Resolving Missing Protein Problems Using Functional Class Scoring

Bertrand Jernhan Wong
Nanyang Technological University
Weijia Kong
Nanyang Technological University
Limsoon Wong
National University of Singapore
Wilson Wen Bin Goh (wilsongoh@ntu.edu.sg)
Nanyang Technological University

Research Article

Keywords: Bioinformatics, Functional Class Scoring (FCS), Networks, Protein Complexes, Proteomics, Statistics

Posted Date: January 5th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1201188/v1

Abstract

Despite technological advances in proteomics, incomplete coverage and inconsistency issues persist, resulting in “data holes”. These data holes cause the missing protein problem (MPP), where relevant proteins are persistently unobserved, or sporadically observed across samples. This hinders biomarker and drug discovery from proteomics data. Network-based approaches are powerful: The Functional Class Scoring (FCS) method using protein complexes was able to easily recover missed proteins with weak or partial support. However, there are limitations: The verification approach (in determining missing protein recovery) is potentially biased as the test data was based on relatively outdated Data-Dependent Acquisition (DDA) proteomics and FCS does not provide a scoring scheme for individual protein components (in significant complexes). To address these issues: First, we devised a more rigorous evaluation of FCS based on same-sample technical replicates. And second, we evaluate using data from more recent Data-Independent Acquisition (DIA) technologies (viz. SWATH).

Although cross-replicate examination reveals some inconsistencies amongst same-class samples, tissue-differentiating signal is nonetheless strongly conserved. This confirms FCS as a viable method that selects biologically meaningful networks. We also report that predicted missing proteins are statistically significant based on FCS p-values. Although cross-replicate verification rates are not spectacular, the predicted missing proteins as a whole, have higher peptide support than non-predicted proteins. FCS also has the capacity to predict missing proteins that are often lost due to weak specific peptide support. As a yet unresolved limitation, we find that FCS cannot assign meaningful probabilities to individual protein components (no relationship between actual probability of verification and FCS-assigned probability) as it only provides a p-value at the level of complexes.

Introduction

Recent advances in hardware have made proteomics increasingly useful for clinical investigation. These include more efficient protein extraction procedures (e.g. PCT [1]), brute-force spectra-capture methods (e.g. SWATH [2]), and improved multiplexing kits [3]. Although proteomics generally falls behind genomics on throughput and adoption, developments in proteomic technologies are of critical importance to biological and clinical/translational research; assaying protein identities/quantities, and their associated post-translational modifications, paints an immediate picture of the underlying functional landscape—a crucial limitation that genomics technologies such as next-generation sequencing cannot address.

Despite recent technological improvements, current proteomics still suffers from incomplete proteome coverage issues (not all proteins in an organism are observed in a single screen), and more critically, inconsistency issues (different screens on the same samples generate different protein sets) [4]. These issues give rise to problems during functional analysis, impeding efforts towards understanding the functional phenotype, unravelling mechanisms, or identifying reproducible biomarkers.

Network-based approaches allow the representation of proteomic samples in terms of their functional modules and pathways despite differing proteome coverage and limited inter-sample consistency. Previous efforts have shown that network-based analyses permit a deeper understanding of biological mechanisms [5] and facilitate reproducible comparative analyses [6–8]. Several network-based analyses methods exist; among these, Functional Class Scoring (FCS) is a notable example previously shown to outperform other network methods [9,10].

To further demonstrate the strength and relevance of network-based methods for proteome profiling and functional analysis, we provide an update on FCS benchmarking and its performance on proteomic data generated with more recent technologies such as Data Independent Acquisition (DIA)/Sequential Window Acquisition of all THeoretical Spectra (SWATH) [2]. Additionally, we develop a more rigorous method of evaluating validity and benchmarking FCS-based missing protein prediction, utilizing cross-verification of technical replicates; this study places particular focus on the challenging tasks of recovering low-support or low-abundance proteins (e.g., proteins not meeting the two-peptide rule, or supported by peptides with low intensities).

Materials And Methods

All methods were performed in accordance with the relevant guidelines and regulations

Renal cancer control (RCC)

The renal cancer control dataset (RCC), from the study of Guo *et al.* [1], comprises 12 SWATH runs originating from a normal human kidney test tissue digested in quadruplicates (x4) and each digest analyzed in triplicates (x3) using a tripleTOF 5600 mass spectrometer (AB Sciex).

1,632 proteins are quantified across the 12 SWATH maps with a peptide and protein false-discovery rate (FDR) of 1%. General details of RCC are shown in Figure 1A. The RCC dataset is less complex (one phenotype class, and no inter-individual variability), but also less data holes (12% missing values), suggesting less inconsistency between samples.

Renal cancer (RC)

The renal cancer (RC) study of Guo *et al.* [1] comprises 24 SWATH runs originating from six pairs of non-tumorous and tumorous clear-cell renal carcinoma (ccRCC) tissues, with two technical replicates (duplicates) each (Figure 1A).

All SWATH maps are analyzed using OpenSWATH [11] against a spectral library containing 49,959 reference spectra for 41,542 proteotypic peptides from 4,624 reviewed SwissProt proteins [1]. 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.

For further details on RC and RCC, please refer to the NetProt package[12].

Colorectal Cancer (CR)

To demonstrate that the significant complexes (e.g. via FCS), and by implication, their associated proteins, present in a sample tend to be more similar if they come from the same tissue (in spite of different proteomics screen), we include a non-kidney dataset (based on 30 normal liver samples from the colorectal dataset of Zhang *et al.* [13,14]) to be compared against RC and RCC.

Network feature vector based on real biological complexes

Biological complexes can recover missing proteins with unmatched sensitivity, and are more effective and practical for analytical use than inferred clusters from protein interaction networks [1–2, 9–10, 25–30] : Firstly, information on biological complexes tends to be highly centralized and easily accessible, e.g. CORUM [31–33] for human complexes and MIPS [34–35] for yeast. Secondly, biological complexes exhibit high signal enrichment [15], over many other sources of data including expressional correlation and predicted subnetworks. Third, a set of complexes can be a standardizable reference, facilitating cross-comparability between different studies [1, 26]. Finally, standardizing 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[36]); and information regarding the exact topological configuration amongst constituent proteins in a complex is not required (except when we want to distinguish between core/peripheral proteins, or classify complexes into families) [1]. For our network feature vector, we use curated protein complexes from the CORUM database (release 2018) [31] As small complexes can cause high fluctuation in test statistics, only protein complexes of at least size 5 are used in analysis (600 out of 1,323) [16].

Functional Class Scoring (FCS)

In FCS[9,10], given a set of observed proteins in a proteomics screen S, and a list of component proteins M_i from a real protein complex C_i (where $C_1 \ldots C_n$ constitute the list of protein complexes in the complex vector), an observed overlap, O_i is expressed as:

$$O_i = \frac{|S \cap M_i|}{\left|M_i\right|}$$

To determine if this overlap O_i is significant, 1,000 randomized complexes of size $|M_i|$ (i.e., the size of C_i) are generated using a reference pool of unique proteins drawn from the complexes $C_i \ldots C_n$. From the 1,000 randomized complexes, a vector of null overlaps, $N_1 \ldots N_{1000}$ is generated. E.g., for the j-th randomized complex, which comprises the set of proteins K_j, we may calculate a null overlap N_j, by comparing K_j against S.

$$N_j = \frac{|S \cap K_j|}{\left|K_j\right|}$$

The empirical p-value is the proportion of null overlaps in $N_1 \ldots N_{1000}$ greater than or equal to the observed overlap O_i (Figure 2). For the i-th complex C_i in the complex vector, its p-value, $pval_i$ is:

$$pval_i = \frac{\sum_{j=1}^{1000} if N_j \geq O_i then 1 else 0}{1000}$$

If $pval_i$ falls below 0.05, the complex C_i is statistically significant. Given the set M_i of proteins in C_i, and the set S of observed proteins, the set of missing proteins that are predicted to be present is defined as M_i - S.

Evaluation of missing protein recovery

Predicted missing proteins are verified based on several scenarios: A/ proteins corresponding to the peptide list consisting of all significant peptide-spectra matches (PSMs) from the sample itself, B/ proteins corresponding to the peptide list consisting of all significant PSMs from the cross-batch replicate, and C/ proteins corresponding to the union of the PSMs from self and cross-batch replicate. Additionally, we also consider the following naïve scenarios: D/ Observed proteins in the cross-batch replicate, and E/ proteins corresponding to FCS-significant complexes in the cross-batch replicate.

To determine whether the total set of recovered proteins is significant, we assume that cross-batch replicates should report the same proteins (In practice they do not, thus leading to the MPP). We run FCS on one replicate, and test whether the missing proteins that are predicted to be present show up in other replicates. Let R be the set of missing proteins predicted to be present, 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 many times, and we determine whether |r|/|R| lie 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.

When comparing for overlaps, e.g., to evaluate whether similar missing proteins are predicted across different samples, we use the Jaccard index. Given two sets X and Y, we may define the Jaccard Index J(X,Y) as:

$$J(X, Y) = \frac{X \cap Y}{X \cup Y}$$

Results

Missing values are widespread amongst analyzed samples

The renal cancer dataset RC is more complex than the single-tissue benchmark dataset RCC. This is because RC comprises two phenotype classes, has higher individual variability (due to more patients), and ~3x more data holes (36%). Although many proteins observed in RC and RCC are shared (Figure 1B), a quick check on the dispersal of missing proteins across samples in RC also indicates that the missing proteins are dispersed across a wider range of proteins (Figure 1C and Figure 1D).

Since we want to check for missing protein recovery across technical replicates, it is important that batch effects do not dominate outcome [17]. Figure 1C/D show the relationships between samples annotated by class and batch (The naming nomenclature is class_sample number_batch; e.g. N2_2 means “Normal” sample 2, batch 2) where we ascertained no obvious batch effects, i.e., the sample do not group broadly by the batch labels (Figure 1D).

FCS-predicted complexes are tissue specific and biologically relevant

Given FCS, each sample can be represented in terms of its statistically significant networks or protein complexes. But is this representation biologically meaningful?

We first consider the distribution of FCS p-values (calculated on protein complexes) across samples in RCC, and the two sample classes of RC (RC_N and RC_C, where N and C refers to normal and cancer classes respectively); cf. Figure 3A. Although many significant complexes are shared amongst samples (blue zones), there is a high degree of obfuscation and uncertainty, as represented by the thick mixed color columns in the middle of the heatmaps. This suggests that different samples are predicting a notable proportion of different complexes even though they belong to the same class (and expected to report the same complexes as significant).

Despite this apparent heterogeneity amongst same-class samples, we are curious whether there is conserved signal amongst significant complexes (FCS p-value below 0.05) reported in the same tissue-type, despite different proteomics screen. Based on the inter-sample agreement for RCC, RC_N and RC_C, we find that the Jaccard indices are relatively high (~0.65 to 0.70), compared to overlaps against significant complexes derived from another tissue (colorectal in this case); cf. Figure 3B. Although overlaps fall when we consider similarity of significant complexes between RCC and RC_N (RC_N <-> RCC), the Jaccard indices are still appreciably higher than when we compare RC_N to CR (RC_N <-> CR) and RCC to CR (RCC <-> CR) (Figure 3B). A two-sample t-test shows that the distribution of Jaccard indices for RC_N <-> RCC is significantly higher against (RC_N <-> CR) and (RCC <-> CR) (p-value << 0.01; ***). This means that despite the apparent heterogeneity (in terms of significant complex agreements) amongst same-class samples, there is conserved signal amongst samples derived from the same tissue type, even across different proteomics screens (as with RCC and RC_N).

For each complex overlap between sample pairs, we may determine a significance measure based on the hypergeometric p-value (Figure 3C). Here, regardless of same tissue on same proteomics screen, same tissue on different proteomics screen, or cross-tissue on different proteomics screen, the hypergeometric

p-values are all generally low (p-value << 0.01). We speculate this is due to high numbers of shared complexes (e.g., housekeepers –– transcriptional, translational and protein degradation machinery, etc.) common to many different tissue types anyway (Supplementary Figure 2). However, it is noteworthy that the p-values for cross-tissue comparisons appear somewhat less significant, possibly due to lower inter-tissue overlaps (Figure 3C).

The proteins corresponding to significant complexes unique to liver and kidney may be tissue discriminatory: Based on the Fragments Per Kilobase Million(FPKM) normalized transcriptome profile across 14 different tissues (Human BodyMap 2.0; http://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-513/) [18], we examined gene expressions corresponding to proteins from significant complexes common to RC_N and CR, and proteins from significant complexes unique to RC_N, and proteins from significant complexes unique to CR. The genes are clustered based on hierarchical clustering (Euclidean distance; average linkage). It appears that when examining shared genes (that code for proteins belonging to common complexes), kidney and liver are closely spaced amongst the various tissue types but when considering unique genes (that code for proteins belonging to tissue-specific complexes), the liver and kidney tissues are more widely spaced apart (Supplementary Figure 2). This observation, together with the earlier observation that significant complexes are conserved with respect to tissue type, allows us to infer that FCS makes biologically relevant predictions, in line with the biological characteristics of the tissue class being examined.

FCS-based cross-examination of technical replicates yields modest recovery of missing proteins

Via FCS, we may determine the extent and significance of recovery based on verification on three strategies: Based on the set of proteins corresponding to all significant PSMs in the same sample (Figure 4A), on the set of proteins corresponding to all significant PSMs in the cross-batch replicate (Figure 4B), and on the union of the set of proteins corresponding to all significant PSMs in the same sample and cross-batch replicate (Figure 4C). The notation in Figure 4, e.g., N T1 -> N T2, means N for normal, T is for technical replicate, the direction of the arrow means we are comparing the proteins recovered based on the significant complexes from sample N T1, and checking them against the proteins identified in N T2. We consider each sample (from patient samples 1 to 6) separately. The results in each cell of Figure 4 are shown as two rows: the top row shows the overlap |r|/|R| and its associated p-value on the left and right respectively (see Materials and Methods). The bottom row shows the total number of predicted missing proteins and the number of verified missing proteins on the left and right respectively.

As additional comparisons, we also verify based on observed proteins (i.e., the finalized set of proteins reported in the proteomics screen for a given sample) in the cross-batch replicate (Supplementary Figure 3A) and verification based on the proteins from significant complexes in the cross-batch replicate (Supplementary Figure 3B). It is useful to discuss these two naïve scenarios first: In the former, recovery is extremely low. Not all recoveries are statistically significant, and verification rate is around 2 to 5% (Supplementary Figure 3A). On the other hand, in the latter where we compare missing proteins predicted

to be present in one replicate against the FCS-significant complexes in the corresponding cross-batch replicate, the overlap shoots up dramatically to ~90% (Supplementary Figure 3B). Although cross-batch replicates do not report the same protein sets, these proteins nonetheless map back generally to the same protein complexes in the same sample. However, both of these recovery verification methods are not robust: In the former, the verification rate is too low to be useful. This is not surprising; otherwise, taking multiple technical replicates would have easily resolved MPP (thus absolving the need for research in this area). Unfortunately, this data tells us that missing proteins tend to be harder to observe/recover generally (see next section). In the latter scenario, we focus on direct verification of significant protein complexes between cross batches, and not on mutually supportive predictions of missing-but-present proteins. But this comparison, naïve as it is, is also useful as it tells us that despite the different reported proteins between technical replicates of the same samples, we nonetheless still predict similar complexes. Although gratifying from the perspective some biological signal is evidently conserved, this does not change the fact that replicates from different samples still report quite a lot of different significant complexes which may not be meaningful (Figure 3A).

For verification of predictions of missing-but-present proteins, the PSM list (where proteins with at least one representative peptide are listed) is used for determining whether there is evidence that a predicted missing protein is indeed present. Interestingly, despite the differences in observed proteins, self-recovery and cross-batch replicate recovery have similar results of ~20% recovery rate (Figures 4A and 4B). The cross-batch replicate recovery rates are slightly higher however.

Taking the union of the PSM lists from self and cross-batch replicate increases verification rates modestly from ~ 20% to ~25%; cf. Figure 4C. Although this gives rise to an appreciable improvement of 25% (i.e., 25 − 20 over 20), verification rates are still low. Apparently, where RC is concerned, most predicted missing proteins (~75%) cannot be verified in this manner due to the lack of any supporting PSMs. However, we think there may be a silver lining. In particular, we expect that given more technical replicates and more support, it is possible to improve recovery beyond 25% (as rarer PSMs become observable), although we cannot say by how much more, whether the recovery proportion can become predictable as a function of replicate size, or whether recovery proportion predicted on one dataset is generalizable to other tissues/datasets.

Peptide support is a stronger contributing component towards missing proteins than low abundance

Low abundance is frequently cited as a cause for MPP [19]. The reasoning for this stems from the semi-stochastic loss of proteins in Data-Dependent Acquisition (DDA) paradigm proteomics screens where smaller signals corresponding to low abundance are more likely overlooked. However, low abundance cannot be attributed as a strong or sole contributing factor for the missing proteins observed in the RC dataset where even at higher abundance levels, missing proteins exist nonetheless (Supplementary Figure 1A). The observation that missing proteins are also frequent at high abundance levels is surprising

but also reported before by Webb-Robertson et al in 2015 [19]. Moreover, for relatively high-abundance proteins (greater than the median expression level), there does not appear to be any difference for missing values below or above the median missing-value level (Supplementary Figure 1A). Hence, an alternative explanation is needed to better understand why missing proteins occur.

In the Data Independent Acquisition (DIA)-derived paradigm, there is no semi-stochastic preselection of precursor peptides based on signal intensity, all spectra are captured if it falls within detection limit. The lack of association between low-abundance proteins and increased missing values (Supplementary Figure 1A) is consistent with the nature of DIA, and perhaps is an artifact associated with the older DDA paradigm (higher-intensity precursor spectra tends to be selected for identification, creating the correlation between low abundance and non-detection). Instead, in DIA, we find that low-confidence PSMs and low peptide support for proteins are generally stronger contributing factors towards MPP. Figure 4D shows the distributions of peptide support for Internal (Observed proteins), Recovered (Verified proteins) and External (Proteins that were neither observed nor predicted to be missing in the cross-batch replicate). Observed proteins tend to have the highest peptide support while predicted missing proteins (expected to be present), has relatively lower peptide support. However, unpredicted proteins not observed in the cross-batch replicate have the least peptide support. It is plausible that proteins with lower peptide support may not consistently meet the statistical threshold required when converting PSMs (based on peptides) to the finalized observed protein list, and this leads towards MPP (and data holes in the observed protein expression matrix).

The results (Figure 4D) also boost credibility for complex-based missing protein prediction, since the recovered proteins based on significant complexes are more enriched for higher peptide support than those not predicted to be recoverable at all.

Unverified predicted missing proteins may not exist in tissue in first place

Without prior knowledge on all protein complex families in CORUM, it is difficult to conclude that highly overlapping but significant complexes are contributing to a good number of non-verifiable proteins. While CORUM is a manually curated database concerned with the annotation of biologically relevant complexes, it does not provide a convenient way of ascertaining which complexes belong to the same major family and have tissue-specific properties. For example, the nBAF and npBAF complexes have many similar components but are found in different tissues [20]. In that regard, condensing complexes based on shared components also does not give rise to biologically coherent entities [21].

In the absence of tissue-specific information allowing us to only consider kidney tissue-specific complexes, we concocted a simple check as we believe that since most proteins at the smallest FCS p-values are considered important, it is possible that tissue specificity (of complexes) may contribute towards some degree of non-verification (i.e., we are considering irrelevant complexes that are significant because of deep sharing of core proteins with a tissue-specific relevant complex). This test is important:

If proven correct, then it means we are severely underestimating the recovery rates based on networks because of tissue-specificity issues.

Using sample N1 and its peptide list derived from both its technical replicates, we have a total of 62 unverified proteins, and 557 (observed + verified) proteins. We mapped each of the 62 unverified proteins to the largest complex it is a component of and generated an observed overlap with median of 0.32. Given 1,000 randomized median overlaps, only 7 times were the randomized medians greater. Thus, the empirical p-value is 0.007, indicating strong support for enrichment of observed + verified proteins in the complexes where the unverified proteins are found. Running the same test on sample N2 also reveals similar results with a p-value of 0.008.

Hence, there is some evidence that these unverified proteins belong to some tissue-specific complex variant absent in the tissue sample. But this evidence may also be a bit circular in the sense that the FCS p-value of a protein is correlated with a high fraction of the complex's member proteins being present as well.

This tells us that incorporating all complexes simultaneously without regard for their tissue specificity or the presence of other same complex family members is giving rise to a large proportion of unverified proteins. It also suggests that we may be underestimating the verification rates of our predicted missing proteins as we are predicting proteins that should not be in the tissue in the first place. This finding suggests that perhaps more work should be put into building tissue-specific complexomes for more powerful network analytics.

Discussions

FCS-significant complexes are biologically meaningful

Although we reported previously that FCS predicted MPs have the highest recovery rates against other network methods, we have not evaluated its use as a single sample profiling method. In this case, whether the set of networks significant to a tissue, are functionally relevant and idiosyncratic to itself. This is the first time we have demonstrated that samples from the same tissue class tend to exhibit higher levels of shared complexes (Figure 3). It is more impressive that this similarity is conserved even when the observed proteins are different.

Correlations within same-class samples are high: Cross-comparisons (as indicated by a bi-directional arrow, <->) with a second tissue type (Colorectal; CR tissue) shows significantly lower Jaccard indices (t-test; p-value << 0.01 ***). C: Many complexes are shared between different tissues. Despite the significantly lowered Jaccard indices indicating lower significant complex agreement, a large proportion of complexes are still shared amongst different tissues resulting in generally low hypergeometric p-values. However, unshared complexes and their constituent proteins appear to exhibit high tissue-specificity

Tissue-specificity amongst complexes necessitates the development of complex families as a refinement for FCS

There is a limit to the proportion of verifiable missing proteins (Figure 4C). Aside from the lack of any PSMs for these unverified proteins, another contributing reason may be biological: Some complexes share common components with one another (the core proteins) and differ by a few peripheral proteins. The peripheral proteins modulate the function of the complex, and in some cases, lead to tissue differentiation. Thus, certain complex forms are found only in certain tissues or cell types [22]. For instance, the BAF (BRG1-associated factors) family of complexes, which acts as a switch in neuronal differentiation, exists in two forms; neural progenitor (npBAF) and matured (nBAF) [23]. As neural progenitor cells exit mitosis and differentiate into neurons, npBAF complexes, which contain ACTL6a and PHF10, are exchanged for ACTL6b and DPF1/DPF3 subunits in nBAF [22,23]. While the core remains constant, the exchange of peripheral proteins is involved in differentiation, and so, cannot be present together in the same tissue. In FCS, there is no distinction between core and peripheral proteins. So, suppose we have a family of complexes with common core, all will be reported as significant if there is substantial overlap of the family's core proteins with the observed protein list. But since not all the complexes in the family can form in the same tissue, we should expect that some of the predicted missing proteins to be non-verifiable (the predicted complex does not form in this tissue; and so, its peripheral proteins do not exist in this tissue as well).

Many unverified predicted MPs may be due to tissue-specificity issues. That is, the complex being considered does not exist in that tissue. We find it essential to develop a classification system of existing complexes into families so that we do not consider complexes in a tissue where we know it is not normally present. This in turn, should improve the verification rate of predicted missing proteins. For example, if 40% of the predicted missing proteins were false positives due to complex families and the use of complex family information was able to eliminate these, the 20-25% recovery rate observed earlier in replicates would become 33-42%.

Limitations of this study and future work

This is a study demonstrating the potential of using networks for missing-protein prediction. We also laid down the strategical frameworks on how to recover these missing proteins using other supporting data, including gene expression, cross-replicate information, and peptide information. However, such strategies are purely inferential, and does not directly validate the existence of these missing proteins. Confirming our predictions using biological validation methods would be a logical next step. Moreover, the protein complex vector and nature of proteome screen imposes limits on which proteins we can predict or verify.

The complex network feature vector is used as is. We trusted its quality based on prior studies. But certainly, future work can be performed to improve its information value, e.g., developing tissue-specific complexomes.

Hence, in our future work, we intend to perform further evaluative work on the impact of network feature vector quality. We would also like to expand the feature vector to include more complexes and generate tissue-specific network feature vectors to see if further improvements to recovery is attainable. We would also like to perform deeper studies into how FCS's parameters can be tweaked towards better performance [24].

Conclusions

Recovery of missing proteins is a persistent problem that remains unsolved in proteomics. Using protein complexes and FCS, we have updated our analysis to include the latest DIA paradigm and designed a more sophisticated recovery-benchmarking scheme based on cross-batch replicates and proteins from the full PSM list.

Declarations

Funding

This research is supported by the Ministry of Education, Singapore, under its Academic Research Fund Tier-1 (RG35/20).

Author contribution

BJHW and WJK implemented the analyses and coordinated the project. LW and WWBG supervised and co-wrote the manuscript.

Competing interests

The authors declare no conflicting interests, financial or otherwise.

References

1. Guo, T. *et al.* Rapid mass spectrometric conversion of tissue biopsy samples into permanent quantitative digital proteome maps. *Nature medicine* **21**, 407–413, doi:10.1038/nm.3807 (2015).
2. Gillet, L. C. *et al.* 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**, O111 016717, doi:10.1074/mcp.O111.016717 O111.016717 [pii] (2012).
3. McAlister, G. C. *et al.* Increasing the multiplexing capacity of TMTs using reporter ion isotopologues with isobaric masses. *Analytical chemistry* **84**, 7469–7478, doi:10.1021/ac301572t (2012).
4. Zhao, P., Zhong, J., Liu, W., Zhao, J. & Zhang, G. Protein-Level Integration Strategy of Multiengine MS Spectra Search Results for Higher Confidence and Sequence Coverage. *J Proteome Res* **16**, 4446–4454, doi:10.1021/acs.jproteome.7b00463 (2017).

5. Li, J. *et al.* Network-assisted protein identification and data interpretation in shotgun proteomics. *Molecular systems biology* **5**, 303, doi:10.1038/msb.2009.54 (2009).
6. Goh, W. W. & Wong, L. Design principles for clinical network-based proteomics. *Drug Discov Today* **21**, 1130–1138, doi:10.1016/j.drudis.2016.05.013 (2016).
7. Goh, W. W. & Wong, L. Integrating Networks and Proteomics: Moving Forward. *Trends in biotechnology* **34**, 951–959, doi:10.1016/j.tibtech.2016.05.015 (2016).
8. Goh, W. W. & Wong, L. Advancing clinical proteomics via analysis based on biological complexes: A tale of five paradigms. *Journal of Proteome Research* **15**, 3167–3179, doi:10.1021/acs.jproteome.6b00402 (2016).
9. Goh, W. W., Sergot, M. J., Sng, J. C. & Wong, L. Comparative network-based recovery analysis and proteomic profiling of neurological changes in valproic Acid-treated mice. *J Proteome Res* **12**, 2116–2127, doi:10.1021/pr301127f (2013).
10. Pavlidis, P., Lewis, D. P. & Noble, W. S. Exploring gene expression data with class scores. *Pac Symp Biocomput*, 474–485 (2002).
11. Rost, H. L. *et al.* OpenSWATH enables automated, targeted analysis of data-independent acquisition MS data. *Nat Biotechnol* **32**, 219–223, doi:10.1038/nbt.2841 (2014).
12. Goh, W. W. & Wong, L. NetProt: Complex-based Feature Selection. *J Proteome Res* **16**(**8**), 3102-3112, doi:10.1021/acs.jproteome.7b00363 (2017).
13. Zhang, B. *et al.* Proteogenomic characterization of human colon and rectal cancer. *Nature* **513**, 382–387, doi:10.1038/nature13438 (2014).
14. Goh, W. W. & Wong, L. Evaluating feature-selection stability in next-generation proteomics. *Journal of bioinformatics and computational biology* **14**, 16500293, doi:10.1142/S0219720016500293 (2016).
15. Fraser, H. B. & Plotkin, J. B. Using protein complexes to predict phenotypic effects of gene mutation. *Genome biology* **8**, R252, doi:10.1186/gb-2007-8-11-r252 (2007).
16. Soh, D., Dong, D., Guo, Y. & Wong, L. Finding consistent disease subnetworks across microarray datasets. *BMC Bioinformatics* **12** S15, doi:10.1186/1471-2105-12-S13-S15 1471-2105-12-S13-S15 [pii] (2011).
17. Goh, W. W., Wang, W. & Wong, L. Why batch effects matter in omics data, and how to avoid them. *Trends in biotechnology* **35**(**6**), 498–507, doi:10.1016/j.tibtech.2017.02.012 (2017).
18. Asmann, Y. W. *et al.* Detection of redundant fusion transcripts as biomarkers or disease-specific therapeutic targets in breast cancer. *Cancer Res* **72**, 1921–1928, doi:10.1158/0008-5472.CAN-11-3142 (2012).
19. Webb-Robertson, B.-J. M. *et al.* Review, Evaluation, and Discussion of the Challenges of Missing Value Imputation for Mass Spectrometry-Based Label-Free Global Proteomics. *J Proteome Res* **14**, 1993–2001, doi:10.1021/pr501138h (2015).
20. Staahl, B. T. *et al.* Kinetic analysis of npBAF to nBAF switching reveals exchange of SS18 with CREST and integration with neural developmental pathways. *The Journal of neuroscience: the*

official journal of the Society for Neuroscience **33**, 10348–10361, doi:10.1523/JNEUROSCI.1258-13.2013 (2013).

21. Wu, M. *et al.* Benchmarking human protein complexes to investigate drug-related systems and evaluate predicted protein complexes. *PLoS One* **8**, e53197, doi:10.1371/journal.pone.0053197 (2013).
22. Goh, W. W., Oikawa, H., Sng, J. C., Sergot, M. & Wong, L. The role of miRNAs in complex formation and control. *Bioinformatics* **28**, 453–456, doi:btr693 [pii] 1093/bioinformatics/btr693 (2012).
23. Yoo, A. S., Staahl, B. T., Chen, L. & Crabtree, G. R. MicroRNA-mediated switching of chromatin-remodelling complexes in neural development. *Nature* **460**, 642–646, doi:10.1038/nature08139 (2009).
24. Zhao, Y., Sue, A. C. & Goh, W. W. B. Deeper investigation into the utility of functional class scoring in missing protein prediction from proteomics data. *Journal of bioinformatics and computational biology* **17**, 1950013, doi:10.1142/S0219720019500136 (2019).

Figures

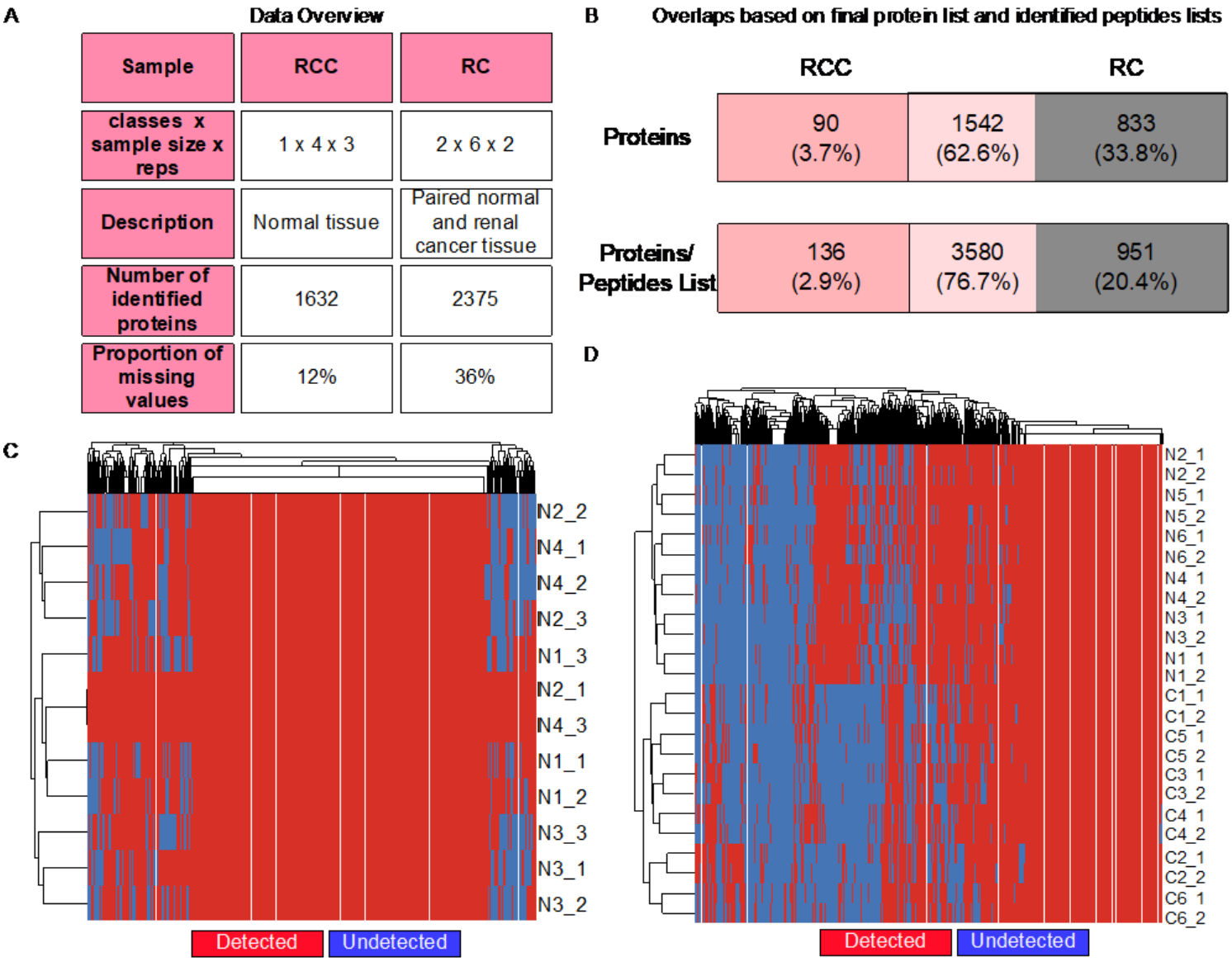

Figure 1

Characteristics of analyzed datasets. A: Broad Overview (Renal Cancer, RC; Renal Cancer Control, RCC). It is noteworthy that there are many more data holes, and therefore missing proteins, in RC than in RCC. B: Protein and peptide agreements between RC and RCC. Each row shows proteins unique to RCC (dark pink), shared between RCC and RC (light pink) and unique to RC (white). The top row shows only proteins in the finalized observed protein list while the bottom includes all proteins with at least one unique peptide. C: Distribution of detected and missing proteins in RCC. Hierarchical clustering suggests that the distribution of data holes across samples is not batch related (samples are named by Class/Biological Replicate/Technical Replicate; e.g., N2_2 means "Normal" class sample 2, batch 2). D: Distribution of detected and missing proteins in RC. Hierarchical clustering suggests that the distribution of data holes across samples is not batch related (Class/Biological Replicate/Technical Replicate). While approximately 20% proteins are consistently observable across samples, most proteins have missing values meaning they are not observed in a subset of samples. Proteins more prone to exhibit missing behavior cannot be solely explained by low abundance (cf. Supplementary Figure 1).

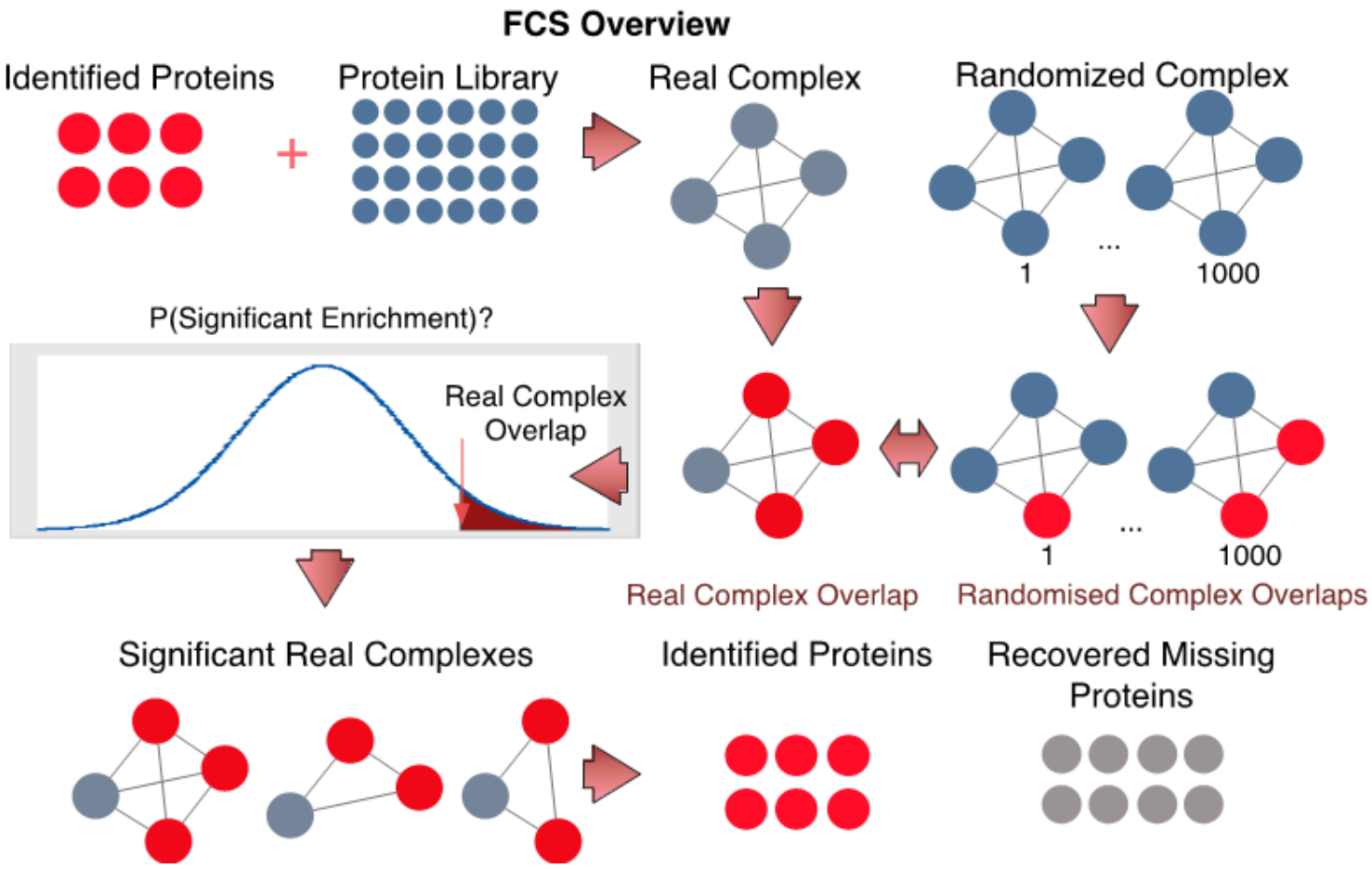

Figure 2

How Functional Class Scoring (FCS) works. In FCS, given a set of observed proteins and a real protein complex, an observed overlap is defined. To determine if this observed overlap is significant, randomized complexes equal to the size of the original complex are generated 1,000 times (using a pool of proteins found in complexes), and a null overlap calculated each time. The null overlaps form a null distribution allowing an empirical p-value to be defined as the proportion of null overlaps greater than or equal to the observed overlap.

Figure 3

FCS-significant complexes are tissue-specific. A: Distribution of FCS p-values across samples for renal cancer control (RCC) and renal cancer normal (RC_N) and cancer (RC_C) tissues. Even in the same tissue class, individual samples report variable complexes as significant. B: Pairwise tissue similarity based on overlaps of significant complexes. Samples from same tissues class tend to exhibit higher levels of shared complexes (Pairwise Jaccard Index) even when the observed proteins are different. Cross-comparisons (as indicated by a bi-directional arrow, <->) with a second tissue type (Colorectal; CR tissue) shows significantly lower Jaccard indices (t-test; p-value << 0.01 ***). C: Many complexes are shared between different tissues. Despite the significantly lowered Jaccard indices indicating lower significant

complex agreement, a large proportion of complexes are still shared amongst different tissues resulting in generally low hypergeometric p-values. However, unshared complexes and their constituent proteins appear to exhibit high tissue-specificity (cf. Supplementary Figure 2)

Figure 4

A: Missing-protein verification for RC based on various strategies. A: Missing protein verification based on full peptide list of self. The predicted missing proteins are verified on a peptide list originating from the sample itself (The notation N T1 -> N T1 refers to a normal class sample of technical replicate 1 on the left side. The arrow sign is the direction of verification and the N T1 on the right side refers to the peptide list from N1 T1 itself). The four elements in each box are the overlaps and its corresponding p-value on the top left and right. The bottom left and right are the total number of predicted missing proteins and the number of verified missing proteins. Significant verifications (p-value ≤ 0.05) are shaded in pink. B: Missing-protein verification based on full peptide list of second technical replicate. Similar to A, but the missing protein verification is performed on the second technical replicate instead. Verification rates are slightly higher but as with A, all verification rates are significant. C: Missing-protein verification based on union of full peptide list from first and second technical replicates. Taking the union of the peptide lists increases verification rates from ~ 20% to ~25% but this is still relatively low. Most missing proteins cannot be verified in this direct manner. D: Peptide support across various protein categories (I; Identified refers to observed proteins in the proteomics screen, R; Recovered refers to predicted missing proteins, E; External refers to proteins that are neither observed nor predicted as present in the cross-batch replicate). Identified (Observed) proteins tend to have higher peptide support across the board compared to predicted missing proteins. Unreported proteins that are not predicted as missing and are not observed in the cross-batch replicate tend to have lowest peptide support generally.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Supportinginformation.docx