CIRFESS: An interactive resource for querying the set of theoretically detectable peptides for cell surface and extracellular enrichment proteomic studies

Matthew Waas (https://orcid.org/0000-0003-4537-1502)[1], Jack Littrell (https://orcid.org/0000-0003-1264-894X)[1], Rebekah L. Gundry (https://orcid.org/0000-0002-9263-833X)[1]

[1]CardiOmics Program, Center for Heart and Vascular Research; Division of Cardiovascular Medicine; and Department of Cellular and Integrative Physiology, University of Nebraska Medical Center, Omaha, NE, 68198, USA

Running Title: CIRFESS informs cell surface proteomics

Address reprint requests to: Rebekah L. Gundry, PhD, University of Nebraska Medical Center, Department of Cellular and Integrative Physiology, 985850 Nebraska Medical Center, Omaha, NE, 68198-5850, USA, Telephone: 402-559-4426, Fax: 402-559-4438, Email: rebekah.gundry@unmc.edu

Abstract

Cell surface transmembrane, extracellular, and secreted proteins are high value targets for immunophenotyping, drug development, and studies related to intercellular communication in health and disease. As the number of specific and validated affinity reagents that target this subproteome are limited, mass spectrometry (MS)-based approaches will continue to play a critical role in enabling discovery and quantitation of these molecules. Given the technical considerations that make MS-based cell surface proteome studies uniquely challenging, it can be difficult to select an appropriate experimental approach. To this end, we have integrated multiple prediction strategies and annotations into a single online resource, Compiled Interactive Resource for Extracellular and Surface Studies (CIRFESS). CIRFESS enables rapid interrogation of the human proteome to reveal the cell surface proteome theoretically detectable by current approaches and highlights where current prediction strategies provide concordant and discordant information. We applied CIRFESS to identify the percentage of various subsets of the proteome which are expected to be captured by targeted enrichment strategies, including two established methods and one that is possible but not yet demonstrated. These results will inform the selection of available proteomic strategies and development of new strategies to enhance coverage of the cell surface and extracellular proteome. CIRFESS is available at www.cellsurfer.net/cirfess.

Introduction

The emergence of proteomics as a major discipline within the life science has been in no small part due to the development and eager adoption of computational strategies to enable the rapid analysis of mass spectrometry (MS) data files and inferred biological results. Since 1994, when the Yates laboratory introduced SEQUEST[1], the first computational tool for fully automated database searching, continued developments in database construction, algorithm design, and software development have propelled the evolution of MS-based proteomics[2–9]. All aspects of MS-based proteomics, including interpretation of raw spectra and database searching, visualization of results, and subsequent biological inferences benefit from advances in bioinformatics. Beyond the analysis of experimental data, data science tools that integrate machine-learning or ontological resources have become increasingly popular for prediction and classification of protein-level information, a subject of recent review[10]. Such approaches rely on experimental data to train or inform predictions and annotations, and in turn, the prediction strategies can benefit experimental design.

To scientists at the bench, perhaps the most exciting and impactful bioinformatic tools are those that can inform the next experiment. To this end, web-based formats have become increasingly popular resources as they often require less setup (*e.g.* installation), avoid operating system compatibility issues, and can be used in a familiar framework. Hundreds of web-based bioinformatics tools are now available for proteomics (*e.g.* www.expasy.org/proteomics). Current tools span a broad range of utility, including systems-level distribution of proteins based on experimental observations, visualization of experimental results, and prediction and cataloging of specific post-translational modifications, interactomes, and subcellular proteomes. Despite the increase in availability of web-based proteomics tools, there are currently relatively few resources designed to specifically assist in experimental design and analysis of the cell surface proteome.

The cell surface and extracellular space contain proteins which play key roles in a wide range of biological processes and can be utilized as valuable markers for immunophenotyping and drug targets. Despite their importance, the cell surface proteome remains relatively poorly characterized compared to the depth that most intracellular proteomes have been described. Given the relative low abundance, presence of hydrophobic transmembrane spanning regions, and dynamic nature of the cell surface proteome owing to continuous cycling of proteins due to internalization, secretion, and stimulus-triggered recruitment to the plasma membrane, specialized techniques are typically required to enhance the detection of cell surface proteins by MS. Such proteomic approaches include enrichment strategies which exploit the biophysical properties of membranes - such as density gradient flotation, differential centrifugation, or silica-bead capture[11–15] - or affinity-based approaches that use proximity labels[16–18], lectins[19], metabolic[20–22] or chemical labels[23–27] to enrich cell surface proteins. Application of these approaches have supported efforts to catalog the cell surface and secretome and have led to large scale efforts in experimentation[28] and collation[29]. As with most proteomic methods, currently available strategies to probe the cell surface and secreted proteome are biased towards proteins that contain specific features (*e.g.* presence of an *N*-glycosylation site or lysine within an extracellular region of the protein that will generate a detectable peptide after trypsin digestion). Also, the implementation of these approaches can be inconsistent among users, resulting in variability in the specificity (*i.e.* cell surface versus non-cell surface) of enrichment observed. Hence, the integration of bioinformatic predictions with experimental data provides orthogonal means to interpret and filter results. Relevant predictions for surface and extracellular proteins include the presence of signal peptides[30–36] and transmembrane domains[30,31,37–39]. Other approaches have applied manual curation, ontological annotations, or machine learning approaches to predict the subset of proteins that are localized to the cell surface and extracellular regions[40–43]. However, not all cell surface proteins contain canonical signal peptides[44]. Also, GPI-anchored and extracellular matrix or secreted proteins do not contain transmembrane domains, and gene ontology annotations may be insufficiently specific *(e.g.* cell surface versus membrane). Thus, filtering a proteomic dataset by these constraints often does not provide the complete picture of the cell surface proteome for a specific cell type.

Based on these limitations, in theory, it remains necessary to rely on experimental data to precisely define the proteins localized to the cell surface in a specific cell type. This leads to the question: *Which experimental approach is the best to use?* No doubt, the answer will be context dependent. If a specific monoclonal antibody is available, flow cytometry can be an effective approach for determining the surface localization of a protein. If antibodies are not available, the MS-based proteomic method of choice will depend on whether a cell surface proteome-wide screen or detection of a particular protein or subclass of proteins is desired. It will depend on the type and availability of the source material (*e.g.* metabolic labeling approaches cannot be used routinely for the analysis of primary human cells). Given the numerous technical considerations that make cell surface proteome studies uniquely challenging, it can be difficult to decide which approach to use. Currently, there is no single bioinformatic tool that can assist the investigator in determining which MS-based method is likely to be the most suitable approach for surface proteome studies.

To address this, we constructed a resource that integrates multiple prediction strategies and annotations relevant for the analysis of cell surface and extracellular proteins by MS and applied it to interrogate the human proteome. The results from these resources were compiled into a single interface and are accessible via a web-application termed Compiled Interactive Resource for Extracellular and Surface Studies (CIRFESS), accessible at www.cellsurfer.net/cirfess. By bringing together key resources used to interrogate the surface and extracellular space, CIRFESS helps to prevent duplication of efforts and continued pinging of separate prediction servers for the same set of proteins. We expect CIRFESS will be informative for a broad range of future applications and will inform the selection or development of proteomic strategies to enhance coverage of the cell surface and extracellular proteome.

Methods

Database and prediction server access

The human reference proteome was downloaded from UniProt (canonical only, 20416 entries, accessed Sept 13, 2019). The proteome was filtered and split to meet the requirements of the individual prediction servers (*e.g.* length of proteins, number of entries). Default scoring settings were applied, and the outputs were collected as specified: TMHMM– 'one line per protein', Phobius – 'Short', PrediSi, – 'Text', Signal P – 'Short output'. For analysis involving the protein-level evidence, the "Protein existence" field for each accession number was retrieved from UniProt. To aid in interpretation, a summary of the different categories is provided in the Table 1 (adapted from https://www.uniprot.org/help/protein_existence).

Assigned Bin	UniProt "Protein existence" value(s)	UniProt Definition
Protein-level evidence	Experimental evidence at protein level	There is clear experimental evidence for the existence of the protein. The criteria include partial or complete Edman sequencing, clear identification by mass spectrometry, X-ray or NMR structure, good quality protein-protein interaction or detection of the protein by antibodies.
Transcript-level evidence	Experimental evidence at transcript level	The existence of a protein has not been strictly proven but that expression data (such as existence of cDNA(s), RT-PCR or Northern blots) indicate the existence of a transcript.
Oher	Protein inferred by homology	The existence of a protein is probable because clear orthologs exist in closely related species.
	Protein predicted	Entries without evidence at protein, transcript, or homology levels.
	Protein uncertain	The existence of the protein is unsure.

Table 1: Summary of UniProt levels of "Protein existence" and the corresponding bins used in the analysis of levels of evidence for the different subsets of proteins.

Integrating prediction outputs

The outputs from the independent prediction servers were parsed and integrated using Python 3.7. The source code is made available as a Jupyter Notebook to enable implementation on batch predictions from TMHMM, Phobius, PrediSi and SignalP[30,36,39,45] for other species with minimal alteration (https://github.com/GundryLab/cirfess). A schematic of the inputs and the generated data structure is shown in Figure 1.

Evaluating peptides, motifs, and topology

Protein sequences were digested *in silico* to generate a list of potential peptides using the canonical tryptic cleavage site, X[R/K] where X is not P, allowing for up to two missed cleavages. This list of peptides is subsequently annotated with the following information: (1) presence of motifs for relevant proteomic capture strategies; N[!P][S/T/C/V] for *N*-glycan based capture, C for cysteine-based capture, K for lysine-based capture, (2) topological information – which residues and motifs are predicted to be intracellular and extracellular, and (3) suitability for a standard bottom-up proteomic experiment – length > 5, *m/z* of 2+ or 3+ charge state peptide < 2000.

CIRFESS Web application

A web application (CIRFESS) for accessing the data structure containing the parsed prediction outputs was developed in R[46] using the Shiny package and is available at www.cellsurfer.net/cirfess. Source code is available at https://github.com/GundryLab/cirfess.

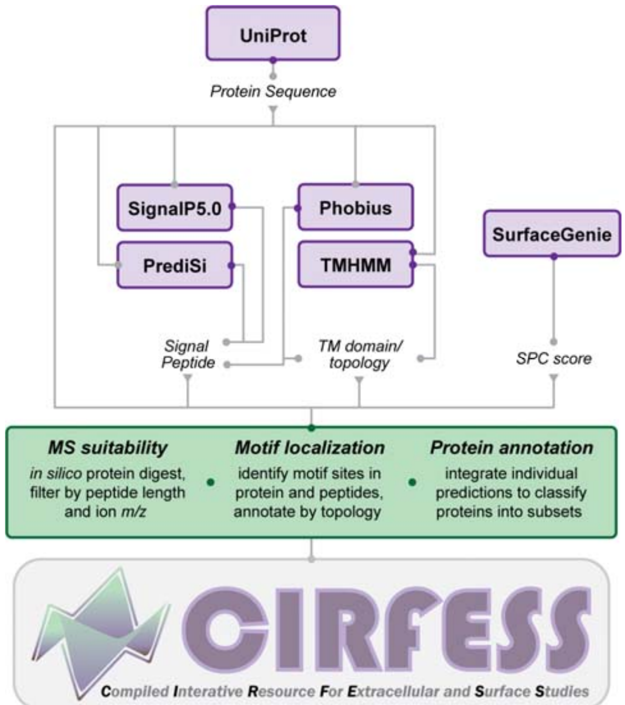


Figure 1: Schematic of the resources used (in purple) and the analyses performed (in green) for the construction of CIRFESS.

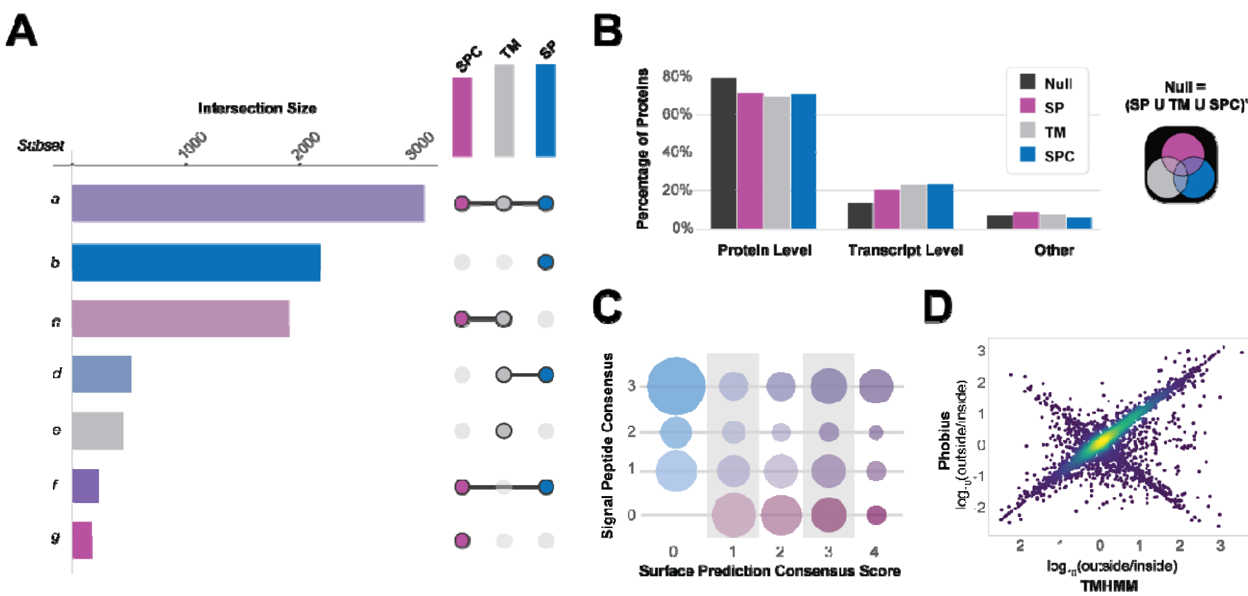
Statistical Analysis

Chi-square analyses were performed using the *chisq.test* function in R (version 3.6.2). Students t-test were performed using the *t.test* function in Excel.

Results and Discussion

Proteome-wide comparison of prediction strategies

There are multiple sources of information to consider for evidence of surface or extracellular localization. Here we utilized transmembrane (TM) predictions[30,36], signal peptide (SP) predictions[31,39], and the Surface Prediction Consensus (SPC)[47] score. To highlight the complementarity of these measures and justification for inclusion in this resource, we performed set analysis of the human proteome using the UpSetR[48] web application. As shown in Figure 2A, the combination of these measures reveals distinct subsets of proteins. For instance, proteins which have a SP, but no TM domain (subsets b,f) are often considered to be the set of secreted proteins. For the set of proteins with predicted TM domains and SP, the SPC score is helpful for distinguishing between organelle membrane proteins (subset d) and cell surface proteins (subset a). Finally, proteins with an SPC score but not TM or SP may contain GPI-



anchored proteins or those without a canonical SP (subset g).

Figure 2: Contrasting views of the human proteome based on prediction strategies relevant for the cell surface proteome. (A) UpSet plot illustrating overlap in human proteins that are classified as containing a SP, TM domain, or SPC > 0. (B) Bar graph depicting the percentage of the human proteome with different levels of evidence, gathered from the "Protein existence" level listed for each accession number in UniProt, that are classified as containing a SP, TM domain, or SPC >0, or none of these features (Null). (C) Relationship between different levels of consensus for SP prediction and SPC score. Here, SP prediction consensus was calculated in a manner analogous to SPC score, where the number of positive SP predictions from SignalP, PrediSI, or Phobius was summed to generate a consensus score ranging from 0 to 3. (D) Plot depicting the $\log_{10}$ ratio of extracellular to intracellular residues predicted by TMHMM and Phobius highlighting that the opposite orientation is predicted for a subset of proteins.

To assess the level of experimental data that currently exists for cell surface proteins, we considered the "Protein existence" annotation within UniProt. For these and further analyses, we compared proteins with positive SP predictions, positive TM predictions, or those with SPC scores > 0, with proteins that were negative for all three analysis – which we term the 'Null set' of proteins (shown visually in Figure 2B). Compared to the Null set, all three classes of proteins have a lower percentage of members with protein-level evidence, 79% for Null proteins compared to 71%, 69%, and 71% for SP, TM, and SPC proteins, respectively (Figure 2B). The difference between the observed frequencies of the different levels of evidence were significant between Null and each other subset of proteins (SP, TM, and SPC), as revealed by Chi-square testing (with p-values of $<2.2 \times 10^{-16}$ for each test). Though mass spectrometry is not the only source of protein-level evidence for UniProt (see Methods), a potential explanation for this discrepancy is a statistical difference in the number of MS-suitable peptides between these subsets, as revealed by Student's t-test summarized in Table 2. Nevertheless, this analysis highlights the need for further experimental investigation of the cell surface proteome as this class is less well-represented by experimental evidence than other subproteomes.

	Null			**SP**			**TM**			**SPC**		
Missed Cleavages	**0**	**≤ 1**	**≤ 2**	**0**	**≤ 1**	**≤ 2**	**0**	**≤ 1**	**≤ 2**	**0**	**≤ 1**	**≤ 2**
Mean # of Peptides / protein	33.1	87.5	146.4	28.3	70.5	112.3	26.2	65.0	103.4	28.1	69.8	111.0
Median # of Peptides / protein	25	66	109	20	50	79	19	48	76	20	50	79
t-test p-value (compared to Null-set)	-	-	-	6×10^{-23}	1×10^{-41}	2×10^{-58}	2×10^{-47}	8×10^{-75}	1×10^{-97}	2×10^{-23}	4×10^{-42}	4×10^{-59}

Table 2: Summary of the average number of peptides per protein that are "ok for MS" in different subsets of proteins. The t-test p-values were calculated by comparing the distribution to the Null-set of numbers of peptides (with the corresponding amount of numbers of missed cleavages).

Another benefit of integrating these disparate predictions into a single analysis is revealed by looking at examples for which they do and do not agree. For example, stratifying proteins with positive SP predictions (6026 proteins) by the number of algorithms for which it was positive reveals that slightly over half (3192, 53%) are predicted by all 3 algorithms. Here, SP prediction consensus was calculated in manner analogous to SPC score, where the number of positive SP predictions from SignalP, PrediSI, or Phobius was summed to generate a consensus score ranging from 0 to 3. Plotting the number of positive SP predictions against SPC score reveals a positive relationship between SPC score and number of positive SP predictions for proteins with SPC score >0 (Figure 2C). However, it also reveals that the majority of proteins with three positive SP predictions has an SPC score of 0 (1648 of 3192, 51.6%). This suggests that secreted proteins may contain signal peptide sequences that are easier to recognize by prediction algorithms than proteins translocated through the membrane.

Focusing on TM proteins, TMHMM and Phobius predict 5353 and 5471 proteins with TM domains, respectively, with 4846 proteins in common and 1132 proteins unique to a single prediction strategy (507 and 625 in TMHMM and Phobius, respectively). While overall there is strong consensus between the two algorithms for predicting which proteins contain TM domains, the number of TM domains predicted differs for 1306 out of the 4846 commonly predicted proteins. Furthermore, the opposite membrane orientation was predicted for a subset of proteins, visualized by plotting the $\log_{10}$ ratio of the predicted extracellular to intracellular residues (Figure 2D). Altogether, these analyses demonstrate the value of integrating data from multiple sources and reveal that no single feature is sufficient to comprehensively predict the set of cell surface and extracellular proteins.

Motif coverage of extracellular and surface predicted proteins

As prediction strategies alone are insufficient to define the set of proteins localized to the cell surface and extracellular space, experimentation is required. To aid in the selection of proteomic strategies that are likely to produce the desired coverage of the cell surface proteome, the SP, TM, and SPC analyses described above were integrated with *in silico* analyses designed to predict which proteins would generate tryptic peptides likely to be detectable by electrospray MS, and of those, which are expected to be captured by application of commonly used biorthogonal enrichment strategies targeting *N*-glycans and lysines[23–25,27,49–54]. We also considered cysteines as they are enriched in surface proteins compared to nonsurface proteins[40] and numerous affinity reagents are available for targeting cysteines[55], although this is not yet a widely described approach for cell surface proteins. Important for the *N*-glycan approach, although strategies that specifically enrich peptides from the extracellular space (glycan biotinylation is performed on cells with intact plasma membranes) provide an additional level of experimental evidence for surface localization, it is possible to capture *N*-glycopeptides from whole cell lysate. In this case, the MS-based evidence for a glycan modifying an asparagine within the consensus motif for *N*-glycosylation is proposed to serve as standalone evidence for surface localization. Canonically, the consensus motif has been described as NXS/T where X is any amino acid except proline. However, more recently, evidence for *N*-glycosylation has been put forth at NXC[56,57] and NXV[58]. Here, we investigated the frequency of the various consensus motifs occurring in SP, TM, SPC and Null (meaning the protein contains no SP, TM, or SPC) sets of proteins. First, the probability of each motif occurring within the subset of proteins was calculated with respect to the amino acid frequencies. The expected frequencies based on amino acid compositions was consistent among the sets of proteins for each motif (0.26 ± 0.003 %, 0.18 ± 0.009 %, 0.09 ± 0.009 %, and 0.21 ± 0.019 % for NXS, NXT, NXC, and NXV respectively). Next, the observed frequency of each motif was calculated for each subset of proteins. The natural log of the odds ratio of observed to expected for each subset of proteins was calculated and plotted for each motif as well as the canonical (NXS/T) and complete consensus motifs (NXS/T/C/V) (Figure 3A). The results reveal that the NXS and NXT occur more frequently than expected and NXC and NXV occur less frequently than expected for SP, TM, and SPC proteins. Whereas NXS and NXT occur at about the expected rate for the Null set of proteins, NXC and NXV occur slightly above the expected frequency. While the complete consensus motif occurs more frequently than expected for SP, TM, and SPC proteins, the canonical motif demonstrates a much higher odds ratio, especially relative to the odds-ratio for the Null set. This analysis suggests that while the mere presence of the consensus motif provides some evidentiary weight to the localization of a surface protein – (1) it should not be considered conclusive, and (2) the canonical motif provides more meaningful information than the complete consensus motif. Based on these analyses, we elected to only consider the canonical motif as potential targets for *N*-glycan capture for subsequent analyses.

By integrating the topology information provided by the TM predictions with the locations of motifs within proteins, we estimated the coverage that each capture strategy would provide for each subset of the proteome (Null, SP, TM and SPC). For this analysis, proteins were categorized based on whether they contained the relevant motif and whether the motif was in a region determined to be extracellular by one or both TM prediction strategies. The percentage of proteins for which a predicted extracellular motif was located within an MS-suitable peptide was recorded (Figure 3B). This analysis revealed that while 72% of Null proteins contain a consensus motif for *N*-glycosylation, none of the glycopeptides are predicted to be in the extracellular domain. In contrast, of the SPC proteins which contain the consensus motif, 86% of those proteins contain at least one peptide contains the consensus motif within the predicted extracellular domain. These results were further summarized by calculating the percentage of each subproteome that is predicted to be covered by each or multiple capture strategies (Figure 3C). Overall, querying the results from this analysis provides a strategy for investigators to rapidly interrogate the human proteome to determine which experimental strategy is most likely to be useful to address their biological question. In summary, 66.4 ± 0.4% of SP, TM and SPC proteins are likely to be captured by any of the three strategies, 17.3 ± 1.8% are detectable by cysteine or lysine capture, but not detectable by *N*-glycan strategies, 7.4 ± 1.4% are detectable by a single strategy, and 5.7 ± 1.2% are not detectable by any of the three strategies considered here. The identity of the proteins within each classification are provided in Supplemental Table

1A-C and these results provide actionable data related to high interest targets. For example, of the 825 human G-protein coupled receptors (GPCR), a striking 65.3% lack protein-level evidence within UniProt. Of these, all but one are predicted to be captured by at least one enrichment strategy and 70.9% of them are predicted to be captured by all three strategies. Supplementary Table 1D contains the identity of the GPCR proteins and which enrichment strategies are predicted to capture them.

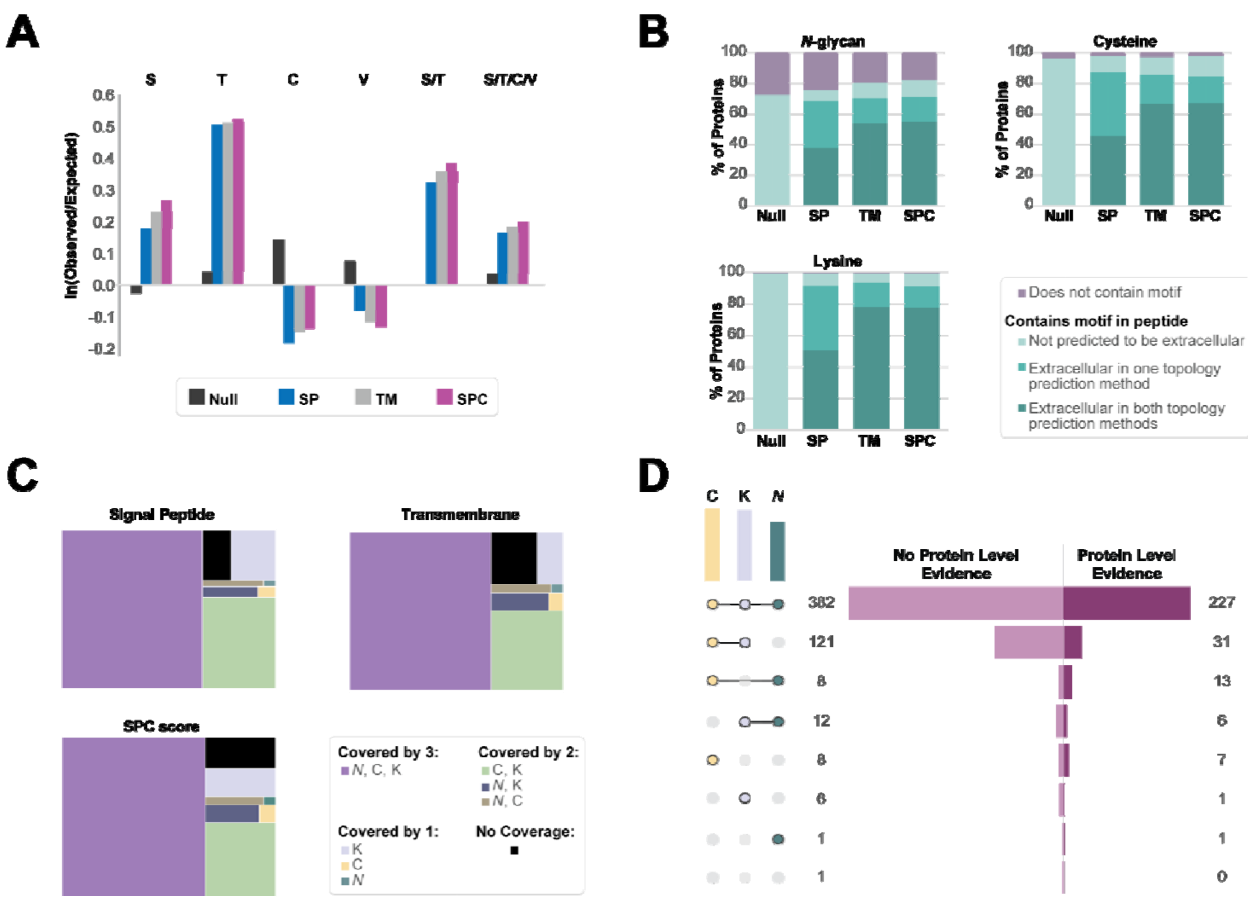


Figure 3. Results of CIRFESS analysis of the human proteome to assess predicted coverage provided by three common cell surface proteomic enrichment strategies. (A) The natural log of the odds ratio for observed-to-expected frequency of each permutation of the *N*-glycan consensus motif along with the canonical (S/T) and complete (S/T/C/V) consensus motif. (B) The expected coverage of the different subsets of proteins for each enrichment strategy broken down by which proteins have peptides with predicted extracellular motifs by one or both TM prediction methods. (C) The makeup of SP, TM, and SPC score proteins based on the

overlapping coverage of the three individual enrichment strategies. (D) The set of human GPCRs based on expected coverage for enrichment strategy and level of evidence in UniProt.

Critical Considerations

Results from the current implementation of CIRFESS are limited to human proteins digested with trypsin and the resulting peptides are detectable in the 2+ or 3+ charge state. These criteria were selected based on common implementation of bottom-up proteomic methods. However, all source files and code are publicly available in the Github repository and a user-specific version of CIRFESS could be generated, requiring minimum alteration to change the *in silico* digestion strategies or criteria for MS-compatible peptide filtering. Implementation on other species would require submission of proteins to the individual prediction servers, but the source code includes scripts to parse and integrate the generated output files. Another critical assumption is related to the *N*-gycan capture strategy where detection depends on the glycosite being occupied by a glycan which is sensitive to the oxidation strategy applied (*e.g.* cis diols for meta-periodate[59]). Currently, as it is not possible to predict which sites will be occupied with specific glycan structures, the peptides predicted to be observable by this strategy should be considered with this caveat in mind. It is possible that post-translational modifications may interfere with the digestion, capture, ionization, and identification of peptides in any of the strategies, and therefore experimental observations may not be fully predictable by this bioinformatics approach. Among the post-translation modification which may interfere with cysteine-based capture are disulfide bridges, which were ignored in this analysis, but a reduction step could be included prior to labeling in such an approach. Moreover, for enrichment strategies which use cleavable linkers, residual portions of the linker that remain after cleavage will increase the mass of the resulting peptide. However, the 2000 *m/z* range used here for predicting detectable peptides should accommodate most commonly used reagents. Finally, it may be beneficial to combine the results from CIRFESS analysis with predictions for peptide detectability[60–62] or proteotypicity[63] to better inform the set of peptides which are most likely to be observable or informative.

Conclusion

CIRFESS is a web-based tool designed to accelerate cell surface proteome studies by eliminating the need to query each bioinformatics source separately and integrating disparate features into a single streamlined resource and output. Within the CIRFESSS interface, users are able to perform single and batch querying of protein accession numbers to extract protein-level and peptide-level annotations as well as information about numbers of motifs and motif-containing peptides. Results may be queried for proteins or protein classes of interest to inform the design of the next experiment. We anticipate that CIRFESS will be broadly applicable for multiple applications across a broad range of biology and disease studies. While there still exist significant technical challenges associated with the implementation of these technologies, particularly on sample-limited systems, these analyses suggest that acquiring protein-level evidence for the majority of predicted cell surface proteins is a matter of applying the right technology to a relevant biological system. Overall, we expect CIRFESS will promote the rational selection of the most apt cell surface proteomic methods and will inspire continued method development (*e.g.* cysteine-targeting) to enable detection of the human proteome not predicted to be accessible by established surface protein enrichment methods.

Author Contributions

R.L.G. and M.W. conceived the study; R.L.G. supervised the study; M.W. and J. L. developed the algorithms and developed the web application; M.W. and R.L.G. analyzed data; M.W.

generated figures; M.W. and R.L.G. co-wrote the manuscript; All authors approved the final manuscript.

Acknowledgements

This work was supported by the National Institutes of Health [R01-HL126785, R01-HL134010, P20GM104320 (as a pilot grant award) to R.L.G.; F31-HL140914 to M.W.]; and JDRF [2-SRA-2019-829-S-B to R.L.G.]; Special thanks to Dr. Christopher Ashwood and Linda Berg Luecke for critical review of the manuscript and insightful discussions and Dr. David Tabb for helpful consultation. Funding sources were not involved in study design, data collection, interpretation, analysis or publication.

Supplementary Information

Supplemental Table 1. A. Coverage of SP proteins. B. Coverage of TM proteins. C. Coverage of SPC Proteins. D. Coverage and evidence of G-protein coupled receptors.

References

(1) Eng, J. K.; McCormack, A. L.; Yates, J. R. An Approach to Correlate Tandem Mass Spectral Data of Peptides with Amino Acid Sequences in a Protein Database. *Journal of the American Society for Mass Spectrometry* **1994**, *5* (11), 976–989.

(2) Eng, J. K.; Jahan, T. A.; Hoopmann, M. R. Comet: An Open-Source MS/MS Sequence Database Search Tool. *Proteomics* **2013**, *13* (1), 22–24.

(3) Gluck, F.; Hoogland, C.; Antinori, P.; Robin, X.; Nikitin, F.; Zufferey, A.; Pasquarello, C.; Fétaud, V.; Dayon, L.; Müller, M.; Lisacek, F.; Geiser, L.; Hochstrasser, D.; Sanchez, J. C.; Scherl, A. EasyProt - An Easy-to-Use Graphical Platform for Proteomics Data Analysis. *Journal of Proteomics* **2013**, *79*, 146–160.

(4) Wenger, C. D.; Coon, J. J. A Proteomics Search Algorithm Specifically Designed for High-Resolution Tandem Mass Spectra. *Journal of Proteome Research* **2013**, *12* (3), 1377–1386.

(5) Dorfer, V.; Pichler, P.; Stranzl, T.; Stadlmann, J.; Taus, T.; Winkler, S.; Mechtler, K. MS Amanda, a Universal Identification Algorithm Optimized for High Accuracy Tandem Mass Spectra. *Journal of Proteome Research* **2014**, *13* (8), 3679–3684.

(6) Kim, S.; Pevzner, P. A. MS-GF+ Makes Progress towards a Universal Database Search Tool for Proteomics. *Nature Communications* **2014**, *5*.

(7) Risk, B. A.; Spitzer, W. J.; Giddings, M. C. Peppy: Proteogenomic Search Software. *Journal of Proteome Research* **2013**, *12* (6), 3019–3025.

(8) Cox, J.; Mann, M. MaxQuant Enables High Peptide Identification Rates, Individualized p.p.b.-Range Mass Accuracies and Proteome-Wide Protein Quantification. *Nature Biotechnology* **2008**, *26* (12), 1367–1372.

(9) Perkins, D. N.; Pappin, D. J. C.; Creasy, D. M.; Cottrell, J. S. Probability-Based Protein Identification by Searching Sequence Databases Using Mass Spectrometry Data. In *Electrophoresis*; Wiley-VCH Verlag, 1999; Vol. 20, pp 3551–3567.

(10) Ching, T.; Himmelstein, D. S.; Beaulieu-Jones, B. K.; Kalinin, A. A.; Do, B. T.; Way, G. P.; Ferrero, E.; Agapow, P. M.; Zietz, M.; Hoffman, M. M.; Xie, W.; Rosen, G. L.; Lengerich, B. J.; Israeli, J.; Lanchantin, J.; Woloszynek, S.; Carpenter, A. E.; Shrikumar, A.; Xu, J.; Cofer, E. M.; Lavender, C. A.; Turaga, S. C.; Alexandari, A. M.; Lu, Z.; Harris, D. J.; Decaprio, D.; Qi, Y.; Kundaje, A.; Peng, Y.; Wiley, L. K.; Segler, M. H. S.; Boca, S. M.; Swamidass, S. J.; Huang, A.; Gitter, A.; Greene, C. S. Opportunities and Obstacles for Deep Learning in Biology and Medicine. *Journal of the Royal Society Interface* **2018**, *15* (141).

(11) Chaney, L. K.; Jacobson, B. S. Coating Cells with Colloidal Silica for High Yield Isolation of Plasma Membrane Sheets and Identification of Transmembrane Proteins. *J. Biol. Chem.* **1983**, *258* (16), 10062–10072.

(12) Kim, Y.; Elschenbroich, S.; Sharma, P.; Sepiashvili, L.; Gramolini, A. O.; Kislinger, T. Use of Colloidal Silica-Beads for the Isolation of Cell-Surface Proteins for Mass Spectrometry-Based Proteomics. In *Immune Receptors: Methods and Protocols*; Rast, J. P., Booth, J. W. D., Eds.; Methods in Molecular Biology; Humana Press: Totowa, NJ, 2011; pp 227–241.

(13) Mulvey, C. M.; Breckels, L. M.; Geladaki, A.; Britovšek, N. K.; Nightingale, D. J. H.; Christoforou, A.; Elzek, M.; Deery, M. J.; Gatto, L.; Lilley, K. S. Using HyperLOPIT to Perform High-Resolution Mapping of the Spatial Proteome. *Nature Protocols* **2017**, *12* (6), 1110–1135.

(14) Jett, M.; Seed, T. M.; Jamieson, G. A. Isolation and Characterization of Plasma Membranes and Intact Nuclei from Lymphoid Cells. *J. Biol. Chem.* **1977**, *252* (6), 2134–2142.

(15) Jones, D. H.; Matus, A. I. Isolation of Synaptic Plasma Membrane from Brain by Combined Flotation-Sedimentation Density Gradient Centrifugation. *Biochimica et Biophysica Acta (BBA) - Biomembranes* **1974**, *356* (3), 276–287.

(16) Rees, J. S.; Li, X.-W.; Perrett, S.; Lilley, K. S.; Jackson, A. P. Selective Proteomic Proximity Labeling Assay Using Tyramide (SPPLAT): A Quantitative Method for the Proteomic Analysis of Localized Membrane-Bound Protein Clusters. *Current Protocols in Protein Science* **2015**, *80* (1), 19.27.1-19.27.18.

(17) Lee, S.-Y.; Kang, M.-G.; Park, J.-S.; Lee, G.; Ting, A. Y.; Rhee, H.-W. APEX Fingerprinting Reveals the Subcellular Localization of Proteins of Interest. *Cell Rep* **2016**, *15* (8), 1837–1847.

(18) Roux, K. J.; Kim, D. I.; Raida, M.; Burke, B. A Promiscuous Biotin Ligase Fusion Protein Identifies Proximal and Interacting Proteins in Mammalian Cells. *J. Cell Biol.* **2012**, *196* (6), 801–810.

(19) Ghosh, D.; Krokhin, O.; Antonovici, M.; Ens, W.; Standing, K. G.; Beavis, R. C.; Wilkins, J. A. Lectin Affinity as an Approach to the Proteomic Analysis of Membrane Glycoproteins. *J. Proteome Res.* **2004**, *3* (4), 841–850.

(20) Breidenbach, M. A.; Gallagher, J. E. G.; King, D. S.; Smart, B. P.; Wu, P.; Bertozzi, C. R. Targeted Metabolic Labeling of Yeast N-Glycans with Unnatural Sugars. *Proc. Natl. Acad. Sci. U.S.A.* **2010**, *107* (9), 3988–3993.

(21) Smeekens, J. M.; Chen, W.; Wu, R. Mass Spectrometric Analysis of the Cell Surface N-Glycoproteome by Combining Metabolic Labeling and Click Chemistry. *J. Am. Soc. Mass Spectrom.* **2015**, *26* (4), 604–614.

(22) Sun, T.; Yu, S.-H.; Zhao, P.; Meng, L.; Moremen, K. W.; Wells, L.; Steet, R.; Boons, G.-J. One-Step Selective Exoenzymatic Labeling (SEEL) Strategy for the Biotinylation and Identification of Glycoproteins of Living Cells. *J. Am. Chem. Soc.* **2016**, *138* (36), 11575–11582.

(23) Kalxdorf, M.; Gade, S.; Eberl, H. C.; Bantscheff, M. Monitoring Cell-Surface N-Glycoproteome Dynamics by Quantitative Proteomics Reveals Mechanistic Insights into Macrophage Differentiation. *Molecular & cellular proteomics : MCP* **2017**, *16* (5), 770–785.

(24) Turtoi, A.; Dumont, B.; Greffe, Y.; Blomme, A.; Mazzucchelli, G.; Delvenne, P.; Mutijima, E. N.; Lifrange, E.; De Pauw, E.; Castronovo, V. Novel Comprehensive Approach for Accessible Biomarker Identification and Absolute Quantification from Precious Human Tissues. *Journal of Proteome Research* **2011**, *10* (7), 3160–3182.

(25) Wollscheid, B.; Bausch-Fluck, D.; Henderson, C.; O'Brien, R.; Bibel, M.; Schiess, R.; Aebersold, R.; Watts, J. D. Mass-Spectrometric Identification and Relative Quantification of N-Linked Cell Surface Glycoproteins. *Nature biotechnology* **2009**, *27* (4), 378–386.

(26) Zhang, H.; Li, X. jun; Martin, D. B.; Aebersold, R. Identification and Quantification of N-Linked Glycoproteins Using Hydrazide Chemistry, Stable Isotope Labeling and Mass Spectrometry. *Nature Biotechnology* **2003**, *21* (6), 660–666.

(27) Castronovo, V.; Kischel, P.; Guillonneau, F.; de Leval, L.; Defféchereux, T.; De Pauw, E.; Neri, D.; Waltregny, D. Identification of Specific Reachable Molecular Targets in Human Breast Cancer Using a Versatile Ex Vivo Proteomic Method. *Proteomics* **2007**, *7* (8), 1188–1196.

(28) Bausch-Fluck, D.; Hofmann, A.; Bock, T.; Frei, A. P.; Cerciello, F.; Jacobs, A.; Moest, H.; Omasits, U.; Gundry, R. L.; Yoon, C.; Schiess, R.; Schmidt, A.; Mirkowska, P.; Härtlová, A.; Van Eyk, J. E.; Bourquin, J.-P.; Aebersold, R.; Boheler, K. R.; Zandstra, P.; Wollscheid, B. A Mass Spectrometric-Derived Cell Surface Protein Atlas. *PLOS ONE* **2015**, *10* (4), e0121314.

(29) Uhlén, M.; Karlsson, M. J.; Hober, A.; Svensson, A.-S.; Scheffel, J.; Kotol, D.; Zhong, W.; Tebani, A.; Strandberg, L.; Edfors, F.; Sjöstedt, E.; Mulder, J.; Mardinoglu, A.; Berling, A.;

Ekblad, S.; Dannemeyer, M.; Kanje, S.; Rockberg, J.; Lundqvist, M.; Malm, M.; Volk, A.-L.; Nilsson, P.; Månberg, A.; Dodig-Crnkovic, T.; Pin, E.; Zwahlen, M.; Oksvold, P.; von Feilitzen, K.; Häussler, R. S.; Hong, M.-G.; Lindskog, C.; Ponten, F.; Katona, B.; Vuu, J.; Lindström, E.; Nielsen, J.; Robinson, J.; Ayoglu, B.; Mahdessian, D.; Sullivan, D.; Thul, P.; Danielsson, F.; Stadler, C.; Lundberg, E.; Bergström, G.; Gummesson, A.; Voldborg, B. G.; Tegel, H.; Hober, S.; Forsström, B.; Schwenk, J. M.; Fagerberg, L.; Sivertsson, Å. The Human Secretome. *Science signaling* **2019**, *12* (609).

(30) Hiller, K.; Grote, A.; Scheer, M.; Münch, R.; Jahn, D. PrediSi: Prediction of Signal Peptides and Their Cleavage Positions. *Nucleic Acids Research* **2004**, *32* (WEB SERVER ISS.).

(31) Käll, L.; Krogh, A.; Sonnhammer, E. L. L. A Combined Transmembrane Topology and Signal Peptide Prediction Method. *Journal of Molecular Biology* **2004**, *338* (5), 1027–1036.

(32) Frank, K.; Sippl, M. J. High-Performance Signal Peptide Prediction Based on Sequence Alignment Techniques. *Bioinformatics* **2008**, *24* (19), 2172–2176.

(33) Signal-3L 3.0: Improving signal peptide prediction through combining attention deep learning with domain rules http://www.csbio.sjtu.edu.cn/bioinf/Signal-3L/ (accessed Jan 19, 2020).

(34) Reynolds, S. M.; Käll, L.; Riffle, M. E.; Bilmes, J. A.; Noble, W. S. Transmembrane Topology and Signal Peptide Prediction Using Dynamic Bayesian Networks. *PLoS Computational Biology* **2008**, *4* (11).

(35) Savojardo, C.; Martelli, P. L.; Fariselli, P.; Casadio, R. DeepSig: Deep Learning Improves Signal Peptide Detection in Proteins. *Bioinformatics* **2018**, *34* (10), 1690–1696.

(36) Almagro Armenteros, J. J.; Tsirigos, K. D.; Sønderby, C. K.; Petersen, T. N.; Winther, O.; Brunak, S.; von Heijne, G.; Nielsen, H. SignalP 5.0 Improves Signal Peptide Predictions Using Deep Neural Networks. *Nature Biotechnology* **2019**, *37* (4), 420–423.

(37) Hirokawa, T.; Boon-Chieng, S.; Mitaku, S. SOSUI: Classification and Secondary Structure Prediction System for Membrane Proteins. *Bioinformatics* **1998**, *14* (4), 378–379.

(38) Bernhofer, M.; Kloppmann, E.; Reeb, J.; Rost, B. TMSEG: Novel Prediction of Transmembrane Helices. *Proteins: Structure, Function and Bioinformatics* **2016**, *84* (11), 1706–1716.

(39) Sonnhammer, E. L.; von Heijne, G.; Krogh, A. A Hidden Markov Model for Predicting Transmembrane Helices in Protein Sequences. *Proceedings / ... International Conference on Intelligent Systems for Molecular Biology ; ISMB. International Conference on Intelligent Systems for Molecular Biology* **1998**, *6*, 175–182.

(40) Bausch-Fluck, D.; Goldmann, U.; Müller, S.; van Oostrum, M.; Müller, M.; Schubert, O. T.; Wollscheid, B. The in Silico Human Surfaceome. *Proceedings of the National Academy of Sciences of the United States of America* **2018**, *115* (46), E10988–E10997.

(41) da Cunha, J. P. C.; Galante, P. A. F.; de Souza, J. E.; de Souza, R. F.; Carvalho, P. M.; Ohara, D. T.; Moura, R. P.; Oba-Shinja, S. M.; Marie, S. K. N.; Silva, W. A.; Perez, R. O.; Stransky, B.; Pieprzyk, M.; Moore, J.; Caballero, O.; Gama-Rodrigues, J.; Habr-Gama, A.; Kuo, W. P.; Simpson, A. J.; Camargo, A. A.; Old, L. J.; de Souza, S. J. Bioinformatics Construction of the Human Cell Surfaceome. *Proceedings of the National Academy of Sciences of the United States of America* **2009**, *106* (39), 16752–16757.

(42) Town, J.; Pais, H.; Harrison, S.; Stead, L. F.; Bataille, C.; Bunjobpol, W.; Zhang, J.; Rabbitts, T. H. Exploring the Surfaceome of Ewing Sarcoma Identifies a New and Unique Therapeutic Target. *Proceedings of the National Academy of Sciences of the United States of America* **2016**, *113* (13), 3603–3608.

(43) Díaz-Ramos, M. C.; Engel, P.; Bastos, R. Towards a Comprehensive Human Cell-Surface Immunome Database. *Immunology letters* **2011**, *134* (2), 183–187.

(44) Kuchler, K.; Thorner, J. Membrane Translocation of Proteins without Hydrophobic Signal Peptides. *Current Opinion in Cell Biology* **1990**, *2* (4), 617–624.

(45) Käll, L.; Krogh, A.; Sonnhammer, E. L. L. A Combined Transmembrane Topology and Signal Peptide Prediction Method. *Journal of Molecular Biology* **2004**, *338* (5), 1027–1036.

(46) *R Core Team (2014). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing,*; Vienna, Austria.

(47) Waas, M.; Snarrenberg, S. T.; Littrell, J.; Lipinski, R. A. J.; Hansen, P. A.; Corbett, J. A.; Gundry, R. L. SurfaceGenie: A Web-Based Application for Prioritizing Cell-Type Specific Marker Candidates. *https://doi.org/10.1101/575969* **2020**, In Revision.

(48) Lex, A.; Gehlenborg, N.; Strobelt, H.; Vuillemot, R.; Pfister, H. UpSet: Visualization of Intersecting Sets. *IEEE transactions on visualization and computer graphics* **2014**, *20* (12), 1983–1992.

(49) Gundry, R. L.; Riordon, D. R.; Tarasova, Y.; Chuppa, S.; Bhattacharya, S.; Juhasz, O.; Wiedemeier, O.; Milanovich, S.; Noto, F. K.; Tchernyshyov, I.; Raginski, K.; Bausch-Fluck, D.; Tae, H.-J.; Marshall, S.; Duncan, S. A.; Wollscheid, B.; Wersto, R. P.; Rao, S.; Eyk, J. E. V.; Boheler, K. R. A Cell Surfaceome Map for Immunophenotyping and Sorting Pluripotent Stem Cells. *Molecular & Cellular Proteomics* **2012**, *11* (8), 303–316.

(50) Ye, X.; Chan, K. C.; Waters, A. M.; Bess, M.; Harned, A.; Wei, B.-R.; Loncarek, J.; Luke, B. T.; Orsburn, B. C.; Hollinger, B. D.; Stephens, R. M.; Bagni, R.; Martinko, A.; Wells, J. A.; Nissley, D. V.; McCormick, F.; Whiteley, G.; Blonder, J. Comparative Proteomics of a Model MCF10A-KRasG12V Cell Line Reveals a Distinct Molecular Signature of the KRasG12V Cell Surface. *Oncotarget* **2016**, *7* (52), 86948–86971.

(51) Ravenhill, B. J.; Soday, L.; Houghton, J.; Antrobus, R.; Weekes, M. P. Comprehensive Cell Surface Proteomics Defines Markers of Classical, Intermediate and Non-Classical Monocytes. *Scientific Reports* **2020**, *10* (1), 1–11.

(52) Chauhan, S.; Danielson, S.; Clements, V.; Edwards, N.; Ostrand-Rosenberg, S.; Fenselau, C. Surface Glycoproteins of Exosomes Shed by Myeloid-Derived Suppressor Cells Contribute to Function. *J. Proteome Res.* **2017**, *16* (1), 238–246.

(53) Chen, W.; Smeekens, J. M.; Wu, R. A Universal Chemical Enrichment Method for Mapping the Yeast N-Glycoproteome by Mass Spectrometry (MS). *Mol. Cell Proteomics* **2014**, *13* (6), 1563–1572.

(54) Langó, T.; Róna, G.; Hunyadi-Gulyás, É.; Turiák, L.; Varga, J.; Dobson, L.; Várady, G.; Drahos, L.; Vértessy, B. G.; Medzihradszky, K. F.; Szakács, G.; Tusnády, G. E. Identification of Extracellular Segments by Mass Spectrometry Improves Topology Prediction of Transmembrane Proteins. *Scientific Reports* **2017**, *7* (1), 1–9.

(55) Elia, G. Biotinylation Reagents for the Study of Cell Surface Proteins. *PROTEOMICS* **2008**, *8* (19), 4012–4024.

(56) Wang, J.; Yang, P.; Tang, B.; Sun, X.; Zhang, R.; Guo, C.; Gong, G.; Liu, Y.; Li, R.; Zhang, L.; Dai, Y.; Li, N. Expression and Characterization of Bioactive Recombinant Human α-Lactalbumin in the Milk of Transgenic Cloned Cows. *Journal of Dairy Science* **2008**, *91* (12), 4466–4476.

(57) Sato, C.; Kim, J.-H.; Abe, Y.; Saito, K.; Yokoyama, S.; Kohda, D. *Characterization of the IV-Oligosaccharides Attached to the Atypical Asn-X-Cys Sequence of Recombinant Human Epidermal Growth Factor Receptor*; 2000; Vol. 127.

(58) Zielinska, D. F.; Gnad, F.; Wiśniewski, J. R.; Mann, M. Precision Mapping of an in Vivo N-Glycoproteome Reveals Rigid Topological and Sequence Constraints. *Cell* **2010**, *141* (5), 897–907.

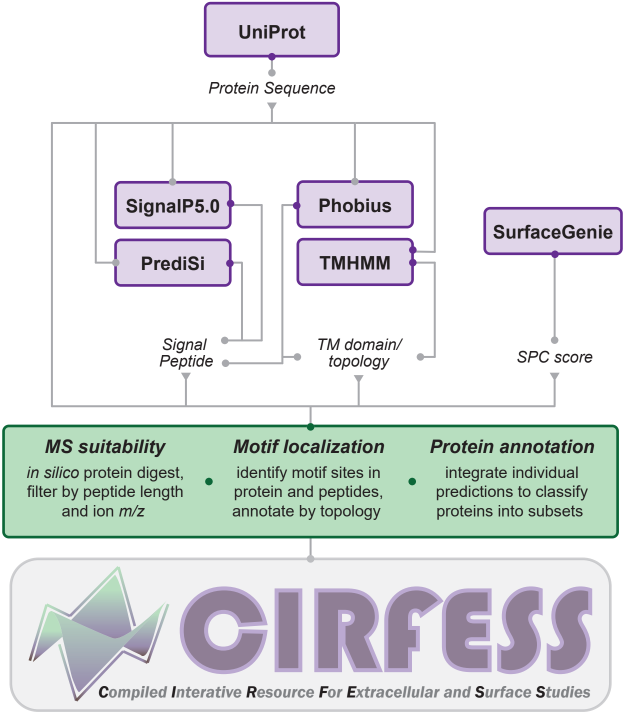
(59) BOBBITT, J. M. Periodate Oxidation of Carbohydrates. *Advances in carbohydrate chemistry* **1956**, *48* (11), 1–41.

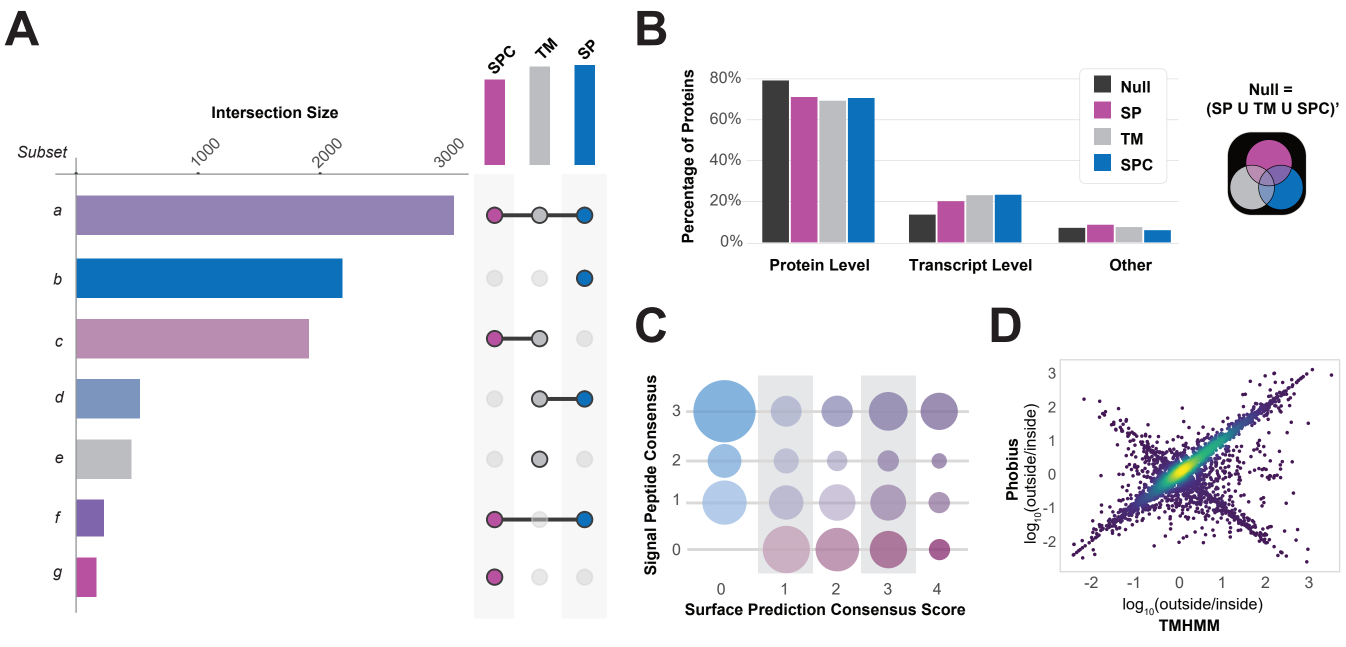
(60) Muntel, J.; Boswell, S. A.; Tang, S.; Ahmed, S.; Wapinski, I.; Foley, G.; Steen, H.; Springer, M. Abundance-Based Classifier for the Prediction of Mass Spectrometric Peptide Detectability upon Enrichment (PPA). *Molecular & cellular proteomics□: MCP* **2015**, *14* (2), 430–440.

(61) Fusaro, V. A.; Mani, D. R.; Mesirov, J. P.; Carr, S. A. Prediction of High-Responding Peptides for Targeted Protein Assays by Mass Spectrometry. *Nature biotechnology* **2009**, *27* (2), 190–198.

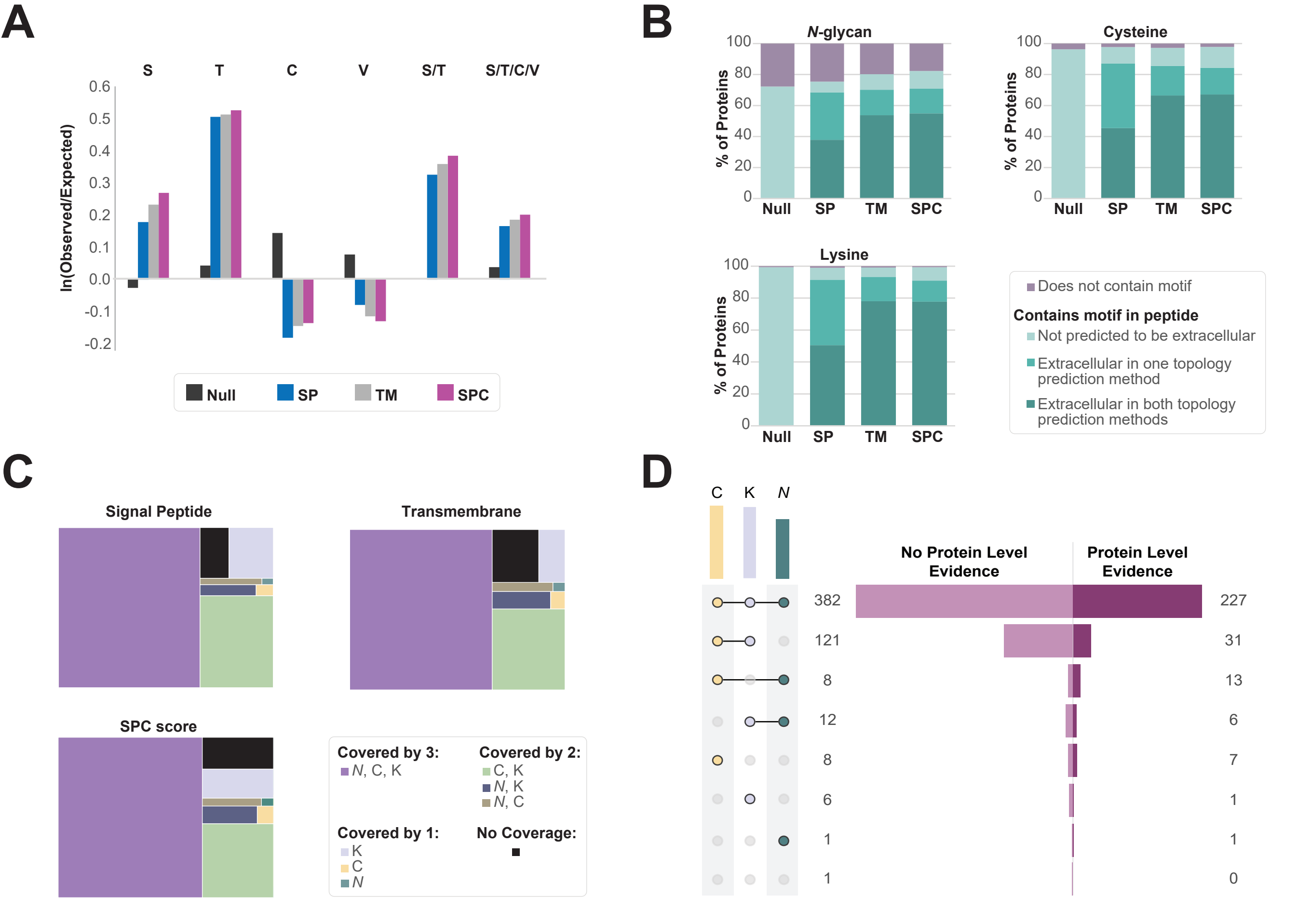
(62) Eyers, C. E.; Lawless, C.; Wedge, D. C.; Lau, K. W.; Gaskell, S. J.; Hubbard, S. J. CONSeQuence: Prediction of Reference Peptides for Absolute Quantitative Proteomics Using Consensus Machine Learning Approaches. *Molecular and Cellular Proteomics* **2011**, *10* (11).

(63) Searle, B. C.; Egertson, J. D.; Bollinger, J. G.; Stergachis, A. B.; MacCoss, M. J. Using Data Independent Acquisition (DIA) to Model High-Responding Peptides for Targeted Proteomics Experiments. *Molecular and Cellular Proteomics* **2015**, *14* (9), 2331–2340.


UniProt
Protein Sequence
SignalP5.0
PrediSi
Phobius
TMHMM
SurfaceGenie
Signal Peptide
TM domain/ topology
SPC score
MS suitability
in silico protein digest, filter by peptide length and ion m/z
Motif localization
identify motif sites in protein and peptides, annotate by topology
Protein annotation
integrate individual predictions to classify proteins into subsets
CIRFESS
Compiled Interative Resource For Extracellular and Surface Studies



A
Intersection Size
Subset
1000
2000
3000
SPC
TM
SP
a
b
c
d
e
f
g
B
Percentage of Proteins
80%
60%
40%
20%
0%
Null
SP
TM
SPC
Protein Level
Transcript Level
Other
Null =
(SP U TM U SPC)'
C
Signal Peptide Consensus
3
2
1
0
0
1
2
3
4
Surface Prediction Consensus Score
D
Phobius
log10(outside/inside)
3
2
1
0
-1
-2
-2
-1
0
1
2
3
log10(outside/inside)
TMHMM



A
S
T
C
V
S/T
S/T/C/V
ln(Observed/Expected)
0.6
0.5
0.4
0.3
0.2
0.1
0.0
-0.1
-0.2
Null
SP
TM
SPC
B
N-glycan
Cysteine
Lysine
% of Proteins
100
80
60
40
20
0
Null
SP
TM
SPC
Does not contain motif
Contains motif in peptide
Not predicted to be extracellular
Extracellular in one topology prediction method
Extracellular in both topology prediction methods
C
Signal Peptide
Transmembrane
SPC score
Covered by 3:
N, C, K
Covered by 2:
C, K
N, K
N, C
Covered by 1:
K
C
N
No Coverage:
D
C
K
N
No Protein Level Evidence
Protein Level Evidence
382
227
121
31
8
13
12
6
8
7
6
1
1
1
1
0