Measuring pathway database coverage of the phosphoproteome

Hannah Huckstep[1,2], Liam G. Fearnley[2,3] and Melissa J. Davis[1,2,4]

[1] Division of Bioinformatics, The Walter and Eliza Hall Institute of Medical Research, Parkville, Victoria, Australia
[2] Department of Medical Biology, Faculty of Medicine, Dentistry and Health Sciences, University of Melbourne, Parkville, Victoria, Australia
[3] Division of Population Health, Walter and Eliza Hall Institute of Medical Research, Parkville, Victoria, Australia
[4] Department of Clinical Pathology, Faculty of Medicine, Dentistry and Health Sciences, University of Melbourne, Melbourne, Victoria, Australia

Corresponding authors
Liam G. Fearnley,
fearnley.l@wehi.edu.au
Melissa J. Davis,
davis.m@wehi.edu.au

ABSTRACT

Protein phosphorylation is one of the best known post-translational mechanisms playing a key role in the regulation of cellular processes. Over 100,000 distinct phosphorylation sites have been discovered through constant improvement of mass spectrometry based phosphoproteomics in the last decade. However, data saturation is occurring and the bottleneck of assigning biologically relevant functionality to phosphosites needs to be addressed. There has been finite success in using data-driven approaches to reveal phosphosite functionality due to a range of limitations. The alternate, more suitable approach is making use of prior knowledge from literature-derived databases. Here, we analysed seven widely used databases to shed light on their suitability to provide functional insights into phosphoproteomics data. We first determined the global coverage of each database at both the protein and phosphosite level. We also determined how consistent each database was in its phosphorylation annotations compared to a global standard. Finally, we looked in detail at the coverage of each database over six experimental datasets. Our analysis highlights the relative strengths and weaknesses of each database, providing a guide in how each can be best used to identify biological mechanisms in phosphoproteomic data.

INTRODUCTION

Phosphorylation is a reversible post-translational modification (PTM) capable of controlling numerous aspects of a protein's function. The effects of phosphorylation are widespread and central to cellular signal transduction, with aberrant protein phosphorylation known to be both a cause and result of many diseases such as cancer and diabetes (*Cohen, 2001*). A network of kinases and phosphatases finely regulate protein phosphorylation and in turn, cellular signalling. These enzymes offer a great opportunity for targeted therapies to treat disease. Currently, more than 28 kinase inhibitors have been approved for clinical use and up to one-third of protein targets being developed in the pharmaceutical industry are directly related to kinases (*Rask-Andersen et al., 2014*; *Bhullar*

et al., 2018). It has been found over years of targeted cancer therapy that the clinical success rate of kinase inhibitors exceeded other cancer therapies demonstrating higher selectivity and lower cytotoxicity (*Walker & Newell, 2009*). These achievements highlight the importance of understanding phosphorylation networks in disease and in normal function.

Continuous advances in mass spectrometry have enabled large-scale identification and profiling of phosphosites across cell lines, tissues and species. The number of protein phosphosites is estimated at 230,000 in humans and 156,000 in mice (*Vlastaridis et al., 2017*). Yet these discoveries are vastly outpacing our understanding of the function and regulation of these PTMs. Phosphorylating kinases are known for only 5% of phosphosites, and a negligible number of phosphosites have recorded the implications of the phosphorylation event (*Needham et al., 2019*). Disagreement exists over the proportion of phosphosites predicted to be functionally important based on being evolutionarily conserved. Studies report anywhere from 35% (*Landry, Levy & Michnick, 2009*) to 65% (*Gray & Kumar, 2011*) of phosphosites are conserved, highlighting the difficulty in estimating the functional proportion of all known phosphosites. Despite this discrepancy, these reports agree that the proportion of functional phosphosites is much higher than the 5% that have been annotated to date.

In principle, the abundance of detailed phosphoproteomics data naturally lends itself to data-driven modelling approaches such as principal component analysis (PCA) or clustering (*Janes & Yaffe, 2006*). Drawing functional conclusions from these methods has not been as simple as anticipated. Many limitations of varying complexity stand in the way of elucidating phosphosite specific mechanisms. First, a protein phosphorylation can have a potential activating or inhibiting effect, which cannot be easily captured through typical high-throughput data alone. Next, operating under the conservative assumption that approximately 65% of phosphosites are not conserved and therefore have no functional consequence, there is an increased difficulty in distinguishing which few remaining phosphosites are functionally relevant (*Lienhard, 2008*; *Levy, Michnick & Landry, 2012*). Finally, it was found that 75% of significantly phosphorylated proteins have at least two phosphosites that are oppositely modulated (*Olsen et al., 2006*) making the use of high-throughput data alone largely unsuitable to determine which phosphosite modulates which function. On the other hand, network-driven methods which make use of previously established literature-derived knowledge of signalling, offer a more straightforward approach (*Sacco, Perfetto & Cesareni, 2018*).

Many databases capable of analysing phosphoproteomics broadly fit into three categories; Pathway databases, Protein–Protein Interaction (PPI) databases, and Phosphoproteomics focused databases. These databases were created for differing purposes and are typically chosen for differing types of analysis. However, each category holds signalling information pertaining to how a phosphorylation may affect signalling cascades. Each category has broad strengths and weaknesses in their capacity to analyse phosphoproteomics data as well as variable information and conclusions that can be drawn from an analysis. For example, a phosphoproteomics database would be best-suited to investigating upstream kinases (either experimentally derived or predicted) of a

phosphosite of interest, while a pathway database can be used to investigate the phosphosite's potential functional role in a particular phenotype or cell of interest. A PPI database captures information on potential interactions that might suggest complexes that phosphoprotein may be involved in, or potential interactions outside of a pathway context. These databases, however, can overlap as some PPI and pathway databases hold kinase-substrate (K-S) information, some pathway databases hold comprehensive complex information, and some phosphorylation databases hold disease and pathway annotations. These databases can be used both individually and in conjunction to provide a comprehensive analysis capable of developing hypotheses around individual phosphorylation signalling mechanisms.

Literature-derived, pathway-oriented databases such as the Kyoto Encyclopedia of Genes and Genomics (KEGG) (*Kanehisa & Goto, 2000*), Reactome (*Joshi-Tope et al., 2005*), WikiPathways (*Slenter et al., 2018*), and the SIGnaling Network Open Resource (SIGNOR) (*Perfetto et al., 2016*) have proven to be immensely useful in providing functional insight for other types of 'omic data. This has been largely driven through the development of enrichment methods using pathway-level information (*Khatri, Sirota & Butte, 2012*). Many tools have been developed to perform such analyses (*Yu et al., 2012*; *Young et al., 2010*; *Subramanian et al., 2005*; *Wu et al., 2010*). Another popular method for analysis and interpretation of 'omics data is to overlay or map data onto pathways. This technique has been successful in analysing data from metabolomics (*Chagoyen & Pazos, 2013*), genomics (*Ramanan et al., 2012*), transcriptomics (*Wang et al., 2019*) and also has been successful in numerous examples with phosphoproteomics data (*Satpathy et al., 2015*; *Humphrey et al., 2013*; *Rudolph et al., 2016*). These database resources are dependent upon curation and are constantly updating molecular interactions making them an invaluable source of current literature-derived knowledge. However, they are not without their own limitations as the curation process tends to result in an overrepresentation of well-studied proteins and a lack of lesser-known proteins, in particular phosphoproteins. In addition, in a comparison of phosphosite databases Sacco et al. found that many phosphosites that had been identified and characterized by low-throughput experiments (approximately 9% of PhosphoSitePlus at the time) could not be found in a typical phosphoproteomics experiment. They hypothesized that this may be due to a number of reasons including abundance, large sample complexity and the stochastic nature of peptide selection (*Sacco, Perfetto & Cesareni, 2018*).

A recent study comparing PPI databases found 375 resources in their literature search and observed that PubMed timeline data indicated a steady increase in research articles on PPIs across the years (*Bajpai et al., 2020*). Resources such as the Human Protein Reference Database (HPRD) (*Peri et al., 2004*; *Keshava Prasad et al., 2009*), the Biological General Repository for Interaction Datasets (BioGRID) (*Stark et al., 2006*), and the Search Tool for Retrieval of Interacting Genes/Proteins (STRING) (*Szklarczyk et al., 2019*) are incredibly useful resources that enable the analysis of interaction data in various contexts. PPI databases tend to be quite large as a result of their derivation from high-throughput experimental data. This may be an advantage when looking for novel associations

highlighted when measurements are mapped onto the data. As with pathway databases, PPIs are not without their limitations, as they typically represent an oversimplified depiction of cellular signalling making interpretation of the functional consequences of molecular perturbations difficult.

Popular phosphorylation focused databases such as PhosphoSitePlus (*Hornbeck et al., 2015*), PHOSIDA (*Gnad et al., 2007*), Phospho.ELM (*Diella et al., 2004*) and qPhos (*Yu et al., 2019*) host many phosphosites and act as repositories for both low and high-throughput data. Some have tools for motif analysis and phosphosite prediction but rarely contain information on the function of a phosphosite. qPhos is amongst the largest such databases and contains quantitative information for almost 200,000 non-redundant phosphorylation sites as well as the cell-type and temporal information. An extensive review of phosphoproteomics resources can be found here (*Savage & Zhang, 2020*). Although these databases can be extremely useful in their own right, the lack of functional annotations makes their use in elucidating phosphosite signalling mechanisms rather limited. Of those databases listed above, the exception is PhosphoSitePlus, which also contains a network of experimentally observed kinase-substrate relationships. Moving away from databases acting as phosphosite repositories, many databases holding signalling information in the form of a K-S networks exist such as RegPhos (*Huang et al., 2014*), PhosphoNet (*Safaei et al., 2011*) and Phosphonetworks (*Hu et al., 2014*). Though they are limited to K-S interactions, they have been proven useful in providing phosphoproteomics mechanistic insight either on their own or while integrated into other databases (*Rohrs et al., 2018*; *McGuire et al., 2017*; *Tong et al., 2019*). One method developed by *Sacco et al. (2016b)* starts by overlaying their data onto a literature-derived, pathway-oriented database, followed by filtering the network to an interpretable size based on a set of rules they developed. This method was successfully implemented in a phosphoproteomics context by mapping their data onto the K-S network from PhosphoSitePlus to discover novel mechanistic insights into phosphorylation mediated insulin signalling (*Sacco et al., 2016a*).

This work aims to systematically investigate the current landscape of resources suitable for the network-based analysis of phosphoproteomic data (as defined by our selection criteria, see Methods and Table S1). Specifically, we are interested in how well equipped each database is to uncover mechanisms of phosphorylation and connect phosphorylation events to a signalling network or pathway. The pathway databases we compare here are Reactome, KEGG, WikiPathways and SIGNOR. The PPI databases we compare are BioGRID and HPRD. PhosphoSitePlus was also included as it contains the largest subset of phosphoproteomic signalling-related proteins. We first analysed the proteome coverage of each database followed by the phosphorylation coverage of a subset of the above databases. Next, we explored the consistency between the database's phosphorylation annotations and the amino acid residue found in UniProt's (*Bateman et al., 2017*) protein sequence. Finally, we assessed the capability of each database in mapping experimental phosphoproteomics datasets of varying backgrounds.

MATERIALS AND METHODS

Systematic literature review of public databases

There is an abundance of pathway databases currently available for public use in interpreting biological data (*Kanehisa & Goto, 2000*; *Joshi-Tope et al., 2005*; *Slenter et al., 2018*; *Zhou et al., 2012*; *Wishart et al., 2006*; *Pratt et al., 2015*; *Szklarczyk et al., 2019*; *Cowley et al., 2012*). However, they tend to differ from each other in several properties. To address this, we systematically compared 29 databases and produced a selection criterion for inclusion in our analysis (Table S1, Supporting Information).

To be included in this comparison, each database had to fit the following criteria:

- Have phosphorylation information
- Have a connected signalling network component
- Be updated yearly, and if not must have been widely used and be widely cited (1,000 minimum)
- Freely accessible
- Downloadable

The pathway databases which met these criteria were HPRD, BioGRID, SIGNOR, Reactome, KEGG, PhosphoSitePlus and WikiPathways.

Data retrieval and processing

Since many of these databases were represented in differing formats and some with unique identifiers, several processing steps were conducted as needed per database to extract protein UniProt accessions and phosphosites as outlined in Fig. S1, Supporting Information. Dates of database access are outlined in Table S2, Supporting Information. For those databases available in a BioPax format, they were first converted into a Neo4j graph database object (*Neo4j, 2020*) (Schema described in Fig. S2, Supporting Information). The Neo4j graph database implementations were built and used to allow for quick and easy extraction of relevant information as described in more detail below.

Protein

For IMEX, qPhos and, UniProt and both Gene Ontology (GO) categories the data was available in TSV format with UniProt accessions easily extractable. Each database downloaded in the BioPax format (Reactome and PhosphoSitePlus) was converted into a Neo4j graph database in Java, where the list of UniProt accessions were extracted. SIGNOR's API was used to download the list of UniProt IDs involved in the signalling network. BioGRID's PPI file was downloaded in TSV format directly from the website with RefSeq IDs from the signalling portion of the database being taken. HPRD legacy files were downloaded from the Integrated Network and Dynamical Reasoning Assembler (INDRA) python package documentation (*Gyori et al., 2017*). For HPRD, RefSeq identifiers were taken, once again only from the signalling portion of the database. We consider the interaction network portion of HPRD and BioGRID, as a source of connected proteins with potential signalling function. For both KEGG and WikiPathways
only the identifiers belonging to a pathway were downloaded as we were only interested in proteins involved in signalling. The KEGG database contained only KEGG identifiers, and the WikiPathways download contained only Entrez IDs. All sets of identifiers were converted to UniProt accessions. No identifier type mapped cleanly to UniProt IDs, so in cases where more than one UniProt accession was found, only the reviewed UniProt accession was taken. If there were multiple reviewed UniProt accessions found, the UniProt accession with the longest protein sequence was selected. If no reviewed UniProt accession was found, the unreviewed UniProt accession with the longest sequence was taken. This was done to maintain a one-to-one ratio of alternate identifiers to UniProt accessions. Because some resources annotate pathways with gene level identifiers, we remove annotated isoforms from other resources in order to maintain a fair comparison.

Once the list of UniProt accessions were obtained for each resource, each was updated to the latest version of UniProt (Jan. 2021). In this instance, 'update' refers to converting all secondary accessions to primary accessions, removing all non-human identifiers and removing identifiers that are classified as obsolete in UniProt.

Phosphosites

In Reactome and PhosphoSitePlus, the phosphosites were extracted from the neo4j graph databases generated for extracting the UniProt accessions mentioned above. For qPhos, the set of all phosphosites were extracted from the main data file downloaded. Phosphosites from HPRD were extracted from the separate PTM file and connected to UniProt IDs converted from RefSeq IDs with only phosphosites belonging to the signalling portion of the database taken. BioGRID's phosphosites were also extracted from a separate PTM file and connected to UniProt IDs converted from RefSeq IDs only from the signalling portion of the database. For SIGNOR, the human PTM file was downloaded directly from the website and processed. Phosphosites from UniProt were extracted directly from the January 2021 downloaded TSV file. Only phosphosites from reviewed UniProt accessions were used in this study. Upon further investigation, phosphorylations listed in KEGG had no specific site annotations. WikiPathways contained multiple phosphosite annotations in multiple formats, making it impossible to extract them all. As with above, since each phosphosite was associated with a single UniProt accession, the UniProt accessions were updated to reflect only human and current UniProt accessions and isoforms were removed.

Consistency analysis

To perform the consistency analysis the neo4j graph database version of Reactome and PhosphoSitePlus were used. For SIGNOR, HPRD and BioGRID the list of unique phosphosites retrieved from each respective database file was used. In each database, all PTM's belonging to a unique UniProt id were gathered into a set. A phosphosite in this manuscript is defined as a single amino acid reported to be phosphorylated and its position as reported in the database. We do not take into surrounding sequence or attempt to standardise or harmonize this information across databases beyond what is described in the methods. Then, for each UniProt ID the canonical sequence and all isoform sequences

were retrieved from UniProt. Next, for each PTM annotation, if the amino acid at the annotated position was the same in the canonical UniProt sequence it was called a match; for example, if Reactome contains an annotation on Protein X that indicates a phosphorylation on a threonine at location 128 and UniProt has a threonine at position 128 in its sequence this is a match. If, however, the annotated residue and the canonical residue didn't match but another phosphorylate-able residue was found there instead, this was called a mismatch; for example, the Reactome annotation on Protein X indicates a phosphorylation on a threonine at location 362 whereas UniProt has a tyrosine at position 362. Finally, if the annotated residue was neither a match or a mismatch, the sequence was searched to find the closest exact match and the offset recorded. Once again using the example above, in Reactome another annotation on Protein X indicates a phosphorylation on a serine at position 495, whereas UniProt lists the closest serine to that location is actually at position 490; this would be counted as an offset of 5. It should be noted that if there was no match to the canonical residue all isoforms were then searched for the same criteria (match, mismatch and offset). Annotated modifications where the position is larger than the canonical sequence length, or where there is no position annotated at all, are considered to have not matched and no mismatch or offset is searched for.

Experimental data

We used phosphoproteomics data from 191 publications across 484 conditions obtained downloaded from the qPhos repository to assess the coverage of phosphoproteomics data by Reactome, PhosphoSitePlus, HPRD, BioGRID and SIGNOR. Data was downloaded from the qPhos download portal in January 2019. Each modification in qPhos was associated with a PubMed ID (PMID) and each PMID was tagged with metadata, such as the 'sample' (cell line) or condition the experiments were performed in. We subset the phosphosites by cell line and tissue type. Only phosphosites (and inherently PMID's) tagged with a single 'sample' type (such as MCF-7 or HeLa) were analysed (in cases where there were multiple conditions it was impossible to discern which modification came from which sample).

Software and data availability

The source code and precompiled binaries are available from github (https://github.com/DavisLaboratory/Db_Compare). The software is compatible with Linux and MacOS operating systems and has been tested on both. The software requires Java version 8, Python 3 and R 3.6. Installation of these and dependencies is managed through the Anaconda package management system; a full list is available in the github repository along with installation instructions.

RESULTS

What proportion of the proteome is covered by different pathway databases?

The full extent of human phosphorylation is unknown and is an active field of research today. One study focused on profiling a single cell line identified approximately 75% of

proteins to be phosphorylated (*Sharma et al., 2014*). Yet, many phosphoproteomics studies perform pathway level analysis at the protein level. This is due to the fact that many databases do not offer phosphosite level mapping. This motivated us to first determine the proteome coverage (and by proxy the coverage of proteins which may potentially be phosphorylated) by each database. This was done in order to provide context around their phosphoproteomic coverage as well as to assess their individual and combined potential for analyzing phosphoproteomic datasets. To do this, Reactome, WikiPathways, KEGG, HPRD, BioGRID, SIGNOR and PhosphoSitePlus were downloaded and their protein lists extracted. We normalised all protein UniProt identifiers to the January 2021 version of UniProt as explained in the methods. Because our focus is on resources that capture signalling reactions, throughout the analysis we used the subset of PhosphoSitePlus that contained kinase-substrate pairs (14,159 pairs) (henceforth referred to as PhosphoSitePlus (K-S) and PSP (K-S) in figures) and did not compare to phosphorylation events with no known kinase (231,645 events).

To characterize each database's coverage of the proteome, a collection of reference knowledgebases and datasets were chosen for comparison (Fig. 1A). This set included qPhos, IMEx (*Orchard et al., 2012*), the Gene Ontology Cell Signalling Category (GO:0023052), the Gene Ontology Protein Kinase Activity Category (GO:0004672) (*Ashburner et al., 2000*) and the subset of UniProt annotated with the term 'Phosphoprotein' (KW-0597). The intersections between each reference dataset and each signalling database were then taken. Intersections with qPhos are expected to represent each knowledgebase's coverage of phosphorylate-able proteins, while the intersections with IMEX are expected to represent each knowledgebase's coverage of global interacting proteins, many of which have potential to be phosphorylated. We found BioGRID to have the highest coverage of each reference dataset and PhosphoSitePlus (K-S) had the lowest (barring the GO Protein Kinase Activity category where it had the 3rd highest coverage). Also of interest is the pattern of decreasing coverage of the all reference datasets across the signalling databases excepting the GO Protein Kinase Activity category. We see that BioGRID contains the most known human kinases, followed by SIGNOR, PhosphoSitePlus (K-S), WikiPathways, HPRD, Reactome and finally KEGG. Taken together these results point to a relative ranking for the analysis of data when mapping on the protein level, where you can expect BioGRID to have a high coverage of a given dataset, followed by Reactome, HPRD, KEGG, WikiPathways, SIGNOR and PhosphoSitePlus (K-S) but a differing coverage of the known human kinome. The same order was found when looking at each database's coverage of SwissProt as well at the entirety of UniProt (Fig. S3, Supporting Information).

To explore the distinctions between each database's proteome coverage we compared the database protein sets with each other using an UpSet plot (*Conway, Lex & Gehlenborg, 2017*). Overlaps of less than 50 were excluded, the entire UpSet plot can be found in Fig. S4, Supporting information. Examining the proteins common to all databases, we see the global intersection is just 5% of the total 19,816 unique proteins held across them (Fig. 1C), indicating a marginal global overlap. Individually, BioGRID consistently has the highest overlap with all other databases (Fig. 1B). Although the global intersection
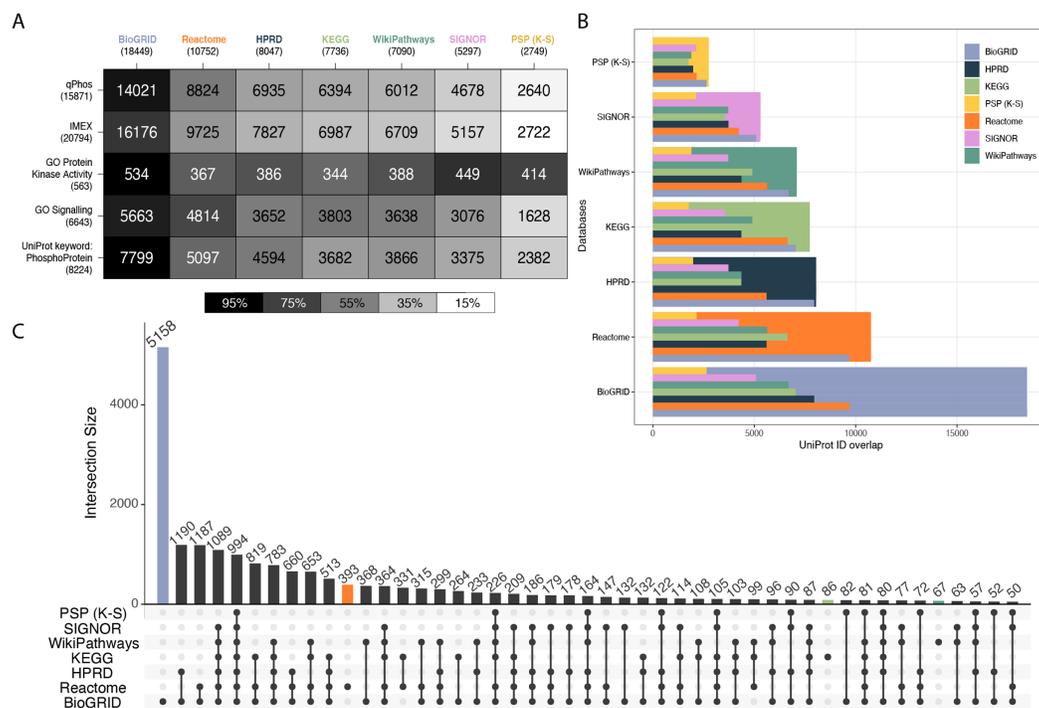
Figure 1 Cross database comparison of proteins. (A) A heatmap of proteins comparing knowledge-bases with reference datasets, where each cell is a pairwise intersection, the number represents the number of proteins found in both the reference dataset (vertical axis) and in the signalling database (horizontal axis); the total number of proteins per database is listed under the database name and the cell shade represents the proportion of coverage of each reference dataset (vertical axis); the darker the colour the higher the coverage of that dataset by its paired signalling database. (B) Each pairwise protein level database intersection displayed where the largest bar represents the total number of proteins in that respective database and the inset bars capture the number of overlapping entities from each resource. (C) An UpSet plot depicting intersections of proteins between the databases Reactome, HPRD, KEGG, WikiPathways, PhosphoSitePlus, BioGRID and SIGNOR; vertical bars indicate the number of UniProt ids found in the intersection of the databases listed with a black dot underneath. Coloured bars capture the number of proteins unique to a resource.

is small, the proportion of each database that is unique to itself ranges from only 2.5–20% (Fig. S4). We see that once again BioGRID consistently has the highest coverage of the other databases followed by Reactome in every database. WikiPathways and KEGG appear to have similar coverage of each other database. Finally, SIGNOR consistently has the 6th highest coverage (in all databases excepting PhosphoSitePlus (K-S)) and PhosphoSitePlus (K-S) the lowest coverage. The balance of proteins found in other databases vs. found only in a single database is an indication of potential information gain and novel insight when choosing which database and type to use. The set differences of each pair were also investigated to understand the scope of uniqueness of each reference dataset and signalling knowledgebase (Fig. S5, Supporting Information). Lastly, of note, PhosphoSitePlus (K-S) always maintains the lowest coverage of the other databases yet is the highest covered itself. This comparison is important as PhosphoSitePlus contains the most comprehensive known K-S network and each database's relative coverage provides further evidence to its ability to connect phosphorylation signalling to a greater context.

As a potential consequence of the differing curation practices each database has adhered to, each database has tended to focus on a different area of biology. Of the number of proteins unique to each database, BioGRID had the most at 5,158, followed by Reactome with 393, KEGG and WikiPathways with 86 and 67 respectively, and then SIGNOR, HPRD and PhosphoSitePlus (K-S) all with less than 25 (Fig. 1C, coloured bars, Fig. S4 coloured bars). This alludes to the potential impact of database choice on conclusions drawn when performing pathway enrichment analysis. To explore this further, we performed an over enrichment analysis using the GO Molecular Function category on the proteins found to be unique to each database (Fig. S6, Supporting Information). We found that the proteins unique to Reactome were mostly involved in transporter activity. This could be because it is the only database in our comparison with abundant subcellular location annotations. BioGRID's unique proteins were enriched for demethylase activity. WikiPathways was enriched for receptor activity. HPRD, SIGNOR, PhosphoSitePlus (K-S) and KEGG's unique proteins did not produce any enrichment.

What proportion of the phospho-proteome is covered by different pathway databases?

We sought to address the question of the level of phosphoproteomic coverage of each database. None of the included databases (excepting PhosphoSitePlus (K-S)) have the ability to input a list of phosphosites and have them mapped on the phosphoproteomic level. To do this, all phosphorylation annotations were taken from each database where possible; neither KEGG nor WikiPathways were included in this section of the analysis as modification annotations in these databases are inaccessible.

In the first level of database characterization of coverage of the phosphoproteome, the databases of interest were compared to qPhos' phosphorylation annotations as well as UniProt's. qPhos contains 39,652 less phosphorylation annotations than the entirety of PhosphoSitePlus. However, qPhos was chosen as the authority on experimentally derived phosphorylations in this analysis, to avoid the self-comparison that would result from using a component of PhosphoSitePlus in evaluating its own utility. Pairwise Intersections were taken between each signalling knowledgebase and each reference dataset (Fig. 2A). Once again, we see a relatively consistent ranking of the databases, but in a differing order than the protein level. Here, HPRD has the highest coverage of qPhos' phosphorylations, followed by PhosphoSitePlus (K-S), BioGRID, SIGNOR and then Reactome. A slightly different pattern occurs in the coverage of UniProt phosphorylations where the pattern follows the number of phosphosites in each database. Although HPRD has the highest number of phosphorylations annotated in the database category (with 43% of its phosphosites also recorded in UniProt) its coverage only represents 33% of all phosphosites recorded in UniProt. qPhos has almost $5\times$ more phosphorylation annotations than UniProt, and we see that HPRD has the highest overlap with 81% of its phosphorylation events found in qPhos, but still only representing 13% of qPhos' total. Next is PhosphoSitePlus (K-S), SIGNOR and then BioGRID with 52%, 54% and 77% of their phosphosites respectively in common with qPhos. Yet, each database only covers a small proportion of qPhos at less than 3% for all three. Reactome has the smallest coverage
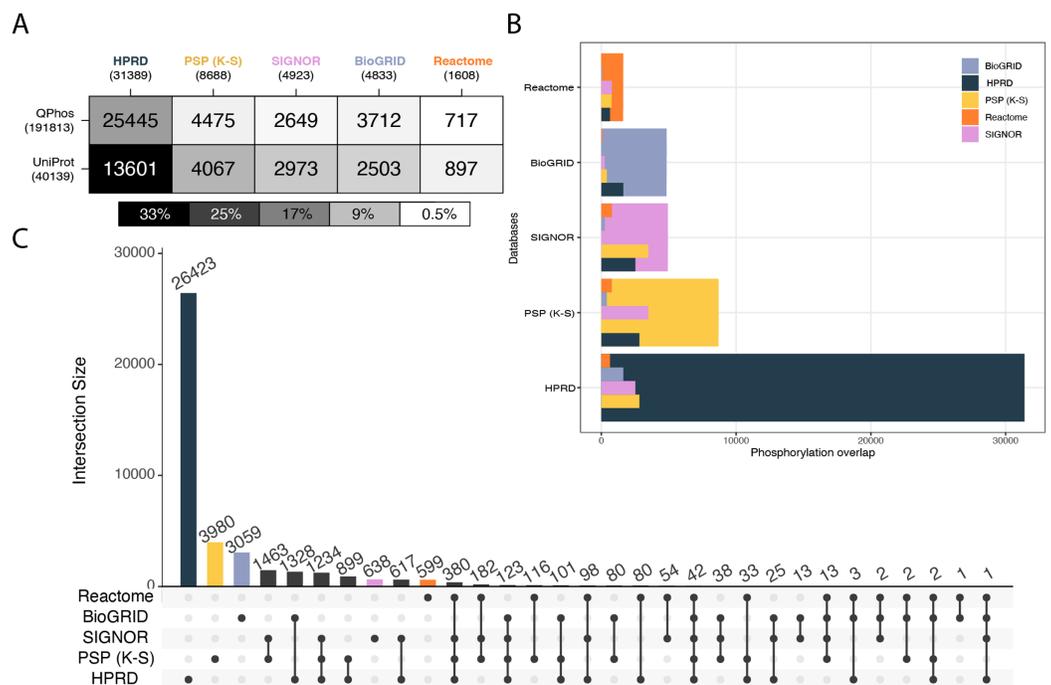
Figure 2 Cross database comparison of phosphorylation annotations. (A) A heatmap where each cell is a pairwise intersection, the number represents the phosphorylations annotated in both the reference dataset (vertical axis) and in the signalling database (horizontal axis); total number of phosphorylations per database is listed under the database name; cell shade represents the proportion of coverage of each reference dataset, the darker the colour the higher the coverage of that dataset by its paired signalling database. (B) Each pairwise phosphorylation level database intersection displayed where the largest bar represents the total number of phosphorylations in that respective database and the inset bars capture the number of overlapping entities from each resource. (C) An UpSet plot depicting all intersections of phosphorylations annotated in BioGRID, HPRD, PhosphoSitePlus, SIGNOR and Reactome; vertical bars indicate the number of phosphorylation annotations found in the intersection of the databases listed with a black dot underneath. Coloured bars capture the number of proteins unique to a resource.

of qPhos at 0.3% but this accounts for 46% of its own database. This means that approximately half of each of the latter 4 databases or more appear in global experimental datasets, but only a tiny proportion of the global experimental data can be interrogated through these knowledgebases. Taken together, this points to the disproportionate amount of data available and the need for tools to assign functional annotations. Added to the knowledge that only 20% of kinases are associated with 87% of currently annotated phosphosites (*Needham et al., 2019*), there is clear need to start to explore lesser known kinases and their contribution to cellular regulation.

Unlike on the proteomic level we see is no clear pattern in the overlap between databases (Fig. 2B). We see that the levels of uniqueness are much higher. The varying levels of overlap once more indicate the potential new information to be gained when mapping to a database. Keeping in mind that the subset of phosphosites used here from PhosphoSitePlus (K-S) all have a kinase or phosphatase linked to the feature. Though mapping events at the phosphorylation level is not yet supported by these databases, manual inspection of proteins identified in a study may reveal the presence of particular phosphorylation events

of interest and their associated reactions, which may be then further explored. This analysis (Fig. 2) can be used as a guide to determining relative likelihood of finding a phosphorylation of interest in each database and the likely level of information that will be associated with it.

When looking at the inter-database coverage we see that the number of phosphosites found in all 5 databases is only 42, a mere 0.1% of the total 41,629 unique phosphorylation annotations across these databases (Fig. 2C) meaning each database covers a very different portion of the phosphoproteome. HPRD has 84% of its phosphosites unique to itself while PhosphoSitePlus (K-S) has 46%, followed closely by BioGRID at 63% Reactome at 37% and lastly SIGNOR at 13%. This is a clear indication of the vastness of the phosphoproteome and further the importance of database choice when it comes to analysing phosphoproteomic data. When looking at the number of unique phosphorylations belonging to each database, HPRD possesses the most at 26,423, PhosphoSitePlus (K-S) at 3,980, followed by BioGRID at 3,059, SIGNOR with 638 and Reactome with 599. The large number of unique phosphosites found in HPRD indicate its dedication to the curation of novel phosphosites (Goel et al., 2012). The set differences of each pair were also investigated on the phosphorylation level to understand the scope of uniqueness of each reference dataset and signalling knowledgebase (Fig. S7, Supporting Information).

Exploring the consistency between annotated phosphorylated residues and UniProt sequences

A major issue often found in phosphoproteomics is the incorrect assignment of substrates to kinases. This phenomenon is detailed by Humphrey et al. where they described that the correction of low-quality annotations is more difficult than an initial assignment due to the amount of evidence needed to refute prior knowledge. A major compounding factor to this problem is the lack of consistency in substrate phosphorylation annotations across databases. These mis-annotations can often be propagated by both human curators as well as text mining systems if left unchecked, contributing to the incorrect kinase assignments (Bachman, Gyori & Sorger, 2019). To explore this aspect of the problem we decided to test the consistency, or accuracy of the identity of annotated phosphorylated residues against canonical protein sequences found in UniProt.
For example, if a database recorded a protein to have a phosphorylation of a serine on residue 456, we wanted to know if a serine would actually be found in the canonical (or isoform) protein sequence at position 456 (described in more detail in methods). Only phosphorylation sites annotated on the canonical protein sequence of a gene were included in Table 1. A match in an isoform was only searched for if a match was not found in the canonical sequence. Further, although a number of phosphorylations have been found on Histidine (Fuhs & Hunter, 2017), here we limited phosphorylate-able residues to Serine, Tyrosine and Threonine.

Mismatched sites can typically occur because their positions were annotated on post-translationally cleaved proteins, non-canonical isoforms, or on non-human

Table 1 Comparing annotated modification position consistency. Statistics of the modified residues annotated in each resource. Number of Modified residues corresponds to the number of unique phosphorylated residues in a protein sequence (only a representative protein per gene is analysed). Percent of consistent residues indicates the percent of annotated modified residues that were consistent with the amino acid residue listed in UniProt at that location. Isoform matches refer to the number of modified residues where the annotated residue was consistent with a protein isoform sequence rather than the canonical protein sequence. Mismatch residue count contains the number of phosphorylated sites where a different phosphorylate-able residue was found in the UniProt sequence rather than the one listed in the named database. The closest exact match offset is the number of modified amino acids found at an offset to the originally annotated modified position.

	qPhos		PSP (K-S)		PSP-full		HPRD		BioGRID		Reactome		SIGNOR	
Number of modified Residues	191,813		8,688		231,465		31,385		4,833		1,608		4,923	
Percent of consistent residues*	189,928 (99.0%)		8,656 (99.6%)		230,694 (99.7%)		30,885 (98.4%)		4,272 (89.5%)		1,250 (77.7%)		4,840 (98.5%)	
Isoform matches	657		3		260		202		151		37		33	
Mismatch residue count	192		2		64		43		47		44		7	
Closest exact match offset	1	424	1	26	1	186	1	144	1	97	1	36	1	14
	2	248	2	2	2	102	2	81	2	64	2	23	2	13
	3	129	3	1	3	63	3	36	3	42	3	14	3	2
	+	792	+	1	+	336	+	152	+	240	+	134	+	36

Note:
* Note the percent of consistent residues does not include isoform matches.

orthologues (*Bachman, Gyori & Sorger, 2019*). It is likely that an offset of −1 is due to initial methionine cleavage. Offsets of larger size and differing direction are likely to be caused by other issues. The closest matching residue may not be the modified residue. In general, resolving inconsistencies like these will require individual investigation, however automated approaches have been developed that may address the most common sources (*Bachman, Gyori & Sorger, 2019*). Alternatively, allowing for small offsets (<2) may improve mapping by accounting for slight inconsistencies in annotations (*Babur et al., 2018*). We report all offsets here for completeness.

We found that the entirety of the PhosphoSitePlus database had the highest sequence consistency with UniProt canonical sequences at 99.7%. We found 260 cases where a match was found in an isoform rather than the canonical sequence, mostly owing to discrepancies with UniProt nomenclature (Table 1). In addition, we found 64 cases where the wrong phosphorylate-able residue was found, labelled a mismatch. Further, looking at PhosphoSitePlus' K-S subnetwork, we see that it is almost the same at 99.6% consistent. We discovered only three phosphosites were found in isoforms as well as only 26 cases where the annotated residue was found next to the annotated location (i.e. offset by one residue). Next looking at qPhos we see a 99% percent consistency followed by SIGNOR with 98.5% consistency. So far each of these databases has shown exceptional curation practices in gathering phosphoproteomic data. Turning to HPRD we found overall phosphosite consistency to be 98.4% in relation to UniProt canonical sequences, followed by BioGRID with a phosphosite consistency of only 89.5%. Finally, we see Reactome which is 77.7% consistent. This lower consistency may be attributed to the manual curation practices of Reactome, or imprecise recording of isoforms in the literature curated for the database.

Figure 3 **Heatmap of proteomic coverage by databases.** A heatmap where each cell represents the pairwise intersection of proteins between the pathway databases being analysed (listed on the horizontal axis), and all proteins found in qPhos that had been tagged with the corresponding cell line (listed on the vertical axis); total number of proteins per database are listed under the database name; shade of each cell refers to the proportion of proteins per cell line that intersect with each pathway knowledgebase.

How well do databases cover experimental phosphoproteomic data?

Using phosphoproteomics to interrogate how cells sense and respond to their environment can lead to important insights into biological mechanisms behind these changes. Searching 'phosphoproteomics' in PubMed reveals over 3,500 results as well as depicting a steady rise in publications related to the search term over the last 20 years with 456 new publications in 2019 alone (Fig. S8, Supporting Information). As a result of the rise in interest and potential of phosphoproteomics, a number of databases have attempted to assemble the vast amounts of phosphoproteomic data being generated. As seen above, qPhos is one such database that provides curated phosphoproteomics datasets across 191 publications under a wide range of conditions. In this analysis we compared data from these cells with each knowledgebase to establish what proportion of the measurements from each cell type would map onto the knowledge contained in pathway databases. To pose the question another way, we sought to establish the coverage of pathway databases over phosphoproteomics experiments. Here we present results for the top six cell lines, ranked by the number of publications archived in qPhos (Table S3, Supporting Information). The full results are presented in Tables S4 and S5, Supporting Information.

Consistent with the coverage analysis we performed against the full proteome and the phosphoproteome, we see the same relative ordering in the databases in terms of coverage over experimental data. When compared at the protein level BioGRID has the highest coverage, followed by Reactome, HPRD, KEGG, WikiPathways, SIGNOR and lastly PhosphoSitePlus (K-S) (Fig. 3). Noting the order in which the cell lines are covered by each database, we see that coverage is highest for HELA cells across all databases, followed by HEK293, HeLa S3, Jurkat, MCF-7 and finally MCF-10. This order follows the number

	BioGRID (4833)	Reactome (1608)	HPRD (31389)	SIGNOR (4923)	PSP (K-S) (8688)
qPhos HeLa (65956)	2819	354	16645	1509	2626
qPhos HEK293 (30820)	2366	273	9659	1023	1663
qPhos HeLa S3 (28944)	1970	243	14442	995	1699
qPhos Jurkat (27147)	2082	320	9496	1071	1827
qPhos MCF-7 (24567)	2398	248	9404	1005	1578
qPhos MCF-10A (8666)	1068	166	3797	525	757

50%	38%	25%	13%	0.5%

Figure 4 Heatmap of phosphosite coverage by databases. A heatmap where each cell is a pairwise intersection, the number represents the phosphorylations between the pathway databases being analysed (listed on the horizontal axis), and all phosphorylations found in qPhos that had been tagged with the corresponding cell line (listed on the vertical axis); total number of phosphorylations per database is listed under the database name; the shade of each cell refers to the proportion of phosphorylations per cell line that intersects with each pathway knowledgebases annotated phosphorylation.

of proteins identified in each cell line in a decreasing order, which likely explains the pattern.

When comparing coverage of individual phosphorylation events, we see that HPRD has the highest coverage, followed by, BioGRID, PhosphoSitePlus (K-S), SIGNOR and finally Reactome (Fig. 4). In contrast to findings on the proteomic level, when looking at the phosphorylation level data we see a different pattern emerging in the cell line coverage by each database, where once again HeLa is the highest covered but Jurkat, HeLa S3, HEK293 and MCF-7 do not remain consistent and MCF-10 is the lowest covered (Fig. 4). Interestingly this pattern does not follow the number of phosphorylations per cell line. This may be due to the curation practices of each database, potentially highlighting underlying biases towards different common cell lines in experimental data from which pathway knowledge is drawn. These results also demonstrate that the proportion of measured phosphosites for which reaction level information is available is relatively small. For the most commonly measured cell line, HeLa, less than 1% of phosphosites map to annotations in Reactome and attach to known signalling pathways. This small number none-the-less corresponds to 22% of the phosphosites in Reactome. While HeLa phosphosite mapping to PhosphoSitePlus (K-S) is higher at around 4%, this knowledge doesn't present the same level of mechanistic insight as is captured in Reactome, with functional information limited to K-S interactions. These results (Fig. 4) demonstrate that the vast majority of measured phosphosites lack pathway level annotation.

The contrast in coverage on these two levels by the same databases depict the consequences of generalizing phosphoproteomic level data to the protein level. Global coverage of qPhos can be compared in Figs. 1A and 2A and is reiterated in Figs. 3 and 4's side-by-side comparison. Through this comparison we see that HPRD has a good coverage of phosphorylate-able proteins and a high coverage of phosphorylations with 56% of its

proteins possessing a phosphorylation annotation. Conversely, BioGRID has a high coverage of phosphorylate-able proteins but many fewer phosphorylations when comparing a PPI of similar size (HPRD) with only 19% of its proteins recording a phosphorylation annotation. Reactome has a high level of phosphorylate-able proteins but a low number of phosphorylations translating to 5% of its proteins holding a phosphorylation annotation, while in PhosphoSitePlus (K-S) we see a low number of phosphorylate-able proteins but a high number of phosphorylations (96% of proteins with a phosphorylation annotation). SIGNOR, finally, appears to be on the lower end of proteins but possess a good number of phosphorylation annotations and has 33% of its proteins holding a phosphorylation annotation. This variation in the typical concentration of phosphorylations per protein emphasises the importance of database choice. It also highlights the different strengths of each database should for analyse phosphoproteomics data, as well as the biases that might result from database choice.

DISCUSSION

We conducted a comprehensive analysis to systematically compare seven widely used knowledgebases, comparing the extent of their coverage over proteome, phosphoproteome and experimental phosphoproteomic data. Recent studies have highlighted the disparity between the amount of phosphosites discovered and the number with known functionality and their impact on cellular signalling (*Needham et al., 2019*). Here we have attempted to provide novel insights into the relative strengths and weakness of popular literature derived signalling knowledgebases and their comparative suitability for the analysis of phosphoproteomic data.

We found that even though HPRD has not been updated since 2010, it maintains a good coverage of proteins and phosphosites identified in proteomics experiments, as well as a high level of accuracy in the annotated sequence positions of the residues on which phosphorylations are observed. However, as this database is no longer updated it will likely continue to decline in relevance as phosphoproteomic data improves and is disseminated across public databases such as UniProt. The downloadable Protein-Protein Interaction network from HPRD, has a high quantity of proteins and potential interactions, but due to the high-throughput nature of PPI's, the quality of each datapoint is much lower, making the task of functionally annotating phosphorylations much more difficult. However, the HPRD website contains many phosphosite specific upstream kinases, isoforms as well as many validated low-throughput interactions making it a highly valuable resource still today. HPRD is a great candidate to be used in phosphoproteomic data analysis as a means of hypotheses generation that can be explored further using a more-detailed signalling knowledgebase. However, as previously stated it is no longer actively maintained, and its usefulness is expected to decline over time.

BioGRID had the highest proteomic coverage of the databases assessed but has relatively low phosphoproteome coverage. It also features the second lowest level of consistency of phosphosites with UniProt. BioGRID is actively maintained and being expanded, but in its current form these issues limit its utility in analysing phosphoproteomics data compared to HPRD.

Reactome on the other hand is highly curated with a good coverage of proteomic data, but a much lower coverage of phosphorylation level data especially when combined with a lower consistency between PTM locations and UniProt sequences. Reactome's high coverage of proteomic data indicates a high mapping potential of phosphorylate-able proteins and although this does not translate to a high mapping potential of phosphorylation sites, it is an important starting point. The few phosphorylation events recorded in Reactome are rich with information. Reactome is the only database in this study to represent proteins in specific phosphorylation states as different entities with differing behaviours, functions and interactions as a direct result of the number and combination of phosphorylations attached. This level of information is critical in gaining functional insights as we know that proteins can be multiply phosphorylated and these phosphorylations can in fact have contradicting or additive effects on a protein's function (*Olsen et al., 2006*). However, due to Reactome's low level of phosphorylation annotations it is not yet comprehensive enough to provide insights to a global phosphoproteomics dataset for more than a very small proportion of the measurements collected in an experiment (see Figs. 2 and 4).

SIGNOR is a database that stores pairwise relationships between biological entities, describing the behaviour of proteins in response to each other (*Perfetto et al., 2016*). Relationships describe up- or down-regulation with information about the magnitude and mechanism of change. SIGNOR describes 5,297 proteins and 4,923 human phosphorylation annotations, intermediate between a typical PPI database and Reactome. SIGNOR has the highest coverage of phosphosites in PhosphoSitePlus (K-S), indicative of its utility for exploring K-S information in the context of interactions. Consistency with UniProt is high, and the associated information regarding regulators of each individual phosphosite is particularly valuable.

PhosphoSitePlus has a vast amount of information directly relating to phosphoproteomics across species, with most data from human and mouse. It also clearly demonstrates the overload of phosphosites known, in comparison to the amount of phosphosites we biologically understand. Only a subset of this database has the ability to provide mechanistic insights with respect to the kinases likely to be responsible for phosphorylation events. However, the phosphorylations belonging to those proteins were well characterized and very consistent. Like Reactome these phosphosites have a wealth of information attached to them such as regulatory roles and associations with diseases making PhosphoSitePlus the best available resource for inferring kinase activity. However, as this is a specialist database, there is a clear lack of functional information outside of protein modifications, making PhosphoSitePlus a useful tool for database mining but less useful in untangling signalling transduction implications outside of the phosphorylation context.

KEGG and WikiPathways were both found to have a moderate coverage of the proteome but could not be investigated in depth for their use in phosphoproteomics data analysis due to the absence of accessible phosphosite-specific annotation. Nonetheless, today both databases continue to be used in pathway analysis of phosphoproteomic data,
largely as a source of gene sets for enrichment type studies (*Abe et al., 2020*; *Park et al., 2015*; *Kwon et al., 2015*; *Kawata et al., 2019*).

It can be seen in Figs. 3 and 4 that the successful mapping of a phosphoprotein does not necessarily translate to the successful mapping of a phosphosite. In some cases, the protein maps to the knowledgebase, but the knowledgebase has no recorded phosphorylations on that protein, or potentially only a single phosphorylation when a different phosphorylation or even multiple were found on that protein experimentally. The implications of generalizing a phosphoprotein to any unphosphorylated version should be carefully considered. An unphosphorylated protein and protein phosphorylated in a specific manner will most likely behave differently.

When it comes to analysing phosphoproteomic data using general signalling knowledgebases, other factors should also be considered. Firstly, how well does the database cover other types of PTM? PTMs such as ubiquitination and acetylation can have strong cooperative or contradictory effects. For example, an association was found between phosphorylation and acetylation in a genome-reduced *Mycoplasma pneumoniae* (*Van Noort et al., 2012*) demonstrating that the presence of one modification might facilitate or hinder modification of a closely positioned residue. Another important consideration is the coverage of protein complexes, as they play crucial roles in signal transduction and their formation is often regulated by phosphorylation (*Day, Sosale & Lazzara, 2016*). Finally, isoforms, which expand the complexity of the proteome, but also present problems for annotation, need to be considered. As shown in Table 1 many modifications are found on non-canonical (alternatively spliced) protein sequences. Isoform specific phosphorylation should be recorded, and the resulting modified isoforms treated as separate entities to their canonical form as they can have a large impact on molecular functions. For example, one study found that Connexin isoform 30 closed upon Protein Kinase C phosphorylation but not connexin isoform 43, completely altering its molecular function (*Alstrøm et al., 2015*). Of the databases characterized above, Reactome has demonstrated all of these properties making it a valuable starting point for phosphoproteomics functional analysis. While PhosphoSitePlus treats isoforms independently and maintains a repository of PTM's other than phosphorylation, its kinase-substrate network treats each phosphosite independently, and does not hold any information for multiply phosphorylated proteins nor complexes. HPRD does hold isoforms as well as alternate modifications but in most cases annotates protein–protein interactions as common to all isoforms and modified proteins (*Keshava Prasad et al., 2009*). Similarly, BioGRID also holds alternate modifications as well as isoforms but does not treat them as separate entities. KEGG holds complexes, and some isoform information (but does not seem to treat them separately), as well as alternate modifications but does not provide the modified residue. Finally, at the time of publication WikiPathways did not have a downloadable BioPax (*Demir et al., 2010*) file and was unable to be investigated, however it does contain many Reactome pathways which do possess all of the above qualities.

Data saturation is occurring and there is a huge need for bioinformatic analysis to aid in the characterisation of phosphosite function. The use of general signalling

knowledgebases has great potential to aid in this endeavour through data browsing, mining and hypothesis generation. No single database analysed in this study emerged as a 'best' option, as each had clear strengths and weaknesses. Instead for now these databases should be used in conjunction with each-other to build on the assets that each maintains and makes up for in the others.

CONCLUSION

We have used global phosphoproteomics datasets to provide insights into the relative strengths and weaknesses across databases and highlighted general concerns in current phosphoproteomics analysis practices. We found that phosphorylation events in Reactome are rare, but typically come with a subcellular location, as well as interpretable functional annotations for a protein in a specific phosphorylation state (whether that be singly or multi-phosphorylated) leading to potential mechanistic insights. In HPRD phosphorylation events are much more abundant but the downloadable PPI has little information associated with phosphosites. BioGRID possessed few annotated phosphorylation events, meaning it is unlikely that mapping phosphoproteomics data onto this database will yield a high level of successful matches and ultimately helpful insights into phosphoproteomic signalling. SIGNOR on the other hand possessed less phosphorylation annotations but each is highly consistent and holds many other valuable annotations. Many known phosphorylation events can be found in the PhosphoSitePlus database, but most of these do not have K-S annotation. Further, PhosphoSitePlus lacks information on cellular signalling outside the K-S context resulting in limited potential for specific mechanistic insights. The coverage, consistency, number of proteins and number of phosphorylations, as well as the level of detail in each database's annotations differ greatly. All aspects are important to characterize in order to get a holistic understanding of how each database is capable of characterizing phosphoproteomics data.

ACKNOWLEDGEMENTS

The authors thank Daniel Esposito, Samuel Lee and Dharmesh Bhuva for technical support.

Funding

This work was supported by the National Health and Medical Research Council funding (Project Grant 1128609 to Melissa J. Davis), Cancer Council Victoria funding (Project grant 1187825 to MJD), and the National Breast Cancer Foundation and Cure Brain Cancer Foundation funding (Project Grant CBCNBCF-19-009 to MJD). Melissa J. Davis is the recipient of the Betty Smyth Centenary Fellowship in Bioinformatics. HAH was supported by the Peter Hall Scholarship. This study was made possible through Victorian State Government Operational Infrastructure Support and Australian Government

NHMRC Independent Research Institute Infrastructure Support scheme. Liam G. Fearnley was supported by the DHB Foundation Centenary Postdoctoral Fellowship in Neurogenetic Systems Biology and philanthropic funding provided through the Walter and Eliza Hall Institute. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Grant Disclosures

The following grant information was disclosed by the authors:
National Health and Medical Research Council: 1128609.
Cancer Council Victoria: 1187825.
National Breast Cancer Foundation and Cure Brain Cancer Foundation: CBCNBCF-19-009.
Betty Smyth Centenary Fellowship.
Peter Hall Scholarship.
Victorian State Government Operational Infrastructure Support and Australian Government NHMRC Independent Research Institute Infrastructure Support scheme.
Walter and Eliza Hall Institute.

Competing Interests

Melissa J. Davis is an Academic Editor for PeerJ.

Author Contributions

- Hannah Huckstep conceived and designed the experiments, performed the experiments, analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the paper, and approved the final draft.
- Liam G. Fearnley conceived and designed the experiments, authored or reviewed drafts of the paper, and approved the final draft.
- Melissa J. Davis conceived and designed the experiments, authored or reviewed drafts of the paper, and approved the final draft.

Data Availability

The source code and precompiled binaries are available at GitHub: https://github.com/DavisLaboratory/Db_Compare.

The software is compatible with Linux and MacOS operating systems and has been tested on both. The software requires Java version 8, Python 3, and R 3.6. Installation of these and dependencies is managed through the Anaconda package management system; a full list is available in the GitHub repository along with installation instructions.

Supplemental Information

Supplemental information for this article can be found online at http://dx.doi.org/10.7717/peerj.11298#supplemental-information.

REFERENCES

Abe Y, Hirano H, Shoji H, Tada A, Isoyama J, Kakudo A, Gunji D, Honda K, Boku N, Adachi J, Tomonaga T. 2020. Comprehensive characterization of the phosphoproteome of gastric cancer from endoscopic biopsy specimens. *Theranostics* **10(5)**:2115–2129 DOI 10.7150/thno.37623.

Alstrøm JS, Hansen DB, Nielsen MS, MacAulay N. 2015. Isoform-specific phosphorylation-dependent regulation of connexin hemichannels. *Journal of Neurophysiology* **114(5)**:3014–3022 DOI 10.1152/jn.00575.2015.

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G. 2000. Gene ontology: tool for the unification of biology. *Nature Genetics* **25(1)**:25–29 DOI 10.1038/75556.

Babur Ö, Luna A, Korkut A, Durupinar F, Siper MC, Dogrusoz U, Aslan JE, Sander C, Demir E. 2018. Causal interactions from proteomic profiles: molecular data meets pathway knowledge. **9**:e1003290 DOI 10.1101/258855.

Bachman J, Gyori B, Sorger P. 2019. Assembling a phosphoproteomic knowledge base using protmapper to normalize phosphosite information from databases and text mining. *BioRxiv* DOI 10.1101/822668.

Bajpai AK, Davuluri S, Tiwary K, Narayanan S, Oguru S, Basavaraju K, Dayalan D, Thirumurugan K, Acharya KK. 2020. Systematic comparison of the protein–protein interaction databases from a user's perspective. *Journal of Biomedical Informatics* **103(Suppl. 5)**:103380 DOI 10.1016/j.jbi.2020.103380.

Bateman A, Martin MJ, O'Donovan C, Magrane M, Alpi E, Antunes R, Bely B, Bingley M, Bonilla C, Britto R, Bursteinas B, Bye-AJee H, Cowley A, Silva AD, Giorgi MD, Dogan T, Fazzini F, Castro LG, Figueira L, Garmiri P, Georghiou G, Gonzalez D, Hatton-Ellis E, Li W, Liu W, Lopez R, Luo J, Lussi Y, MacDougall Al, Nightingale A, Palka B, Pichler K, Poggioli D, Pundir S, Pureza L, Qi G, Rosanoff S, Saidi R, Sawford T, Shypitsyna A, Speretta E, Turner E, Tyagi N, Volynkin V, Wardell T, Warner K, Watkins X, Zaru R, Zellner H, Xenarios I, Bougueleret L, Bridge A, Poux S, Redaschi N, Aimo L, ArgoudPuy G, Auchincloss A, Axelsen K, Bansal P, Baratin D, Blatter MC, Boeckmann B, Bolleman J, Boutet E, Breuza L, Casal-Casas C, Castro ED, Coudert E, Cuche B, Doche M, Dornevil D, Duvaud S, Estreicher A, Famiglietti L, Feuermann M, Gasteiger E, Gehant S, Gerritsen V, Gos A, Gruaz-Gumowski N, Hinz U, Hulo C, Jungo F, Keller G, Lara V, Lemercier P, Lieberherr D, Lombardot T, Martin X, Masson P, Morgat A, Neto T, Nouspikel N, Paesano S, Pedruzzi I, Pilbout S, Pozzato M, Pruess M, Rivoire C, Roechert B, Schneider M, Sigrist C, Sonesson K, Staehli S, Stutz A, Sundaram S, Tognolli M, Verbregue L, Veuthey AL, Wu CH, Arighi CN, Arminski L, Chen C, Chen Y, Garavelli JS, Huang H, Laiho K, McGarvey P, Natale DA, Ross K, Vinayaka CR, Wang Q, Wang Y, Yeh LS, Zhang J. 2017. UniProt: the universal protein knowledgebase. *Nucleic Acids Research* **45**:D158–D169 DOI 10.1093/nar/gkw1099.

Bhullar KS, Lagarón NO, McGowan EM, Parmar I, Jha A, Hubbard BP, Rupasinghe HPV. 2018. Kinase-targeted cancer therapies: progress, challenges and future directions. *Molecular Cancer* **17(1)**:859 DOI 10.1186/s12943-018-0804-2.

Chagoyen M, Pazos F. 2013. Tools for the functional interpretation of metabolomic experiments. *Briefings in Bioinformatics* **14(6)**:737–744 DOI 10.1093/bib/bbs055.

Cohen P. 2001. The role of protein phosphorylation in human health and disease: delivered on June 30th 2001 at the FEBS Meeting in Lisbon. *European Journal of Biochemistry* **268(19)**:5001–5010 DOI 10.1046/j.0014-2956.2001.02473.x.

Conway JR, Lex A, Gehlenborg N. 2017. UpSetR: an R package for the visualization of intersecting sets and their properties. *Bioinformatics* **33(18)**:2938–2940 DOI 10.1093/bioinformatics/btx364.

Cowley MJ, Pinese M, Kassahn KS, Waddell N, Pearson JV, Grimmond SM, Biankin AV, Hautaniemi S, Wu J. 2012. PINA v2.0: mining interactome modules. *Nucleic Acids Research* **40(D1)**:D862–D865 DOI 10.1093/nar/gkr967.

Day EK, Sosale NG, Lazzara MJ. 2016. Cell signaling regulation by protein phosphorylation: a multivariate, heterogeneous, and context-dependent process. *Current Opinion in Biotechnology* **40(August)**:185–192 DOI 10.1016/j.copbio.2016.06.005.

Demir E, Cary MP, Paley S, Fukuda K, Lemer C, Vastrik I, Wu G, D'Eustachio P, Schaefer C, Luciano J, Schacherer F, Martinez-Flores I, Hu Z, Jimenez-Jacinto V, Joshi-Tope G, Kandasamy K, Lopez-Fuentes AC, Mi H, Pichler E, Rodchenkov I, Splendiani A, Tkachev S, Zucker J, Gopinath G, Rajasimha H, Ramakrishnan R, Shah I, Syed M, Anwar N, Babur Ö, Blinov M, Brauner E, Corwin D, Donaldson S, Gibbons F, Goldberg R, Hornbeck P, Luna A, Murray-Rust P, Neumann E, Reubenacker O, Samwald M, Van Iersel M, Wimalaratne S, Allen K, Braun B, Whirl-Carrillo M, Cheung K-H, Dahlquist K, Finney A, Gillespie M, Glass E, Gong L, Haw R, Honig M, Hubaut O, Kane D, Krupa S, Kutmon M, Leonard J, Marks D, Merberg D, Petri V, Pico A, Ravenscroft D, Ren L, Shah N, Sunshine M, Tang R, Whaley R, Letovksy S, Buetow KH, Rzhetsky A, Schachter V, Sobral BS, Dogrusoz U, McWeeney S, Aladjem M, Birney E, Collado-Vides J, Goto S, Hucka M, Le Novère N, Maltsev N, Pandey A, Thomas P, Wingender E, Karp PD, Sander C, Bader GD, Ruebenacker O, Samwald M, Van Iersel M, Wimalaratne S, Allen K, Braun B, Whirl-Carrillo M, Cheung K-H, Dahlquist K, Finney A, Gillespie M, Glass E, Gong L, Haw R, Honig M, Hubaut O, Kane D, Krupa S, Kutmon M, Leonard J, Marks D, Merberg D, Petri V, Pico A, Ravenscroft D, Ren L, Shah N, Sunshine M, Tang R, Whaley R, Letovksy S, Buetow KH, Rzhetsky A, Schachter V, Sobral BS, Dogrusoz U, McWeeney S, Aladjem M, Birney E, Collado-Vides J, Goto S, Hucka M, Le Novère N, Maltsev N, Pandey A, Thomas P, Wingender E, Karp PD, Sander C, Bader GD. 2010. The BioPAX community standard for pathway data sharing. *Nature Biotechnology* **28(9)**:935–942 DOI 10.1038/nbt.1666.

Diella F, Cameron S, Gemünd C, Linding R, Via A, Kuster B, Sicheritz-Pontén T, Blom N, Gibson TJ. 2004. Phospho.ELM: a database of experimentally verified phosphorylation sites in eukaryotic proteins. *BMC Bioinformatics* **5(1)**:79 DOI 10.1186/1471-2105-5-79.

Fuhs SR, Hunter T. 2017. pHisphorylation: the emergence of histidine phosphorylation as a reversible regulatory modification. *Current Opinion in Cell Biology* **45**:8–16 DOI 10.1016/j.ceb.2016.12.010.

Gnad F, Ren S, Cox J, Olsen JV, Macek B, Oroshi M, Mann M. 2007. PHOSIDA (Phosphorylation Site Database): management, structural and evolutionary investigation, and prediction of phosphosites. *Genome Biology* **8(11)**:R250 DOI 10.1186/gb-2007-8-11-r250.

Goel R, Harsha HC, Pandey A, Prasad TSK. 2012. Human protein reference database and human proteinpedia as resources for phosphoproteome analysis. *Molecular BioSystems: Royal Society of Chemistry* **8(2)**:453–463 DOI 10.1039/c1mb05340j.

Gray VE, Kumar S. 2011. Rampant purifying selection conserves positions with posttranslational modifications in human proteins. *Molecular Biology and Evolution* **28(5)**:1565–1568 DOI 10.1093/molbev/msr013.

Gyori BM, Bachman JA, Subramanian K, Muhlich JL, Galescu L, Sorger PK. 2017. From word models to executable models of signaling networks using automated assembly. *Molecular Systems Biology* **13**:954.

Hornbeck PV, Zhang B, Murray B, Kornhauser JM, Latham V, Skrzypek E. 2015. PhosphoSitePlus, 2014: mutations, PTMs and recalibrations. *Nucleic Acids Research* **43(D1)**:D512–DD20 DOI 10.1093/nar/gku1267.

Hu J, Rho HS, Newman RH, Zhang J, Zhu H, Qian J. 2014. Phospho networks: a database for human phosphorylation networks. *Bioinformatics* **30(1)**:141–142 DOI 10.1093/bioinformatics/btt627.

Huang KY, Wu HY, Chen YJ, Lu CT, Su MG, Hsieh YC, Tsai CM, Lin KI, Huang HD, Lee TY, Chen YJ. 2014. RegPhos 2.0: an updated resource to explore protein kinase-substrate phosphorylation networks in mammals. *Database* **2014**:1–12 DOI 10.1093/database/bau034.

Humphrey SJ, Yang G, Yang P, Fazakerley DJ, Stöckli J, Yang JY, James DE, Ckli JS, Yang JY, James DE. 2013. Dynamic adipocyte phosphoproteome reveals that Akt directly regulates MTORC2. *Cell Metabolism* **17(6)**:1009–1020 DOI 10.1016/j.cmet.2013.04.010.

Rudolph JD, De Graauw M, Van de Water B, Geiger T, Sharan R, Rudolph JD, De Graauw M, Van de Water B, Geiger T, Sharan R. 2016. Elucidation of signaling pathways from large-scale phosphoproteomic data using protein interaction networks. *Cell Systems* **3(6)**:585–593.e3 DOI 10.1016/j.cels.2016.11.005.

Janes KA, Yaffe MB. 2006. Data-driven modelling of signal-transduction networks. *Nature Reviews Molecular Cell Biology* **7(11)**:820–828 DOI 10.1038/nrm2041.

Joshi-Tope G, Gillespie M, Vastrik I, D'Eustachio P, Schmidt E, De Bono B, Jassal B, Gopinath GR, Wu GR, Matthews L, Lewis S, Birney E, Stein L. 2005. Reactome: a knowledgebase of biological pathways. *Nucleic Acids Research* **33(Database issue)**:D428–D432 DOI 10.1093/nar/gki072.

Kanehisa M, Goto S. 2000. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Research* **28(1)**:27–30 DOI 10.1093/nar/28.1.27.

Kawata K, Yugi K, Hatano A, Kokaji T, Tomizawa Y, Fujii M, Uda S, Kubota H, Matsumoto M, Nakayama KI, Kuroda S. 2019. Reconstruction of global regulatory network from signaling to cellular functions using phosphoproteomic data. *Genes to Cells* **24(1)**:82–93 DOI 10.1111/gtc.12655.

Keshava Prasad TS, Goel R, Kandasamy K, Keerthikumar S, Kumar S, Mathivanan S, Telikicherla D, Raju R, Shafreen B, Venugopal A, Balakrishnan L, Marimuthu A, Banerjee S, Somanathan DS, Sebastian A, Rani S, Somak Ray CJHK, Kanth S, Ahmed M, Kashyap MK, Riaz Mohmood YIR, Krishna V, Rahiman BA, Mohan S, Ranganathan P, Ramabadran S, Chaerkady R, Pandey A. 2009. Human protein reference database—2009 Update. *Nucleic Acids Research* **37(Suppl. 1)**:767–772 DOI 10.1093/nar/gkn892.

Khatri P, Sirota M, Butte AJ. 2012. Ten years of pathway analysis: current approaches and outstanding challenges. *PLOS Computational Biology* **8(2)**:e1002375 DOI 10.1371/journal.pcbi.1002375.

Kwon OK, Sim JH, Kim SJ, Sung E, Kim JY, Jeong TC, Lee S. 2015. Comprehensive analysis of in vivo phosphoproteome of mouse liver microsomes. *Journal of Proteome Research* **14(12)**:5215–5224 DOI 10.1021/acs.jproteome.5b00812.

Landry CR, Levy ED, Michnick SW. 2009. Weak functional constraints on phosphoproteomes. *Trends in Genetics* **25(5)**:193–197 DOI 10.1016/j.tig.2009.03.003.

Levy ED, Michnick SW, Landry CR. 2012. Protein abundance is key to distinguish promiscuous from functional phosphorylation based on evolutionary information. *Philosophical Transactions of the Royal Society B: Biological Sciences* **367(1602)**:2594–2606 DOI 10.1098/rstb.2012.0078.

Lienhard GE. 2008. Non-functional phosphorylations? *Trends in Biochemical Sciences* **33(8)**:351–352 DOI 10.1016/j.tibs.2008.05.004.

McGuire JL, Depasquale EA, Funk AJ, O'Donnovan SM, Hasselfeld K, Marwaha S, Hammond JH, Hartounian V, Meador-Woodruff JH, Meller J, McCullumsmith RE. 2017. Abnormalities of signal transduction networks in chronic schizophrenia. *Npj Schizophrenia* **3(1)**:30 DOI 10.1038/s41537-017-0032-6.

Needham EJ, Parker BL, Burykin T, James DE, Humphrey SJ. 2019. Illuminating the dark phosphoproteome. *Science Signaling* **12(565)**:eaau8645 DOI 10.1126/scisignal.aau8645.

Neo4j. 2020. *Available at* https://neo4j.com/.

Van Noort V, Seebacher J, Bader S, Mohammed S, Vonkova I, Betts MJ, Kühner S, Kumar R, Maier T, O'Flaherty M, Rybin V, Schmeisky A, Yus E, Stülke J, Serrano L, Russell RB, Heck AJR, Bork P, Gavin A-C. 2012. Cross-talk between phosphorylation and lysine acetylation in a genome-reduced bacterium. *Molecular Systems Biology* **8(1)**:571 DOI 10.1038/msb.2012.4.

Olsen JV, Blagoev B, Gnad F, Macek B, Kumar C, Mortensen P, Mann M. 2006. Global vivo, and site-specific phosphorylation dynamics in signaling networks. *Cell* **127(3)**:635–648 DOI 10.1016/j.cell.2006.09.026.

Orchard S, Kerrien S, Abbani S, Aranda B, Bhate J, Bidwell S, Bridge A, Briganti L, Brinkman FSL, Cesareni G, Chatr-aryamontri A, Chautard E, Chen C, Dumousseau M, Goll J, Hancock REW, Hannick LI, Jurisica I, Khadake J, Lynn DJ, Mahadevan U, Perfetto L, Raghunath A, Ricard-Blum S, Roechert B, Salwinski L, Stümpflen V, Tyers M, Uetz P, Xenarios I, Hermjakob H. 2012. Protein interaction data curation: the international molecular exchange (IMEx) consortium. *Nature Methods* **9(4)**:345–350 DOI 10.1038/nmeth.1931.

Park J-M, Park J-H, Mun D-G, Bae J, Jung JH, Back S, Lee H, Kim H, Jung H-J, Kim HK, Lee H, Kim KP, Hwang D, Lee S-W. 2015. Integrated analysis of global proteome, phosphoproteome and glycoproteome enables complementary interpretation of disease-related protein networks. *Scientific Reports* **5(1)**:18189 DOI 10.1038/srep18189.

Perfetto L, Briganti L, Calderone A, Cerquone Perpetuini A, Iannuccelli M, Langone F, Licata L, Marinkovic M, Mattioni A, Pavlidou T, Peluso D, Petrilli LL, Pirrò S, Posca D, Santonico E, Silvestri A, Spada F, Castagnoli L, Cesareni G. 2016. SIGNOR: a database of causal relationships between biological entities. *Nucleic Acids Research* **44(D1)**:D548–D554 DOI 10.1093/nar/gkv1048.

Peri S, Navarro JD, Kristiansen TZ, Amanchy R, Surendranath V, Babylakshmi Muthusamy TKBG, Chandrika KN, Deshpande N, Shubha Suresh BPR, Shanker K, Padma N, Vidya Niranjan HCH, Naveen Talreja BMV, Ramya MA, Yatish AJ, Mary Joy HNS, Kavitha MP, Menezes M, Choudhury DR, Neelanjana Ghosh RS, Chandran S, Mohan S, Chandra Kiran Jonnalagadda CKP, Kumar-Sinha C, Deshpande KS, Pandey A. 2004. Human protein reference database as a discovery resource for proteomics. *Nucleic Acids Research* **32**:497–501 DOI 10.1093/nar/gkh070.

Pratt D, Chen J, Welker D, Rivas R, Pillich R, Rynkov V, Ono K, Miello C, Hicks L, Szalma S, Stojmirovic A, Dobrin R, Braxenthaler M, Kuentzer J, Demchak B, Ideker T. 2015. NDEx, the network data exchange. *Cell Systems* **1(4)**:302–305 DOI 10.1016/j.cels.2015.10.001.

Ramanan VK, Shen L, Moore JH, Saykin AJ. 2012. Pathway analysis of genomic data: concepts, methods, and prospects for future development. *Trends in Genetics* **28(7)**:323–332 DOI 10.1016/j.tig.2012.03.004.

Rask-Andersen M, Zhang J, Fabbro D, Schiöth HB. 2014. Advances in kinase targeting: current clinical use and clinical trials. *Trends in Pharmacological Sciences* **35(11)**:604–620 DOI 10.1016/j.tips.2014.09.007.

Rohrs JA, Zheng D, Graham NA, Wang P, Finley SD. 2018. Computational model of chimeric antigen receptors explains site-specific phosphorylation kinetics. *Biophysical Journal* **115(6)**:1116–1129 DOI 10.1016/j.bpj.2018.08.018.

Sacco F, Humphrey SJ, Cox J, Mischnik M, Schulte A, Klabunde T, Schäfer M, Mann M. 2016a. Glucose-regulated and drug-perturbed phosphoproteome reveals molecular mechanisms controlling insulin secretion. *Nature Communications* **7(1)**:13250 DOI 10.1038/ncomms13250.

Sacco F, Perfetto L, Cesareni G. 2018. Combining phosphoproteomics datasets and literature information to reveal the functional connections in a cell phosphorylation network. *Proteomics* **18(5–6)**:1700311 DOI 10.1002/pmic.201700311.

Sacco F, Silvestri A, Posca D, Pirrò S, Gherardini PF, Castagnoli L, Mann M, Cesareni G. 2016b. Deep proteomics of breast cancer cells reveals that metformin rewires signaling networks away from a pro-growth state. *Cell Systems* **2(3)**:159–171 DOI 10.1016/j.cels.2016.02.005.

Safaei J, Maňuch J, Gupta A, Stacho L, Pelech S. 2011. Prediction of 492 human protein kinase substrate specificities. *Proteome Science* **9(Suppl. 1)**:S6 DOI 10.1186/1477-5956-9-S1-S6.

Satpathy S, Wagner SA, Beli P, Gupta R, Kristiansen TA, Malinova D, Francavilla C, Tolar P, Bishop GA, Hostager BS, Choudhary C. 2015. Systems-wide analysis of BCR signalosomes and downstream phosphorylation and ubiquitylation. *Molecular Systems Biology* **11(6)**:810 DOI 10.15252/msb.20145880.

Savage SR, Zhang B. 2020. Using phosphoproteomics data to understand cellular signaling: a comprehensive guide to bioinformatics resources. *Clinical Proteomics* **17(1)**:27 DOI 10.1186/s12014-020-09290-x.

Sharma K, D'Souza RCJ, Tyanova S, Schaab C, Wiśniewski JR, Cox J, Mann M. 2014. Ultradeep human phosphoproteome reveals a distinct regulatory nature of Tyr and Ser/Thr-based signaling. *Cell Reports* **8(5)**:1583–1594 DOI 10.1016/j.celrep.2014.07.036.

Slenter DN, Kutmon M, Hanspers K, Riutta A, Windsor J, Nunes N, Mélius J, Cirillo E, Coort SL, Digles D, Ehrhart F, Giesbertz P, Kalafati M, Martens M, Miller R, Nishida K, Rieswijk L, Waagmeester A, Eijssen LMT, Evelo CT, Pico AR, Willighagen EL. 2018. WikiPathways: a multifaceted pathway database bridging metabolomics to other omics research. *Nucleic Acids Research* **46(D1)**:D661–D667 DOI 10.1093/nar/gkx1064.

Stark C, Breitkreutz BJ, Reguly T, Boucher L, Breitkreutz A, Tyers M. 2006. BioGRID: a general repository for interaction datasets. *Nucleic Acids Research* **34(90001)**:D535–D539 DOI 10.1093/nar/gkj109.

Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP. 2005. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America* **102(43)**:15545–15550 DOI 10.1073/pnas.0506580102.

Szklarczyk D, Gable AL, Lyon D, Junge A, Wyder S, Huerta-Cepas J, Simonovic M, Doncheva NT, Morris JH, Bork P, Jensen LJ, Mering CV. 2019. STRING V11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Research* **47(D1)**:D607–D613 DOI 10.1093/nar/gky1131.

Tong M, Yu C, Zhan D, Zhang M, Zhen B, Zhu W, Wang Y, Wu C, He F, Qin J, Li T. 2019. Molecular subtyping of cancer and nomination of kinase candidates for inhibition with phosphoproteomics. *EBioMedicine* **40(563–575)**:305–317 DOI 10.1016/j.ebiom.2018.12.039.

Vlastaridis P, Kyriakidou P, Chaliotis A, Van de Peer Y, Oliver SG, Amoutzias GD. 2017. Estimating the total number of phosphoproteins and phosphorylation sites in eukaryotic proteomes. *Gigascience* **6(2)**:1–11 DOI 10.1093/gigascience/giw015.

Walker I, Newell H. 2009. Do molecularly targeted agents in oncology have reduced attrition rates? *Nature Reviews Drug Discovery* **8(1)**:15–16 DOI 10.1038/nrd2758.

Wang J, Wang H, Deng T, Liu Z, Wang X. 2019. Time-coursed transcriptome analysis identifies key expressional regulation in growth cessation and dormancy induced by short days in Paulownia. *Scientific Reports* **9(1)**:1–14 DOI 10.1038/s41598-019-53283-2.

Wishart DS, Knox C, Guo AC, Shrivastava S, Hassanali M, Stothard P, Chang Z, Woolsey J. 2006. DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Research* **34(90001)**:D668–D672 DOI 10.1093/nar/gkj067.

Wu D, Lim E, Vaillant F, Asselin-Labat M-L, Visvader JE, Smyth GK. 2010. ROAST: rotation gene set tests for complex microarray experiments. *Bioinformatics* **26(17)**:2176–2182 DOI 10.1093/bioinformatics/btq401.

Young MD, Wakefield MJ, Smyth GK, Oshlack A. 2010. Gene ontology analysis for RNA-Seq: accounting for selection bias. *Genome Biology* **11(2)**:R14 DOI 10.1186/gb-2010-11-2-r14.

Yu G, Wang L-G, Han Y, He Q-Y. 2012. ClusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS A Journal of Integrative Biology* **16(5)**:284–287 DOI 10.1089/omi.2011.0118.

Yu K, Zhang Q, Liu Z, Zhao Q, Zhang X, Wang Y, Wang Z-X, Jin Y, Li X, Liu Z-X, Xu R-H. 2019. QPhos: a database of protein phosphorylation dynamics in humans. *Nucleic Acids Research* **47(D1)**:D451–D458 DOI 10.1093/nar/gky1052.

Zhou H, Jin J, Zhang H, Yi B, Wozniak M, Wong L. 2012. IntPath-an integrated pathway gene relationship database for model organisms and important pathogens. *BMC Systems Biology* **6(Suppl. 2)**:S2 DOI 10.1186/1752-0509-6-S2-S2.