Journal of Proteomics 250 (2022) 104392

Contents lists available at ScienceDirect

Journal of Proteomics

journal homepage: www.elsevier.com/locate/jprot

PROTREC: A probability-based approach for recovering missing proteins based on biological networks

Weijia Kong [a,b,1], Bertrand Jern Han Wong [a], Huanhuan Gao [c,d], Tiannan Guo [c,d], Xianming Liu [e], Xiaoxian Du [e], Limsoon Wong [b,*], Wilson Wen Bin Goh [a,f,**]

[a] School of Biological Sciences, Nanyang Technological University, Singapore
[b] Department of Computer Science, National University of Singapore, Singapore
[c] Zhejiang Provincial Laboratory of Life Sciences and Biomedicine, Key Laboratory of Structural Biology of Zhejiang Province, School of Life Sciences, Westlake University, Zhejiang, China
[d] Institute of Basic Medical Sciences, Westlake Institute for Advanced Study, Zhejiang Province, China
[e] Bruker (Beijing) Scientific Technology Co., Ltd, Shanghai, China
[f] Lee Kong Chian School of Medicine, Nanyang Technological University, Singapore

ARTICLE INFO

Keywords:
Bioinformatics
Protein complexes
Proteomics
Missing proteins
Networks
Statistics

ABSTRACT

A novel network-based approach for predicting missing proteins (MPs) is proposed here. This approach, PROTREC (short for PROtein RECovery), dominates existing network-based methods – such as Functional Class Scoring (FCS), Hypergeometric Enrichment (HE), and Gene Set Enrichment Analysis (GSEA) – across a variety of proteomics datasets derived from different proteomics data acquisition paradigms: Higher PROTREC scores are much more closely correlated with higher recovery rates of MPs across sample replicates. The PROTREC score, unlike methods reporting *p*-values, can be directly interpreted as the probability that an unreported protein in a proteomic screen is actually present in the sample being screened.

Significance: Mass spectrometry (MS) has developed rapidly in recent years; however, an obvious proportion of proteins is still undetected, leading to missing protein problems. A few existing protein recovery methods are based on biological networks, but the performance is not satisfactory. We propose a new protein recovery method, PROTREC, a Bayesian-inspired approach based on biological networks, which shows exceptional performance across multiple validation strategies. It does not rely on peptide information, so it avoids the ambiguity issue that most protein assembly methods face.

1. Introduction

Despite technological advances in proteomics, proteome coverage and protein identification consistency issues persist, resulting in "data holes" corresponding to missing proteins (MPs). Although MPs are sometimes defined as proteins that are persistently unobservable in proteomics, a more inclusive definition is used here: MPs are defined here, with respect to a proteomic screen of a given sample, as those proteins which are not observed in that proteomics screen of that sample, but which are actually present in that sample. This definition, therefore, also includes proteins that are sporadically or inconsistently observed in a given tissue [1]. This definition of MPs is also more relevant should proteomic screens be used in a day-to-day clinical context. Collectively, the issue of MPs in this scenario is referred as the missing protein problem (MPP), where relevant proteins are persistently unobserved or sporadically or inconsistently observed across samples [1]. MPP hampers reproducible functional analysis and impedes the task of identifying biomarkers and novel drug targets from proteomics data [2].

MPP may be resolved via network-based analysis methods [3]. In the literature, several approaches already exist, and have been demonstrated to aid in the recovery of MPs (i.e., identifying proteins which are not reported in a proteomic screen of a sample but are present in the sample). These approaches include Functional Class Scoring (FCS)

* Correspondence to: L. Wong, Department of Computer Science, National University of Singapore, 13 Computing Drive, 117417, Singapore.
** Correspondence to: W. W. B. Goh, School of Biological Sciences, Nanyang Technological University, 60 Nanyang Drive, 637551, Singapore.
E-mail addresses: wongls@comp.nus.edu.sg (L. Wong), wilsongoh@ntu.edu.sg (W.W.B. Goh).
[1] First Author(s).

https://doi.org/10.1016/j.jprot.2021.104392
Received 13 July 2021; Received in revised form 30 August 2021; Accepted 2 September 2021
Available online 7 October 2021

[4–6], Maxlink [7,8], Gene Set Enrichment Analysis (GSEA) [9,10] based on the Kolmogorov-Smirnov test [4], and the hypergeometric enrichment (HE) test [11]. Previously, we demonstrated that resolving MPP using FCS based on protein complexes provides unmatched performance over other network-based approaches [12].

However, network-based analysis methods can be further improved and evaluated more widely. In particular, prior methods such as HE, GSEA and FCS depend on *p*-values, which are reportedly unstable [13]. The *p*-value also only provides information on the likelihood that an observed result is not due to chance and does not provide a direct indicator on effect size; thus, a user is prone to fallaciously mistaking a significant *p*-value as evidence of the presence of an effect or signal.

Here, we develop a method, PROTein RECovery (PROTREC), using improved contextual reasoning for MP recovery. We perform a rigorous comparative evaluation of PROTREC and several known computational methods for MP recovery across several proteomic acquisition paradigms, including Data Dependent Acquisition-Parallel Accumulation Serial Fragmentation (DDA-PASEF) [14],-Parallel Accumulation-Serial Fragmentation combined with data-independent acquisition (diaPASEF) [15] and Sequential window acquisition of all theoretical fragment ion spectra (SWATH) [16]. We demonstrate that PROTREC is superior to other methods by convincing margins.

2. Materials and methods

2.1. Datasets

2.1.1. Renal cancer (RC) acquired by DIA-SWATH

The renal cancer (RC) study of Guo et al. [17] comprises 24 SWATH runs originating from six pairs of non-tumorous and tumorous clear-cell renal carcinoma (ccRCC) tissues, with two technical replicates each. RC has two phenotype classes, RC normal (RC_N) and RC cancer (RC_C), contains high amounts of individual variability, and has 36% data holes; cf. Supplementary Fig. 1.

All SWATH maps are analyzed using OpenSWATH(version 10.5) [18] against a spectral library containing 49,959 reference spectra for 41,542 proteotypic peptides from 4624 reviewed SwissProt proteins [17]. The library is compiled via library search of spectra captured in DDA mode (linking spectra *m/z* and rt. coordinates to a library peptide). Protein isoforms and protein groups are excluded from this analysis. Proteins are quantified based on the intensities of the top two most abundant peptides.

2.1.2. HeLa and SiHa data acquired by DDA-PASEF

The HeLa and SiHa DDA datasets have only one phenotype class with three technical replicates. The detailed data acquisition process is shown in Supplementary Method 1. Both datasets were analyzed by Peaks Studio (Version10.5 build on April 15th, 2020, Bioinformatics Solution Inc.) to search against the reference library obtained from SwissProt Human (20,421 sequences, downloaded on May 8th, 2019).

For HeLa datasets, 310,277 PSMs, 57,856 peptides and 6090 proteins were identified, on average, across the three replicates at the peptide FDR of 1%. For SiHa datasets, 351,782 PSMs, 74,658 peptides and 7298 proteins were obtained on average across three replicates at the peptide FDR of 1%. There are 7% and 5% missing values in HeLa and SiHa respectively, the heatmap of the datasets is shown in Supplementary Fig. 1.

2.1.3. HeLa and SiHa data acquired by diaPASEF

A second HeLa dataset was acquired by diaPASEF with two technical replicates. A second SiHa dataset was similarly acquired with three replicates. Details on this process can be found in Supplementary Method 1.

The project-specific library from 24 high-pH reversed-phase peptide fractions of a HeLa digest with DDA-PASEF consisted of 301,353 target precursors and 16,578 target proteins on average. The SiHa library from 6 high-pH reversed-phase peptide fractions consisted of 153,771 target precursors and 9774 target proteins on average. Protein inference was performed via ID-Picker, with the PSM-protein and precursor-protein FDR set to 1%. There are on average 9821 proteins quantified for HeLa DIA data and 8774 proteins for SiHa DIA data. Compared with its DDA counterpart, diaPASEF has higher consistency (~2% data holes) and protein coverage. However, due to software constraints, we cannot obtain full peptide information prior to protein inference. Hence, we use the diaPASEF data as a verification reference (for predicted missing proteins).

2.2. Network-based methods

We compare our new method PROTREC (described at the end of this section) against three network-based methods. These are Functional Class Scoring (FCS), the Hypergeometric Enrichment (HE) test and Gene Set Enrichment Analysis (GSEA). Since these are used widely, we will not provide full explanations here. This information is available in Supplementary Method 2.

2.3. S-value transformation of p-values

p-values are distributed on an inverse-exponential scale and are thus, non-linear and non-intuitive to interpret. So, following a common practice, we take the negative log (base 2) of the *p*-value, which yields the Shannon information value or surprisal value (S-value). Where required, an S-value cutoff of 4.32 is used (equivalent to a *p*-value cutoff of 0.05) [19]. S-value transformation is applied on the *p*-values obtained from FCS, HE and GSEA.

2.4. Network feature vector based on real complexes

Protein complexes are a special case of biological networks. Performance-wise, protein complexes can recover MPs with unmatched sensitivity [12], effectiveness and practicality [3,4,11,12,20–25]. Another advantage of using protein complex information is that information pertaining to protein complexes are highly centralized and easily accessible. For example, the CORUM [26,27] and MIPS [28,29] databases for human and yeast protein complexes, respectively, serve as dedicated species-specific repositories. Prior works demonstrate that protein complexes exhibit superior signal enrichment over many other sources of data including expressional correlation and predicted subnetworks [11,12]. Since protein complexes are established independently of the proteomics data, a set of complexes can serve as a standardized reference, facilitating cross-comparability between different proteomic studies [3,21]. Finally, standardizing protein complex representation is easy: A complex is simply a list of its constituent proteins (where stable identifiers for proteins, e.g. UniProtKB accessions, already exist [30]); and information regarding the exact topological configuration among constituent proteins in a complex is not required (except in situations where we want to distinguish between core/peripheral proteins, or further classify complexes into topological families) [3].

Here, we use curated protein complexes from CORUM (release 2018) [26]. As small complexes can cause high fluctuation in test statistics, only protein complexes with accessions of Human and have at least a size of 5 are used in the analysis (611 out of 2916) [11]. Detailed information regarding the protein complexes can be found in Supplementary Table 1.

2.5. Evaluation of missing protein recovery based on replicates

We define a recovered missing protein as one that is not observed in an initial proteomics screen (i.e., the protein is missing) but is predicted to be present (based on e.g., PROTREC) and subsequently verified to be present (using a variety of scenarios; see next paragraph). Prediction

without verification does not constitute recovery.

Predicted MPs are verified based on the following scenarios: A/ proteins corresponding to the peptide list consisting of all significant peptide-spectra matches (PSMs) from a cross-batch replicate with fixed threshold cutoff, B/ proteins corresponding to the peptide list consisting of all significant PSMs from a cross-batch replicate using only top N proteins, C/ proteins corresponding to the assembled protein list based on the two-peptide rule, and D/ proteins obtained by other data acquisition methods of the same tissue or cell line.

To determine whether the total set of recovered proteins is significant, we first assume that cross-batch replicates should report the same proteins, we then make predictions on one replicate, and test whether the predicted MPs show up in other replicates. Let R be the set of predicted MPs, and r be the members of R that show up in other replicates. We generate a random set R' of the same size as R, and let r' be the members of R' that show up in other replicates. This randomization is repeated 1000 times. We use this randomization to determine whether $|r|/|R|$ is at the extreme right end of the $|r'|/|R'|$ null distribution. If so, we say that this set of recovered proteins is significant and relevant towards the samples being studied. We express the recovery rate *p*-value (pval) as:

$$pval = \frac{\sum_{i=1}^{n} \frac{|r_i'|}{|R_i'|} \geq \frac{|r|}{|R|}}{n}$$

where n is the number of randomization rounds. In this case, it is 1000.

2.6. PROTein RECovery (PROTREC)

PROTREC is a novel probability-based scoring scheme that estimates the probability of a protein being present in a screen. It is based on the reasonable postulate that the probability of a protein being present in a sample being screened is dependent on the joint probability of it being present if its complex is formed (based on the fraction of constituent proteins being correctly detected), and the probability it is present if its parent complex is not formed. This is different from FCS-based inheritance where a protein inherits its probability of being present from the probability that the enrichment of the complex in the screen is non-random (see Supplementary Method 2).

To calculate the probability above for a protein x, we first find all the protein complexes containing protein x. Let z denote this set of complexes. Then, we calculate the probability of a complex $z_i \in z$ being present, in the sample being screened, by the following equation:

$$p(z_i) = \frac{\sum(x_i \in L)^*(1 - FDR)}{|z_i|}$$

x_i denotes a protein inside the complex z_i. L denotes the set of proteins reported by the proteomic screen. If x_i is reported by the proteomics screen L, then its prior probability is $(1 - FDR)$, where FDR is the false discovery rate of L. If x_i is not reported in L, its prior probability is set to 0. Thus, the complex probability $p(z_i)$ is calculated by the average prior probability of all its components.

To determine the score of a protein x, which is a member of multiple protein complexes, we use the complex with the highest probability. Thus, PROTREC computes the probability of a protein x being present in a sample being screened using each of the complexes that the protein x is a member of and returns the maximum. That is, the PROTREC score that determines whether a protein x is present in the sample is:

$$p(x) = max_{z_i \in z} \{p(x|z_i)\, p(z_i) + p(x|\bar{z}_i)\, p(\bar{z}_i)\}$$

Since the presence of a protein complex implies the presence of all its constituent proteins, $p(x|z_i) = 1$. Conversely, if x is reported, $p(x|\bar{z}_i) = (1 - FDR)$; i.e., when a complex is absent, the probability of a reported constituent protein being present is simply the complement of the false-discovery rate of the proteomic screen. Also, although $p(x|\bar{z}_i)$ is unknown for an unreported protein x, to make the result more conservative, we set $p(x|\bar{z}_i)$ as 0.

This way, we can sort all proteins by their PROTREC score, and predict unreported proteins above a given PROTREC score threshold as predicted MPs. However, a threshold is only useful if the PROTREC probabilities are meaningful (i.e., higher PROTREC probabilities are associated with higher verification rates). And so, we shall evaluate and verify this in our analyses.

The PROTREC score thus describes the probability of the existence of the protein in the given sample. In this manuscript, we use 0.95 as the cutoff. This can be taken to mean, given a PROTREC score of 0.95, we are at least 95% confident this is a correct prediction.

3. Results

3.1. PROTREC dominates in missing protein recovery across various verification strategies

3.1.1. PSM verification using a fixed threshold cutoff

Recall that the missing protein (MP) recovery rate is the proportion of verified proteins over all predicted MPs. We compare MP recovery rates of PROTREC and existing methods (viz. FCS, HE and GSEA) across four datasets (HeLa, SiHa, RC_N and RC_C), representing a variety of proteomic data acquisition strategies. We first verify predicted MPs by cross-replicate verification, where verified MPs are based on the list of PSMs from cross-batch replicates. The recovery rate and the number of MPs inferred are shown in Fig. 1.

Among the compared methods, PROTREC has the highest recovery rate. Using a 0.95 score cut off for PROTREC and 0.05 *p*-value cut off for other methods, which is their default cut off, the cross-verification for PROTREC generally exceeds 80%, while other methods all scored below 60%. For SiHa and HeLa datasets, PROTREC also achieved good performance: Where other methods achieve generally less than 40% recovery rates, PROTREC achieved more than 95%. For RC, even though FCS predicts more MPs, with 23.6% more in RC_N and 15.8% more in RC_C, PROTREC has substantially more recovered proteins: 32.5% more in RC_N and 38.4% more in RC_C. This indicates that most FCS predictions are low quality and likely false positives. We note both HE and GSEA make less predictions which also corresponded to lower recovery rates. Compared against HE, PROTREC has 26.7% more predicted missing proteins in RC_N and 12.6% in RC_C; PROTREC also has 67.4% more validated proteins in RC_N and 68.6% more in RC_C. Compared against GSEA, PROTREC has 51.7% more MPs in RC_N and 54.5% in RC_C; also, PROTREC has 70.5% more validated proteins in RC_N and 59.6% more in RC_C. Detailed validation information can be found in Supplementary Table 2.

We evaluate the recovery rate *p*-value for the four methods. For PROTREC, all derived recovery rate *p*-values are below 0.05, which means the recovered proteins are significant and cannot be emulated by randomization. However, for FCS, HE and GSEA, there are many

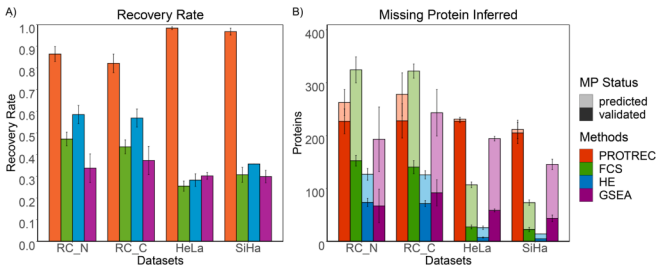

Fig. 1. Verification based on PSMs from cross-batch replicates. A). Recovery rate of the four methods across four datasets. B). Missing protein recovery status are shown as predicted missing proteins (in light shading) and validated missing proteins (in dark shading).

samples with recovery rate *p*-value exceeding 0.05, which means the recovery rates for their predicted missing proteins are not significant and can be exceeded by chance (Supplementary Table 2).

3.1.2. PSM verification using top N

For ease of comparisons, earlier, we used PROTREC score > 0.95 and *p*-value <0.05 (for HE, GSEA and FCS) which are their default cutoffs. This comparison is simply meant to reflect the performance of each method given a typical thresholding strategy. For *p*-value based methods, using a cutoff of *p*-value <0.05 is generally acceptable (provided that the theoretical null distribution used for computing the *p*-value is a good fit for our data). However, this may not be a truly valid comparison as the 0.95 PROTREC score is in fact, not equal to a *p*-value of 0.05 in other methods. In null hypothesis statistical testing (NHST) frameworks, the *p*-value is the chance of rejecting the null hypothesis when it is true; it provides no information when the null hypothesis is false and its magnitude also reflects no information on effect size. Various approaches for converting *p*-values into estimates of effect size and probability that the null hypothesis is false do not seem feasible or introduce yet more unknowns [31].

Instead, we sort all proteins, including reported proteins of the proteomic screen, in descending order based on PROTREC score; this is called the PROTREC list. Then, we count the number of proteins (including reported proteins of the proteomic screen) whose *p*-value is less than 0.05 in a given method and select the same number of proteins in the PROTREC list. Next, we check the recovery rates given those significant proteins based on FCS/HE/GSEA, against the corresponding selected proteins in the PROTREC list. This approach sets PROTREC at a disadvantage, as we are comparing PROTREC against a competing method's optimal performance threshold. Yet, by comparing PROTREC with other methods in turn, PROTREC still dominates, with highest recovery rate and highest number of validated proteins recovered. Detailed information can be found in Supplementary Table 3. For HeLa and SiHa, the top N proteins in HE comparison does not contain predicted MPs in PROTREC (all of the proteins are observed in the original screen), so we mark the recovery rate as 1.

3.1.3. Protein verification based on two-peptide assembly

In earlier comparisons, we used PSMs as evidence of protein presence. But to constitute a strict validated recovery, this may not be robust enough. Suppose if we were to check for verification using assembled proteins instead, would PROTREC still dominate?

To check this, instead of relying on sophisticated protein assemblers such as Percolator, we opted to use a stringent verification by assembling proteins based on the rather conservative two-peptide rule. For RC, to bolster sensitivity, we used all PSMs from the same class for protein assembly, this aggregated information can help expand coverage on those proteins likely to exist. For HeLa and SiHa datasets, we used both de novo only peptide list and data base search PSM list for protein assembly. The recovery rate is calculated by taking the ratio of verified proteins against the number of predicted missing proteins from the two-peptide assembled protein list.

Comparing these three ways of verification, PROTREC dominates protein recovery not only in terms of recovery rates, but also in terms of absolute numbers of verified proteins (Fig. 3). For RC dataset, PROTREC has 73.2% recovery rate in RC_N and 77.2% in RC_C, while other methods scored all below 60%. For HeLa and SiHa dataset, compared to less than 30% recovery rate in other methods, PROTREC has 54.4% recovery rate in HeLa and 51.5% in SiHa.

PROTREC has around 200 validated missing proteins in RC datasets, 2 times more than other methods. It also has more than 100 validated missing proteins in Hela and SiHa, 4 times more than FCS and GSEA, and 10 times more than HE. Detailed information can be found in Supplementary Table 4.

3.1.4. Protein verification based on comparisons against other data acquisition methods

DIA acquisition methods are increasingly used in proteomic profiling. In our HeLa and SiHa datasets, DIA methods identified more proteins while also boasting higher consistency (less data holes). Since DDA and DIA acquisition methods were performed on the same cell line, it makes sense to use predictions made with the weaker acquisition method (less proteins, less consistency) and validate it on the strong method (more proteins, more consistency). Hence, we checked PROTREC's robustness by using DDA-PASEF protein list to predict MPs and verify based on diaPASEF reported protein list. If a predicted missing protein is validated by the diaPASEF reported protein list, it constitutes strong supporting evidence. We compare the predicted MPs, validated MPs and recovery rate for each network method. The recovery rate is calculated by comparing verified MPs against the total number of predicted MPs.

In this comparison, PROTREC dominates (Fig. 4). Compared against other methods which all reported less than 60% recovery rates, PROTREC's recovery rate reached 87.4% in HeLa and 91.4% in SiHa. PROTREC also predicted significantly more validated MPs than other methods. PROTREC has around 176 validated MPs in HeLa and 187 validated MPs in SiHa, several times more than other methods, (see Supplementary Table 5).

3.2. PROTREC provides meaningful scoring functions

The score distribution graph (Fig. 5) is built based on the S-value for FCS, HE and GSEA and the PROTREC score. The S-value is normalized to the same scale as PROTREC score, between 0 and 1. For ease of visualization, Fig. 6 shows the score distribution of four methods for the first sample of RC_N (An example of four datasets' score distribution can be found in Supplementary Fig. 2. Score distribution for every sample in the four datasets can be obtained in Supplementary Table 6). The PROTREC approach for assigning protein probabilities is much better than the other three methods, with all observed proteins in the original screen aggregated on the top as highest probability. Similarly, a high proportion of verified MPs are also aggregated at the top, suggesting that a higher probability assignment correlates with higher verification rates.

3.3. PROTREC performance can be improved by reducing protein complex redundancies

A few protein complexes are formed by protein isotypes that belong to the same protein family or constitute part of bigger protein complex families (e.g., some protein complexes have the same core protein set but different peripheral proteins). Members of such protein complex families may exhibit tissue-specific behaviors [32]. We may improve PROTREC performance if we retain complexes where they are more likely to exist in the tissue being analyzed. Unfortunately, there are no gold-standard tissue-specific "complexome" databases (a complexome being the full complement of protein complexes specific to a particular tissue or sample). And so, we propose a get-around: First, we link complexes that have mutual high similarities (> 75%). This similarity is determined by the number of shared proteins divided by the total number of proteins inside the smaller protein complex. Next, we only consider the complex pair containing at least one protein that appeared in the original screen or validated by cross-verification. Finally, among those linked complexes, we retain the complex with the highest PROTREC score and trash the others. If two complexes tie on their PROTREC scores, both are retained.

Due to our high original PROTREC cut off (0.95), there is no significant improvement for complex filtering as only a few complexes are involved. Thus, we used a 0.50 threshold for illustration here. The recovery rates are shown in Fig. 6. Reducing protein complex redundancy results in a notable improvement for PROTREC on recovery rates for all four datasets. For HeLa and SiHa, both of them exhibited >15%

improvement; while for RC, > 5% improvement. It is worth noting that even though the recovery rate improved, protein complex filtering will result in information loss. Supplementary Table 7 shows there are fewer total and validated proteins recovered. For the other three methods, since the filtering is based on PROTREC score, there is no significant improvement.

3.4. PROTREC protects from information loss

Given a lossy screening where few proteins are reported, PROTREC is the best way to recover the lost information. To demonstrate this, we predicate on one data from a replicate pair, and randomly drop a certain number of proteins in its original protein list. We then predict using the four methods and then evaluate which approach recovers the most dropped proteins. Since each technical replicate contains slightly different numbers of proteins, we choose to drop the same number of proteins for each replicate to maintain result consistency. For HeLa and SiHa, we drop 2000 proteins whereas for RC, we drop 800 proteins. The performance is measured by sensitivity, recovery rate and precision. Sensitivity is calculated as the fraction of dropped proteins which are predicted as MPs. Recovery rate is calculated as the fraction of predicted MPs which are in the original protein list. Precision is calculated by using the self PSM list as the reference.

According to Fig. 7, PROTREC has the highest sensitivity compared to other methods, with at least two times higher than other methods. This indicates PROTREC has the best ability to recover the dataset from information loss without using other reference information from replicate samples.

Given HeLa and SiHa, we can see from Fig. 8, among the proteins that are predicted by PROTREC, it also has the highest recovery rate across the four methods. For RC datasets it is around 70% and in HeLa and SiHa it reaches 90%. Notably, PROTREC precision is more than 95% across all datasets indicating most predicted proteins are validated. Furthermore, PROTREC has the highest number of validated proteins. PROTREC correctly predicts 2–10 times more MPs than other methods. The result shows that where there is pronounced information loss, PROTREC can tell us which unreported proteins should indeed be present.

3.5. PROTREC predictions are verifiable on other independent data platforms

Previously, we used proteomic data from DIA-PASEF platform, to validate PROTREC predictions made from its DDA counterpart for HeLa and SiHa. While this constitutes some form of cross-platform independent support, it be further strengthened. And so, we cross checked gene and protein expression data from similar transcriptomic datasets and on Protein Atlas (a database of tissue-specific protein expression profiles).

We checked the predicted MPs for RC_C and RC_N on kidney tissue-based datasets (RC_C are histologically cancerous tissue while RC_N are histologically normal-like tissue derived from cancer patients). Since we are searching additional evidence manually to corroborate our predicted missing proteins, it is unfeasible to check sample-by-sample. We thus added the following conditions: PROTREC score > 0.95 and not reported in any dataset sample.

First, we cross-checked with transcriptomic evidence. We select the set of proteins that can be mapped to the GSE168845 gene expression dataset on the GEO website [33]. GSE168845 contains normal and cancer microarray expression profiles corresponding to normal kidney tissue and cancer clear-cell renal carcinoma, same tissue types as RC_C and RC_N. We simplistically assert that if the gene corresponding to the missing protein is present in the gene expression profile, then there is gene-level evidence. We cross map gene-protein identifiers using the UniProt database. We cross-validated 99.0% of proteins (96/97) for RC_N and 98.4% of proteins (61/62) for RC_C using the microarray expression profile. In both cases, only 1 protein was missed. Detailed information may be found in Supplementary Table 8.

Next, we also rechecked the predicted MPs in the Tissue Atlas (ver 20.1) inside Protein Atlas [34], which contains protein information across multiple tissues. Tissue Atlas discretizes protein expression into 4 levels of tissue confidence, 'Enhanced', 'Supported', 'Approved' and 'Uncertain'. Excluding 'Uncertain', the other three levels are reliable; so, we filter the Tissue Atlas to confidence level above 'Uncertainty'. Beyond confidence levels, Tissue Atlas also uses a discretized ordinal scale for tissue abundance level, 'High', 'Medium', 'Low', and 'Not detected'. We assign numeric values of 3, 2, 1 and 0 according to the abundance level respectively. If a protein has a score more than 0 in the specific tissue, there is evidence of that it has the corresponding tissue abundance. Among PROTREC predicted MPs, 87/97 in RC_N and 50/62 in RC_C of proteins are verifiable in the Tissue Atlas bank. Among them, 90.8% (79/87) of the proteins in RC_N and 94.0% (47/50) of the proteins in RC_C are validated with protein expression in kidney tissue. Detailed information may be found in Supplementary Table 9.

Finally, even though 8 proteins in RC_N and 3 proteins in RC_C have no supporting evidence in Tissue Atlas, we rechecked these on the DIA-SWATH protein and peptide lists from the other technical replicates. First, we do a cross-check in the protein list. Interestingly, all proteins (8/8) of RC_N are found in the reported list of RC_C and vice versa. We also checked the peptide list. 7/8 proteins in RC_N and 3/3 proteins in RC_C found peptide evidence in the self-peptide list. 1/8 protein in RC_N did not find peptide evidence in self RC_N but find evidence in RC_C. We guess most of these predicted missing proteins failed to be reported in the respective proteomic screens due to their low abundance and were filtered during the assembly process. In addition, Protein Atlas may not be an objective confirmation check for clinical tissue reflecting idiosyncratic development fates. Protein Atlas reflects typical expression profiles of normal tissue. But our tissues, both RC_C and RC_N come from cancer patients. That is, RC_N and RC_C are obtained from the different kidney regions in the same patients so that cancer might affect protein expression in normal cells. If we check these 8 proteins in RC_N on Protein Atlas website, we can see that 6/8 proteins are regarded as prognostic markers in renal cancer. We guess these proteins are unusually expressed in both RC_N and RC_C due to cancer. However, their protein expression in RC_N is much smaller than other reported proteins, which explains why they failed to be detected during the data acquisition process. Another plausible explanation is that these proteome profiles are tissue/cell mixtures, and perhaps, during the sample acquisition process, some cancer cells may have been mixed in.

In summary, checking the transcriptomic and Protein Atlas data gives us more robust evidence of our MP predictions. Most predicted MPs seem to have reason to exist, given both independent protein and gene expression information.

3.6. PROTREC distinguishes ambiguous proteins

Mass spectrometry-based proteomics may suffer from ambiguous protein identification issues: A protein identification from a screen is said to be ambiguous if all the reported peptides used for supporting the identification of the said protein from the screen are ambiguous (i.e. each of these reported peptides occur in at least two reference proteins). PROTREC is able to single out ambiguous protein identifications that are reliable; i.e., it is able to distinguish recoverable ambiguous proteins from those that are not. To prove this, we perform an experiment on HeLa and SiHa dataset. For HeLa and SiHa DDA-PASEF, we have two types of peptide list: de novo peptide list and database PSM list. De novo peptide list contains mostly ambiguous peptide while database PSM list contains more unique peptide. First, we use the de novo peptide list as reference, we collect all purely ambiguous proteins having a PROTREC score above 0.95 and appearing only once in the protein complex list after reducing protein complex redundancies. Then, for each selected ambiguous protein A, we search for other proteins B with the same or a non-empty subset of the ambiguous peptides that support the selected

ambiguous protein. Next, we compare the complex that each protein belongs to and calculate the overlap of the two complexes. If the overlap is low, it means proteins A and B have different neighbors and different supporting information, which means both of the proteins should exist.

We found for both HeLa and SiHa data, all A and B protein pairs have low complex overlap, even though they are supported by same ambiguous peptide. Thus, they are both likely to exist. To further prove our idea, we use the database PSM list and rechecked the mapped proteins. The database PSM list shows both protein A and B are mostly supported by at least one unique peptide, which is further biological evidence that each protein is there (Supplementary Table 10).

4. Discussion

4.1. PROTREC design is sound because it considers contextual information

Network-based methods for MPP are sound because they build on the fundamental understanding that proteins do not work alone, but as higher order aggregates such as protein complexes, pathways and modules. For such aggregates to function, all member proteins need be present. And so, the need for all components to be present for function to exist imposes a tight constraint, presenting an exploitable opportunity.

Therefore, network-based analysis methods add an independent information layer on proteomics data. Because it is independent, network overlays can still work, even if a protein is very sporadically observed given repeated proteomics screens on a given tissue. Hence, network-based methods are useful for dealing with persistent MPs, which are difficult to observe with MS-based proteomics due to a lack of unique sequences, low-abundances, etc. We rechecked the proteomic feature difference in observed, validated and un-validated proteins by cross replicate PSM list. The most obvious proteomic feature is the peptide support, which is calculated by taking the number of supporting peptides for a given protein. Across all datasets, validated MPs have no significant difference in terms of supporting peptide information compared with the un-validated MPs with PROTREC scores above the 0.95 threshold (Supplementary Fig. 3). This may not mean that the unvalidated MPs are not present in the sample being screened. This non-validation may also come about due to a lack of resolution of the validation technique. Hence, we also compare with the supporting information of low PROTREC score proteins. For RC_C and RC_N, compared to the low score proteins (lower than 0.3 PROTREC score), validated and unvalidated high score proteins (at least 0.95 PROTREC score) has much higher peptide support. However, if the low score proteins are set to have score below 0.95 PROTREC score, then there is no significant difference between the supporting information of the three types of proteins (Supplementary Fig. 4). This suggests that recovering missing proteins may not be readily achieved purely by spectra analysis without using additional contextual information. Pure spectra analysis can only differentiate the proteins with very low peptide support. For protein with similar peptide support, it is hard to tell which one should be present or not. Using protein complexes as context, however, provides additional information that augments acquired spectral information on distinguishing missing proteins.

Before PROTREC is introduced here, FCS was the most powerful method for identifying MPs, attaining the highest accuracy when benchmarked on the same proteomics data. Although FCS performs well in our analyses, according to our result discussed earlier, it is notably inferior to PROTREC. From Figs. 1–4, we can see an obvious superiority for PROTREC in terms of recovery rate and number of validated missing proteins. From Fig. 5, it is clear that PROTREC has more meaningful score distribution; and from Fig. 7, it shows PROTREC is much better in preventing information loss.

Moreover, while the FCS test statistic is simple (it is simply the overlap between a network's components and reported proteins in the proteomics screen), our results reveal that it cannot be used to proxy our

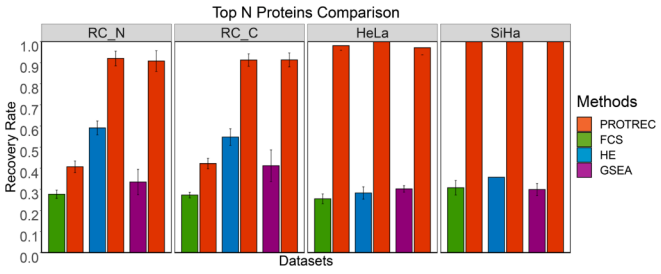

Fig. 2. Comparison by selecting top N proteins in PROTREC against significant proteins in other methods. We count the number of proteins with p-value <0.05 in one method (FCS, HE and GSEA) and select the exact number of top N proteins in PROTREC list. The recovery rate is calculated based on cross-verification.

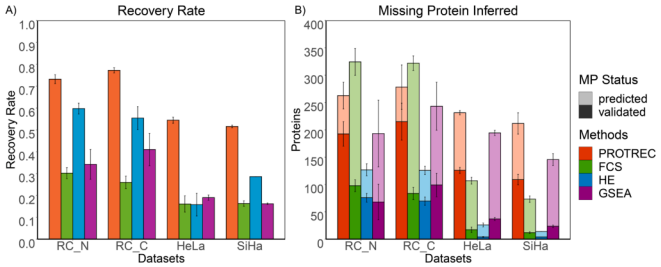

Fig. 3. Verification based on proteins assembled by two peptide rule. A). Recovery rate of the four methods across four datasets. B). Missing protein recovery status are shown as predicted missing proteins (in light shading) and validated missing proteins (in dark shading).

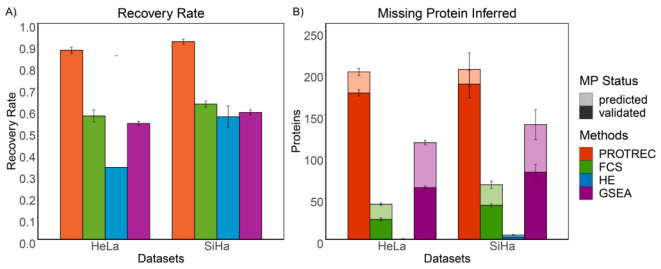

Fig. 4. Verification based on other data acquisition method. A). Recovery rate of the four methods in HeLa and SiHa datasets. B). The number of predicted missing proteins (in light color) and validated missing proteins (in dark color).

confidence in the verification rate of a prediction. FCS seems to have estimated p-values incorrectly: It uses random sets of proteins to form pseudo complexes to estimate the significance of the observed overlap of reported proteins in a real protein complex; this randomization procedure unreasonably assumes that all proteins have an equal and independent chance to form a complex of that size with each other. This FCS p-value is related to HE: Although FCS's p-value is empirically generated, it converges to the hypergeometric distribution used by HE when sample sizes are large enough. Hence, HE has similar difficulty as FCS.

PROTREC recognizes that that test statistics and p-values are insufficient. The PROTREC perspective suggests that we should consider evidence in totality, including looking at the same evidence that suggests support for non-true effect. For example, when a protein is not reported by the original screen, we do not immediately draw the conclusion that the protein is not there. We would instead judge by analyzing the presence of a relevant protein complex (containing that protein) and consider the protein's presence by the joint probability. This perspective turns the situation from whether "the protein is not present" to "the network is not present".

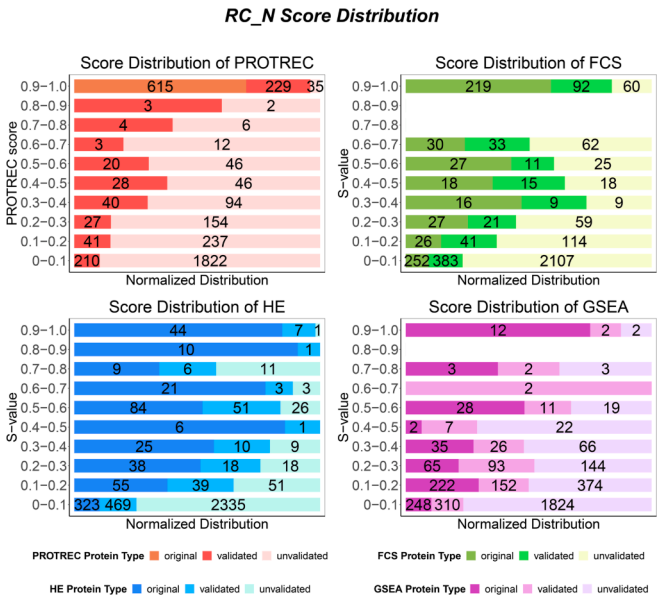

Fig. 5. Score distribution of RC_N. The"original"label means the protein is found in the reported protein list. The "validated"label means the protein is missing but found in the cross replicate PSM list. "Unvalidated" means the predicted missing protein was non-verifiable.

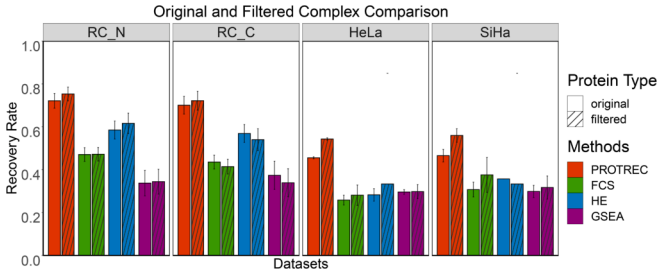

Fig. 6. Recovery rate before and after complex filtering with PROTREC threshold 0.50. The original bars show the recovery rate before complex filtering and the spotted bars show the recovery rate after complex filtering.

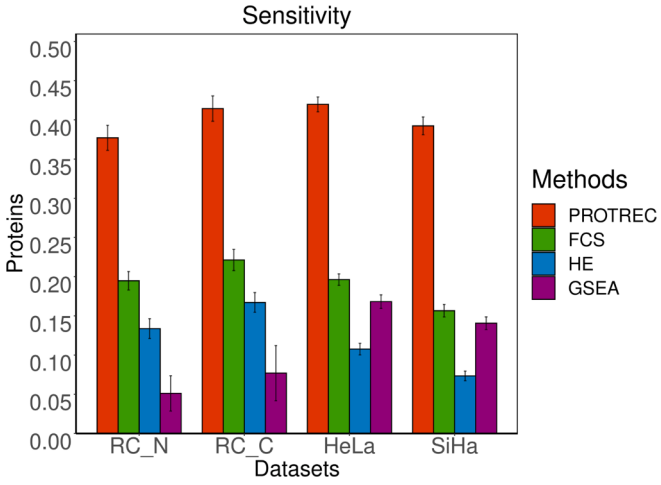

Fig. 7. Performance of sensitivity from information loss.

Therefore, the PROTREC score is calculated based on the compound probability of the complex probability and the protein itself. This calculation allows us to abandon *p*-values in favor of likelihoods on

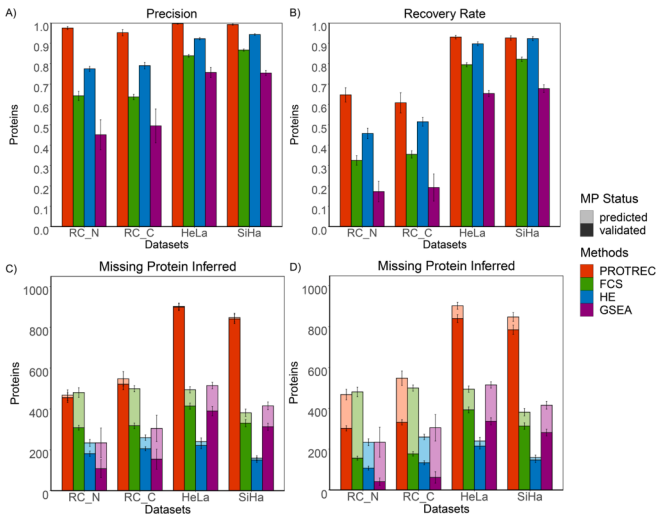

Fig. 8. Performance of recovery from information loss. A). Precision of the four methods across four datasets. B). Recovery rate of the four methods across four datasets. C). The number of predicted missing proteins (in light color) and validated missing proteins (in dark color) by taken the self PSM list as verification. D). The number of predicted missing proteins (in light color) and validated missing proteins (in dark color) by taken the original protein list as verification.

whether a reported protein's parent network exists. Our result in Figs. 1–4 and 7 suggests no matter what aspect we look at, the recovery rate and the number of validated proteins of PROTREC are much superior to other methods.

4.2. *PROTREC predicts rare proteins that are difficult to observe typically*

Regarding protein assembly, the conventional Human Proteome Project guideline suggests using the two-peptide rule to define if a protein exists [35]. The two-peptide rule defines a protein as detected in a sample when two non-nested detected peptides of length more than nine amino acids uniquely map to the protein. However, mass spectrometry-based proteomics already suffers from incomplete proteome coverage and consequent inconsistency issues, and there are quite a few peptides with low detectability and highly ambiguous proteins. Thus, this rule while stringent, does result in widespread data loss. Some approaches have tried using alternative ways (e.g., PEAKS Studio tends to use one supported unique peptide to define the existence of a protein), but they still face the same problem. Especially for purely ambiguous proteins, there is no way these methods can recover them.

By considering the contextual information, PROTREC can infer proteins based on the networks; even when a protein is ambiguous, it can still be present in a sample and be detected by PROTREC. PROTREC does not use the two-peptide rule and thus it can be used to find proteins that cannot be traditionally verified. In Supplementary Table 10, we show that even when two sets of pure ambiguous proteins have inclusive peptide support, their protein complex neighbors are different. According to PROTREC's algorithm, a protein's existence is determined by its associated protein complex's probability. Since the proteins inside the two complexes are different and the two complexes all have a high PROTREC score, it means both complexes should have the right to exist independently. As a result, even though the two proteins have inclusive ambiguous peptide support, they have different surrounding information, so they should both exist.

4.3. *Limitations and future plans*

Although PROTREC performs well, there are some limitations. First, there is no direct validation of MPs— in some cases, we inferred based

on PSMs from the sample and also PSMs from cross-replicate comparisons. Although we showed that using cross replicate protein assemblies still work well, these are still not direct evidence of existence. Although we performed a reasonable inference on ambiguous proteins with low support (lack of unique peptides), MPs whose peptides are entirely invisible and never detected in any sample cannot be verified directly. Moreover, the reliance on contextual information means that if a protein cannot be linked to some reference protein complex or network, it will remain invisible.

Network-based methods like PROTREC are reliant on the quality and extensiveness of the network feature vector: If the network feature vector lacks coverage (where many important complexes are not included), then many MPs cannot be predicted. Similarly, if the network feature contains biases and errors, then use of these may lead towards poor quality predictions and mislead the experimentalist. Our benchmarks have been performed on CORUM, a high-quality curated database of protein complexes. These are known to exhibit high biological coherence but may not provide interesting nor groundbreaking insight. Conversely, predicting synthetic complexes from network data has the potential to unveil new discoveries, but such synthetic complexes need to be carefully analyzed and processed to minimize wastage of resources on validation efforts.

Although we have picked a few representative methods for comparison, we acknowledge we cannot test all network-based approaches in this study. However, these represent a wide variety of network methods: While HE and GSEA are classic examples of over-representation analysis and direct-group analysis respectively; and although FCS is similar to HE, it uses an entirely different *p*-value generation strategy.

Our results suggest that augmenting the network feature vector and reducing redundancies might improve performance. Indeed, the idea of tissue-specific complexomes have been floated some years back, and new algorithms now exist for making such predictions [36]. However, dedicated databases for tissue-specific complexes are still lacking. In future work, we would like to explore this further, as we feel that the quality and comprehensiveness of the list of complexes used are paramount for practical deployment of PROTREC.

In addition, it is arguable that PROTREC is limited in value if it only predicts presence of an MP but not quantity. Our work showed that there is immense value in binary prediction: PROTREC's noise resistance quality means that should a screen severely under-reports protein identifications, PROTREC can be used as a rescue algorithm. PROTREC can reliably expand the set of proteins that should in theory be present. Knowing such information in advance, provides clear guidelines on which proteins we can try to attempt quantitation, whether based on existing information from the screen itself, or transferred from other screens of the same tissue. Leveraging on the former would be an immense upgrade for PROTREC. Now that we know it is extremely powerful in predicting the right proteins, this also gives future impetus on how we can use this for quantitation. Foreseeable challenges include how to attribute information from ambiguous peptides and how to resolve high variability signal from supporting peptides.

5. Conclusion

PROTREC is a novel probability-based approach for estimating the existence probability of a missing protein. Via PROTREC, we can recover proteins that are previously not found in the original proteomic screen. Comparing PROTREC against FCS, HE and GSEA across four datasets, PROTREC dominates in recovery and validation, score distribution and information protection. Since PROTREC relies on protein complexes to make predictions, we also show that PROTREC can be further improved by optimizing the protein complex set. PROTREC is a quantum leap for missing protein prediction algorithms. Given its efficacy across different proteomic acquisition platforms, it can be widely applied.

Declaration of competing interest

None.

Data availability

Data will be made available on request.

Acknowledgements

The graphical abstract is Created with BioRender.com. This work is supported in part by a Singapore Ministry of Education tier-2 grant (MOE2019-T2-1-042).

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jprot.2021.104392.

References

[1] M.S. Baker, S.B. Ahn, A. Mohamedali, M.T. Islam, D. Cantor, P.D. Verhaert, S. Fanayan, S. Sharma, E.C. Nice, M. Connor, Accelerating the search for the missing proteins in the human proteome, Nat. Commun. 8 (1) (2017) 1–13.
[2] L. Zhou, L. Wong, W.W.B. Goh, Understanding missing proteins: a functional perspective, Drug Discov. Today 23 (3) (2018) 644–651.
[3] W.W.B. Goh, L. Wong, Integrating networks and proteomics: moving forward, Trends Biotechnol. 34 (12) (2016) 951–959.
[4] W.W.B. Goh, L. Wong, Advancing clinical proteomics via analysis based on biological complexes: a tale of five paradigms, J. Proteome Res. 15 (9) (2016) 3167–3179.
[5] W.W.B. Goh, L. Wong, NetProt: complex-based feature selection, J. Proteome Res. 16 (8) (2017) 3102–3112.
[6] Y. Zhao, A.C. Sue, W.W.B. Goh, Deeper investigation into the utility of functional class scoring in missing protein prediction from proteomics data, J. Bioinforma. Comput. Biol. 17 (2) (2019) 1950013.
[7] W.W.B. Goh, Y.H. Lee, Z.M. Ramdzan, M.C. Chung, L. Wong, M.J. Sergot, A network-based maximum link approach towards MS identifies potentially important roles for undetected ARRB1/2 and ACTB in liver cancer progression, Int. J. Bioinforma. Res. Appl. 8 (3) (2012) 155–170.
[8] D. Guala, E. Sjolund, E.L. Sonnhammer, Maxlink: network-based prioritization of genes tightly linked to a disease seed set, Bioinformatics 30 (18) (2014) 2689–2690.
[9] A. Subramanian, P. Tamayo, V.K. Mootha, S. Mukherjee, B.L. Ebert, M.A. Gillette, A. Paulovich, S.L. Pomeroy, T.R. Golub, E.S. Lander, J.P. Mesirov, Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles, Proc. Natl. Acad. Sci. U. S. A. 102 (43) (2005) 15545–15550.
[10] J. Zyla, M. Marczyk, T. Domaszewska, S.H.E. Kaufmann, J. Polanska, J. Weiner, Gene set enrichment for reproducible science: comparison of CERNO and eight other algorithms, Bioinformatics 35 (24) (2019) 5146–5154.
[11] W.W.B. Goh, T. Guo, R. Aebersold, L. Wong, Quantitative proteomics signature profiling based on network contextualization, Biol. Direct 10 (1) (2015) 71.
[12] W.W.B. Goh, M.J. Sergot, J.C. Sng, L. Wong, Comparative network-based recovery analysis and proteomic profiling of neurological changes in valproic acid-treated mice, J. Proteome Res. 12 (5) (2013) 2116–2127.
[13] L.G. Halsey, D. Curran-Everett, S.L. Vowler, G.B. Drummond, The fickle P value generates irreproducible results, Nat. Methods 12 (3) (2015) 179–185.
[14] F. Meier, A.D. Brunner, S. Koch, et al., Online Parallel Accumulation–Serial Fragmentation (PASEF) with a Novel Trapped Ion Mobility Mass Spectrometer, Mol. Cell Proteomics. 17 (12) (2018) 2534–2545.
[15] F. Meier, A.-D. Brunner, M. Frank, A. Ha, I. Bludau, E. Voytik, S. Kaspar-Schoenefeld, M. Lubeck, O. Raether, R. Aebersold, B. Collins, H. Röst, M. Mann, Parallel accumulation – serial fragmentation combined with data-independent acquisition (diaPASEF): bottom-up proteomics with near optimal ion usage, bioRxiv (2019), https://doi.org/10.1101/656207.
[16] L.C. Gillet, P. Navarro, S. Tate, H. Rost, N. Selevsek, L. Reiter, R. Bonner, R. Aebersold, Targeted data extraction of the MS/MS spectra generated by data-independent acquisition: a new concept for consistent and accurate proteome analysis, Mol. Cell. Proteomics 11 (6) (2012). O111 016717.
[17] T. Guo, P. Kouvonen, C.C. Koh, L.C. Gillet, W.E. Wolski, H.L. Rost, G. Rosenberger, B.C. Collins, L.C. Blum, S. Gillessen, M. Joerger, W. Jochum, R. Aebersold, Rapid mass spectrometric conversion of tissue biopsy samples into permanent quantitative digital proteome maps, Nat. Med. 21 (4) (2015) 407–413.
[18] H.L. Rost, G. Rosenberger, P. Navarro, L. Gillet, S.M. Miladinovic, O.T. Schubert, W. Wolski, B.C. Collins, J. Malmstrom, L. Malmstrom, R. Aebersold, OpenSWATH enables automated, targeted analysis of data-independent acquisition MS data, Nat. Biotechnol. 32 (3) (2014) 219–223.
[19] C.E.J.T. Shannon, A mathematical theory of communication, Bell Syst. Tech. J. 27 (3) (1948) 379–423.

[20] W.W.B. Goh, L. Wong, Evaluating feature-selection stability in next-generation proteomics, J. Bioinforma. Comput. Biol. 14 (5) (2016) 16500293.
[21] W.W.B. Goh, L. Wong, Design principles for clinical network-based proteomics, Drug Discov. Today 21 (7) (2016) 1130–1138.
[22] W.W.B. Goh, L. Wong, Computational proteomics: designing a comprehensive analytical strategy, Drug Discov. Today 19 (3) (2014) 266–274.
[23] W.W.B. Goh, L. Wong, Networks in proteomics analysis of cancer, Curr. Opin. Biotechnol. 24 (6) (2013) 1122–1128.
[24] W.W.B. Goh, M. Fan, H.S. Low, M. Sergot, L. Wong, Enhancing the utility of Proteomics Signature Profiling (PSP) with Pathway Derived Subnets (PDSs), performance analysis and specialised ontologies, BMC Genomics 14 (2013) 35.
[25] W.W.B. Goh, Y.H. Lee, Z.M. Ramdzan, M.J. Sergot, M. Chung, L. Wong, Proteomics signature profiling (PSP): a novel contextualization approach for cancer proteomics, J. Proteome Res. 11 (3) (2012) 1571–1581.
[26] A. Ruepp, B. Brauner, I. Dunger-Kaltenbach, G. Frishman, C. Montrone, M. Stransky, B. Waegele, T. Schmidt, O.N. Doudieu, V. Stumpflen, H.W. Mewes, CORUM: the comprehensive resource of mammalian protein complexes, Nucleic Acids Res. 36 (Database issue) (2008) D646–D650.
[27] A. Ruepp, B. Waegele, M. Lechner, B. Brauner, I. Dunger-Kaltenbach, G. Fobo, G. Frishman, C. Montrone, H.W. Mewes, CORUM: the comprehensive resource of mammalian protein complexes – 2009, Nucleic Acids Res. 38 (Database issue) (2009) D497–D501.
[28] H.W. Mewes, C. Amid, R. Arnold, D. Frishman, U. Guldener, G. Mannhaupt, M. Munsterkotter, P. Pagel, N. Strack, V. Stumpflen, J. Warfsmann, A. Ruepp, MIPS: analysis and annotation of proteins from whole genomes, Nucleic Acids Res. 32 (Database issue) (2004) D41–D44.
[29] H.W. Mewes, D. Frishman, K.F. Mayer, M. Munsterkotter, O. Noubibou, P. Pagel, T. Rattei, M. Oesterheld, A. Ruepp, V. Stumpflen, MIPS: analysis and annotation of proteins from whole genomes in 2005, Nucleic Acids Res. 34 (Database issue) (2006) D169–D172.
[30] J. Griss, M. Martin, C. O'Donovan, R. Apweiler, H. Hermjakob, J.A. Vizcaino, Consequences of the discontinuation of the International Protein Index (IPI) database and its substitution by the UniProtKB "complete proteome" sets, Proteomics 11 (22) (2011) 4434–4438.
[31] D.J. Benjamin, J.O. Berger, Three recommendations for improving the use of p-values, Am. Stat. 73 (sup1) (2019) 186–191.
[32] W.W.B. Goh, H. Oikawa, J.C.G. Sng, M. Sergot, L. Wong, The role of miRNAs in complex formation and control, Bioinformatics 28 (4) (2012) 453–456.
[33] T. Barrett, D.B. Troup, S.E. Wilhite, P. Ledoux, D. Rudnev, C. Evangelista, I.F. Kim, A. Soboleva, M. Tomashevsky, K.A. Marshall, NCBI GEO: archive for high-throughput functional genomic data, Nucleic Acids Res. 37 (suppl_1) (2009) D885–D890.
[34] P.J. Thul, C. Lindskog, The Human Protein Atlas: a spatial map of the human proteome, Protein Sci. 27 (1) (2018) 233–244.
[35] R.A. Bradshaw, A.L. Burlingame, S. Carr, R. Aebersold, Reporting protein identification data: the next generation of guidelines, Mol. Cell. Proteomics 5 (5) (2006) 787–788.
[36] S. Rizzetto, P. Moyseos, B. Baldacci, C. Priami, A. Csikász-Nagy, Context-dependent prediction of protein complexes by SiComPre, NPJ Syst. Biol. Appl. 4 (1) (2018) 1–9.