveDriver(Seqs,dis,m,p)

capture.output(phosresults.pan, file = "phosresults.pan.txt")
edisummary(phosresults.pan.txt)
phosresults
head(m)
head(p)
ActiveDriver




###################################
msk_2017=read.csv("D:/fif_data/Desktop/data_save_msk_2017_26.10.2017.csv",header=
TRUE,sep=",")
dim(msk_2017)
#Finding common genes###
msk_2017_gene_table=table(msk_2017$gene_symbol);
r=row.names(msk_2017_gene_table);r
length(r)
gene_sym=read.table("D:/fif_data/Desktop/gene_symbol_to_refseq.tab",header=T)
head(gene_sym)
```

```
length(gene_sym$gene)
gene_sym_table=table(gene_sym$gene)
r2=rownames(gene_sym_table)
length(r2)
common_genes=intersect(r,r2)
common_genes=as.vector(common_genes)
length(common_genes)
#####1200+#####
data_common_genes=msk_2017[is.element(msk_2017$gene_symbol,common_genes),]
head(data_common_genes)
write.csv2(data_common_genes, "D:/fif_data/Desktop/data_common_genes.csv")
####################

msk_2017=read.csv("D:/fif_data/Desktop/Thesis/data_save_msk_2017_22.11.2017.csv",h
eader=TRUE,sep=",")
head(msk_2017)
colnames(m)

colnames(msk_2017)


head(msk_2017)


m0=subset(msk_2017,
select=c("gene_symbol","mutation_status","case_id","position","wt_residue","mt_residue"))
head(m0)
colnames(m0)<-c("gene","cancer_type","sample_id","position","wt_residue", "mut_residue")
head(m0)



phosresults_0 = ActiveDriver(x,y,m0,p)

names(p)
m1=m0[2:9,]
m1



####  GNAS data #############
seq_temp<-sequences[which(names(sequences) == "GNAS")]
dis_temp<-disorder_values[which(names(disorder_values) == "GNAS")]
names(dis_temp) <- c("GNAS_O95467","GNAS_P63092","GNAS_Q5JWF2")
names(seq_temp) <- c("GNAS_O95467","GNAS_P63092","GNAS_Q5JWF2")
m1=m0[grep("GNAS_",as.vector(m0$gene)),]

grep("GNAS_",as.vector(m0$gene),"GNAS_O95467")
m1=m0[grep("GNAS_",as.vector(m0$gene)),]
which(as.vector(p1$gene)=="GNAS")
p1[which(as.vector(p1$gene)=="GNAS"),]
p2=p1[which(as.vector(p1$gene)=="GNAS"),]
p2$gene=as.vector(p2$gene)
p2$gene
p2$gene[1]="GNAS_P63092"
```

```
ActiveDriver(seq_temp,dis_temp,m1,p1)

######Breast_cancer_2016 study####

breast_cancer_01.03.2017 = ActiveDriver(sequences,disorder_values,m2,p1)

write.csv(breast_cancer_01.03.2017$merged_report, file =
"breast_cancer_01.03.2017_merged_report.csv")
################## Find commpon gene#########
library(ggplot2)
library(magrittr)
library(ggpubr)
library(ggrepel)
library(tidyverse)

trp1 <- read_csv("phosresults_20.02.2017_merged_report.csv",col_names= TRUE)
tr1 <- read_csv("phosresults_20.02.2017_merged_report.csv",col_names= TRUE) %>%
select(gene)
tr1
tn1 <- read_csv("Dephsopsho_db_signi - Copy.csv",col_names= TRUE) %>% select(gene)

tn1

trn1 <- intersect(tr1,tn1) %>% as.data.frame
dim(trn1)
dim(trp1)

write.csv(trn1, file = "int_th17vstreg.txt",row.names=TRUE)

trn2 <- left_join(trn1,trp1,by="gene") %>% mutate(gene_comm=gene) %>% as.data.frame

trn3 <- inner_join(trn1,trp1,by="gene") %>% as.data.frame
write.csv(trn3, file = "common_gene_pancancer_2.csv",row.names=TRUE)
trn2
dim(trn2)
trn4 <-na.omit(trn3$active_region)

############ active driver on PP2A dephosphorylome  (P1)  with MSK##### ####

setwd("D:/fif_data/Desktop/Thesis/")
all_seq_dis<-read.table("prdos_scores_human_final_seq_binary.txt",header=TRUE)

cc<-load(file="prdos_scores_human_final_seq_binary.rav") ##dis_data_final
gene_sym=read.table("D:/fif_data/Desktop/Thesis/hugoformat_uniprot_16.02.17.csv",head
er=TRUE,",")
hug_set<-gene_sym[,c(1,4)]

hug_set
colnames(all_seq_dis)


colnames(hug_set)<-c("Uniprot","HuGO_gene")


all_seq_dis_names<-merge(dis_data_final,hug_set, by="Uniprot", x.all=TRUE)
```

```
head(all_seq_dis_names)
sequences<-as.character(all_seq_dis_names[,3])


names(sequences)<-as.vector(all_seq_dis_names$HuGO_gene)

sequences[1:3]

disorder_values<-as.character(all_seq_dis_names[,5])
names(disorder_values)<-as.vector(all_seq_dis_names$HuGO_gene)
disorder_values[1:3]

msk_2017_mut=read.csv("D:/fif_data/Desktop/Thesis/mutation_new_msk_2017_22.11.201
7_COPY.csv",header=TRUE,sep = ",")
colnames(msk_2017_mut)
m0=subset(msk_2017_mut,
select=c("gene_symbol","case_id","position","wt_residue","mt_residue"))
colnames(m0)<-c("gene","sample_id","position","wt_residue", "mut_residue")
head(m0)

p_PP2A=
read.csv("D:/fif_data/Desktop/Thesis/Dephsopsho_db_signi_24.09.18.csv",header=TRUE,s
ep = ",")
colnames(p_PP2A)
p0=subset(p_PP2A, select=c("gene_name","position","residue","kinase"))
colnames(p0)<-c("gene","position","residue","kinase")
dephospho_MSK2017_24.09.18= ActiveDriver(sequences,disorder_values,m0,p0)
write.csv(dephospho_MSK2017_15.03$merged_report, file =
"dephospho_MSK2017_15.03_merged_report.csv")




###############cosmic data#######
cosmic_mut <-read.csv("D:/fif_data/Desktop/Thesis/cosmic_mutant.csv",header=TRUE,sep
= ",")

colnames(cosmic_mut)
m2=subset(cosmic_mut,
select=c("Gene.name","Sample.name","position","wt_residue","mt_residue"))
colnames(m2)<-c("gene","sample_id","position","wt_residue", "mut_residue")
head(m2)

p_PP2A=
read.csv("D:/fif_data/Desktop/Thesis/Dephsopsho_db_signi.csv",header=TRUE,sep = ",")
colnames(p_PP2A)
p0=subset(p_PP2A, select=c("gene_name","position","residue","kinase"))
colnames(p0)<-c("gene","position","residue","kinase")
dephospho_cosmic= ActiveDriver(sequences,disorder_values,m2,p0)

####### t test and fdr test####
require(graphics)

a= read.csv("C:/Users/mfifar.UTU/Desktop/c.csv",header=TRUE,sep=",")
colnames(a)
```

```
pv<-matrix(NA, nrow = dim(a)[1], ncol = 1)
for (i in 1:dim(a)[1])
{
  xx<-t.test(a[i, 2:4], a[i, 5:7], var.equal = FALSE, paired = FALSE, alternative = "two.sided")
  pv[i]<-xx$p.value
}

p=round(pv,3)
p
write.csv(p, file = "pv_B56_dataset_3.csv")



padj<-round(p.adjust(p, method = "BH"),3)
padj

write.csv(padj, file = "fdr_B56.csv")

a$pvalue<-pv
a$padj<-padj

###olcano plot##########
##Identify the genes that have a p-value < 0.05
a$threshold = as.factor(a$p.adjusted < 0.05)

##Construct the plot object
g <- ggplot(data=a,
            aes(x=log2(Fold.Change), y =-log10(P.value) ,
                scale_color_manual(values=c("green", "red"))   )) +
  geom_point(alpha=0.4, size=1.75) +
  xlim(c(-1.5, 1.5)) +

  xlab("log2 fold change") + ylab("-log10 q.value") +
  theme_bw() +
  theme(legend.position="none")
scale_color_manual(values=c("green", "red"))
g+geom_text_repel(data=head(a, 10), aes(label=Gene))
g

###### heatmap#### ###
a1= as.vector(a)
heatmap3(a1,

         RowSideLabs = FALSE,
         showRowDendro=FALSE,
         showColDendro=FALSE,
         main = "Heatmap of Significant genes")
dev.off()
```

```
###### histogram of MSK_2017#####

z= read.csv("C:/Users/mfifar.UTU/Desktop/z.csv",header=TRUE,sep=",")
z1=z$active_region_p
as.vector(z)
hist(z1)
hist(z1,

    main = paste("Histogram of Active region" ),

    xlab = "Active region p value", ylab= "frequency",
    axes = TRUE, plot = TRUE, labels = FALSE
)

dev.off()
hist(z1,

    +   main="Old Faithful Eruptions", # the main title
    +   xlab="Duration minutes")
############ heatmap##########
library(RColorBrewer)
library(gplots)




j1=read.table("C:/Users/mfifar.UTU/Downloads/c.csv",header=T,sep=",")
Genes <- as.vector(j1[,1])
data<-as.matrix(j1[,-1])
colnames(data)<-gsub("SCR.*","SCR",colnames(data))
colnames(data)<-gsub("A.*","A.sub",colnames(data))

heatmap.2(data,
        dendrogram="none",
        Rowv=T,
        Colv=T,
        scale="row",
        key=T,
        trace="none",
        col = colorRampPalette(c("red","white","blue"))(512),
        #breaks=col_breaks,
        na.rm=F,
        labRow = Genes,
        keysize=1.5,
        density.info="none")
```