PLOS ONE

RESEARCH ARTICLE

Reproducible Analysis of Post-Translational Modifications in Proteomes—Application to Human Mutations

Alex S. Holehouse[1,2], **Kristen M. Naegle**[2,3]*

1 Division of Biology and Biomedical Sciences, Washington University, St. Louis, MO, United States of America, **2** The Center for Biological Systems Engineering, Washington University, St. Louis, MO, United States of America, **3** Biomedical Engineering, Washington University, St. Louis, MO, United States of America

* knaegle@wustl.edu

OPEN ACCESS

Citation: Holehouse AS, Naegle KM (2015) Reproducible Analysis of Post-Translational Modifications in Proteomes—Application to Human Mutations. PLoS ONE 10(12): e0144692. doi:10.1371/journal.pone.0144692

Editor: Frederique Lisacek, Swiss Institute of Bioinformatics, SWITZERLAND

Received: August 7, 2015

Accepted: November 23, 2015

Published: December 14, 2015

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: The authors have no support or funding to report.

Competing Interests: The authors have declared that no competing interests exist.

Abstract

Background

Protein post-translational modifications (PTMs) are an important aspect of protein regulation. The number of PTMs discovered within the human proteome, and other proteomes, has been rapidly expanding in recent years. As a consequence of the rate in which new PTMs are identified, analysis done in one year may result in different conclusions when repeated in subsequent years. Among the various functional questions pertaining to PTMs, one important relationship to address is the interplay between modifications and mutations. Specifically, because the linear sequence surrounding a modification site often determines molecular recognition, it is hypothesized that mutations near sites of PTMs may be more likely to result in a detrimental effect on protein function, resulting in the development of disease.

Methods and Results

We wrote an application programming interface (API) to make analysis of ProteomeScout, a comprehensive database of PTMs and protein information, easy and reproducible. We used this API to analyze the relationship between PTMs and human mutations associated with disease (based on the 'Clinical Significance' annotation from dbSNP). Proteins containing pathogenic mutations demonstrated a significant study bias which was controlled for by analyzing only well-studied proteins, based on their having at least one pathogenic mutation. We found that pathogenic mutations are significantly more likely to lie within eight amino acids of a phosphoserine, phosphotyrosine or ubiquitination site when compared to mutations in general, based on a Fisher's Exact test. Despite the skew of pathogenic mutations occurring on positively charged arginines, we could not account for this relationship based only on residue type. Finally, we hypothesize a potential mechanism for a pathogenic mutation on RAF1, based on its proximity to a phosphorylation site, which represents a subtle regulation difference that may explain why its biochemical effect has failed to be uncovered previously. The combination of the API and a dynamically expanding PTM database will make the reanalysis of this question and other systems-level questions easier in the future.

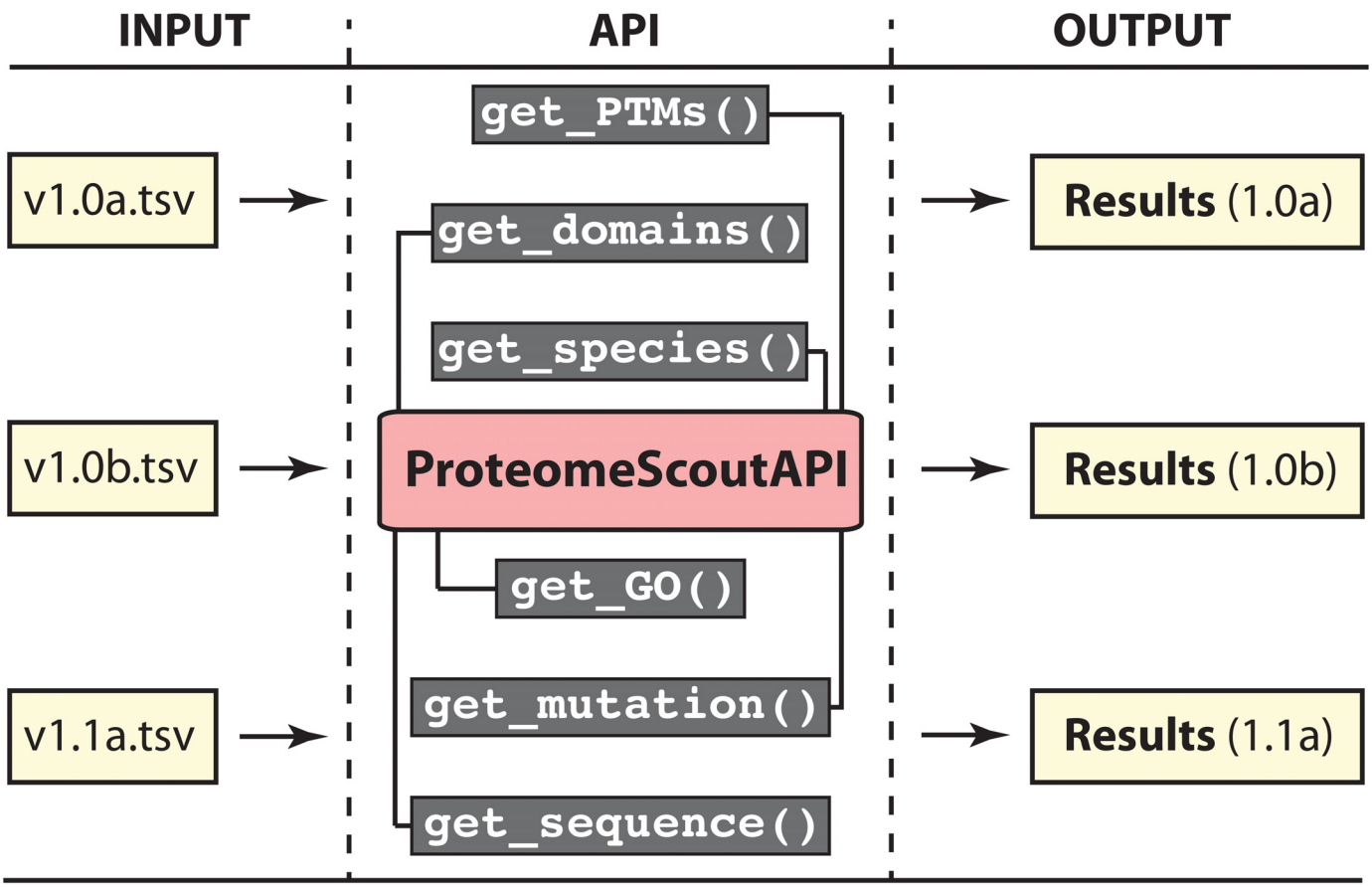
Introduction

The development of high-throughput measurement technologies, such as mass spectrometry based proteomics, has led to a rapid expansion in the discovery of post-translational modifications (PTMs) [1]. In order to understand how these PTMs regulate protein turnover, activity, interactions, localization, and other aspects of protein function, a large body of research has developed based on the analysis of PTMs relative to other protein features [2–6] and the relationship between PTMs and protein evolution [7–12]. Several similar systems-level studies have also been focused on how the gain, loss, or dysregulation of PTMs may be involved in human disease [13–15]. To perform these studies, researchers have manually combined a variety of PTM resources—a process which poses a significant challenge from a data acquisition, scrubbing, and warehousing perspective. Beyond this, such an analysis is based on a single snapshot of a collection of resources, a perspective which does not easily allow a longitudinal analysis to explore how a relationships of interest may change as new information becomes available and the PTM datasets evolve. In this study we examine how the number of identified PTMs has changed over time, a result which suggests there could be serious limitations in the conclusions drawn from systems-level analysis of PTMs in previous years, when compared to analysis done using data available today.

To overcome the challenge of data curation and enable easy updates and future re-analysis, we wrote an application program interface (API) for the expansive and dynamically growing database of PTMs, ProteomeScout [16] (Fig 1). The Python-based API makes it easy to interact with a ProteomeScout database download. The ProteomeScout database has a stable release every six months, as well as weekly updates to reflect dynamics snapshots of the current state of knowledge of PTMs across organisms. We have used this API, in conjunction with the current stable release of the ProteomeScout database, to analyze the relationship between PTMs and mutations. Specifically, we, and others [14], hypothesize that mutations within the recognition sequence of PTMs used by enzymes and binding partners would be likely to cause protein dysregulation and manifest as human-disease related mutations.

To test the hypothesis of whether disease-causing mutations, based on the 'Clinical Significance' category within dbSNP [17], are more likely to be near sites of PTMs, we found that we first had to correct for study bias. Proteins with at least one pathogenic mutation were consistently more likely to have other annotations, including PTMs and Gene Ontology terms [18]. Therefore, we first controlled for study bias by only analyzing the set of proteins with at least one pathogenic mutation. We found that pathogenic mutations were more likely to be within eight amino acids of a PTM of several types, including ubiquitination and phosphorylation of serine and tyrosine residues. The limited amount of data currently available on other modifications limits the analysis of these PTMs, although this may change in the future as the associated datasets grow as a result of community-based deposition of PTMs in ProteomeScout or the major compendia that ProteomeScout incorporates. Since pathogenic mutations were much more likely to occur on arginine residues, we tested for the possibility that PTMs and pathogenic mutations were coincidentally together based on surface accessibility. However, in control tests we could not explain the enrichment of PTMs near pathogenic mutations based on charge alone. Importantly, the API, the specific database snapshot that was used in this analysis, and the open-source analysis scripts, are available as supplementary information on our website, making the reproducibility of this analysis easy and certain. Additionally, this means that we and others can continuously update these proteome-wide statistical relationships between mutations and PTMs as the database expands and a new database file is used in the analysis pipeline.

After discovering that globally, the nearness of mutations to a PTM may be indicative of a likelihood of being related to disease, we asked whether we could use this to develop specific




```
from proteomeScoutAPI import ProteomeScoutAPI
PTM_API = ProteomeScoutAPI("everything_20150712.tsv")
PTM_API.get_PTMs("P04637") # get all p53 PTMs
```

Fig 1. ProteomeScout API and its application to reproducible analyses. The API block gives examples of functions that operate on a tab-separated file, which can be downloaded from ProteomeScout [16]. The analysis that can be done on the PTM-centric information in ProteomeScout can take on many forms given the flexibility of the API. Example code is given for retrieving all PTMs associated with the protein p53 (UniProtKB accession P04637). As the ProteomeScout dataset evolves and grows through new external data the same analyses can be re-run in the future.

doi:10.1371/journal.pone.0144692.g001

hypotheses of how pathogenic mutations may alter protein function. The serine/threonine kinase RAF1 is heavily mutated, and several of these mutations are linked to dysregulation of MAPK activity [19, 20]. Despite a link to disease, the RAF1 V263A mutation has failed to demonstrate disruption of 14-3-3 protein binding, which leads to increased RAF1 activity [21, 22], unlike other nearby mutations [23]. By understanding the PTMs within the region and using the data available on ProteomeScout from a quantitative phosphoproteomics study of regulation of this region during stem cell differentiation [24], we hypothesize that the V263A mutation may be affecting regulation important to differentiation, which results in the development of MAPK-related developmental disorders where V263A has been observed [19, 20]. These explorations demonstrate that the under-appreciation of nearby PTMs may be playing a role in RAF1 regulation, particularly during development, and explain how biochemical assays may have failed to uncover the effect of the V263A mutation via 14-3-3 misrecognition alone.

Materials and Methods

PTM growth

Numbers of PTMs for 1999 and 2004 were taken from early PhosphoBase and PhosphoSite papers, [25] and [26], respectively. For information from latter years, we parsed database downloads of files from PhosphoSite [27]. Database downloads were performed by the authors on the following dates: May 2007, September 2009, July 2013, January 2014, and July 2015. The data file and the iPythonNotebook that analyzed these data are available on ProteomeScout's documentation page (https://www.assembla.com/spaces/proteomescout/wiki). The file of all PTMs and numbers is also available as S1 Table.

Implementation of the ProteomeScout API and mutations analysis

The ProteomeScoutAPI was written in Python and is available in a Mercurial repository on the ProteomeScout Assembla project page. The current stable release (v1.0b, November, 2015) of ProteomeScout mammalian PTM file was downloaded from the ProteomeScout stable release FTP site ftp://ftp.seas.wustl.edu/pub/ProteomeScout_DbF/current_stable_release/. All calculations and analyses were performed in Python, specifically using iPython [28] notebooks and the following open-source projects: Pandas [29], NumPy [30], and Matplotlib [31]. Testing for enrichment was done using a one-sided Fisher's Exact test. We counted amino acids uniquely for having either a mutation or a modification within the specified window of 0 (on the amino acid) or 8 (within +/- 8 amino acids of the residue). If more than one mutation exists on the same amino acid and at least one of the mutations was known to be pathogenic or disease-related, then we assigned that residue a pathogenic/disease phenotype during all analyses. False discovery rate [32] was used as a multiple hypothesis correction technique, where denoted.

In accordance with recommendations on best practices [33] for developing bioinformatics software, the ProteomeScoutAPI has a full software testing suite, built using the Python unittest framework. This series of tests ensures that updates and changes to the code do not inadvertently lead to the introduction of software bugs elsewhere. Importantly, extending and growing this suite to accommodate new features is simply a few lines of additional code, ensuring that as the ProteomeScoutAPI grows and new functionality is added, a formal testing framework can be built in parallel.

The analysis code is available on GitHub and can be visualized on nbviewer at: http://nbviewer.ipython.org/github/knaegle/MutationsNotebooks/tree/master/. All calculations and graphs for study bias, PTM enrichment, and resampling for charge distributions are available in the iPython notebooks. SVG exports for RAF1 studies and data from the study by Rigbolt et al. [24] were taken from ProteomeScout [16].

Definition of disease mutations

This study focuses on mutations which were identified as being disease-related based on having a dbSNP Clinical Significance annotation of 'pathogenic'. However, we also performed an identical analysis on mutations identified as being disease-related based on their UniProKB annotation, as defined by the Human polymorphisms and disease mutations index file (http://www.uniprot.org/docs/humsavar, S3 Table). The dbSNP annotations yield 784 proteins with at least one pathogenic mutation, while the UniProtKB annotations yield 2273 such proteins. The dbSNP annotations are taken directly from the ProteomeScout database annotations using the ProteomeScoutAPI. The UniProtKB annotations were pulled from the humsavar file and then mapped to ProteomeScout database records. All the code for performing this analysis is provided.

Controlling for amino acid content of the pathogenic set

We created random foregrounds the same size and with the same distribution of amino acid types as the real foreground, the pathogenic set, which is enriched for arginines in particular. We then tested these random foregrounds for enrichment of nearby PTMs using the same analysis as outlined above. We ran many sets of 10 random foregrounds and never observed enrichment between mutations in random foregrounds and PTMs above what was expected by random chance alone.

Results and Discussion

Growth of the PTM proteome (PTMome)

Our understanding of post-translational modifications in a wide range of organisms is rapidly expanding. We explored the growth of the most well-studied PTMs in the human proteome by looking at a single resource of PTMs across time, shown in Fig 2. Specifically, we examined the PhosphoSite database [27], obtaining the number of phosphosites from the PhosphoSite papers in 1999 [25] and 2004 [26] as well as the datasets downloaded by the PTMScout and ProteomeScout authors from PhosphoSite between 2007 and 2015. The number of identified phosphosites grows exponentially during this decade and a half, with just tens of sites in 1999 growing to thousands of known sites by 2015. The first acetylation, sumoylation, and ubiquitination database entries appear in 2010, when high-throughput purification methods were

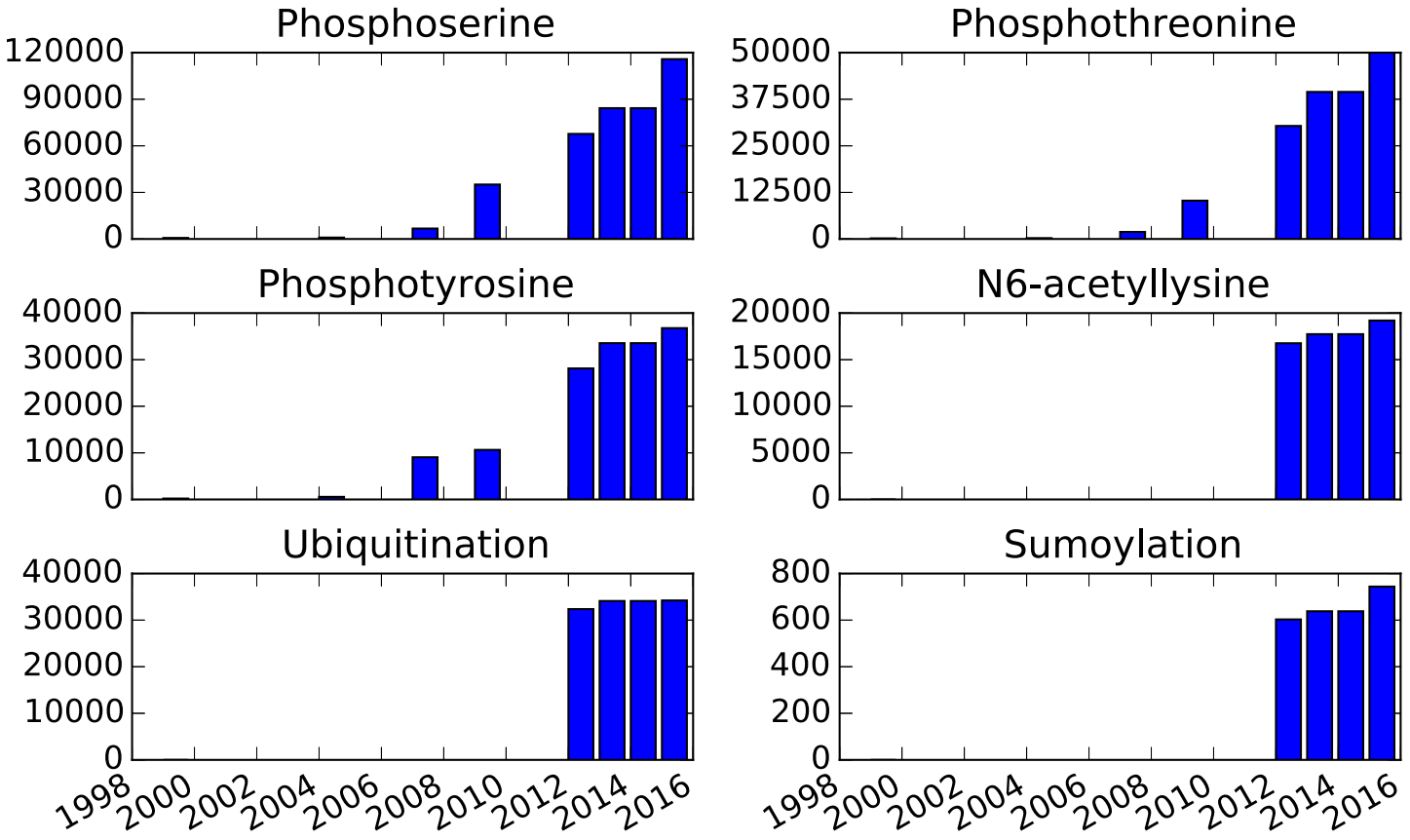


Fig 2. Growth of identified human PTMs across time. Data is based on PhosphoBase publications in 1999 [25] and 2004 [26] and PhosphoSite [27] downloads from 2007, 2009, 2012, 2013, 2014, and 2015.

doi:10.1371/journal.pone.0144692.g002

developed and tied to mass spectrometry [1]. The slowing of growth in the number of newly identified phosphorylation sites in the latter years may indicate that the number of novel phosphorylation sites discovered in human proteins from standard cell lines is approaching saturation. However, it is expected that the identification of phosphosites in new organisms, as well as tissue- and disease-specific sites, will take over as the primary contributors to novel measurements in the coming years. While phosphorylation has, historically, been the most well understood class of PTMs, coverage of other modification types may be expected to grow robustly based on these trends.

The growth of information associated with PTMs—as well as the surrounding information regarding protein sequence annotations—motivates the necessity for a sustainable database of post-translational modifications and protein annotations such that analysis and research can be updated easily. Certainly, these results indicate that conclusions drawn as recently as 2010 would no longer reflects our understanding in 2015, and that the shift in our understanding based on newly available data alters the set of relevant questions surrounding PTMs. In the remainder of this work we will introduce our solution to performing reproducible analyses of the PTMome and use this tool to explore the relationship between PTMs and human mutations.

Reproducible analysis of PTM-centric studies

To address the challenge of a rapidly changing PTMome, we built an application programming interface (API) to interact with ProteomeScout database files. ProteomeScout's database files are updated weekly and every six months a stable release is created. The weekly updates may include changes since the last week as a result of users uploading data. Uploading data triggers an update of annotations associated with those specific protein records in the database. In the stable release, all protein annotations, such as Gene Ontology terms [18] and mutations [17, 34] are updated. Additionally, the major compendia, such as UniProtKB and Phosphosite PTM datasets are updated. The weekly releases are available directly on the ProteomeScout website and the stable releases are available via FTP hosting, as described in the methods section.

In flat text files generated by ProteomeScout, all of the annotations for proteins are written in column-wise fashion with multiple annotations of a certain type being separated by semicolons. The database files are described in detail on ProteomeScout's wiki (available at https://www.assembla.com/spaces/proteomescout/wiki). However, the ProteomeScoutAPI removes the necessity of understanding the exact formatting of the database file by automatically parsing the file into Python objects that can be interacted with in a straightforward manner. For example, using the get_mutations(<ACC>) command, one can retrieve all mutations for a protein record, based on a particular protein accession. The ProteomeScoutAPI code and help documentation are available on the freely-hosted ProteomeScout project page (https://www.assembla.com/spaces/proteomescout). Fig 1 demonstrates the usage of the API and how, when analyses are based on a ProteomeScout database download, the analysis written in Python with the API can be used for the re-analysis of the data with the same database download or updated with future database downloads to reflect the growing knowledge of PTM and other protein information.

Study bias confounds proteome analysis

In this study, we wish to test the hypothesis that disease-related point mutations are more likely to be near or at sites of post-translational modifications. To do this we used the non-synonymous mutation annotations within ProteomeScout, which are harvested from NCBI's dbSNP [17]. A subset of these mutations are annotated with 'Clinical Significance', which includes the categories 'Pathogenic' and 'Non-pathogenic'. Prior to testing mutations from all protein

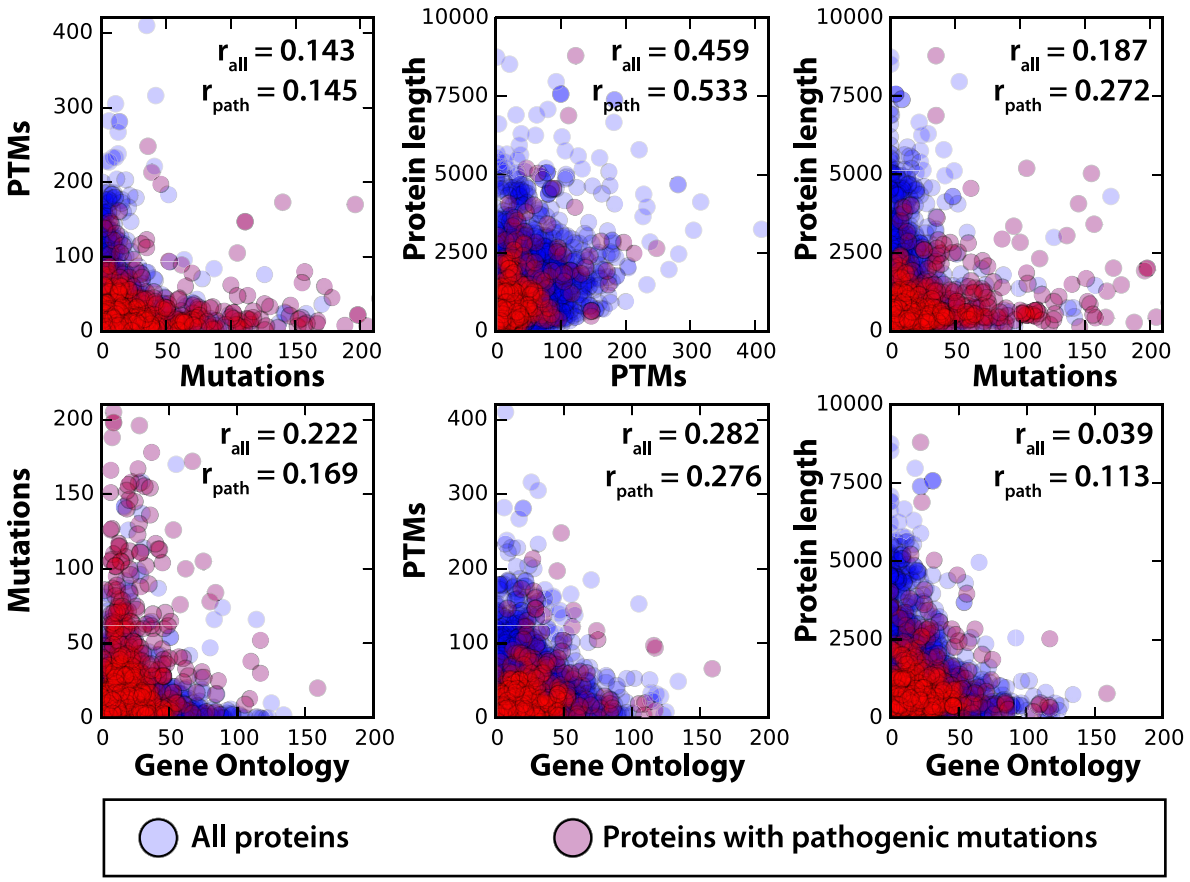


Fig 3. Correlation between protein annotations. Scatter plots for all comparisons of the number of annotations per protein or the length of the protein. Points in blue represent the number of annotations on a protein that does not contain a pathogenic mutation. Red represents a protein with at least one pathogenic mutation, which becomes the set of proteins studied in subsequent analyses. Correlations between numbers of labels on a per protein basis are given as well as the correlation between annotations on the pathogenic set. All correlations were significant with a p-value less than 1E-08. These plots and correlation values, broken down by PTM type, are available in S1 Fig.

doi:10.1371/journal.pone.0144692.g003

records for their proximity to known PTMs, we first explored for the possibility of study bias, i.e. the possibility that proteins with pathogenic mutations are more likely to have annotations of other types simply because they are well studied.

To test for study bias, we first looked at the correlation between the number (per protein) of two different protein annotations, or the number of protein annotations versus the total length of the protein (Fig 3). Specifically, the annotations we considered were GO terms, mutation records, and PTM records. In total, 21,910 human protein records existed in the stable release considered for this study, 15,480 of which have at least one known mutation. The p-values of all correlation values are significant, and there is some degree of positive correlation between every pairing. The highest correlation exists between PTMs and sequence length, indicating that the longer a sequence is, the more likely it is to have a larger number of PTMs. To a lesser extent this also holds true for mutations.

The positive correlation of all mutations with PTMs and with pathogenic mutations and PTMs (Fig 3) led us to test for specific study bias of proteins containing records of pathogenic mutations. To test for this, we performed a Fisher's Exact Test to compare the significance of label distributions based on proteins having at least one pathogenic mutation. There were 784 proteins with a pathogenic mutation and these proteins are significantly more likely to have a

large number of GO terms and PTMs in addition to having many annotated mutations in general. For example, when we tested for enrichment of proteins having at least ten GO annotations or PTMs we find pathogenic mutation containing proteins to be significantly enriched (p-values of 2e-101 and 2e-19, respectively). Study bias appears to occur on the level of the full protein, i.e. a protein with a known pathogenic mutation is much more likely to have more GO annotations, PTM annotations, and mutations in general. Therefore, in the following work we control for study bias by only considering those proteins that have at least one known pathogenic mutation, and are therefore more likely to have more annotations and mutations overall.

Pathogenic mutations are enriched for nearby PTMs

To test for the relationship between mutations and PTMs, controlling for study bias, we used the set of human proteins that have at least one pathogenic mutation. This included 795 human proteins, which all together contain 21,085 mutations (1,896 are pathogenic) and 16,701 total PTMs. For each of the human PTMs with a sufficiently large number of annotations (phosphoserine, phosphothreonine, phosphotyrosine, acetylation, ubiquitination, and N-Glycosylation), we tested for the significance of having a pathogenic mutation on the site of modification or within a possible recognition area of the site of modification using a one-sided Fisher's Exact test. We found that pathogenic mutations were not significantly more likely occur on sites of modification than other sites in the proteome. Only about 10% of all mutations (1,896) occur on a specific site of modification with relatively small numbers for the individual modification types. Therefore, we cannot rule out that lack of significance is not meaningful based on the sample sizes. However, most modifications demonstrated a significant relationship with pathogenic mutations when a regulatory window was considered (Table 1). In particular, ubiquitination occurring near a site of mutation means the mutation is much more likely to have a known relationship with human disease, whereas N6-acetylysine and N-Glycosylation were not significantly related to pathogenic mutations (possibly also due to limited set size). To a lesser extent than ubiquitination, phosphorylation of serines, threonines and tyrosines were significantly related to pathogenic mutations. These results indicate there is a strong relationship between the likelihood that mutations will be associated with pathogenicity and their proximity to known protein modifications, and that this relationship is dependent on the type of modification.

The relationship between modifications and pathogenic mutations indicates that misrecognition of PTM sites by enzymes or binding domains could lead to functional disruption which is more likely to cause disease than other mutations in general. However, an alternate hypothesis is that pathogenic mutations and sites of modification are coincident based on some other property they have in common. For example, if mutations on the protein surface are more

Table 1. Significance between dbSNP pathogenic or disease mutations and PTMs.

Modification	dbSNP 'pathogenic' p-value	UniProtKB 'Disease' p-value
Ubiquitination	2E-07	4E-03
Phosphotyrosine	3E-02	1E-05
Phosphoserine	2E-02	1E+00
Phosphothreonine	2E-02	1E+00
N6-acetyllysine	3E-01	1E-01
N-Glycosylation	1E+00	7E-01

Significance calculated on one-sided Fisher's exact test for 'pathogenic' or 'disease' mutations having the designated PTM within eight amino acids, compared to all mutations in the background set (corrected for study bias).

doi:10.1371/journal.pone.0144692.t001

likely to be detrimental and modifications are more likely to occur on the surface of the protein, then their coincidental location on the protein alone might explain the significant relationship we observed. Indeed, we found a skew in the types of amino acids that have known pathogenic mutations. Specifically, there is a significant bias towards arginines in the pathogenic set of mutations (p-value 2E-15), which might indicate a higher likelihood of surface accessibility. To rule out the effect of the distribution of amino acid types, and possible location of amino acids within a protein structure, we created random foregrounds that contained the same number of mutations as the pathogenic set and the same distribution of amino acids. In ten randomized trials, we never observed a significant enrichment of these foregrounds having nearby modifications. These results indicate that globally there is a significant relationship between pathogenic human mutations and modifications that cannot be accounted for by co-incidental charge of the mutations alone.

Extending to a larger set of disease-related mutations

We used dbSNP's 'pathogenic' mutations for our initial tests since these labels are available in the ProteomeScout database. However, there is a significantly larger number of 'Disease'-related variants available in the UniProtKB database. Therefore, we also tested the relationship between the likelihood of being labeled as disease-related in UniProtKB [35] with PTMs using the same analysis by cross-referencing from this set to the ProteomeScout database file using the API. We observed similar trends with regards to study bias—disease-mutation containing proteins were much more likely to have more Gene Ontology terms and PTMs. Therefore, we controlled for study bias in this set in the same manner, by creating a background consisting of only the mutations from proteins that contain at least one disease-annotated mutation. Of the 68,819 mutations in UniProtKB, 21,999 are labeled with the annotation 'Disease' (32%). However, when we identified the subset of proteins with at least one disease-labeled mutation and re-evaluate the annotation associated with mutations in that subset of proteins the proportion of disease mutations goes from 32% to 68%. This suggests two possibilities: 1) certain proteins are hubs for pathogenicity or 2) there is bias in the annotation of disease mutations. In this second "rich-get-richer" scenario, the identification of a disease-related mutation on a protein may lead to focused efforts on identification of other mutations on that protein that also result in a disease annotation. There tends to be clustering of disease annotated mutations on proteins (S3 Table), which supports both hypotheses.

Despite the inherent differences of the two mutation datasets, with regard to their size and distribution of pathogenic/disease to total mutation ratio (68% vs. 10%), we performed the same statistical analysis of testing for enrichment of disease mutations set near known PTMs (Table 1). Ubiquitination and tyrosine phosphorylation continue to be enriched—they are much more likely to be near a disease-related mutation than would be expected by random chance. However, there is a change in their significance, relative to the findings in the dbSNP pathogenic set and no other modifications are enriched. Considering this, the exact conclusions are highly dependent on the set of mutations and their disease annotations. As a corollary, these results are also dependent on the current state of PTM knowledge. In particular, most PTMs such as sumoylation, myristoylation, and methylation, have too few annotations for sufficient statistical testing. As PTM measurement and discovery expands, particularly for those PTMs with few ProteomeScout annotations, we (and others) can test for a relationship with disease-related mutations in the future. The dependence on the data of mutations used and state of PTM knowledge on the findings highlights the importance for a framework that enables the easy reproducibility of current results, extensibility to different analysis types, and the re-analysis at a future date when knowledge has changed.

Using PTMs to predict the effect of mutations

Given that there is consistent insight to pathogenicity based on proximity to PTMs, we sought to identify whether knowing the proximity or number of modifications near a mutation could be helpful in determining the impact of a mutation. S2 Table lists the sites of human dbSNP mutations. Not surprisingly, 21 of the top 28 mutations with the largest number of nearby modifications are on the tumor suppressor protein TP53. Also, in the top 100 of mutations, ranked by number of nearby modifications, are a series of RAF1 mutations occurring between positions R256 and V263, Fig 4A and 4B. When phosphorylated, S259 is recognized by 14-3-3, whose binding negatively regulates RAF1 activity and subsequently decreases MAPK activation [21, 22]. Multiple genetic studies have identified RAF1 mutations near S259 in patients with Noonan and LEOPARD Syndrome, both syndromes which show a characteristic increase in MAPK activation [19, 20]. Multiple independent studies have verified that a number of these mutations reduce binding to 14-3-3 and increase MAPK activation, including R256S, S257L, S259F, T260I/R, and P261A/S/L [19, 20, 23]. Based on these results, it is clear that these disease-related mutations act through the disruption of the recognition and regulation of PTMs and highlights the utility of understanding the relationship between mutations and modifications.

The ProteomeScout protein view [16] of this region of RAF1 highlights the density of phosphorylation sites and known mutations. In addition to S259 phosphorylation, which has a well-understood role in regulating RAF1 activity, phosphorylation has been observed on other nearby phosphorylation sites including S257, T258, and T260, Fig 4A. ProteomeScout currently

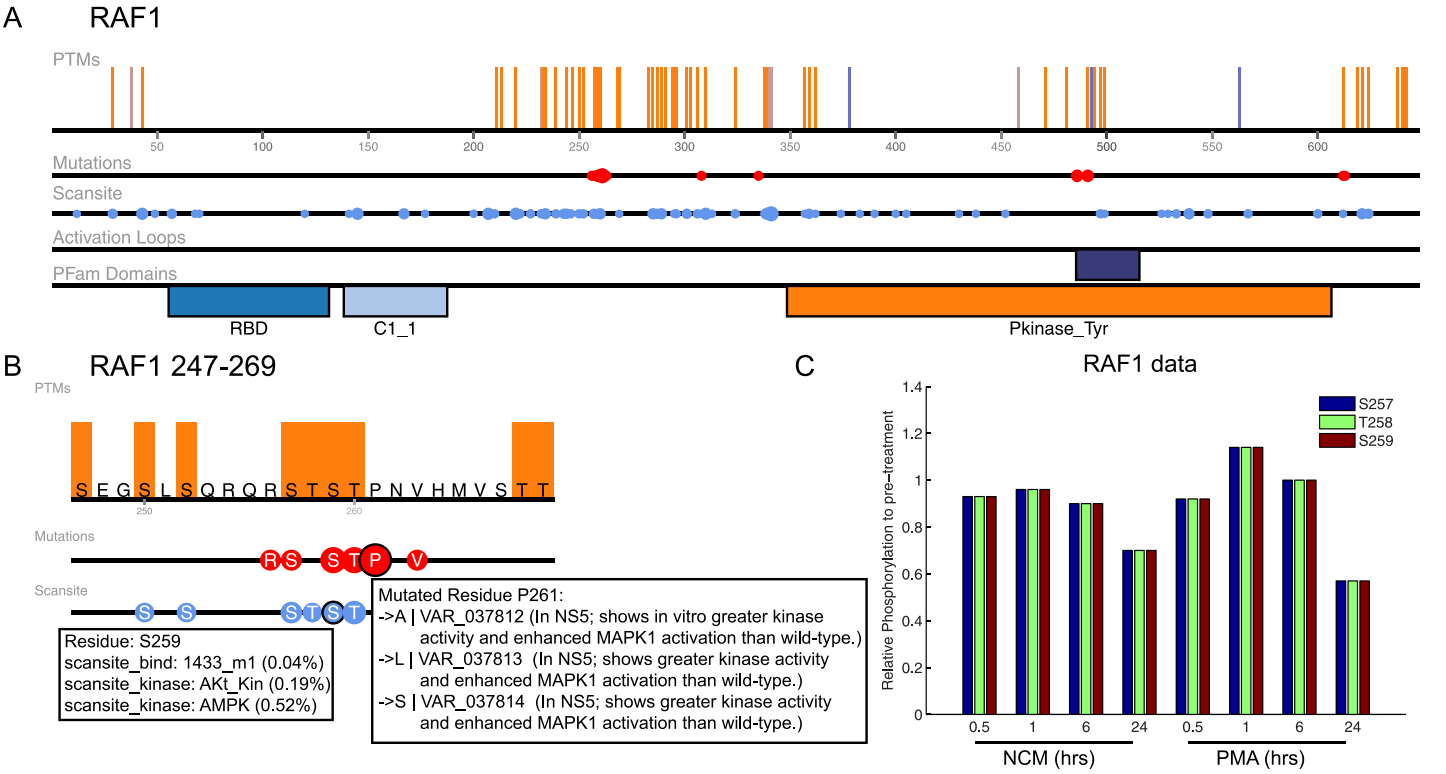


Fig 4. Predicting effect of RAF1 mutations. (A) Full length RAF1 from ProteomeScout [16] with PTM annotations, domains, mutations from dbSNP, and Scansite predictions. (B) RAF1 in the area of interest near S259, which is involved in 14-3-3 recognition. (C) Quantitative measurements of phosphorylation on S257, T258 and S259 from a study of human embryonic stem cells without conditioning and with conditioning for stem cell differentiation (PMA treatment) [24]. Data from study downloaded from ProteomeScout [16].

doi:10.1371/journal.pone.0144692.g004

includes one experiment with quantitative measurements of phosphorylation in this region. Rigbolt et al. identified phosphorylation of S257, T258, and S259 in human embryonic stem cells (hESCs) and quantitatively measured their relative phosphorylation levels in response to differentiation. Fig 4C contains the plot of ProteomeScout data from [24] for this region of phosphorylation sites in changing to non-conditioned media (NCM) or in response to phorbol 12-myristate 13-acetate (PMA), treatments that initiate stem cell differentiation, compared to the pre-treated samples. The three sites in the region of interest, which are in the concentrated region of mutations found in Noonan and LEOPARD syndrome samples, follow the same pattern and exhibit no relative change until after 24 hours of treatment, where they decrease in phosphorylation. We were surprised the quantitative data amongst the three sites were identical. Upon revisiting the original supplementary information on this dataset in Rigbolt et al., the data is faithfully represented in ProteomeScout, but the assignment score suggests the specific site of modification was not accurately identified. However, given multiple sources of identification for these phosphorylation sites [23, 27, 35–43] it is likely that this experiment indicates that at least in hESCs, phosphorylation occurs on some subset of these sites and that they demonstrate a dynamic response to initiation of stem cell differentiation. Phosphorylation on these alternate sites is not currently appreciated as playing a role in regulating 14-3-3 activity, yet this may represent a process by which traditional S259/14-3-3 recognition is altered. It also expands the possibility of mechanisms by which these mutations affect protein function and regulation, and may help lead to hypotheses of how V263A plays a role in development of Noonan Syndrome, despite having no measurable effect on 14-3-3 binding [23].

Supporting Information

S1 Table. The number of all modifications collected from PhosphoSite from 1998 to 2015.
(CSV)

S2 Table. List of human mutations and PTMs from dbSNP.
(XLSX)

S3 Table. List of human mutations and PTMs from UniProtKB.
(TXT)

S1 Fig. Correlation plots and values for all metrics broken down by PTM type.
(PDF)

Acknowledgments

The authors wish to thank members of the Center for Biological Systems Engineering and the Naegle Lab for support and feedback of this work.

Author Contributions

Conceived and designed the experiments: AH KMN. Performed the experiments: AH KMN. Analyzed the data: KMN. Contributed reagents/materials/analysis tools: AH KMN. Wrote the paper: AH KMN.

References

1. Olsen JV, Mann M. Status of Large-scale Analysis of Post-translational Modifications by Mass Spectrometry. Mol Cell Proteomics. 2013 dec; 12(12):3444–52. Available from: http://www.ncbi.nlm.nih.gov/pubmed/24187339. doi: 10.1074/mcp.O113.034181 PMID: 24187339

2. Uyar B, Weatheritt RJ, Dinkel H, Davey E, Gibson TJ. Proteome-wide analysis of human disease mutations in short linear motifs: neglected players in cancer? Mol Biosyst. 2014 doi: 10.1039/c4mb00290c PMID: 25057855
3. Tatarova Z, Brabek J, Rosel D, Novotny M. SH3 Domain Tyrosine Phosphorylation—Sites, Role and Evolution. PLoS One. 2012; 7(5):1–8.
4. Beltrao P, Albanèse V, Kenner L, Swaney D, Burlingame A, Villén J, et al. Systematic Functional Prioritization of Protein Posttranslational Modifications. Cell. 2012 jul; 150(2):413–425. Available from: http://linkinghub.elsevier.com/retrieve/pii/S0092867412007064.PMID: 22817900
5. Minguez P, Parca L, Diella F, Mende DR, Kumar R, Helmer-citterich M, et al. Deciphering a global network of functionally associated post-translational modifications. Mol Syst Biol. 2012; 8(599):1–14. Available from: http://dx.doi.org/10.1038/msb.2012.31.
6. Schweiger R, Linial M. Cooperativity within proximal phosphorylation sites is revealed from large-scale proteomics data. Biol Direct. 2010; 5(6):1–17.
7. Huyck L, Troys MV, Ampe C. Phosphosite conservation in single domain orthologs versus paralogs: A way to combine differential regulation with redundant core functions. FEBS Lett. 2012; 586(4):296–302.doi: 10.1016/j.febslet.2012.01.018 PMID: 22265693
8. Freschi L, Courcelles M, Thibault P, Michnick SW, Landry CR. Phosphorylation network rewiring by gene duplication. Mol Syst Biol. 2011 jan; 7(504):504. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3159966&tool = pmcentrez&rendertype = abstract. doi: 10.1038/msb.2011.43 PMID: 21734643
9. Landry CR, Levy ED, Michnick SW. Weak functional constraints on phosphoproteomes. Trends Genet. 2009 may; 25(5):193–7. Available from: http://www.ncbi.nlm.nih.gov/pubmed/19349092. doi: 10.1016/j.tig.2009.03.003 PMID: 19349092
10. Pearlman SM, Serber Z, JEF Jr. Theory A Mechanism for the Evolution of Phosphorylation Sites. Cell. 2011; 147(4):934–946. Available from: http://dx.doi.org/10.1016/j.cell.2011.08.052.
11. Wang Z, Ding G, Geistlinger L, Li H, Liu L, Zeng R, et al. Evolution of protein phosphorylation for distinct functional modules in vertebrate genomes. Mol Biol Evol. 2011 mar; 28(3):1131–40. Available from: http://www.ncbi.nlm.nih.gov/pubmed/20956806. doi: 10.1093/molbev/msq268 PMID: 20956806
12. Gnad F, Forner F, Zielinska DF, Birney E, Gunawardena J, Mann M. Evolutionary constraints of phosphorylation in eukaryotes, prokaryotes, and mitochondria. Mol Cell Proteomics. 2010 dec; 9(12):2642–53. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3101853&tool = pmcentrez&rendertype = abstract. doi: 10.1074/mcp.M110.001594 PMID: 20688971
13. Radivojac P, Baenziger PH, Kann MG, Mort ME, Hahn MW, Mooney SD. Gain and loss of phosphorylation sites in human cancer. Bioinformatics. 2008; 24:241–247. doi: 10.1093/bioinformatics/btn267
14. Reimand J, Bader GD. Systematic analysis of somatic mutations in phosphorylation signaling predicts novel cancer drivers. Mol Syst Biol. 2013; 9(637):637. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3564258&tool = pmcentrez&rendertype = abstract. doi: 10.1038/msb.2012.68 PMID: 23340843
15. Tan CSH, Bodenmiller B, Pasculescu A, Jovanovic M, Hengartner MO, Jørgensen C, et al. Comparative analysis reveals conserved protein phosphorylation networks implicated in multiple diseases. Sci Signal. 2009 jan; 2(81):ra39. Available from: http://www.ncbi.nlm.nih.gov/pubmed/19638616. doi: 10.1126/scisignal.2000316 PMID: 19638616
16. Matlock MK, Holehouse AS, Naegle KM. ProteomeScout: a repository and analysis resource for post-translational modifications and proteins. Nucleic Acids Res. 2015 nov; 43(D1):D521–D530. Available from: http://www.ncbi.nlm.nih.gov/pubmed/25414335. doi: 10.1093/nar/gku1154 PMID: 25414335
17. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, et al. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. 2001 jan; 29(1):308–11. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=29783&tool = pmcentrez&rendertype = abstract. doi: 10.1093/nar/29.1.308 PMID: 11125122
18. Consortium" TGO. Gene Ontology: tool for the unification of biology. Nat Genet. 2000; 25(may):25–29. doi: 10.1038/75556
19. Pandit B, Sarkozy A, Pennacchio LA, Carta C, Oishi K, Martinelli S, et al. Gain-of-function RAF1 mutations cause Noonan and LEOPARD syndromes with hypertrophic cardiomyopathy. Nat Genet. 2007; 39(8):1007–1012.PMID: 17603483
20. Kobayashi T, Aoki Y, Niihori T, Cavé H, Verloes A, Okamoto N, et al. Molecular and clinical analysis of RAF1 in Noonan syndrome and related disorders: dephosphorylation of serine 259 as the essential mechanism for mutant activation. Hum Mutat. 2010 mar; 31(3):284–94. Available from: http://www.ncbi.nlm.nih.gov/pubmed/20052757. doi: 10.1002/humu.21187 PMID: 20052757

21. Wilker E, Yaffe MB. 14-3-3 Proteins–a focus on cancer and human disease. J Mol Cell Cardiol. 2004 sep; 37(3):633–42. Available from: http://www.ncbi.nlm.nih.gov/pubmed/15350836.PMID: 15350836

22. Light Y, Paterson H, Marais R. 14-3-3 Antagonizes Ras-Mediated Raf-1 Recruitment to the Plasma Membrane To Maintain Signaling Fidelity. Mol Cell Biol. 2002; 22(14):4984–4996. doi: 10.1128/MCB.22.14.4984-4996.2002 PMID: 12077328

23. Molzan M, Schumacher B, Ottmann C, Baljuls A, Polzien L, Weyand M, et al. Impaired binding of 14-3-3 to C-RAF in Noonan syndrome suggests new approaches in diseases with increased Ras signaling. Mol Cell Biol. 2010 oct; 30(19):4698–711. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2950525&tool = pmcentrez&rendertype = abstract. doi: 10.1128/MCB.01636-09 PMID: 20679480

24. Rigbolt KTG, Prokhorova TA, Akimov V, Henningsen J, Johansen PT, Kratchmarova I, et al. System-wide temporal characterization of the proteome and phosphoproteome of human embryonic stem cell differentiation. Sci Signal. 2011 jan; 4(164):rs3. Available from: http://www.ncbi.nlm.nih.gov/pubmed/21406692. doi: 10.1126/scisignal.2001570 PMID: 21406692

25. Kreegipuu a, Blom N, Brunak S. PhosphoBase, a database of phosphorylation sites: release 2.0. Nucleic Acids Res. 1999 jan; 27(1):237–9. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=148144&tool = pmcentrez&rendertype = abstract. doi: 10.1093/nar/27.1.237 PMID: 9847189

26. Hornbeck PV, Chabra I, Kornhauser JM, Skrzypek E, Zhang B. PhosphoSite: A bioinformatics resource dedicated to physiological protein phosphorylation. Proteomics. 2004; 4:1551–1561. doi: 10.1002/pmic.200300772 PMID: 15174125

27. Hornbeck PV, Kornhauser JM, Tkachev S, Zhang B, Skrzypek E, Murray B, et al. PhosphoSitePlus: a comprehensive resource for investigating the structure and function of experimentally determined post-translational modifications in man and mouse. Nucleic Acids Res. 2012 jan; 40(Database issue):D261–70. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3245126&tool = pmcentrez&rendertype = abstract. doi: 10.1093/nar/gkr1122 PMID: 22135298

28. Perez F, Granger BE. IPython: A System for Interactive Scientific Computing. Comput Sci Eng. 2007; (9:):21–29. doi: 10.1109/MCSE.2007.53

29. McKinney W. Data Structures for Statistical Computing in Python. In: van der Walt, S, Millman, J, editors. Proc. 9th Python Sci. Conf.; 2010. p. 51–56.

30. Walt SVD, Colbert SC, Varoquaux G. The NumPy Array: A structure for efficient numerical computation. Comput Sci Eng. 2011; 13(2):22–30. doi: 10.1109/MCSE.2011.37

31. Hunter JD. Matplotlib: A 2D Graphics Environment. Comput Sci Eng. 2007;(9:):90–95. doi: 10.1109/MCSE.2007.55

32. Benjamini Y, Hochberg Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. J R Stat Soc Ser B. 1995; 57(1):289–300.

33. Leprevost FdV, Barbosa VC, Francisco EL, Perez-Riverol Y, Carvalho PC. On best practices in the development of bioinformatics software. Front Genet. 2014; 5:1–3. Available from: http://www.genome.org/cgi/doi/10.1101/gr.4086505. doi: 10.3389/fgene.2014.00199

34. Coordinators NR. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 2014 jan; 42(Database issue):D7–17. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3965057&tool = pmcentrez&rendertype = abstract. doi: 10.1093/nar/gkt1146

35. Consortium TU. Activities at the Universal Protein Resource (UniProt). Nucleic Acids Res. 2014 jan; 42 (1):D191–8. Available from: http://www.ncbi.nlm.nih.gov/pubmed/24253303. doi: 10.1093/nar/gkt1140

36. Lu CT, Huang KY, Su MG, Lee TY, Bretaña NA, Chang WC, et al. dbPTM 3.0: an informative resource for investigating substrate site specificity and functional association of protein post-translational modifications. Nucleic Acids Res. 2013 jan; 41(Database issue):D295–305. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3531199&tool = pmcentrez&rendertype = abstract. doi: 10.1093/nar/gks1229 PMID: 23193290

37. Prasad TSK, Goel R, Kandasamy K, Keerthikumar S, Kumar S, Mathivanan S, et al. Human Protein Reference Database–2009 update. Nucleic Acids Res. 2009 jan; 37(Database issue):D767–72. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2686490&tool = pmcentrez&rendertype = abstract.

38. Diella F, Gould CM, Chica C, Via A, Gibson TJ. Phospho.ELM: a database of phosphorylation sites—update 2008. Nucleic Acids Res. 2008; 36(October 2007):240–244.

39. Zhou H, Di Palma S, Preisinger C, Peng M, Polat AN, Heck AJR, et al. Toward a comprehensive characterization of a human cancer cell phosphoproteome. J Proteome Res. 2013; 12(1):260–71. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23186163. doi: 10.1021/pr300630k PMID: 23186163

40. Moritz A, Li Y, Guo A, Villén J, Wang Y, Kornhauser J, et al. Akt–RSK– S6 Kinase Signaling Networks Activated by Oncogenic Receptor Tyrosine Kinases. Sci Signal. 2014; 3(136):1–12.

41. Cantin GT, Yi W, Lu B, Park SK, Xu T, Lee Jd, et al. Combining Protein-Based IMAC, Peptide-Based IMAC, and MudPIT for Efficient Phosphoproteomic Analysis Greg. J Proteome Res. 2008; 7:1346–1351.PMID: 18220336

42. Chen Rq, Yang Qk, Lu Bw, Yi W, Cantin G, Chen Yl, et al. CDC25B Mediates Rapamycin-Induced Oncogenic Responses in Cancer Cells. Cancer Res. 2009; 69(6):2663–2668. Available from: http://cancerres.aacrjournals.org/cgi/doi/10.1158/0008-5472.CAN-08-3222. doi: 10.1158/0008-5472.CAN-08-3222 PMID: 19276368

43. Wang Yt, Tsai Cf, Hong TC, Tsou Cc, Lin Py, Pan SH, et al. An informatics-assisted label-free quantitation strategy that depicts phosphoproteomic profiles in lung cancer cell invasion. J Proteome Res. 2010; 9(11):5582–97. Available from: http://www.ncbi.nlm.nih.gov/pubmed/20815410. doi: 10.1021/pr100394u PMID: 20815410