Evaluation of Different Methods For False Discovery Rate Estimation In Large-Scale Proteomics Data

Anabel Yong, B164439

University of Edinburgh Biochemistry Honours Dissertation

(under supervision of Dr. Georg Kustatcher)

15th May 2023

Abstract

Proteomics is a powerful tool for identifying proteins in complex biological samples, providing valuable insights into protein function and regulation. Recently, machine learning approaches applied to proteomics have shown great promise for predicting protein function based on peptide sequences[1][2]. The accuracy of these predictions depends on the quality of input data and the ability to accurately identify and quantify proteins. In this study, we refined the data processing pipeline for ProteomeHD2 curation to improve false discovery rate (FDR) estimation, maximizing detection of true proteins while minimizing false-positives using the newly proposed FDR strategy[3]. This new FDR strategy shows a baseline improvement of 56.2% for protein identification and 32.8% for isoform identification when compared to all previously known FDR methods. Practically, the strategy has identified several functionally uncharacterized microproteins in our newly curated ProteomeHD2 dataset. These findings strongly suggest that this new FDR strategy should be the preferred method for all future proteomics datasets.

Dissertation Word Count: 4993 words

Abstract: 153 words

Abbreviations:

FDR: False Discovery Rate; **MS/MS**: Tandem Mass Spectrometry; **SILAC**: Stable Isotope Labelling by Amino acids in Cell culture Quantitation; **TDS**: Target-Decoy Competition Strategy; **PSM**: Peptide Spectrum Match; **pgFDR**: Picked Protein Group FDR (picked_protein_group_no_remap); **ppFDR**: Picked Protein FDR/Savitski FDR (savitski_no_remap); **cFDR**: Classic FDR (classic_grouping_no_remap); **smORFs**: small open reading frames; **nC**: Near Cognate ORFs detected from Ribosome-profiling(Ribo-Seq data); **pI**: smORFs supported by multiple Ribo-Seq data; **sR**: smORFs detected supported by single Ribo-Seq data

Contents

1. Introduction

1.1 Proteomics, ProteomeHD & ProteomeHD2

Proteomics involves the comprehensive investigation of the structure and function of proteins in complex biological samples. This approach holds great promise for gaining insights into the intricate workings of living organisms [1][4][5]. Proteomic methods enable rapid large-scale analysis of the proteome, allowing for detection, identification and functional characterization of proteins[4][5]. Recent advancements in protein sample fractionation and labelling techniques in mass spectrometry (MS) have enhanced the identification of low-abundance proteins[6][7], such as microproteins[2]. Proteomics research aims to identify proteins and protein complexes found in cells grown under different conditions. To accomplish this, tandem mass spectrometry (MS/MS) is commonly applied to determine the protein constituents of complex mixtures[1][6][7]. Briefly, the process involves digesting proteins into smaller peptides with an enzyme, such as trypsin and using chromatography to separate them. The resulting peptides are ionized and fragmented to produce characteristic MS/MS spectra that are used to identify these peptides[8].

ProteomeHD is a recent development in proteomics that aims to create a comprehensive map of the human proteome[2]. This goal is to identify and annotate all protein-coding genes in the human genome and provide a detailed understanding of the functions of encoded proteins. This database of 10,323 human proteins provides high-quality standardized data for use in proteomics research[9]. This resource is particularly useful for large-scale studies that require consistent and accurate data across multiple MS experiments. By using this database, proteomics research can more reliably elucidate protein structure and function in a variety of biological systems. Continuing in this direction, the curation of a ProteomeHD2 database with newly discovered 7264 small open reading frames (smORFs) [10] is currently underway. smORFs are open reading frames with length of less than 100 amino acids, which can be translated to microproteins [10]. This will provide researchers with a comprehensive resource for studying microproteins and their roles in cellular processes. With the ongoing advances in omics technologies and identification of novel microproteins, ProteomeHD2 database is expected to continue expanding. In addition to identifying these new smORFs, an important aspect of our work is ensuring the accuracy and reliability of our findings.

1.2 Target-Decoy Competition Strategy

The current challenges in high-throughput proteomics is the derivation of a list of identified peptides and their corresponding proteins from a large number of MS/MS spectra generated by database search programs such as X!Tandem [11] and SEQUEST[12]. The main task involved is to differentiate between accurate peptide

assignments and false positive identifications among the results of database search. For small datasets, this can be achieved through manual verification of peptide assignments by researchers with relevant expertise. However, this approach is not feasible for large datasets containing more than 10000 mass spectra. To address this challenge, the target-decoy competition strategy (TDS) has been widely used to empirically estimate the false discovery rate (FDR) at a pre-determined threshold (1%)[13][14]. The approach ensures that at most 1% of reported identifications are false positives. FDR is the rate at which false identifications are made among total identifications reported in a proteomics experiment [15]. Despite the critical role of FDR estimation in accuracy of enormous lists of peptide identifications, there is no global consensus in the protoemics field on how to apply TDS to minimize biased FDR estimates [16][17][3].

In TDS strategy, a decoy database is created by either reversing or shuffling the amino acid sequences of the target database[13][18]. The two databases are then searched together and resulting peptide identifications are assigned a score based on their statistical significance[18]. The target and decoy hits are compared and a threshold is set to control the FDR. This holds numerous advantages- including that it provides a way to estimate the number of false positive identifications, which is critical for interpreting results of mass spectrometry experiments. This has been the gold standard for controlling the number of false positive identifications in proteomics[13]. The reliability of protein identification is crucial in high-throughput proteomics, and accuracy of FDR estimation determines the credibility of reported results. Therefore, developing and improving FDR estimation methods has been a substantial focus of many recent studies[16][3][19]. To date, three primary FDR estimation methods have been proposed, explained below: classic FDR (cFDR)[16], picked protein-level FDR/ Savitski FDR(ppFDR) [17] and picked protein group FDR (pgFDR)[3]. However, it is still unclear which of these methods is the most suitable for high-throughput proteomics data analysis.

1.3 False Discovery Rate (FDR) Estimation

TDS has paved the way for the development of various FDR control strategies in proteomics research. FDR control is implemented at peptide-spectrum match (PSM)-level[20], peptide-level[21], or protein-level[21][22]. Briefly, at PSM level, where each identification is an observed spectrum linked to a peptide that is inferred to be responsible for generating the spectrum. At peptide-level, in which multiple PSMs for the same peptide sequence are considered jointly. Third, at protein-level FDR, where evidence to control FDR is by accumulating all peptides associated with the same protein are considered[22]. These 3 methods are graphically explained in Appendix 1.1. For the purpose of this study, we are particularly interested in how the 3 FDR methods, perform at 1% protein-level FDR as protein-level FDR are generally considered to be less biased than at PSM-level or peptide-level[23].

False discovery rate (FDR) is measure of proportion of false positives among all identified peptide-spectrum matches (PSMs) or protein identifications[15]. In cFDR method, all PSMs are considered together, and FDR is calculated based on the number of false identifications compared to total number of identifications. This method holds the assumption that all peptides are equally likely to be identified, regardless of the protein they are assigned to[16]. PpFDR takes into consideration that peptides are not equally to be identified. Instead, it considers only proteins that have at least one peptide confidently identified with 1% FDR [17][22]. This concept, whose benefits were demonstrated in [17] and [22], is to perform a direct competition between each target protein and respective decoy, keeping only the higher score of the two[21][22]. Conceptually, pgFDR groups proteins that share at least one peptide and considers them as a single entity (i.e "protein group"). FDR is estimated based on the number of decoy protein groups among those identified at a 1% FDR. This method takes into account the fact that proteins in the same group are more likely to be identified if one of them is confidently identified[3]. Thus, we compared the performance of these 3 FDR methods at on our manually curated dataset, ProteomeHD2, for the accurate retrieval of true proteins at 1% protein-level FDR to aid in ProteomeHD2 curation essential for future novel protein chracterization and protein-protein interactions by proteomics.

2. Methodology

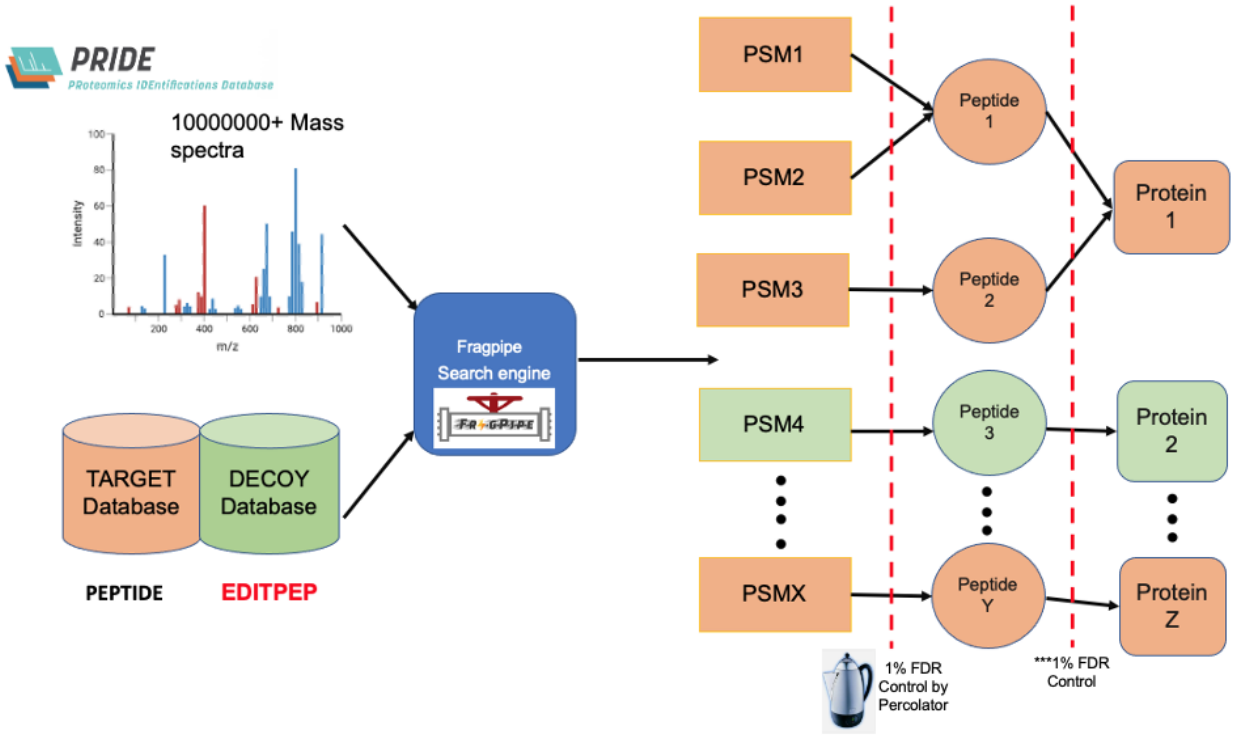


Figure 1: **Experimental Workflow of ProteomeHD2 curation and ***FDR method evaluation:** 170 MS-based proteomics projects were downloaded from PRIDE[24], and searched against a concatenated target-decoy database to FragPipe software [25]. PSMs were then collected and rescored at 1% PSM-level FDR with Percolator software. Peptides are subsequently collected and controlled with 3 aforementioned FDR methods implemented for evaluation at 1% protein-level FDR.

2.1 FragPipe Data Processing & ProteomeHD2 Curation

ProteomeHD2 dataset was curated from a comprehensive set of 170 MS projects, containing a large number of raw files (spectral files from mass spectrometry imaging) downloaded from the publicly available repository of MS experiments, PRIDE[24]. Raw files were as followed in the ProteomeHD protocol[2]. 22796 raw files were processed through FragPipe [25], a software used to identify peptide fragments from MS2-labelled samples. The workflow implemented in FragPipe was based on the SILAC (Stable Isotope Labeling by Amino acids in Cell Culture) protocol[25]. SILAC is a MS-based proteomics technique that involves labelling cells or tissues with different isotopic forms of amino acids[26], allowing for protein identification with high precision across multiple samples throughout the 170 projects. The workflow involves processing raw MS data to identify labelled peptides and matching them to protein sequences in a reference target database as illustrated in Figure 1.

Publicly available 7264 smORFs translated sequences[10] were added into the reference target peptide

database. For each peptide in the database, a custom python script was implemented to create reversed decoy sequences. These were appended to the original ProteomeHD target-decoy peptide database[2]. The curated database was then used as the input for the FragPipe software to search for matching observed PSMs of PRIDE MS runs to the concatenated reference human target-decoy peptide database. These are subsequently collated for Percolator rescoring.

2.2 Percolator PSM rescoring

Percolator is a semi-supervised learning algorithm for peptide identification from shotgun proteomics datasets[27]. Illustrated in Figure 1, Percolator is implemented to control the FDR at 1% for assignment of PSMs to peptide sequences from Fragpipe software. Reliable statistical confidence measures such as q-values and posterior error probabilities (PEP) were assigned to matched PSMs from FragPipe. Subsequently, PSMs were ranked by their score at 1% PSM-level FDR filtering by Percolator. Q-values are adjusted p-values that consider the multiple hypothesis testing problem as identifying PSM matches involves testing multiple hypothesis simultaneously [28][29]. For each spectrum, there are potentially many candidate peptide sequences that could match, and for each candidate sequence, multiple possible modifications and charge states must be considered. PEP measures the probability the observed PSM is incorrect[29][30]. For the purpose of this study, a q-value of 0.01, for example, indicates that the identification is expected to have a 1% protein-level FDR which is the threshold researchers can confidently claim that a particular protein has been identified in the sample.

Percolator was applied separately to each MS project to improve accuracy of PSM assignments because the properties of respective PSMs can vary between different projects. These include properties such as length of matched peptide sequences as shorter peptides may be more difficult to identify accurately compared to longer peptides. Mass accuracy of measured fragment ions can vary, with lower mass accuracy resulting in more false positive identifications. Percolator results were extracted on 1% PSM-level FDR and only the best scoring PSMs per peptide sequence were retained. The resulting peptide lists were combined into a single list based on the -log10(q-value) of the peptide score, and only the best PSM for each peptide was kept, as shown in Appendix 1.2. These PSMs were filtered at 1% FDR and ranked accordingly. PSMs were generated in the format of a Percolator output file, which the respective columns: PSM IDs, scores, q-value and posterior error probability (PEP) and the protein IDs.

2.3 Protein-Level Filtering

Protein-level FDR estimation methods often involve providing a list of PSMs without applying a peptide-level FDR threshold[17][31]. Many current studies adapt both PSM-level filtering and protein-level FDR filtering only[17][3][32] without combination of peptide-level filtering. This approach enables the estimation of protein-level FDRs even for proteins with limited evidence, such as protein isoforms that typically have a small number of unique peptides[3]. These PSMs controlled at 1% FDR with Percolator, were subsequently analysed by the 3 respective FDR estimation methods: cFDR, ppFDR, pgFDR respectively.

Before executing the python package for 3 FDR methods, a custom R script was developed to create a concatenated database of target and decoy sequences of matched PSMs detected and generated by Percolator. Consequently, with the same R script, the curation of peptide-to-protein mapping database was to investigate protein-level FDR filtering. This provided a manual list of protein IDs (i.e. sp| P15289) and their respective peptide sequences for both target and decoy peptide sequences. sp represents to their unique fasta IDs[33], and P15289 refers to a UniProt accession IDs [34]. Decoys were annotated as REV_sp. This protein mapping database was implemented as Percolator does not automatically map the peptides observed by PSMs to proteins [27].

A python package [3], deposited in Github ((Picked Group FDR)), was implemented in order to run the FDR estimation methods through Terminal command line. 3 FDR methods were available in this python package, by manipulating the command line to call the respective methods which were available in .toml format. FDR methods: cFDR, ppFDR and pgFDR are named as classic_grouping_no_remap, savitski_no_remap and picked_protein_group_no_remap respectively[3]. The methods were called "no_remap" due to the curation of peptide-to-protein mapping database. This package generates a "proteinGroups.output" formatted showing different Uniprot Protein IDs [35] identified from PSMs generated form Percolator, and relevant features such as best peptide sequence matched to Protein IDs, q-value and PEP score. These output files were subsequently analysed for post-processing analysis in regards to evaluating FDR method performance. Additionally, new methods were generated by manipulating the python scripts (.toml files) in this package, in order to investigate reasoning behind each FDR method's performance at 1% protein-level FDR.

2.4 Computation, Data Visualisation & Statisical Validation

Manipulated parameters and command line options specified for each FDR method available in Appendix 1.3 and 1.4 respectively. Raw files were subsetted to show cumulative FDR performance in ProteomeHD2 dataset. Graphs were reproduced via (Jupyter Notebook). All python and R scripts for FDR method eval-

uation are accessible through Github repository(FDR EVALUATION WITH PROTEOMEHD2). In order to validate the performance of FDR methods, random sampling was used by extracting random Percolator output files in triplicates. Each method was executed in triplicates for increased reliability.

3 Results

3.1 pgFDR retrieves more target proteins at 1% protein-level FDR

With ProteomeHD2 dataset, it is illustrated in Figure 2D, that pgFDR retrieves the most number of target proteins at 1% FDR. It scaled exponentially more than ppFDR and cFDR at lower raw files, and the saturation of protein detection maintained approximately 2-fold of the performance of the other two FDR methods (see Appendix 1.6). Simultaneously, when extracting 1000 random MS files generated, it can be illustrated in Figure 2E that the average target protein detection at 1% FDR is significantly higher than ppFDR and cFDR, where it demonstrated a 56.2% and 66.2% increase in the detection of proteins at 1% FDR respectively.

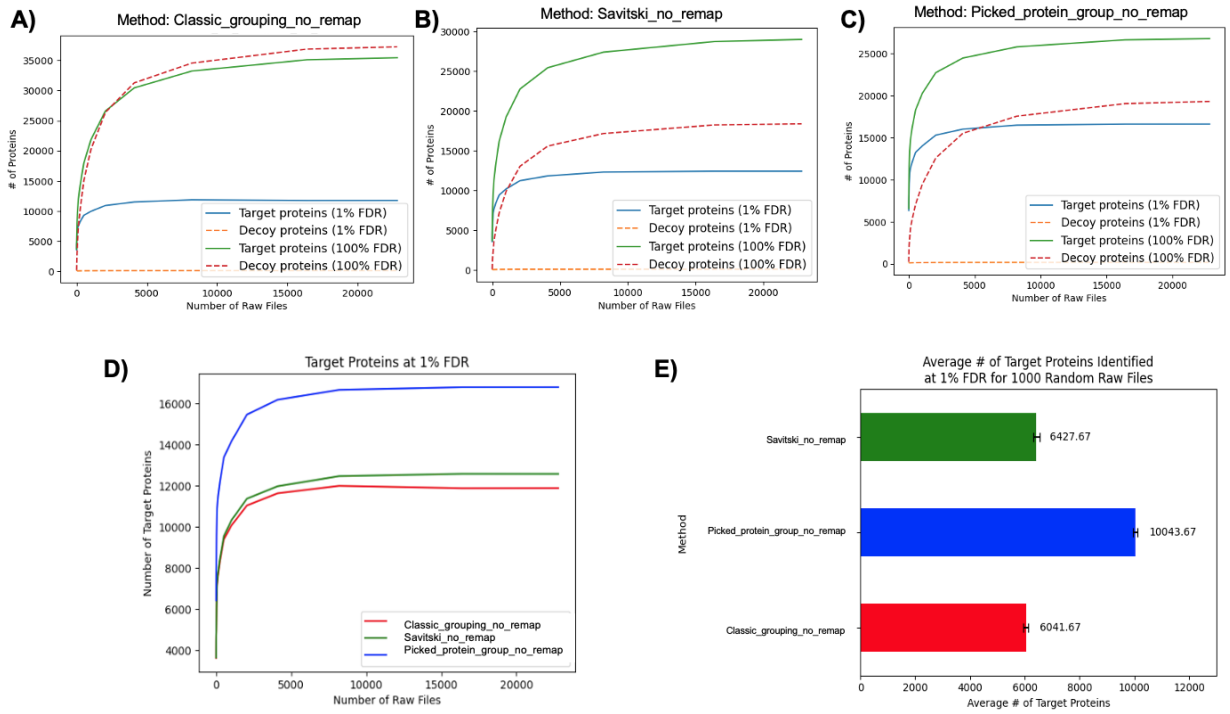


Figure 2: **pgFDR method's performance is greater than cFDR and ppFDR in detecting target proteins at 1% and 100% protein-level FDR after cumulatively aggregating ProteomeHD2;** Breakdown of **A) cFDR**, **B) ppFDR** and **C)pgFDR** we cumulatively aggregated (22796 raw files) respectively filtered at 1% Protein-level FDR and calculated number of target and decoy proteins at 1% & 100% FDR at each step **D)** Number of target proteins identified at 1% FDR over ProteomeHD2 dataset of respective FDR methods, number of target proteins saturated quickly at around 5000 raw files **E)** Average target proteins detected at 1% FDR of 1000 MS raw files

At 1% FDR, pgFDR method identified the highest number of target proteins, followed by ppFDR and cFDR at 22796 raw files of ProteomeHD2 as shown in Figure 2D. This result demonstrates the superior

performance of pgFDR in retrieving more target proteins at 1% FDR.

3.2 pgFDR scales to large proteomic datasets

Investigating the number of target protein detection was not sufficient to conclude the performance of the respective methods. Therefore, we evaluated if pgFDR would scale to larger proteomic datasets. We applied the cFDR, ppFDR, and pgFDR to the cumulatively aggregated ProteomeHD2, leading to approximately 250 million PSMs at 22796 raw files. The observation shown in Figure 2A was consistent with a recent proteomics paper by Savitski et al (2015)[17], which utilized a smaller proteomics dataset. Figure 2A demonstrates that as more experiments are appended, number of target proteins detected by cFDR approach saturates quickly, with more false positives detected than true positives at 100% FDR. In contrast, number of decoys detected by ppFDR and pgFDR never exceeds the number of target proteins at 100% FDR, as shown in Figure 2B-C, respectively. At 1% FDR, pgFDR identified the highest number of target proteins, followed by ppFDR and cFDR as shown in Figure 2D. These results suggest that pgFDR method can relatively scale to larger proteomic datasets, compared to the other FDR methods.

3.3 Why does pgFDR outperform other FDR strategies substantially?

Instinctively, we investigated the reasoning behind why pgFDR performs approximately 2-fold better than cFDR and ppFDR. After benchmarking the python package for the pgFDR method, we standardised the parameters for each of the 3 respective FDR methods. The 3 FDR methods vary in 2 different parameters which are the grouping [36][[37] and target-decoy competition (TDS) strategies [13][14][15]. There are two uniform parameters which are ScoreType and sharedPeptides [3].

Grouping Strategy	**TDS Strategy**	**Method Name (.toml)**	**FDR Method**
Subset grouping	Classic TDS	classic_grouping_no_remap	cFDR[3]
No grouping	Picked TDS	savitski_no_remap	ppFDR [3]
Rescued subset grouping	Picked group TDS	picked_protein_group_no_remap	pgFDR[3]

Table 1: **Different Constellation of Parameters for respective FDR methods:** Grouping strategy represents the different common practices to group proteins as peptides have shared identifications [36][37][38] and TDS strategy indicates target-decoy competition step before protein-level FDR estimation[13][38][39][40]

During investigation of different Grouping Strategy and TDS Strategy parameters, it was found that there are 3 types of grouping methods: no grouping, subset grouping, and rescued subset grouping(rsG). While grouping small peptides from the same protein can help in protein identification (subset grouping)[36], low-quality peptides can cause a protein group to split into multiple groups, which makes it difficult to

accurately identify proteins[37]. To address this issue, the rsG strategy was developed[3], which involves a two-step procedure. Briefly, rsG strategy involves a two-step procedure to filter out low-quality peptides based on a % FDR threshold at the protein-group level FDR. This produces a list of protein groups that is supplemented with any groups from the first list that did not contain proteins in the second list[3]. As for TDS Strategy, there were the classic TDS, picked (protein-level) TDS and picked group TDS. Various combinations of the 2 parameters developed six additional methods which aid in the investigation of why pgFDR outperforms other methods. The classic TDS strategy in cFDR only considers the protein group with the highest score, while the picked group TDS strategy utilizes all proteins in a group that have an FDR less than the 1% threshold [17][3].

These methods and parameter combinations were tested to evaluate the reasoning behind why pgFDR detects more target proteins and identify the most effective parameter and method combination for accurate protein identification. When evaluating each method's performance, it is essential to strike a balance between number of target protein identification at 1% FDR and decoy accumulation that occurs with the increase in number of raw files. As shown in Figure 3A-C, for classic TDS strategies, it is reasonable to infer that performance is poor as the number of decoy identifications increases as the number of raw files in ProteomeHD2 cumulative aggregated, as compared to Figure 3E-I with picked TDS and picked group TDS. When comparing different grouping strategies for pciekd group TDS strategy, shown in Figure 3G-I, it is reasonable to infer the number of targets at 1% FDR identified is relatively higher with the rsG strategy as the number of targets at 1% protein-level FDR is approximately 17000 proteins, more than the subset grouping and no grouping strategy. This behaviour is also observed with Figure 3D-F and Figure 3A-C respectively. Therefore, it is plausible to deduce that the grouping strategy plays a significant role in FDR method performance.

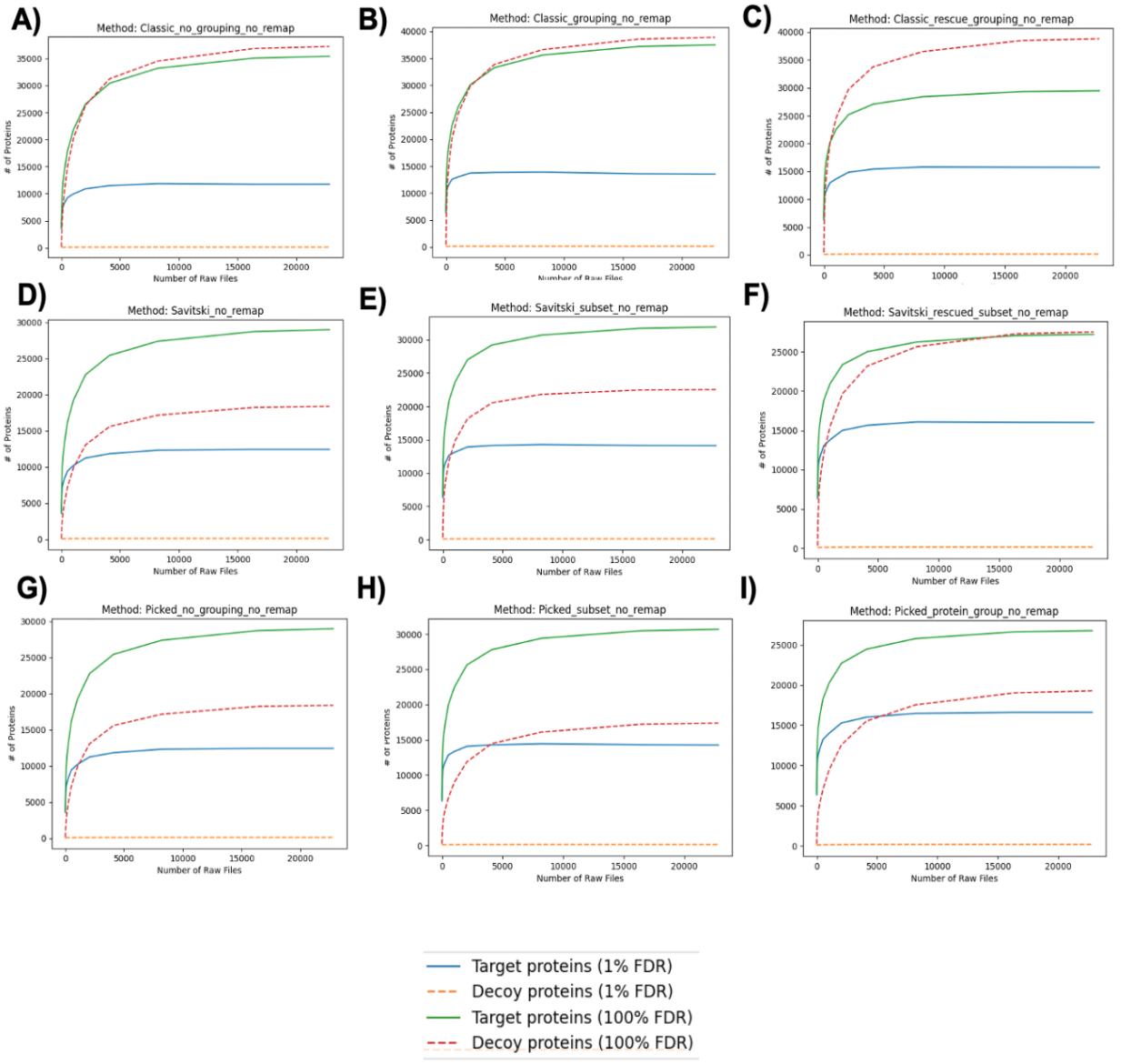


Figure 3: **Performance of B) cFDR, D) ppFDR, I) pgFDR and 6 additional combinatorial methods in detecting target proteins at 1% FDR and 100% protein-level FDR after cumulatively aggregating ProteomeHD2; Breakdown of Classic TDS-(Fig 3A-C), Picked TDS-(Fig 3D-F), Picked group TDS- (Fig 3G-I) and their respective grouping strategies** A) Classic TDS-no grouping (Classic_no_grouping_no_remap) B) Classic FDR (Classic_grouping_no_remap) C) Classic TDS-rsG (Classic_rescue_grouping_no_remap) D) PpFDR (savitski_no_remap) E) Picked TDS-subset grouping (picked_subset_no_remap) F) Picked TDS-rsG (savitski_rescued_subset_no_remap) G) Picked group TDS- no grouping (picked_no_grouping_no_remap) H) Picked group TDS- subset grouping(picked_subset_no_remap) I) pgFDR (picked_protein_group_no_remap)

3.4 pgFDR shows higher sensitivity to ProteomeHD2 containing isoforms

Additionally, as omics technology has advanced, the issue of protein isoforms has become more challenging, with multiple peptide sequences being shared between multiple proteins isoforms as a result of alternative splicing[41][42]. To address this issue, protein grouping that share identified peptides into a single entity has become a popular approach[36][38]. Isoforms are protein variants which have the same amino acid sequence but differ in their post-translational modifications (PTMs) or splicing patterns[41]. These differences can lead to variations in protein function, protein localisation and interactions with other molecules[42]. From Figure 4A below, it is illustrated that pgFDR shows an increase in number of isoforms retrieved as raw files aggregate. It has constantly maintained its performance in the protein isoform detection at 1% FDR and contributes to why pgFDR detects more target proteins. In Figure 4B, pgFDR retrieves the highest number of isoforms when compared to cFDR, ppFDR method and 6 respective differential combination of methods. Additionally, from Figure 4C-D, Classic TDS- and Picked TDS- rsG methods retrieved less isoforms than Classic TDS- and Picked TDS- subset grouping/no grouping methods. However, in Figure 4E, pgFDR, with rsG strategy, retrieved most isoforms, when compared to its respective different grouping methods. Therefore, it is reasonable to infer that pgFDR, from Table 1 and Figure 4E, is the optimized method in retrieving isoforms at 1% FDR.

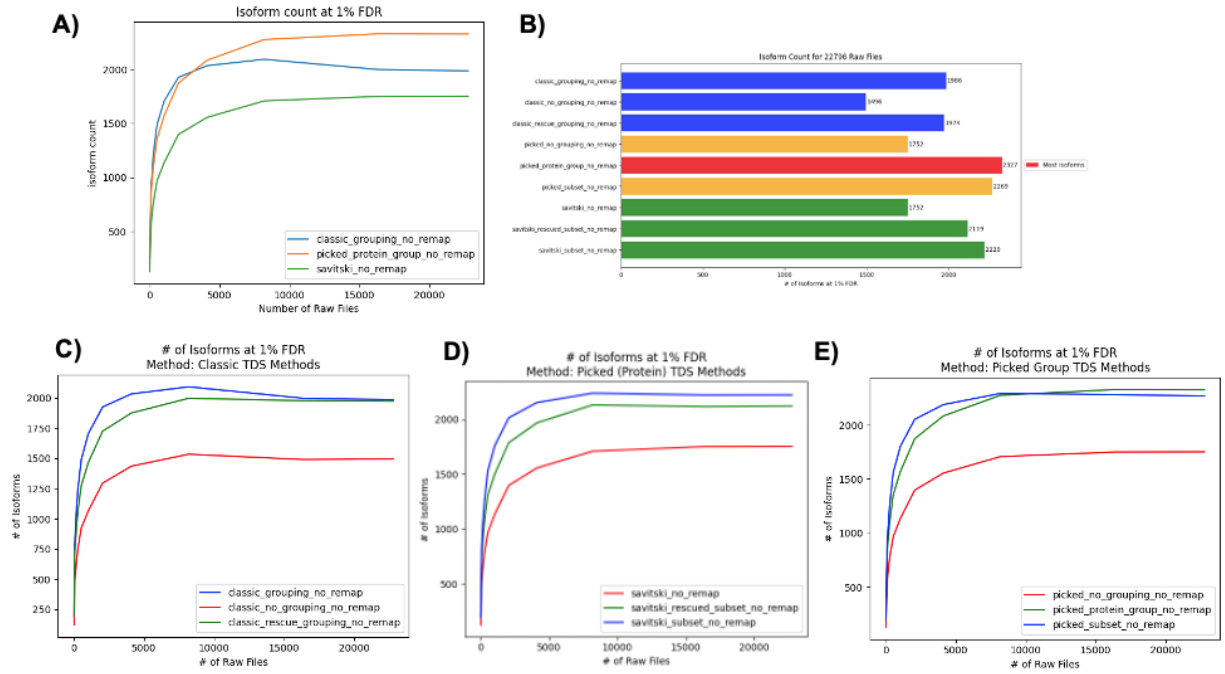


Figure 4: **pgFDR is the optimal method at detecting isoforms at 1% FDR.** A) Number of true isoforms retrieved by cFDR, ppFDR and pgFDR at 1% FDR, B) Average count of isoforms detected at 1% FDR for whole Proteome HD2. 9 TDS-grouping strategies performance at retrieving isoforms at 1% FDR for C)3 Classic TDS-grouping strategies, D) 3 Picked TDS-grouping strategies E) 3 Picked group TDS- grouping strategies respectively

3.5 RSG grouping strategy plays a significant role in pgFDR performance

In Figure 4C-E, the methods containing either "subset grouping" and "rescued subset grouping" are able to retrieve more isoforms. Conceptually, by grouping together proteins with shared peptides, protein grouping approaches increase the sensitivity of protein identification and provide more confidence in the identification of specific isoforms[41][42]. Therefore, we wanted to validate if the grouping strategy played a significant role in retrieval of protein groups. For the purpose of this study, protein groups are defined as a collection of proteins that share at least one peptide sequence. Protein groups are used to deal with the issue of shared peptides between multiple proteins, as identifying peptides to specific proteins can be difficult[3][41][42]. Therefore, we investigated whether pgFDR collected more protein isoforms because there are more target protein groups retrieved at 1% FDR.

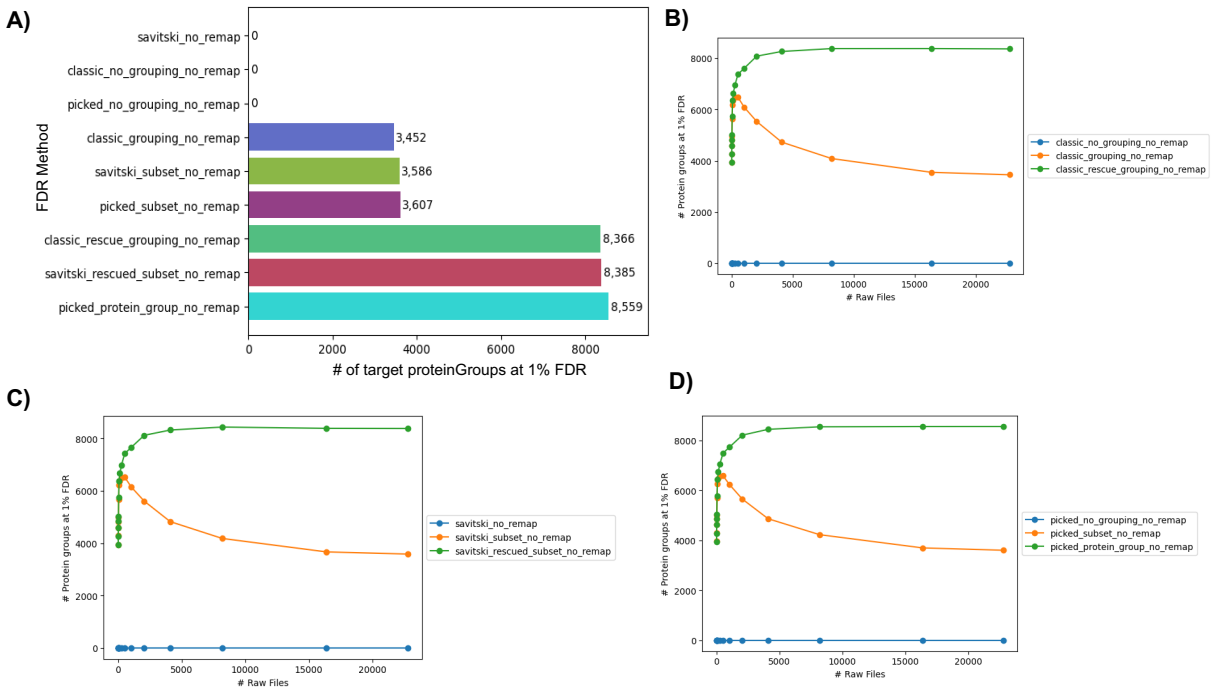


Figure 5: **pgFDR retrieves most number of target protein groups at 1% FDR due to rSG strategy. A)** Bar chart illustrating average number of protein groups detected at 1% FDR for 9 methods Performance at target protein groups retrieval at 1% protein-level FDR illustrated in **B) Classic TDS- C) Picked TDS- D) Picked Group TDS-** and their respective grouping strategies with accumulation of ProteomeDH2 files.

From Figure 5A, pgFDR retrieves most protein groups at 1% FDR. This potential might contribute to the reasoning behind why pgFDR retrieves more isoforms. Contrastingly,in Figure 5B-D, as raw files increase, for subset grouping strategy (yellow lines), these methods behave in a similar way, where these methods start retrieving less protein groups. Hence, this could contribute to why cFDR performs less well than pgFDR. Therefore, it is reasonable to infer that the rsG strategy makes a significant contribution to retrieving target proteins and protein isoforms with larger proteomic datasets.

3.6 pgFDR retrieves more microproteins at 1% FDR

Since rsG strategy of pgFDR has shown to be crucial in retrieving target proteins, we can infer that retrieving more protein groups can help detect smaller proteins by combining information from multiple peptides that map to the same protein group [3][37][38]. For smaller proteins, it may be difficult to identify them based on a single peptide alone, as the peptide may be present in other proteins or may be too short to provide a unique identification[43]. However, by considering all peptides that map to a particular protein group, it may be possible to detect smaller proteins that would otherwise be missed[41][42]. We asked if pgFDR scaled to microproteins which are translated from smORFs found by Mudge et al (2022)[10]. Microproteins are small

proteins encoded by smORFs which are often overlooked in traditional protein identification methods[44]. Despite their small size, microproteins are increasingly being recognized as important regulators of cellular processes and pathways[44][45].

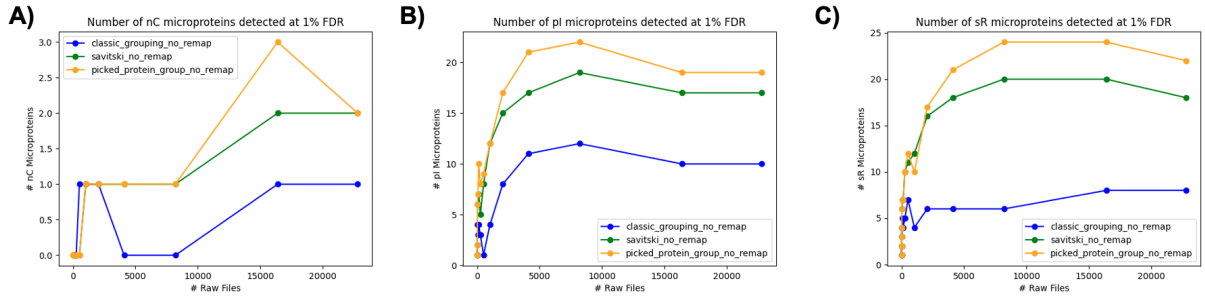


Figure 6: pgFDR retrieves more microproteins in ProteomeHD2 (yellow lines): FDR Method Retrieval of different microproteins **(A) nC microproteins[46] B) pI microproteins [10][45] C) sR microproteins[47]** at 1% protein-level FDR; cFDR, ppFDR and pgFDR represented by blue, green and yellow lines respectively

From Figure 6A-C, pgFDR performs slightly better at retrieving microproteins at 1% FDR, as raw files accumulate. nC microproteins detected represent near cognate ORFs which are translated smORFs [46]; pI microproteins, are translated smORFs identified from multiple ribosome profiling (Ribo-Seq) sequences detected by Mudge et al[10][45]; sR microproteins are translated smORFs which are single Ribo-Seq sequences which are less than 10 amino acids[47]. Methods behave in a similar way, except only a small increase in detected of microprotein with pgFDR than ppFDR. cFDR maintained performance at retrieval at the lowest level as raw files increase. Additionally, microproteins found in this study, have not been functionally characterized and annotated through proteomics as they have only been recently identified[10].

3.7 FDR methods demonstrate no bias to protein mass

With the slight increase in microproteins detected, we asked if there was a relationship between protein mass and proteins surviving the Q-value threshold of 0.01 (1% protein-level FDR). Was there a bias in protein mass and q-value calculation by pgFDR?

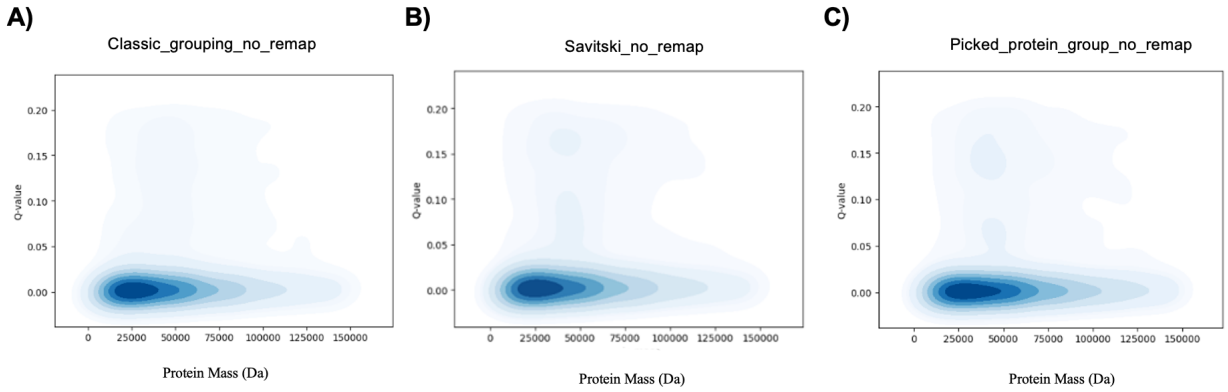


Figure 7: **FDR methods (A) cFDR B) ppFDR C) pgFDR show no particular bias towards protein mass** as 2D density plots which illustrate the relationship between proteins' mass against their Q-values maintained similar.

In Figure 6, pgFDR performs better than ppFDR and cFDR in retrieving smaller proteins. Therefore, we investigated whether there was a bias towards smaller target proteins. Through mapping the protein IDs[35] found in the "proteinGroups.output" file with a downloaded UniProtKB[35] Human Proteome database with respective masses, we found that there was no significant bias in retrieving smaller proteins in Figure 7A-C, as density plots show that a large proportion of the proteins, interestingly with smaller masses, have a range of q-values regardless of mass. Figure 7 demonstrates there is no significance between the q-value, protein mass and FDR method, as proteins, regardless of mass, did not show a correlation (see Appendix 1.7).

4 Discussion

In larger datasets, there is a higher chance of false positives, and cFDR approach potentially has a more significant impact on removing these false positives, leading to fewer identified targets as it is conceptually known to assume each PSM hit is independent. However, pgFDR managed to outperform cFDR, where it retrieved approximately 4000 more target proteins at 1% protein-level FDR than cFDR in the same ProteomeHD2 dataset. It has outperformed approximately both by 2-fold when compared to ppFDR and cFDR. The introduction of rsG strategy and investigating proteins at 1% protein-level FDR has immensely contributed to the performance of pgFDR. This study also identified microproteins with fasta identifiers of nC, pI and sR(see Appendix 1.5), which are detected cumulatively when number of raw files increase. They can potentially now functionally characterised through further proteomics research as pgFDR is powerful enough to identify these translated smORFs. However, it is important to note that the small number of microproteins detected in ProteomeHD2 also indicates that it cannot be confidently inferred that pgFDR outperforms the other methods for identifying these microproteins.

In regards to the experimental design, Percolator software was applied for rescoring PSMs on Fragpipe for FDR control on a PSM-level. This was employed instead of MaxQuant as previous studies have demonstrated that MaxQuant software, another alternative to Percolator in PSM rescoring, generates PSM-level PEPs that are less well-calibrated[3][48], which consequently will lead to anti-conservative protein-level FDR estimates. Anti-conservative protein-level FDR refers to estimated FDR for proteins is lower than the actual FDR[[37], which might increase the number of false positive hits in a dataset. For sharedPeptides strategies (omitted from Table 1), there was another option termed RazorPeptides, instead of Discard strategy. This was not implemented as previous studies have shown that the RazorPeptides principle, derivative of Occam's razor principle, leads to anti-conservative FDR estimation for large datasets. Occam's razor heuristic algorithm, most common in MaxQuant software platform, assigns shared peptides to the protein group with the most unique peptides, but is known to result in incorrect PSM assignments in large-scale proteomics data, leading to a loss of FDR control [3][37][38]. Additionally, it is important to note that peptide-level FDR filtering are usually more stringent around 0.12% FDR, usually leading to conservative FDR estimation and loss of true positive identifications[3].

It is also important to note, Picked TDS strategy cannot be directly applied to protein groups[3]. Therefore, Picked TDS-subset grouping and Picked TDS-rsG method are conceptually impossible as their decoy counterparts of target proteins in a protein group might not be grouped together. Respective decoy proteins have their own set of unique peptides and shared peptides upon which grouping is based[3]. Therefore, aforementioned methods in Figure 3E-F are not representative of reliable protein identification methods,

merely just executed for comparison of method parameters. Further improvements to this study would be the performance of entrapment searches where briefly, the target database is extended by an entrapment database, containing only false sequences[49]. This can be utilized to assess if pgFDR is well-calibrated for protein-level FDR estimation. Conclusively, results of this study suggest that pgFDR is the most effective FDR estimation method for identifying target proteins in large and complex biological datasets such as ProteomeHD2 but further investigation is required to fully evaluate the potential of pgFDR.

5 Conclusion

Overall, pgFDR is the most effective FDR estimation method for identifying target proteins in large biological datasets to date. With pgFDR, the manually curated ProteomeHD2 is now further refined with precision and accuracy where these new translated smORFs can be functionally characterised in the future, such as demonstrated in the previously curated ProteomeHD[2] which identified co-regulation partners of microproteins, inferring functionality. Microproteins identified in this study, can ultimately be further investigated to contribute to the complex understanding of the human proteome.

6 Acknowledgements

I would love to express my sincerest gratitude to Georg Kustatscher, and Savvas Kourtis for giving me the opportunity to work with the proteomics as part of my dissertation. Thank you Savvas, for helping me relentlessly through every step of the way- from processing the ProteomeHD2 dataset to being there for me when I needed guidance about programming in R. Their insights and feedback have been instrumental in helping me to refine my research questions and methodologies. Furthermore, their commitment to advancing the field of proteomics and their passion for scientific discovery have been a constant source of inspiration for me. I feel incredibly fortunate to have had the opportunity to learn from such exceptional scientists and mentors, and I am grateful for their contributions to my academic and professional development. Additionally, I would love to thank my parents for their endless support, and my friends in Appleton Tower for cheering me on through my dissertation journey.

References

1. Viéitez C, Busby BP, Ochoa D, Mateus A, Memon D, Galardini M, Yildiz U, Trovato M, Jawed A, Geiger AG, Oborská-Oplová M, Potel CM, Vonesch SC, Szu Tu C, Shahraz M, Stein F, Steinmetz LM, Panse VG, Noh KM, Savitski MM, Typas A, and Beltrao P. High-throughput functional characterization of protein phosphorylation sites in yeast. Nature Biotechnology 2021 Oct; 40:382–90. DOI: 10.1038/s41587-021-01051-x

2. Kustatscher G, Grabowski P, Schrader TA, Passmore JB, Schrader M, and Rappsilber J. Co-regulation map of the human proteome enables identification of protein functions. Nature Biotechnology 2019 Nov; 37:1361–71. DOI: 10.1038/s41587-019-0298-5

3. The M, Samaras P, Kuster B, and Wilhelm M. Reanalysis of ProteomicsDB Using an Accurate, Sensitive, and Scalable False Discovery Rate Estimation Approach for Protein Groups. Molecular & Cellular Proteomics 2022 Dec; 21:100437. DOI: 10.1016/j.mcpro.2022.100437

4. Lachén-Montes M, Mendizuri N, Ausín K, Pérez-Mediavilla A, Azkargorta M, Iloro I, Elortza F, Kondo H, Ohigashi I, Ferrer I, Torre R de la, Robledo P, Fernández-Irigoyen J, and Santamaría E. Smelling the Dark Proteome: Functional Characterization of PITH Domain-Containing Protein 1 (C1orf128) in Olfactory Metabolism. Journal of Proteome Research 2020 Nov; 19:4826–43. DOI: 10.1021/acs.jproteome.0c00452

5. Neuman BW, Joseph JS, Saikatendu KS, Serrano P, Chatterjee A, Johnson MA, Liao L, Klaus JP, Yates JR, Wüthrich K, Stevens RC, Buchmeier MJ, and Kuhn P. Proteomics Analysis Unravels the Functional Repertoire of Coronavirus Nonstructural Protein 3. Journal of Virology 2008 Mar; 82:5279–94. DOI: 10.1128/jvi.02631-07

6. Yates III JR. Recent technical advances in proteomics. F1000Research 2019 Mar; 8:351. DOI: 10.12688/f1000research.16987.1. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6441878/

7. Tiambeng TN, Roberts DS, Brown KA, Zhu Y, Chen B, Wu Z, Mitchell SD, Guardado-Alvarez TM, Jin S, and Ge Y. Nanoproteomics enables proteoform-resolved analysis of low-abundance proteins in human serum. Nature Communications 2020 Aug; 11. DOI: 10.1038/s41467-020-17643-1

8. Chandramouli K and Qian PY. Proteomics: Challenges, Techniques and Possibilities to Overcome Biological Sample Complexity. Human Genomics and Proteomics 2009 Jan; 1. Ed. by Godwin AK. DOI: 10.4061/2009/239204

9. Kustatscher G, Collins T, Gingras AC, Guo T, Hermjakob H, Ideker T, Lilley KS, Lundberg E, Marcotte EM, Ralser M, and Rappsilber J. Understudied proteins: opportunities and challenges for functional proteomics. Nature Methods 2022 May; 19:774–9. DOI: 10.1038/s41592-022-01454-x

10. Mudge JM, Ruiz-Orera J, Prensner JR, Brunet MA, Calvet F, Jungreis I, Gonzalez JM, Magrane M, Martinez TF, Schulz JF, Yang YT, Albà MM, Aspden JL, Baranov PV, Bazzini AA, Bruford E, Martin MJ, Calviello L, Carvunis AR, Chen J, Couso JP, Deutsch EW, Flicek P, Frankish A, Gerstein M, Hubner N, Ingolia NT, Kellis M, Menschaert G, Moritz RL, Ohler U, Roucou X, Saghatelian A, Weissman JS, and Heesch S van. Standardized annotation of translated open reading frames. Nature Biotechnology 2022 Jul; 40:994–9. DOI: 10.1038/s41587-022-01369-0

11. Bjornson RD, Carriero NJ, Colangelo C, Shifman M, Cheung KH, Miller PL, and Williams K. X!!Tandem, an Improved Method for Running X!Tandem in Parallel on Collections of Commodity Computers. Journal of Proteome Research 2007 Sep; 7:293–9. DOI: 10.1021/pr0701198

12. Diament BJ and Noble WS. Faster SEQUEST Searching for Peptide Identification from Tandem Mass Spectra. Journal of Proteome Research 2011 Sep; 10:3871–9. DOI: 10.1021/pr101196n

13. Elias JE and Gygi SP. Target-Decoy Search Strategy for Mass Spectrometry-Based Proteomics. Methods in Molecular Biology 2009 Dec :55–71. DOI: 10.1007/978-1-60761-444-9_5. Available from: https://dx.doi.org/10.1007%5C%2F978-1-60761-444-9_5

14. Keich U, Tamura K, and Noble WS. Averaging Strategy To Reduce Variability in Target-Decoy Estimates of False Discovery Rate. Journal of Proteome Research 2018 Dec; 18:585–93. DOI: 10.1021/acs.jproteome.8b00802

15. Korthauer K, Kimes PK, Duvallet C, Reyes A, Subramanian A, Teng M, Shukla C, Alm EJ, and Hicks SC(. A practical guide to methods controlling false discoveries in computational biology. Genome Biology (Online) 2019 Jun; 20. DOI: 10.1186/s13059-019-1716-1. Available from: https://www.osti.gov/biblio/1618920-practical-guide-methods-controlling-false-discoveries-computational-biology

16. Keller A, Nesvizhskii AI, Kolker E, and Aebersold R. Empirical Statistical Model To Estimate the Accuracy of Peptide Identifications Made by MS/MS and Database Search. Analytical Chemistry 2002 Sep; 74:5383–92. DOI: 10.1021/ac025747h

17. Savitski MM, Debrauwer L, Hahne H, Rolain JM, and Bantscheff M. A Scalable Approach for Protein False Discovery Rate Estimation in Large Proteomic Data Sets. Molecular & Cellular Proteomics 2015 Sep; 14:2394–404. DOI: 10.1074/mcp.m114.046995

18. Paulo JA. Practical and Efficient Searching in Proteomics: A Cross Engine Comparison. WebmedCentral 2013 Oct; 4:20–30. DOI: 10.9754/journal.wplus.2013.0052. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4208621/

19. Pounds S and Cheng C. Improving false discovery rate estimation. Bioinformatics 2004 Feb; 20:1737–45. DOI: 10.1093/bioinformatics/bth160. [Accessed on: 2021 Apr 12]

20. Reiter L, Claassen M, Schrimpf SP, Jovanovic M, Schmidt A, Buhmann JM, Hengartner MO, and Aebersold R. Protein Identification False Discovery Rates for Very Large Proteomics Data Sets Generated by Tandem Mass Spectrometry. Molecular & Cellular Proteomics 2009 Jul; 8:2405–17. DOI: 10.1074/mcp.m900317-mcp200

21. Rosenberger G, Bludau I, Schmitt U, Heusel M, Hunter CL, Liu Y, MacCoss MJ, MacLean BX, Nesvizhskii AI, Pedrioli PGA, Reiter L, Röst HL, Tate S, Ting YS, Collins BC, and Aebersold R. Statistical control of peptide and protein error rates in large-scale targeted data-independent acquisition analyses. Nature Methods 2017 Aug; 14:921–7. DOI: 10.1038/nmeth.4398

22. The M, Tasnim A, and Käll L. How to talk about protein-level false discovery rates in shotgun proteomics. PROTEOMICS 2016 Sep; 16:2461–9. DOI: 10.1002/pmic.201500431

23. Lin A, Short T, Noble WS, and Keich U. Detecting more peptides from bottom-up mass spectrometry data via peptide-level target-decoy competition. 2022 May :22–30. DOI: doi:https://doi.org/10.1101/2022.05.11.491571. Available from: https://www.biorxiv.org/content/10.1101/2022.05.11.491571v1

24. Perez-Riverol Y, Bai J, Bandla C, García-Seisdedos D, Hewapathirana S, Kamatchinathan S, Kundu DJ, Prakash A, Frericks-Zipper A, Eisenacher M, Walzer M, Wang S, Brazma A, and Vizcaíno JA. The PRIDE database resources in 2022: a hub for mass spectrometry-based proteomics evidences. Nucleic Acids Research 2022 Jan; 50:D543–D552. DOI: 10.1093/nar/gkab1038. Available from: https://pubmed.ncbi.nlm.nih.gov/34723319/

25. Veiga Leprevost F da, Haynes SE, Avtonomov DM, Chang HY, Shanmugam AK, Mellacheruvu D, Kong AT, and Nesvizhskii AI. Philosopher: a versatile toolkit for shotgun proteomics data analysis. Nature Methods 2020 Jul; 17:869–70. DOI: 10.1038/s41592-020-0912-y

26. Chen X, Wei S, Ji Y, Guo X, and Yang F. Quantitative proteomics using SILAC: Principles, applications, and developments. PROTEOMICS 2015 Jul; 15:3175–92. DOI: 10.1002/pmic.201500108

27. Spivak M, Weston J, Bottou L, Käll L, and Noble WS. Improvements to the Percolator Algorithm for Peptide Identification from Shotgun Proteomics Data Sets. Journal of Proteome Research 2009 Jul; 8:3737–45. DOI: 10.1021/pr801109k

28. Benjamini Y and Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society: Series B (Methodological) 1995 Jan; 57:289–300. DOI: 10.1111/j.2517-6161.1995.tb02031.x

29. Käll L, Storey JD, MacCoss MJ, and Noble WS. Posterior Error Probabilities and False Discovery Rates: Two Sides of the Same Coin. Journal of Proteome Research 2007 Dec; 7:40–4. DOI: 10.1021/pr700739d

30. Yi X, Gong F, and Fu Y. Transfer posterior error probability estimation for peptide identification. BMC Bioinformatics 2020 May; 21. DOI: 10.1186/s12859-020-3485-y. [Accessed on: 2023 May 14]

31. The M, MacCoss MJ, Noble WS, and Käll L. Fast and Accurate Protein False Discovery Rates on Large-Scale Proteomics Data Sets with Percolator 3.0. Journal of the American Society for Mass Spectrometry 2016 Aug; 27:1719–27. DOI: 10.1007/s13361-016-1460-7

32. Nakayasu ES, Gritsenko M, Piehowski PD, Gao Y, Orton DJ, Schepmoes AA, Fillmore TL, Frohnert BI, Rewers M, Krischer JP, Ansong C, Suchy-Dicey AM, Evans-Molina C, Qian WJ, Webb-Robertson BJM, and Metz TO. Tutorial: best practices and considerations for mass-spectrometry-based protein biomarker discovery and validation. Nature Protocols 2021 Aug; 16:3737–60. DOI: 10.1038/s41596-021-00566-6. Available from: https://www.nature.com/articles/s41596-021-00566-6#Sec37

33. FASTA Format for Nucleotide Sequences. www.ncbi.nlm.nih.gov. Available from: https://www.ncbi.nlm.nih.gov/genbank/fastaformat/

34. Accession. www.uniprot.org, 2022 Oct. Available from: https://www.uniprot.org/help/accession_numbers

35. UniProt: the universal protein knowledgebase. Nucleic Acids Research 2016 Nov; 45:D158–D169. DOI: 10.1093/nar/gkw1099

36. Nesvizhskii AI and Aebersold R. Interpretation of Shotgun Proteomic Data. Molecular & Cellular Proteomics 2005 Jul; 4:1419–40. DOI: 10.1074/mcp.r500012-mcp200

37. Serang O, Moruz L, Hoopmann MR, and Käll L. Recognizing Uncertainty Increases Robustness and Reproducibility of Mass Spectrometry-based Protein Inferences. Journal of Proteome Research 2012 Nov; 11:5586–91. DOI: 10.1021/pr300426s

38. Noble W and Serang O. A review of statistical methods for protein identification using tandem mass spectrometry. Statistics and Its Interface 2012; 5:3–20. DOI: 10.4310/sii.2012.v5.n1.a2

39. Cerqueira FR, Graber A, Schwikowski B, and Baumgartner C. MUDE: A New Approach for Optimizing Sensitivity in the Target-Decoy Search Strategy for Large-Scale Peptide/Protein Identification. Journal of Proteome Research 2010 May; 9:2265–77. DOI: 10.1021/pr901023v

40. Elias JE and Gygi SP. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. Nature Methods 2007 Mar; 4:207–14. DOI: 10.1038/nmeth1019. Available from: https://www.nature.com/articles/nmeth1019

41. Reixachs-Solé M and Eyras E. Uncovering the impacts of alternative splicing on the proteome with current omics techniques. WIREs RNA 2022 Jan; 13. DOI: 10.1002/wrna.1707

42. Miller RM, Jordan BT, Mehlferber MM, Jeffery ED, Chatzipantsiou C, Kaur S, Millikin RJ, Dai Y, Tiberi S, Castaldi PJ, Shortreed MR, Luckey CJ, Conesa A, Smith LM, Deslattes Mays A, and Sheynkman GM. Enhanced protein isoform characterization through long-read proteogenomics. Genome Biology 2022 Mar; 23. DOI: 10.1186/s13059-022-02624-y

43. Dupree EJ, Jayathirtha M, Yorkey H, Mihasan M, Petre BA, and Darie CC. A Critical Review of Bottom-Up Proteomics: The Good, the Bad, and the Future of This Field. Proteomes 2020 Jul; 8. DOI: 10.3390/proteomes8030014. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7564415/

44. Leong AZX, Lee PY, Mohtar MA, Syafruddin SE, Pung YF, and Low TY. Short open reading frames (sORFs) and microproteins: an update on their identification and validation measures. Journal of Biomedical Science 2022 Mar; 29. DOI: 10.1186/s12929-022-00802-5

45. Kruusvee V and Wenkel S. Microproteins — lost in translation. Nature Chemical Biology 2022 Apr; 18:581–2. DOI: 10.1038/s41589-022-01007-5

46. Mudge JM, Ruiz-Orera J, Prensner JR, Brunet MA, Gonzalez JM, Magrane M, Martinez T, Schulz JF, Yang YT, Albà MM, Baranov PV, Bazzini A, Bruford E, Martin MJ, Carvunis AR, Chen J, Couso JP, Flicek P, Frankish A, Gerstein M, Hubner N, Ingolia NT, Menschaert G, Ohler U, Roucou X, Saghatelian A, Weissman J, and Heesch S van. A community-driven roadmap to advance research on translated open reading frames detected by Ribo-seq. 2021 Jun. DOI: 10.1101/2021.06.10.447896

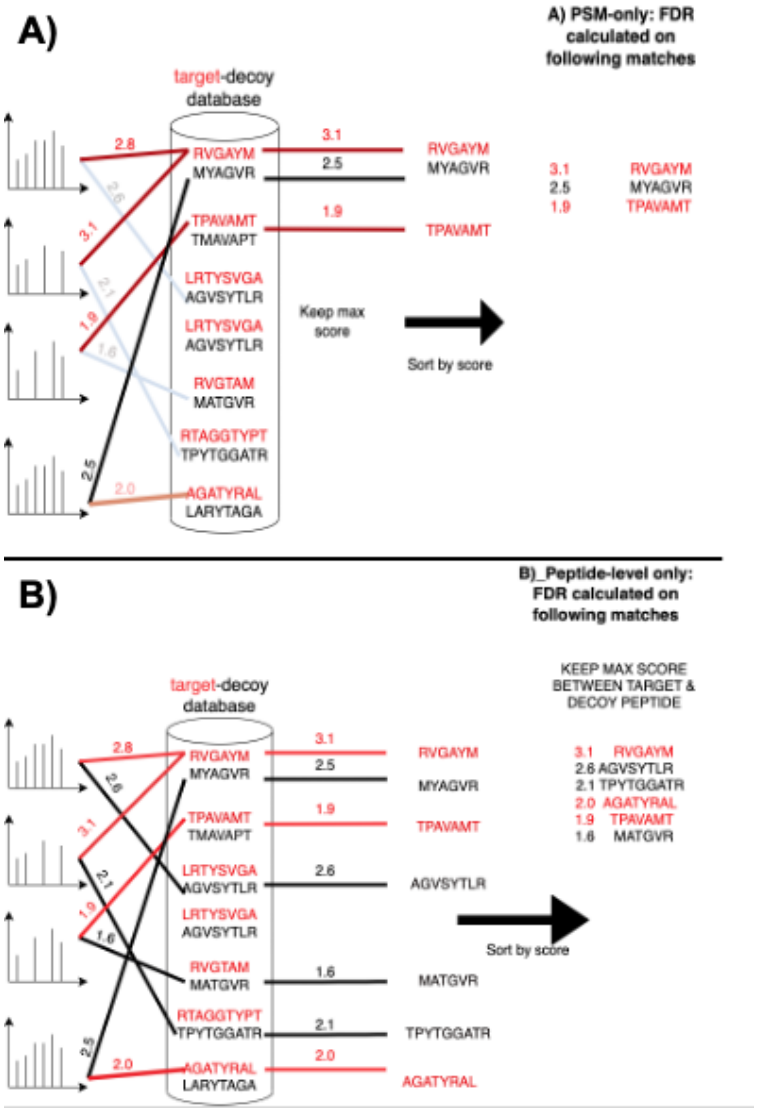
47. Bartholomäus A, Kolte B, Mustafayeva A, Goebel I, Fuchs S, Benndorf D, Engelmann S, and Ignatova Z. smORFer: a modular algorithm to detect small ORFs in prokaryotes. Nucleic Acids Research 2021 Jun. DOI: 10.1093/nar/gkab477

48. Palomba A, Abbondio M, Fiorito G, Uzzau S, Pagnozzi D, and Tanca A. Comparative Evaluation of MaxQuant and Proteome Discoverer MS1-Based Protein Quantification Tools. Journal of Proteome Research 2021 May; 20:3497–507. DOI: 10.1021/acs.jproteome.1c00143

49. Granholm V, Navarro JF, Noble WS, and Käll L. Determining the calibration of confidence estimation procedures for unique peptides in shotgun proteomics. Journal of Proteomics 2013 Mar; 80:123–31. DOI: 10.1016/j.jprot.2012.12.007

Appendix

Appendix 1.1 Graphical Illustration of PSM-level filtering, Peptide-level Filtering, & Protein-level Filtering



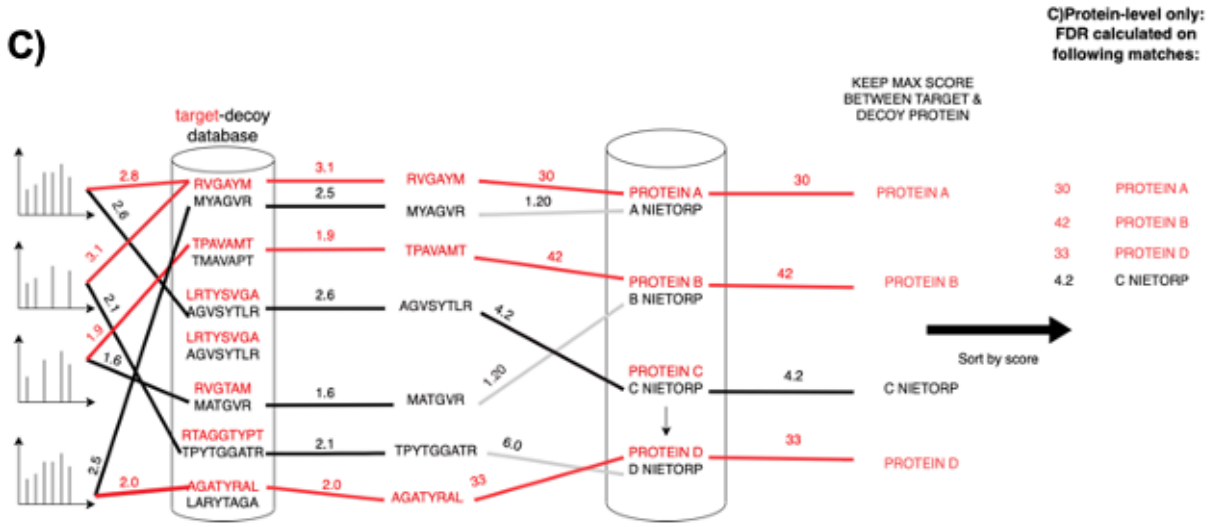


Figure 8: Detailed overview of different types of FDR filtering levels for 1% protein-level FDR control A) PSM-level filtering prodcedure which is implemented by Percolator algorithm, which carries out direct competition of target and reversed decoys on PSM-level B) Peptide-level filtering does noy employ PSM-level competition between targets and decoys, but separately scans each spectrum against target and decoy database, and employs a direct competition between each target and reversed decoy peptide, where the higher scoring peptide between the two is kept. C) Protein-level filtering does not employ PSM-level and Peptide-level competition between targets and decoys, but separately scans each spectrum against target and decoy sequence and protein database, which employs a direct competition between each target and reversed decoy protein, where the higher scoring between them is kept.

Appendix 1.2: Percolator Output File

```
PSMId,score,q-value,posterior_error_prob,peptide,proteinIds
366_6__H_Koll_1.7788.7788.2_1,0.498612,0.000350631,0.0113172,R.IPILVAR.M,rev_sp|Q17RN3|FA98C_HUMAN
366_6__H_Koll_1.6984.6984.3_1,0.45261,0.000523104,0.0154603,K.DAITSNLEITK[6.0201].F,rev_sp|Q9Y6V0-6|PCLO_;rev_sp|Q9Y6V0|PCLO_HUMAN
366_6__H_Koll_1.8694.8694.4_1,0.444234,0.000694203,0.0163604,R.WHHNELVSMNQYLNALHHNTK.I,rev_sp|Q9NP56|PDE7B_HUMAN
366_6__H_Koll_1.6746.6746.2_1,0.437302,0.000865501,0.0171437,K.LGALEEELAR.L,rev_sp|O95347-2|SMC2_;rev_sp|O95347|SMC2_HUMAN
366_6__H_Koll_1.3350.3350.3_1,0.430924,0.00100688,0.0178967,R.VC[57.0215]SALDLGEAK[6.0201]RR.F,rev_sp|Q8NHQ9|DDX55_HUMAN
366_6__H_Koll_1.11045.11045.2_1,0.383053,0.00116183,0.0246715,R.QEAQVLQAELEAELER[10.0083].S,rev_sp|Q9GZM8-2|NDEL1_;rev_sp|Q9GZM8-3|NDEL1_;rev_sp|
Q9GZM8|NDEL1_HUMAN
366_6__H_Koll_1.13129.13129.3_1,0.368106,0.00131148,0.0272551,K.QLINETAMDISATAEDMILISSK[6.0201].R,rev_sp|O60706-2|ABCC9_;rev_sp|O60706|ABCC9_HUMAN
366_6__H_Koll_1.5551.5551.2_1,0.347106,0.00146175,0.0313296,K.LHLLLENK.S,rev_sp|P50749|RASF2_HUMAN
366_6__H_Koll_1.4608.4608.2_1,0.334103,0.00162048,0.0341396,R.LVVFKDK.K,rev_sp|Q13472-2|TOP3A_;rev_sp|Q13472-3|TOP3A_;rev_sp|Q13472|TOP3A_HUMAN
366_6__H_Koll_1.1704.1704.2_1,0.327439,0.00175691,0.0356717,K.TIISNEK.V,rev_sp|O95391|SLU7_HUMAN
366_6__H_Koll_1.5087.5087.3_1,0.305143,0.00189783,0.041289,K.VEKNSTVVIDLEGGSK[6.0201].I,rev_sp|Q96M27-2|PRRC1_;rev_sp|Q96M27-3|PRRC1_;rev_sp|
Q96M27-4|PRRC1_;rev_sp|Q96M27-5|PRRC1_;rev_sp|Q96M27|PRRC1_HUMAN
366_6__H_Koll_1.10999.10999.2_1,0.282288,0.00204306,0.0479134,R.QEAQVLQAELEAELER[10.0083].S,rev_sp|Q9GZM8-2|NDEL1_;rev_sp|Q9GZM8-3|NDEL1_;rev_sp|
Q9GZM8|NDEL1_HUMAN
366_6__H_Koll_1.8531.8531.4_1,0.262818,0.00218613,0.054335,K.SLEATSAEDANEDFVTIR[10.0083].S,rev_sp|O60566-2|BUB1B_;rev_sp|O60566-3|BUB1B_;rev_sp|
O60566|BUB1B_HUMAN
366_6__H_Koll_1.6291.6291.2_1,0.24913,0.00232414,0.0593224,R.EVSDSILR[10.0083].E,rev_sp|Q5T200-2|ZC3HD_;rev_sp|Q5T200|ZC3HD_HUMAN
366_6__H_Koll_1.3483.3483.2_1,0.224128,0.00242461,0.0695461,R.LDHDNADSIGPSVPR.I,rev_sp|A8MZ59-1|LEUTX_;rev_sp|A8MZ59|LEUTX_HUMAN
366_6__H_Koll_1.9520.9520.2_1,0.163294,0.00257615,0.10148,K.GALTDGIAVIAR.Q,rev_sp|Q96JE7-2|SC16B_;rev_sp|Q96JE7-3|SC16B_;rev_sp|Q96JE7|SC16B_HUMAN
366_6__H_Koll_1.4953.4953.3_1,0.162858,0.00272521,0.101749,R.IAESFTLK[6.0201]MR[10.0083]K.A,rev_sp|P48058|GRIA4_HUMAN
366_6__H_Koll_1.6438.6438.3_1,0.16117,0.00287443,0.1028,K.DALMNISIC[57.0215]QHIR.C,rev_sp|P60228|EIF3E_HUMAN
366_6__H_Koll_1.11824.11824.2_1,0.158992,0.00302526,0.104168,R.EIK[6.0201]EPVSSSMIIPVSVFR[10.0083].Q,rev_sp|Q86UW6-2|N4BP2_;rev_sp|Q86UW6|
N4BP2_HUMAN
366_6__H_Koll_1.4681.4681.2_1,0.157967,0.00317556,0.104819,R.LVVFKDK.K,rev_sp|Q13472-2|TOP3A_;rev_sp|Q13472-3|TOP3A_;rev_sp|Q13472|TOP3A_HUMAN
366_6__H_Koll_1.9398.9398.2_1,0.15541,0.00332226,0.106455,K.DFLLFHSNVGEK[6.0201].F,rev_sp|Q8N7B1|HORM2_HUMAN
366_6__H_Koll_1.7383.7383.2_1,0.152914,0.00347117,0.108075,K.IAIGLGK.V,rev_sp|Q6NVV3|NIPA3_;rev_sp|Q8N8Q9-2|NIPA2_;rev_sp|Q8N8Q9|NIPA2_HUMAN
366_6__H_Koll_1.6742.6742.2_1,0.151719,0.00362155,0.108857,K.LGALEEELAR[10.0083].L,rev_sp|O95347-2|SMC2_;rev_sp|O95347|SMC2_HUMAN
366_6__H_Koll_1.6198.6198.3_1,0.150448,0.00375488,0.109695,K.MLKSTDGSVIADVLK[6.0201].K,rev_sp|P57078-2|RIPK4_;rev_sp|P57078|RIPK4_HUMAN
366_6__H_Koll_1.4712.4712.3_1,0.134974,0.00389981,0.120353,R.LTTITAPDR[10.0083]LR[10.0083].G,rev_sp|Q00005-2|2ABB_;rev_sp|Q00005-3|2ABB_;rev_sp|
Q00005-4|2ABB_;rev_sp|Q00005-5|2ABB_;rev_sp|Q00005-6|2ABB_;rev_sp|Q00005-7|2ABB_;rev_sp|Q00005|2ABB_HUMAN
366_6__H_Koll_1.10618.10618.3_1,0.129345,0.00404858,0.124445,R.TPVASPAVGGGGPPEEIFVLR[10.0083].P,rev_sp|Q9NRU3-2|CNNM1_;rev_sp|Q9NRU3|CNNM1_HUMAN
366_6__H_Koll_1.9671.9671.2_1,0.12812,0.00419539,0.125351,M.QFLLAGPSTLK.G,rev_sp|O60566-2|BUB1B_;rev_sp|O60566-3|BUB1B_;rev_sp|O60566|BUB1B_HUMAN
366_6__H_Koll_1.5205.5205.3_1,0.123084,0.00433937,0.129135,K.HSLTINYSDVNGPPR[10.0083].Q,rev_sp|P23467-2|PTPRB_;rev_sp|P23467-3|PTPRB_;rev_sp|
P23467-4|PTPRB_;rev_sp|P23467|PTPRB_HUMAN
366_6__H_Koll_1.2664.2664.3_1,0.117406,0.00448498,0.133516,K.ETADGVIIKGGR.K,rev_sp|Q96I25|SPF45_HUMAN
366_6__H_Koll_1.5557.5557.2_1,0.110992,0.00462618,0.138614,K.AGGISIIR[10.0083].Y,rev_sp|Q9NSC5-2|HOME3_;rev_sp|Q9NSC5-3|HOME3_;rev_sp|Q9NSC5-4|
HOME3_;rev_sp|Q9NSC5-5|HOME3_;rev_sp|Q9NSC5|HOME3_HUMAN
366_6__H_Koll_1.8322.8322.2_1,0.103674,0.00477469,0.144629,R.LIEAGSPFPLK.R,rev_sp|Q9UKN7|MYO15_HUMAN
366_6__H_Koll_1.9493.9493.3_1,0.102732,0.00491437,0.145418,K.PDSVLPIM[15.9949]NPGLGGQK[6.0201].K,rev_sp|Q06830|PRDX1_HUMAN
366_6__H_Koll_1.10744.10744.2_1,0.0933443,0.00505877,0.153484,K.EFEPLVINLEK.I,rev_sp|Q71F23-2|CENPU_;rev_sp|Q71F23-3|CENPU_;rev_sp|Q71F23|
CENPU_HUMAN
366_6__H_Koll_1.5511.5511.2_1,0.0900873,0.00520678,0.156367,K.AGGISIIR[10.0083].Y,rev_sp|Q9NSC5-2|HOME3_;rev_sp|Q9NSC5-3|HOME3_;rev_sp|Q9NSC5-4|
HOME3_;rev_sp|Q9NSC5-5|HOME3_;rev_sp|Q9NSC5|HOME3_HUMAN
366_6__H_Koll_1.5605.5605.2_1,0.0899321,0.00535316,0.156505,R.QIIPLGR[10.0083].G,rev_sp|Q9Y3Y2-3|CHTOP_;rev_sp|Q9Y3Y2-4|CHTOP_;rev_sp|Q9Y3Y2|
CHTOP_HUMAN
366_6__H_Koll_1.3235.3235.3_1,0.0846535,0.00549777,0.161273,R.VPASM[15.9949]PSSPRPQPGR.S,rev_sp|Q9UGJ0-3|AAKG2_;rev_sp|Q9UGJ0|AAKG2_HUMAN
366_6__H_Koll_1.6579.6579.2_1,0.080461,0.00564636,0.165143,R.ILFAGTANGEAQLR.R,rev_sp|Q13882|PTK6_HUMAN
366_6__H_Koll_1.4142.4142.4_1,0.0800255,0.00578979,0.165549,K.HSNVDLFRHFR.F,rev_sp|Q96RT8-2|GCP5_;rev_sp|Q96RT8|GCP5_HUMAN
366_6__H_Koll_1.5398.5398.2_1,0.0755044,0.00593208,0.169813,R.QIIPLGR[10.0083].G,rev_sp|Q9Y3Y2-3|CHTOP_;rev_sp|Q9Y3Y2-4|CHTOP_;rev_sp|Q9Y3Y2|
CHTOP_HUMAN
366_6__H_Koll_1.9429.9429.2_1,0.0711944,0.00607587,0.17396,K.LAEK[6.0201]YQVFSITGR.G,rev_sp|A6NFE3|EFC10_HUMAN
366_6__H_Koll_1.5555.5555.3_1,0.0682922,0.0062213,0.176797,R.LTSAQSEDAPAALAETTR.P,rev_sp|Q96FC9-2|DDX11_;rev_sp|Q96FC9-3|DDX11_;rev_sp|Q96FC9-4|
DDX11_;rev_sp|Q96FC9|DDX11_HUMAN
366_6__H_Koll_1.9821.9821.2_1,0.0603433,0.00636754,0.184754,K.LIETVSAVALK.R,rev_sp|O75165|DJC13_HUMAN
```

Figure 9: Example of a Percolator Output File filtering a subset of ProteomeHD2 at 1% PSM-level FDR, percolator output files uploaded onto GitHub (see Data Availability section of Appendix); PSM.IDs correspond to a PSM and the MS project ID of PRIDE, followed by statistical confidence measures, q-value, score and posterior_error_prob (PEP), peptide sequence and protein IDs (Uniprot IDs) which are mapped in the pep-to-prot-mapping file programmatically

Appendix 1.3: Differential methods produced by combination of different parameters

Grouping Strategy	TDS Strategy	Method Name
No grouping	Classic TDS	*classic_no_grouping_no_remap*
Subset grouping	Classic TDS	*classic_grouping_no_remap*
Rescued subset grouping	Classic TDS	*classic_rescue_grouping_no_remap*
No grouping	Picked TDS	*savitski_no_remap*
Subset grouping	Picked TDS	*savitski_subset_no_remap*
Rescued subset grouping	Picked TDS	*savitski_rescued_subset_no_remap*
No grouping	Picked_group TDS	$picked\_no\_grouping\_no_remap$
Subset grouping	Picked_group TDS	*picked_subset_no_remap*
Rescued subset grouping	Picked_group TDS	$picked\_protein\_group_no\_remap$

Table 2: Different methods produced from a combination of various Grouping strategy settings and TDS strategy settings; note that classic_grouping_no_remap = cFDR; savitski_no_remap =ppFDR; pciked_protein_group_no_remap= pgFDR

Appendix 1.4: Command line execution of Picked Group FDR Tool and 3 FDR methods

```
#modified shell script picked group FDR package. Modify them according to range of
subset.
#cFDR method execution below
cat subset_100_target_rawfiles.tsv <(tail -n+2 subset_100_decoy_rawfiles.tsv) >
combined_targets+decoys_100.tsv
sed -i ''-e 's/,/\t/g;s/rev_/REV__/g;s/_;/_HUMAN\t/g'
combined_targets+decoys_100.tsv
python -um picked_group_fdr --perc_evidence combined_targets+decoys_100.tsv --
protein_groups_out proteinGroups.txt --method classic_grouping_no_remap   --
peptide_protein_map pep_to_prot_mapping_100.txt --special-aas '' --enzyme trypsinp
| tee proteinGroups.log

#ppFDR method
cat subset_100_target_rawfiles.tsv <(tail -n+2 subset_100_decoy_rawfiles.tsv) >
combined_targets+decoys_100.tsv
sed -i ''-e 's/,/\t/g;s/rev_/REV__/g;s/_;/_HUMAN\t/g'
combined_targets+decoys_100.tsv
python -um picked_group_fdr --perc_evidence combined_targets+decoys_100.tsv --
protein_groups_out proteinGroups.txt --method savitski_no_remap   --
peptide_protein_map pep_to_prot_mapping_100.txt --special-aas '' --enzyme trypsinp
| tee proteinGroups.log

#pgFDR method
cat subset_100_target_rawfiles.tsv <(tail -n+2 subset_100_decoy_rawfiles.tsv) >
combined_targets+decoys_100.tsv
sed -i ''-e 's/,/\t/g;s/rev_/REV__/g;s/_;/_HUMAN\t/g'
combined_targets+decoys_100.tsv
python -um picked_group_fdr --perc_evidence combined_targets+decoys_100.tsv --
protein_groups_out proteinGroups.txt --method picked_protein_group_no_remap   --
peptide_protein_map pep_to_prot_mapping_100.txt --special-aas '' --enzyme trypsinp
| tee proteinGroups.log
```

Figure 10: Terminal Command Line of Picked Group FDR tool: method is specified as illustrated in the figure, 3 commands where the 1) cat is used to concatenate the target and decoy peptides for protein-level filtering 2) sed is used for gaining reversed decoy protein database for generation of peptide-to-protein mapping file 3) python command to run the tool on Terminal where –method, –enzyme of interest (trypsin),–peptide protein map can be specified according when subsetting number of raw files in ProteomeHD2

Appendix 1.5: Example list of translated smORFs (microproteins) detected by pgFDR at 1% FDR which are uncharacterized

Q-value	Fasta Header	Protein Name
0.000081	sR	single_ribo2506_HUMAN
0.000433	sR	single_ribo2691_HUMAN
0.000943	sR	single_ribo3540_HUMAN
0.000943	sR	single_ribo2603_HUMAN
0.001122	sR	single_ribo3004_HUMAN
0.001184	sR	single_ribo1737_HUMAN
0.002272	REV__sR	single_ribo2365
0.002332	REV__sR	single_ribo1485
0.002641	sR	single_ribo355_HUMAN
0.003405	sR	single_ribo2945_HUMAN
0.003697	REV__sR	single_ribo1927
0.003817	REV__sR	single_ribo709
0.004054	REV__sR	single_ribo4149
0.004744	REV__sR	single_ribo507
0.005146	REV__sR	single_ribo719
0.005252	sR	single_ribo3103_HUMAN
0.005367	REV__sR	single_ribo3173
0.006352	sR	single_ribo2927_HUMAN
0.006800	sR	single_ribo560_HUMAN
0.006910	sR	single_ribo2030_HUMAN
0.006910	sR	single_ribo2687_HUMAN
0.007429	REV__sR	single_ribo302
0.007489	REV__sR	single_ribo1530
0.007720	sR	single_ribo2233_HUMAN
0.008128	REV__sR	single_ribo192
0.008355	sR	single_ribo1273_HUMAN;pI
0.008468	REV__sR	single_ribo2214
0.008468	sR	single_ribo2906_HUMAN
0.008524	sR	single_ribo3536_HUMAN
0.009142	sR	single_ribo3735_HUMAN
0.009142	sR	single_ribo1190_HUMAN
0.009199	sR	single_ribo1637_HUMAN
0.009311	sR	single_ribo3326_HUMAN
0.009874	sR	single_ribo4056_HUMAN

Figure 11: **List of uncharacterized sR microproteins detected by pgFDR tool at 1% FDR of 22796 raw files in ProteomeHD2:** Code generated to extract different fasta headers retrieving translated smORFs, Q-value score shows our confidence levels in protein being present where a threshold of $Q < 0.01$ ensures confidence in researchers that it is present in protein sample; Protein Name refers to label given by Mudge et al (2022)[10]

Appendix 1.6: Difference in Target Protein Detection of 3 FDR methods

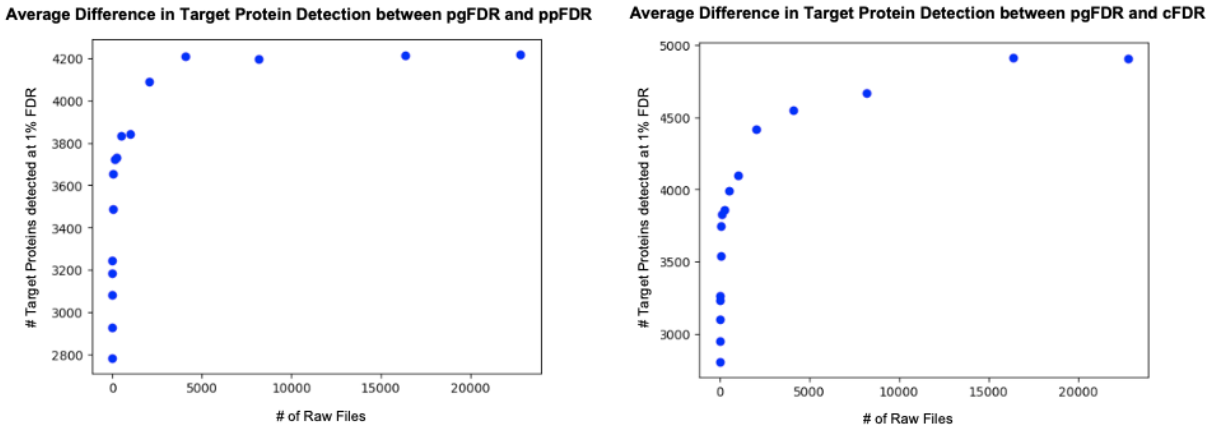


Figure 12: **Average Difference between Target Protein Detection at 1%FDR between pgFDR and two respective methods: ppFDR and cFDR:** Methods were executed in triplicates obtaining the averages, and FDR performance was measured by extracting number of target proteins detected at 22796 raw files by pgFDR over target proteins detected at 22796 raw files by ppFDR (56.2% difference) and cFDR (66.2% difference) respectively

Appendix 1.7: Relationship of Protein Mass and Q-value calculation between 3 FDR methods

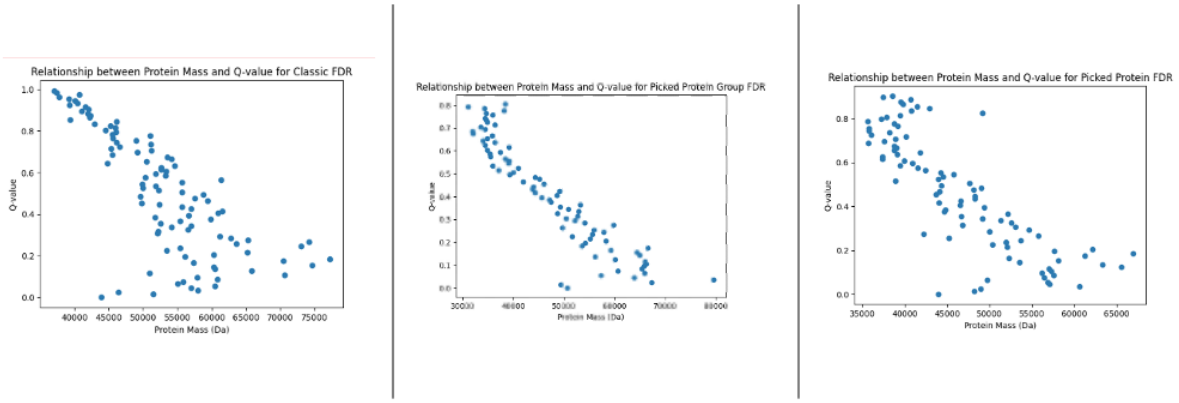


Figure 13: **Relationship between protein mass and q-value calculation by A) cFDR, B) ppFDR, C) pgFDR respectively; this relationship is confidently similar to the concept of smaller proteins having a higher q-value indicating a smaller probability of existing in the file, but the distribution quickly scatters around the plot as proteins increase in mass**

Appendix 1.8: Data availability

Raw files for the ProteomeHD2 dataset are available on PRIDE. Example PRIDE identifier used is: PXD008888, such as linked here (PXD008888). Fragpipe search results and result files of all the methods analysis are available on GitHub (FDR EVALUATION WITH PROTEOMEHD2). This includes the scripts in order to reproduce the figures in this dissertation as well as for the concatenation of target and decoy files, generation of peptide to protein mapping text, and analysis of all different results mentioned in this dissertation.