Biological Sciences

SurfaceGenie: A web-based application for integrating predictive and experimental data for rational candidate surface marker prioritization

Matthew Waas (https://orcid.org/0000-0003-4537-1502)[1], Shana T. Snarrenberg (https://orcid.org/0000-0003-1439-0313)[1], Jack Littrell (https://orcid.org/0000-0003-1264-894X)[1], Rachel A. Jones Lipinski (https://orcid.org/0000-0003-4586-0445)[1], Polly A. Hansen (https://orcid.org/0000-0002-2095-2493)[1], John A. Corbett (https://orcid.org/0000-0002-1134-4664)[1], Rebekah L. Gundry (https://orcid.org/0000-0002-9263-833X)[1,2*]

[1]Department of Biochemistry, Medical College of Wisconsin, Milwaukee, WI 53226, USA

[2]Center for Biomedical Mass Spectrometry Research, Medical College of Wisconsin, Milwaukee, WI 53226, USA

***Corresponding author:**

Rebekah L. Gundry, PhD
Department of Biochemistry
Medical College of Wisconsin
8701 Watertown Plank Road
Milwaukee, WI, 53226
Telephone: 414-955-2825
Fax: 414-955-6568
Email: rgundry@mcw.edu

Keywords: surfaceome, marker discovery, candidate prioritization, cell surface markers

Abstract (250 words)

Cell surface proteins play critical roles in a wide range of biological functions and disease processes through mediation of adhesion and signaling between a cell and its environment. Owing to their biological significance and accessibility, cell surface proteomes (*i.e.* surfaceomes) are a rich source of targets for developing tools and strategies to identify, study, and manipulate specific cell types of interest, from immunophenotyping and immunotherapy to targeted drug delivery and *in vivo* imaging. Despite their relevance, the unique combination of molecules present at the cell surface are not yet described for most cell types. While modern mass spectrometry approaches have proven invaluable for generating discovery-driven, empirically-derived snapshot views of the surfaceome, significant challenges remain when analyzing these often-large datasets for the purpose of identifying candidate markers that are most applicable for downstream applications. To overcome these challenges, we developed SurfaceGenie, a web-based application that integrates a consensus-based prediction of cell surface localization with user-input data to prioritize candidate cell type specific surface markers. Here, we outline the development of the strategy and demonstrate its utility for analyzing human and rodent data from proteomic and transcriptomic workflows. An easy-to-use web application is freely available at www.cellsurfer.net/surfacegenie.

Introduction

Cell surface proteins play critical roles in a wide range of biological functions and disease processes through mediation of adhesion and signaling between a cell and its environment. Owing to their biological significance and accessibility, cell surface proteomes (*i.e.* surfaceomes) are a rich source of targets for developing tools and strategies to identify, study, and manipulate specific cell types of interest, from immunophenotyping and immunotherapy to targeted drug delivery and *in vivo* imaging. A growing interest in cell type specific data has fueled the generation of the Cell Surface Protein Atlas (1), Human Protein Atlas (2), Human Cell Atlas Project (3), and related efforts. However, the unique combination of molecules present specifically at the cell surface are not yet described for most cell types or disease states, and thus continued discovery and annotation efforts are needed.

Mass spectrometry (MS) based workflows can be applied to identify and quantify hundreds to thousands of cell surface proteins (1, 4-14). Particularly, chemoproteomic methods to specifically label and subsequently affinity enrich cell surface proteins can provide experimental evidence of a protein's subcellular location and therefore enable the generation of discovery-driven, empirically-derived snapshot views of the surfaceome (10, 15, 16). These approaches offer significant advantages over transcriptomic approaches, which cannot directly inform protein abundance or localization, and antibody-based strategies which are limited to molecules for which high quality reagents are available. As such, these MS-based chemoproteomic approaches are well-suited to defining cell type specific surfaceomes and serve as a useful first step in defining the cellular phenotype, enabling the development of marker combinations (*i.e.* barcodes) that are cell type specific (17, 18).

Despite their advantages, these chemoproteomic methods generally require >50 million cells, on average, to produce high quality results, which may preclude their application to sample-limited cell types such as primary cells. Although a recent study suggests these methods can be applied to smaller numbers of cells (15), methods that enable routine discovery on very low numbers of cells are not yet widely available. Furthermore, to ensure the results from these approaches provide empirical evidence of surface localization, the initial chemical labeling must be applied to cells with intact plasma membranes, which can pose challenges for certain cell types. For these reasons, more general proteomic approaches that accurately identify and quantify proteins will continue to be useful in the search for cell surface proteins that are informative for a particular cell type or disease status, albeit with the caveat that they offer less inherent specificity for cell surface proteins. Independent of the discovery strategy employed, bioinformatic predictions can serve as an important complement to experimental approaches by providing a means to filter data and prioritize the focus on proteins that are predicted to be localized to the cell surface (19-22).

Though MS is well-suited to the identification of cell-type specific proteins, ultimately, antibodies (Ab) or other affinity reagents that recognize specific epitopes on cell surface proteins are required for most downstream applications such as live cell sorting, imaging, and drug targeting by Ab-drug conjugates. Considering the significant cost and time required to generate and validate affinity reagents for these purposes, it is prudent that the candidate marker prioritization is as selective as possible prior to reagent generation. Specifically, candidate selection should consider whether a marker is likely to be accessible to and detectable by affinity reagents in a manner that allows cell types of interest to be discriminated from non-target cells. Moreover, these assessments should be objective and suited to the analysis of large datasets such as those provided in proteomic and transcriptomic studies. To address these outstanding needs, we developed *GenieScore*, a mathematical strategy that integrates a consensus-based prediction of cell surface localization with user-input data to prioritize candidate cell type specific surface markers. Here, we outline the development of the strategy and demonstrate its utility for analyzing data from proteomic workflows that specifically identify cell surface proteins (*e.g.* CSC) and more general strategies (*e.g.* whole-cell lysate proteomics and transcriptomics). To facilitate its implementation for a broad range of study and data types, we developed SurfaceGenie, an easy-to-use web application that calculates the *GenieScore* for user-input data and further annotates the data with ontology information relevant for cell surface proteins. SurfaceGenie is freely available at www.cellsurfer.net/surfacegenie.

Results

Generation of a surface prediction consensus (SPC) dataset for predictive localization

Based on first principles, three features of a protein predominate its capacity to serve as a cell surface marker capable of distinguishing among cell types (Figure 1A). These include (1) presence at the cell surface, (2) difference in abundance among cell types, and (3) sufficient abundance for antibody-based detection. Whereas features concerning the abundance must be determined empirically, a consensus-based predictive approach was adopted to represent whether a protein is capable of being present at the cell surface, as this feature is largely a function of its primary sequence. To this end, four previous bioinformatic-based constructions of the human cell surface proteome were compiled into a single, surface prediction consensus (SPC) dataset resulting in 5,407 protein accession numbers (Dataset S1, 4.1). The strategies used to generate these predicted human surface protein datasets varied markedly, from manual curation to machine learning, and resulted in datasets ranging 1090-4393 surfaceome proteins each. Overall, the dataset sizes are a primary determinant as to how the datasets intersect (Figure S1). For example, the number of proteins exclusive to a prediction strategy is positively correlated to the size of the original dataset, albeit not in a linear manner, comprising 1.7%, 4.4%, 9.6%, and 26.5% for the Diaz-Ramos, Bausch-Fluck, Town, and Cunha datasets, respectively. Despite these differences, there was considerable overlap among these predictions, with 69% and 41% of proteins in the SPC dataset occurring in ≥ 2 or ≥ 3 individual prediction sets, respectively. To stratify the proteins in the SPC dataset according to how likely they are to be truly present at the cell surface, each protein was assigned one point for each of the individual predicted datasets in which that protein appeared, termed *SPC score* - any protein not present in the dataset is assigned a score of 0 (Dataset S1, 4.1). The distribution of *SPC score*s in the compiled dataset is shown in the histogram in Figure 1B where 1671, 1507, 1497, and 732 proteins are assigned a score of 1, 2, 3, and 4, respectively. (Figure S1). To enable more widespread application, homologous accession numbers were mapped between human and mouse using the Mouse Genome Informatics database (http://www.informatics.jax.org) and human and rat using the Rat Genome Database (https://rgd.mcw.edu) (Dataset S1, 4.2-3).

Benchmarking the SPC dataset against other annotations

The SPC dataset was compared to three established strategies for determination of cell surface localization – Gene Ontology Cellular Component (GO-CC) Annotations, annotations within the Cell Surface Protein Atlas (CSPA), and annotations generated through application of HyperLOPIT(23). Comparisons to GO-CC were consistent with expectations as 'nucleus' and 'cytoplasm' were the two most common terms for proteins with an *SPC scores* of 0, 'integral component of membrane' and 'membrane' for *SPC scores* of 1, and 'integral component of membrane' and 'plasma membrane' for *SPC score*s of 2-4 (Figure S2A). The 'confidence' assignment to proteins in the CSPA correlated well with *SPC score* for both human and mouse, with the notable outlier of ~17% of proteins assigned 'high confidence' having an *SPC score* of 0 (Figure S2B). However, upon closer inspection, 95% these proteins are predicted to be secreted or extracellular matrix proteins (Secretome P, (24)), which can be captured by CSC but are not integral membrane proteins. HyperLOPIT annotations agreed with *SPC score* to a lesser extent, with the most common annotations in proteins with *SPC scores* of 3 or 4 being 'plasma membrane'. However, 'ER/Golgi apparatus' was the most common annotation in proteins with *SPC scores* of 1 or 2 (Figure S2C). Though these comparisons demonstrated agreement overall, the SPC dataset provides unique and specific information in addition to assigning the predictions in a non-binary manner. As the *SPC score* is not dependent on experimental observation, it is more comprehensive in coverage than the CSPA and HyperLOPIT. These differences offer significant advantages for mathematically assigning the likelihood that a protein is present at the cell surface in a predictive manner.

Applying the SPC dataset to compare two proteomic approaches for surface protein identification

The concept of specificity as it relates to cell surface markers is always context dependent, meaning a protein or set of proteins may be useful for identifying a particular cell type in one context, but not another (*e.g.* a protein that is specific to a single cell type within an organ may not be specific to that organ when all other tissues in the body are considered). Therefore, prioritization of cell surface proteins that are likely capable of serving as informative markers should consider experimental data from relevant cell types, including the target and non-target cell types that are to be discriminated. We previously demonstrated that the Cell Surface Capture Technology (CSC) applied to 100 million cells can yield proteins capable of distinguishing among four human lymphocyte cell lines (25). Here, we performed whole-cell lysate (WCL) digestion of 5 million cells of these same cell lines to determine whether a generic proteomic approach coupled with *SPC score* and *GenieScore* analysis could identify cell surface proteins sufficient to distinguish among these cell lines. Compared to the CSC analysis which identified 470 proteins, the WCL approach identified 3858 proteins (≥2 unique peptides). While the majority, 73% (343), of the CSC-identified proteins are predicted to be cell surface localized (*i.e. SPC score*s of 1-4), only 13% (485) of the WCL proteins (Figure 2) met this criterion. This trend is expected due to the high specificity of CSC for cell surface proteins (10, 11, 13, 25). Though predicted surface proteins were identified by both proteomic approaches, the distributions of *SPC scores* suggest more confidence in the surface localization of CSC proteins compared to WCL. This is exemplified by the number of cluster of differentiation (CD) molecules in each SPC-scoring subset, where 109 of 343 proteins from CSC and 50 of 485 proteins from WCL are annotated as CD molecules (Dataset S1 4.4-5). Despite these differences, applying a hierarchical clustering approach to the peptide spectrum matches (PSMs) assigned to individual biological replicates for the subset of proteins in each dataset with an *SPC scores* of 1-4 recapitulated the clustering predicted based on the entire dataset for both proteomic approaches (Figure 2). Although these datasets were collected on the same cell lines, only 127 proteins with *SPC score*s 1-4 were observed in both datasets, which represent 37% and 26% of the CSC and WCL predicted surface proteins, respectively. These data highlight that despite the challenges in identifying cell surface proteins when using generic proteomic strategies that do not specifically enrich for them, application of the

SPC-scoring approach can provide a statistical strategy for determining whether the data are sufficient to differentiate among cells lines.

Testing two label-free quantitation strategies as input data for SurfaceGenie

The *GenieScore* was calculated for each protein in the CSC and WCL datasets using PSMs as inputs for the two terms based on experimental data - *signal dispersion* and *signal strength* (Figure 1). *GenieScores* were plotted against the rank-order - according to *GenieScore* - for CSC and WCL data resulting in a rectangular-hyberbola-like shape, namely, a subset of higher-scoring proteins that trail off into a majority of proteins that are lower-scoring (Figure 2). Although the range of *GenieScores* was similar for both proteomics approaches (6.59 and 6.16 for CSC and WCL, respectively) there are significant differences in the average and distribution, due to the statistical differences between CSC and WCL for each of the terms used to calculate *GenieScore* – *SPC scores*, *signal dispersion*, and *signal strength* (Figure S3). These differences are likely consequences of the highly-selective nature of CSC for identifying cell surface proteins. Although CSC provides empirical evidence of surface localization, unlike WCL, the laborious sample processing involved in selective enrichment of *N*-glycopeptides can introduce more experimental variability compared to the simple WCL digestion. Moreover, CSC results in fewer peptides identified per protein owing to the restriction to tryptic *N*-glycosylated peptides. Despite the differences between these two proteomic approaches, the *GenieScores* for the 127 proteins identified in both proteomic approaches were relatively well correlated (R = 0.66) (Dataset S1 4.6, Figure S3). Recognizing the potential challenges of relying on PSMs for quantitative comparisons, peak areas for selected proteins were calculated using Skyline to provide an alternative type of experimental data for calculating the *GenieScore*. Selection criteria for peptides analyzed in Skyline are provided in the Supporting Information Methods section. The *GenieScores* calculated using MS1 peak areas correlated well with the *GenieScores* using PSMs (R = 0.79 and 0.86 for CSC and WCL, respectively (Figure 2)). As the calculation of *GenieScore* relies on averages (as opposed to individual replicate measurements) the relationship between the product of the *GenieScore* experimental terms (*signal dispersion* and *signal strength*) and the statistical difference (which considers variability in measurement) between cell lines was investigated. A positive relationship was observed, with correlations of 0.47 and 0.73 for CSC and WCL, respectively. The positive relationship suggests that the equation for the *GenieScore* is likely to be prioritizing proteins for which there is a statistical difference (Dataset S1 4.7-8). Overall the *GenieScore* is a robust prioritization metric, demonstrating similar rank ordering for proteins common to CSC and WCL data and for proteins within CSC or WCL using the different quantitative measurements (PSMs or MS1 peak area).

Benchmarking GenieScore against a published study of surface proteins in cancer cell lines

Though the *GenieScore* appears to be a valid metric insofar as it produced similar rank ordering independent of the type of input data, we sought to benchmark it against a published study that validated markers which were originally selected based on experimental proteomic and transcriptomic data. In the test dataset, seven antibodies were generated to surface proteins upregulated on RAS-driven cancer cells compared to a control cell line (26). As the CSC results in this study were reported as a log-fold change without individual values, the signal strength component of the *GenieScore* was calculated using the FPKM values from the RNA-Seq dataset. Of the 122 proteins found to be more abundant in the MCF10A KRAS^{G12V} cells relative to empty vector control, the proteins selected for antibody development ranked 1,2,3,8, 28, and 30 in our *GenieScore* analysis (Figure 3A). The rank-order by *GenieScore* was compared to the rank-order of $\log_2$ fold change in abundance (a metric denoted as selection criteria in the original manuscript) (Figure 3A). The *GenieScore* also performed well using the RNA-Seq data as a starting point, with the SPC analysis rapidly reducing the candidate list from 1139 upregulated proteins to 330 with *SPC scores* of 1-4. The proteins selected for validation by antibody-based analysis in the manuscript are among the top candidates when rank-ordered by *GenieScore* (3, 4, 9, 10, 36) with four of the five genes in the top 3% of the 330 SPC-scoring upregulated proteins. These rank-orders perform favorably compared to using $\log_2$ fold change in transcript levels (25, 37, 43, 50, 115) (Figure 3B). Based on these results, the *GenieScore* is a powerful metric for selection of cell surface proteins that can serve as markers for immunodetection applications, and in this example highlights additional proteins of interest that were not targeted in the original study.

Integrating GenieScores of proteomic and transcriptomic data to reveal candidate markers for Mouse Islet Cell Types

As the *GenieScore* produced useful rank-ordering of potential protein markers from both RNA-Seq and CSC data that were consistent with published results, we sought to determine if it would be a useful metric for integrating data from disparate studies for marker discovery. To this end, we performed CSC on mouse alpha and beta cell lines and compared the results to published RNA-Seq data acquired on primary alpha and beta cells from dissociated mouse islets (27). The datasets shared 321 predicted surface proteins in common, but when the *GenieScores* from CSC data were plotted against the *GenieScores* from the RNA-Seq data, they revealed a poor correlation (R = 0.25) (Figure 4A). This could be due to the fact that the CSC dataset was acquired on cell lines and the RNA-Seq was on primary cells. However, in the context of marker discovery, each of these approaches offers advantages, namely, the CSC data provides experimental evidence regarding abundance at the cell surface and the RNA-Seq analysis of primary cells avoids possible artifacts introduced by culturing cells *ex vivo*. Recognizing the benefits of these complementary approaches, the data were combined in a manner that weighs them equally. Specifically, the *GenieScores* were normalized to the maximum value from each dataset and then the scores were averaged (Figure 4B). The top candidate markers for alpha and beta cells revealed by this combined approach are provided in Figure 4C. Several of these have been studied in the context of islet biology (*e.g.* GLP1R (28), LRP1 (29), CRHR1 (29)) and most (26/30) were identified in a proteomic study of intact human islets, suggesting potential utility across species (30). Altogether, *GenieScore* calculations provide a rapid method for integrating proteomic and transcriptomic data for surface marker prioritization

SurfaceGenie: a web-based application for integrating GenieScore and relevant annotations

SurfaceGenie, a shinyApp written in R, was developed to enable calculation of the *GenieScores* for user input data. In this interface, users upload data as a csv file and can view the distribution of *GenieScores* and *SPC scores* for their data. Proteins are annotated with ontological information including CD and HLA molecule annotations. The plots and data generated are available for download, including the results for individual terms used to calculate *GenieScore*. Additional functionality includes the ability to query accession numbers in single or batch mode, independent of data type, to obtain *SPC Scores*. SurfaceGenie is freely available at http://www.cellsurfer.net/surfacegenie.

Discussion

Despite the central role cell surface proteins play in maintaining cellular structure and function, the cell surface is not well documented for most human cell types. There is currently no comprehensive reference repository of experimentally determined cell surface proteins cataloged by individual human cell types that can be used for comparison to experimental or diseased phenotype. Although specialized proteomic approaches allow for probing the occupancy of the cell surface, the sample requirements and technical sophistication often preclude widespread application, and quantitation is challenging. To overcome these challenges, predictions of surface localization can enable insights from more easily implemented proteomic and transcriptomic approaches, which can be performed on smaller sample sizes. Here, we describe the development of *GenieScore,* a calculation that integrates a predictive metric regarding surface localization with experimental data to prioritize proteins which may be useful as cell surface markers. We demonstrate that *GenieScore* is compatible with CSC, WCL, and RNA-Seq data and is a useful framework by which to integrate multiple sources of data for marker discovery. A web-based application, *SurfaceGenie*, was generated to enable the calculation of *SPC-scores* and *GenieScores* on user-input data and annotation of datasets with functional annotations relevant for cell surface proteins.

It is anticipated that *SurfaceGenie* will enable prioritization of cell surface markers to support a broad range of applications, including immunophenotyping, immunotherapy, and drug targeting for a range of research questions, from mechanistic studies to those in search of markers for disease. However, whether an expressed protein is localized to the cell surface on a specific cell type in a specific experimental or biological condition remains difficult to predict. This is especially true for proteins that do not fit the canonical model (*e.g.* lack a signal peptide) or are only trafficked to the cell surface upon ligand binding (*e.g.* glucose transporter). For these reasons, experimental workflows that provide capabilities for discovery (*i.e.* not limited to available affinity reagents) while providing experimental evidence of cell surface localization on a particular cell type of interest with a specific context (*e.g.* experimental condition, disease state) will remain invaluable.

Methods

All experimental details are provided in Supporting Information.

Cell culture

Human lymphocyte cell lines (Ramos, HG-3, RCH-ACV, Jurkat) were cultured and passaged as previously described (25). Alpha TC1 clone 6 (ATCC CRL-2934) and beta-TC-6 (ATCC CRL-11506) cells were maintained at 37˚C and 5% CO_2, cultured in Dulbecco's Modified Eagle's Medium (Gibco #11885-084) supplemented with 10% heat-inactivated fetal bovine serum containing 16.6 mM or 5.5 mM glucose, respectively.

Cell Lysis, Protein Digestion, and Peptide Cleanup

For WCL analysis of lymphocytes, cell pellets were lysed in 100mM Ammonium Bicarbonate containing 20% acetonitrile and 40% Invitrosol (ThermoFisher Scientific), digested with trypsin overnight, and cleaned by SP2 following the standard operating protocol as described (31). Peptides were quantified using Pierce Quantitative Fluorometric Peptide Assay (ThermoFisher Scientific) according to manufacturer's instructions on a Varioskan LUX Multimode Microplate Reader and SkanIt 5.0 software (ThermoFisher Scientific). For CSC analysis of mouse islet cell lines, samples were prepared as previously described (11, 13, 25).

Label Free Quantitation by Mass Spectrometry

Lymphocyte peptides and CSC samples of mouse islet cell types were analyzed by LC-MS/MS using a Dionex UltiMate 3000 RSLCnano system (ThermoFisher Scientific) in line with a Q Exactive (ThermoFisher Scientific). Lymphocyte samples were prepared as 50 ng/μL total sample peptide concentration with Pierce Peptide Retention Time Calibration Mixture (PRTC, Thermo) spiked in at a final concentration of 2 fmol/μL PRTC, and then blocked and randomized with two technical replicates analyzed per sample. CSC samples of mouse islet cell types were analyzed as described (32, 33). MS data were analyzed using Proteome Discoverer 2.2 (ThermoFisher Scientific) and SkylineDaily.

Construction of a consensus dataset of predicted surface proteins

Four published surfaceome datasets (19-22), each of which used a distinct methodology to bioinformatically predict the subset of the proteome which can be surface localized, were concatenated into a single consensus dataset. In this process, the UniProt retrieve/mapping ID tool (www.uniprot.org) was used to convert the gene names provided in the published surfaceomes to UniProt Accession numbers. Ambiguous matches were clarified by any supplementary information provided in the datasets in addition to gene name (*i.e.* alternate name, molecule name, chromosome). To stratify the proteins within the consensus dataset, each was assigned a surface prediction consensus score (*SPC score*), a summed value whereby one point was awarded for each of the prediction strategies in which the protein appeared.

GenieScore – A mathematical representation of surface marker potential

An equation was developed to mathematically represent key features deemed relevant when considering whether a protein has high potential to be useful as a cell surface marker for distinguishing between cell types or experimental groups. The equation, which returns a metric termed the *GenieScore*, is the product of 1) the *SPC scores* (described above); 2) *signal dispersion*, a measure of the disparity in observations among investigated samples and is mathematically equivalent to the square of the normalized Gini coefficient; and 3) *signal strength*, a logarithmic transformation of the experimental data (*e.g.* number of peptide spectral matches, MS1 peak area, FKPM, or RKPM). A thorough definition and rationalization of the individual equation terms is provided in Supporting Information.

$$GenieScore = (SPC{-}Score)\cdot\left(\frac{G}{G_{Max}}\right)^2\cdot\log(Signal_{Max})$$

SurfaceGenie Web application

A web application for accessing SurfaceGenie was developed as an interactive Shiny app written in R and is available at www.cellsurfer.net/surfacegenie.

Supporting Information

1. Figure S1 – Visualization of the intersections between datasets used to generate *SPC score*
2. Figure S2 – Benchmarking the SPC score against GO terms, CSPA, and HyperLOPIT
3. Figure S3 – Distributions of *GenieScore* terms in WCL and CSC lymphocyte data

4. Dataset S1 – (1) Human SPC dataset, (2) Mouse SPC dataset, (3) Rat SPC dataset, (4) lymphocyte WCL data with *GenieScores*, (5) lymphocyte CSC data with *GenieScores*, (6) *GenieScores* for proteins common to CSC and WCL, (7) ANOVA test statistics for WCL data, (8) ANOVA test statistics for CSC data
5. Supplemental Methods

Author Contributions

R.L.G. and M.W. conceived the study; R.L.G. supervised the study; M.W. developed the algorithms and designed and performed MS experiments; S.S. developed the python code; S.S. and J. L. developed the web application; R.A.J.L., P.A.H., J.A.C., performed analyses of mouse islet cell lines, 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 and R01-HL134010 to R.L.G.; F31-HL140914 to M.W.]. Funding sources had no involvement in study design, data collection, interpretation, analysis or publication.

Figure Legends

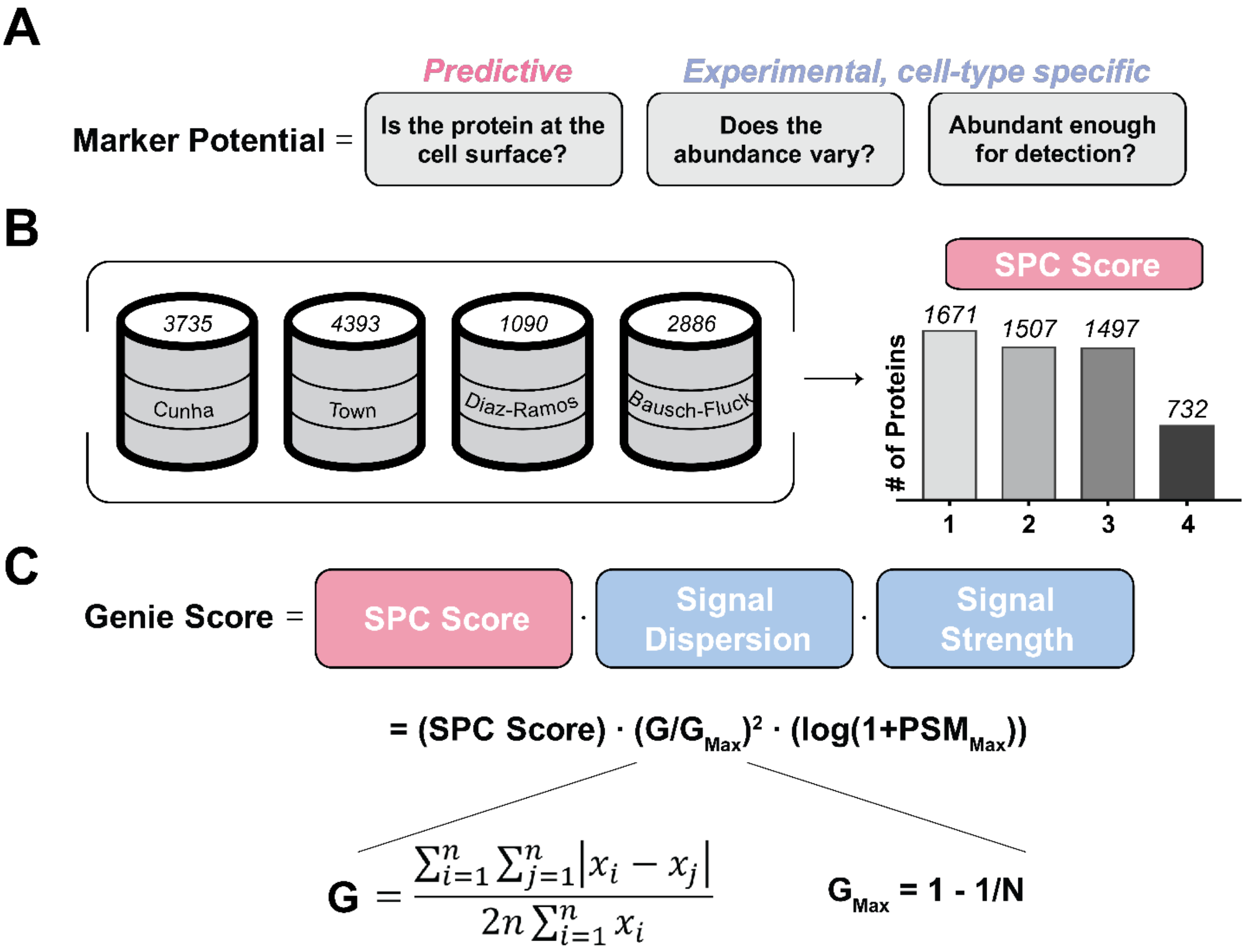
Figure 1: Overview of Surface Prediction Consensus (SPC) score and *GenieScore*. (A) The first principles hypothesized to be correlated to cell surface marker potential. (B) The first author and number of unique accession numbers in is shown for the four bioinformatic predictions used to generate *SPC score* (left) and the overall distribution of *SPC Scores* (right)*.* The full dataset is provided in the Supporting Information (Dataset S1, 4.1) (C) The names of the terms and mathematical equation used to calculate *GenieScore*.

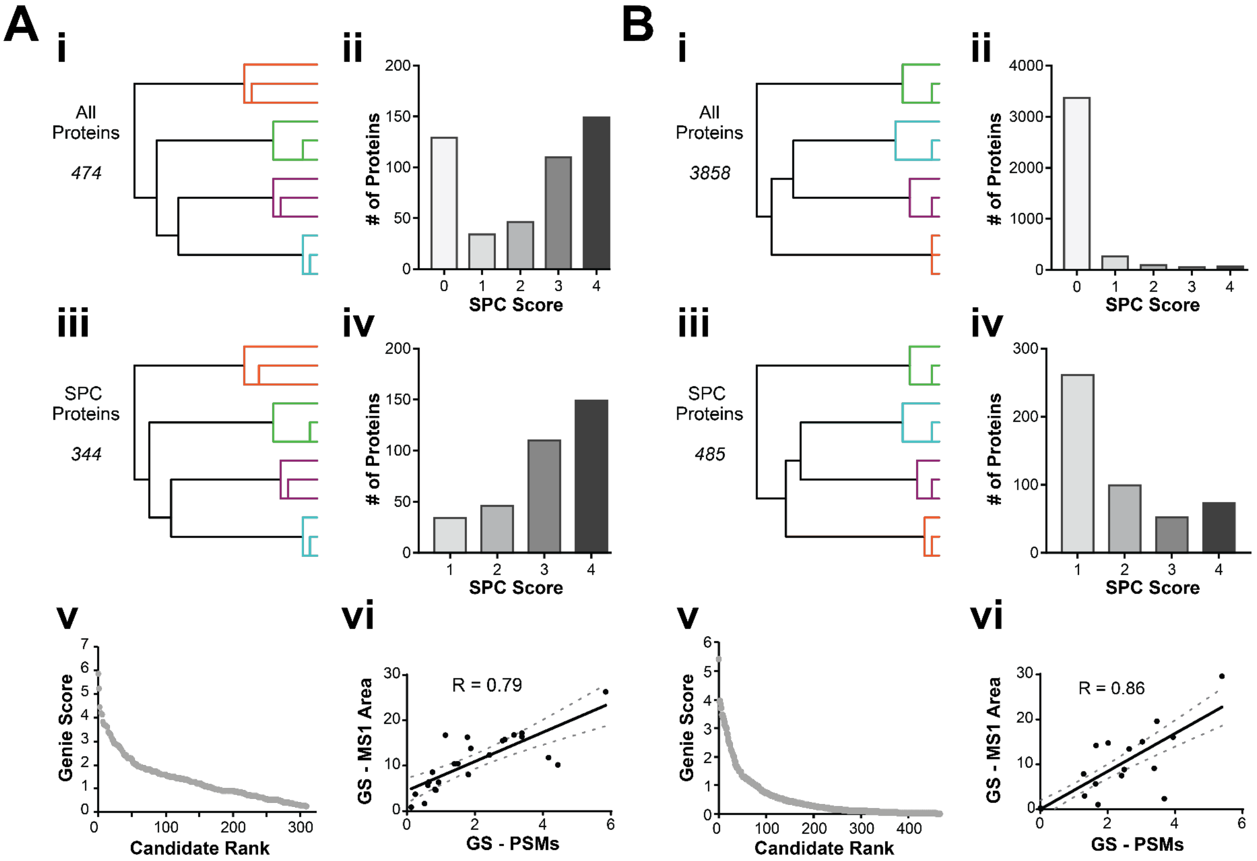
Figure 2. *SPC scores* and *GenieScores* for whole-cell lysate and cell surface capture lymphocyte data. (A) cell surface capture and (B) whole cell lysate data from the analysis of four lymphocyte lines. (i) Hierarchical clustering using all identified proteins. (ii) Distribution of *SPC scores* using all identified proteins. (iii) Hierarchical clustering using proteins predicted to be surface-localized by *SPC scores*. (iv) Distribution of *SPC scores* for only the proteins predicted to be surface-localized. (v) Plot of *GenieScore* against rank-order of candidate cell surface markers. (vi) *GenieScore* calculated using MS1 peak area against *GenieScore* calculated using peptide spectral matches.

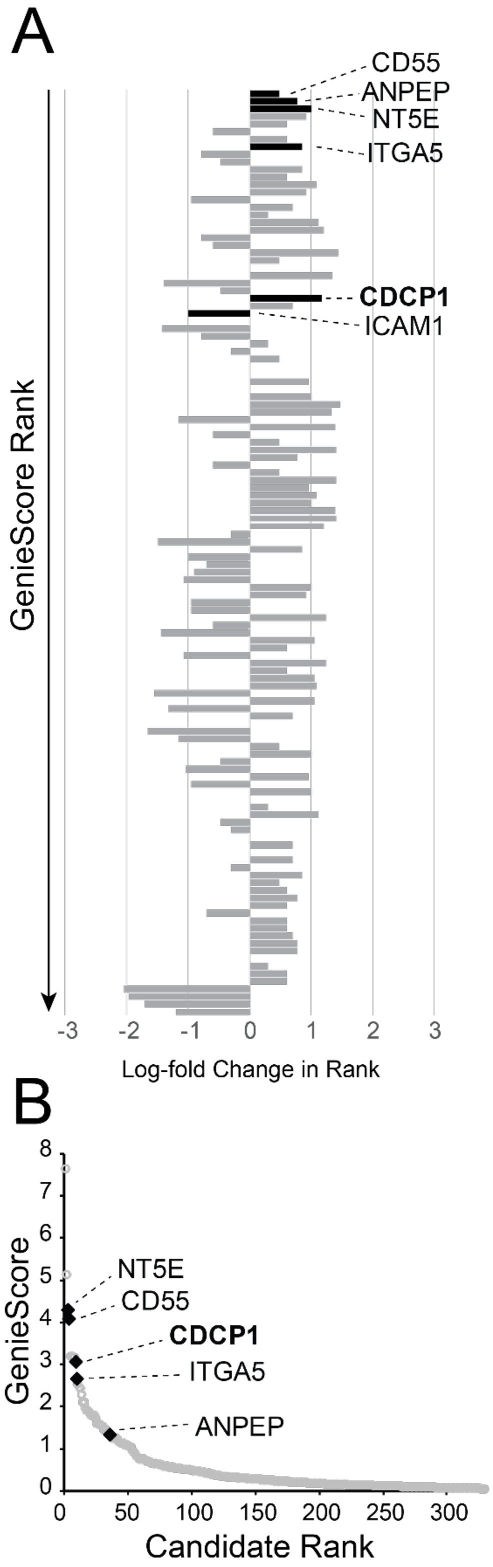
Figure 3. Benchmarking *GenieScore* against a published surface marker study. (A) The *GenieScore* rank-order of proteins identified by CSC that were upregulated in MCF10A KRAS^{G12V} cells relative to empty vector control plotted against the relative difference in rank-order of proteins based on $\log_2$(fold change). (B) The *GenieScores* calculated using RNA-Seq data plotted against rank-order for predicted surface proteins. Labeled proteins were selected for antibody development - the protein in bold, CDCP1, being the focus of the study.

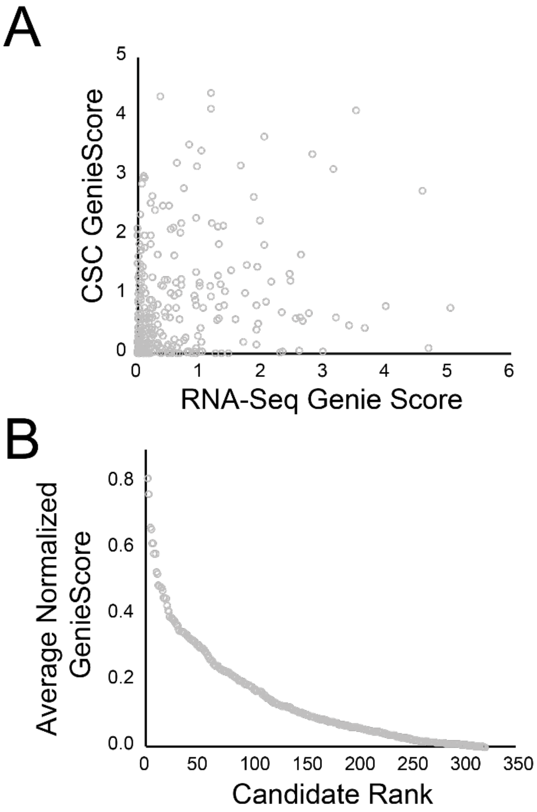
Figure 4. Application of a combined *GenieScore* for islet cell-type marker discovery*.* (A) *GenieScore* calculated using CSC data on mouse alpha and beta cell lines plotted against *GenieScore* calculated using RNA-Seq data on primary mouse alpha and beta cells. (B) The average normalized GenieScore calculated from integration of CSC and RNA-Seq data plotted against rank-order. (C) A table of top marker candidates for alpha and beta cells (colored in green and blue, respectively) shown with the *SPC scores*, average RKPM and PSMs, and combined *GenieScores*.

Figures:


A
Predictive
Experimental, cell-type specific
Marker Potential =
Is the protein at the cell surface?
Does the abundance vary?
Abundant enough for detection?
B
3735
Cunha
4393
Town
1090
Diaz-Ramos
2886
Bausch-Fluck
SPC Score
# of Proteins
1671
1507
1497
732
1
2
3
4
C
Genie Score =
SPC Score
·
Signal Dispersion
·
Signal Strength
$= (\text{SPC Score}) \cdot (G/G_{Max})^2 \cdot (\log(1+PSM_{Max}))$
$G = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}|x_i - x_j|}{2n\sum_{i=1}^{n} x_i}$
$G_{Max} = 1 - 1/N$


A
i
All Proteins
474
ii
# of Proteins
SPC Score
iii
SPC Proteins
344
iv
# of Proteins
SPC Score
v
Genie Score
Candidate Rank
vi
R = 0.79
GS - MS1 Area
GS - PSMs
B
i
All Proteins
3858
ii
# of Proteins
SPC Score
iii
SPC Proteins
485
iv
# of Proteins
SPC Score
v
Genie Score
Candidate Rank
vi
R = 0.86
GS - MS1 Area
GS - PSMs


A
CD55
ANPEP
NT5E
ITGA5
CDCP1
ICAM1
GenieScore Rank
-3
-2
-1
0
1
2
3
Log-fold Change in Rank
B
8
7
6
5
4
3
2
1
0
NT5E
CD55
CDCP1
ITGA5
ANPEP
GenieScore
0
50
100
150
200
250
300
Candidate Rank



C

Gene Name	SPC	Average RPKM		Average PSMs		Combined GenieScore
		alpha	beta	alpha	beta	
Gabbr2	4	0.0	9.6	2.0	19.9	0.81
Glp1r	3	0.0	7.2	15.3	163.4	0.77
Igf1r	4	0.0	5.0	8.1	44.2	0.67
Ptgfrn	4	7.2	0.0	5.2	0.6	0.62
Egfr	4	0.0	11.6	1.2	5.1	0.62
Ceacam1	4	0.0	9.8	0.9	4.1	0.59
Igsf1	4	0.0	5.2	2.2	9.9	0.52
Lrp1	4	5.4	1.8	19.1	1.3	0.49
Mfi2	3	0.0	6.6	0.9	8.4	0.48
Ece1	4	29.2	134.6	85.6	160.9	0.48
Ntrk2	4	22.8	29.4	2.4	34.5	0.48
Nrcam	3	1.2	20.4	2.1	7.8	0.45
Cd14	3	2.6	0.0	13.6	1.0	0.45
Galr1	3	0.0	4.6	0.6	7.1	0.45
Slc2a3	4	5.6	2.6	8.5	0.2	0.41
Alcam	4	28.8	8.2	55.2	16.5	0.41
Ptprk	4	1.4	0.4	10.1	0.5	0.39
Sorcs2	3	0.6	0.0	14.5	0.4	0.39
Slc4a10	3	1.6	13.4	3.8	15.6	0.38
Prnp	4	16.4	45.4	176.5	438.8	0.36
Crhr1	3	0.0	4.8	0.0	1.2	0.35
Kcnk3	2	1.2	0.0	28.8	0.5	0.35
Cdh4	4	0.6	8.6	2.7	3.9	0.35
Emp1	3	5.6	0.4	12.0	2.8	0.34
Cd48	3	0.6	0.0	6.6	0.0	0.33
Sema4c	4	5.0	0.2	4.1	2.3	0.32
Ednra	3	4.4	0.0	1.6	0.2	0.32
Ly75	4	3.0	0.4	3.2	0.4	0.31
Slc7a2	3	16.6	5.8	152.9	43.2	0.30
App	3	2.4	0.4	192.7	59.7	0.28

References

1. Bausch-Fluck D*, et al.* (2015) A mass spectrometric-derived cell surface protein atlas. *PloS one* 10(3):e0121314.
2. Uhlen M*, et al.* (2005) A human protein atlas for normal and cancer tissues based on antibody proteomics. *Molecular & cellular proteomics : MCP* 4(12):1920-1932.
3. Regev A*, et al.* (2017) The Human Cell Atlas. *Elife* 6.
4. Patterson NL*, et al.* (2013) Using proteomics to uncover extracellular matrix interactions during cardiac remodeling. *Proteomics Clin Appl* 7(7-8):516-527.
5. Wang H & Hanash S (2015) Mass spectrometry based proteomics for absolute quantification of proteins from tumor cells. *Methods*.
6. Niehage C*, et al.* (2011) The cell surface proteome of human mesenchymal stromal cells. *PLoS One* 6(5):e20399.
7. Nagano K*, et al.* (2011) Distinct cell surface proteome profiling by biotin labeling and glycoprotein capturing. *J Proteomics* 74(10):1985-1993.
8. Choksawangkarn W*, et al.* (2013) Enrichment of plasma membrane proteins using nanoparticle pellicles: comparison between silica and higher density nanoparticles. *J Proteome Res* 12(3):1134-1141.
9. Zeng Y, Ramya TN, Dirksen A, Dawson PE, & Paulson JC (2009) High-efficiency labeling of sialylated glycoproteins on living cells. *Nature methods* 6(3):207-209.
10. Wollscheid B*, et al.* (2009) Mass-spectrometric identification and relative quantification of N-linked cell surface glycoproteins. *Nature biotechnology* 27(4):378-386.
11. Boheler KR*, et al.* (2014) A human pluripotent stem cell surface N-glycoproteome resource reveals markers, extracellular epitopes, and drug targets. *Stem cell reports* 3(1):185-203.
12. Gundry RL*, et al.* (2009) The mouse C2C12 myoblast cell surface N-linked glycoproteome: identification, glycosite occupancy, and membrane orientation. *Molecular & cellular proteomics : MCP* 8(11):2555-2569.
13. Gundry RL*, et al.* (2012) A cell surfaceome map for immunophenotyping and sorting pluripotent stem cells. *Molecular & cellular proteomics : MCP* 11(8):303-316.
14. Kropp EM*, et al.* (2014) N-glycoprotein surfaceomes of four developmentally distinct mouse cell types. *Proteomics. Clinical applications* 8(7-8):603-609.
15. Kalxdorf M, Gade S, Eberl HC, & Bantscheff M (2017) Monitoring Cell-surface N-Glycoproteome Dynamics by Quantitative Proteomics Reveals Mechanistic Insights into Macrophage Differentiation. *Molecular & cellular proteomics : MCP* 16(5):770-785.
16. Turtoi A*, et al.* (2011) Novel comprehensive approach for accessible biomarker identification and absolute quantification from precious human tissues. *J Proteome Res* 10(7):3160-3182.
17. Boheler KR & Gundry RL (2017) Concise Review: Cell Surface N-Linked Glycoproteins as Potential Stem Cell Markers and Drug Targets. *Stem Cells Transl Med* 6(1):131-138.
18. Fujinaka CM, Waas M, & Gundry RL (2018) Mass Spectrometry-Based Identification of Extracellular Domains of Cell Surface N-Glycoproteins: Defining the Accessible Surfaceome for Immunophenotyping Stem Cells and Their Derivatives. *Methods Mol Biol* 1722:57-78.
19. Bausch-Fluck D*, et al.* (2018) The in silico human surfaceome. *Proc Natl Acad Sci U S A* 115(46):E10988-E10997.
20. da Cunha JP*, et al.* (2009) Bioinformatics construction of the human cell surfaceome. *Proc Natl Acad Sci U S A* 106(39):16752-16757.
21. Town J*, et al.* (2016) Exploring the surfaceome of Ewing sarcoma identifies a new and unique therapeutic target. *Proc Natl Acad Sci U S A* 113(13):3603-3608.

22. Diaz-Ramos MC, Engel P, & Bastos R (2011) Towards a comprehensive human cell-surface immunome database. *Immunol Lett* 134(2):183-187.
23. Christoforou A*, et al.* (2016) A draft map of the mouse pluripotent stem cell spatial proteome. *Nat Commun* 7:8992.
24. Bendtsen JD, Kiemer L, Fausboll A, & Brunak S (2005) Non-classical protein secretion in bacteria. *BMC Microbiol* 5:58.
25. Haverland NA*, et al.* (2017) Cell Surface Proteomics of N-Linked Glycoproteins for Typing of Human Lymphocytes. *Proteomics* 17(19).
26. Martinko AJ*, et al.* (2018) Targeting RAS-driven human cancer cells with antibodies to upregulated and essential cell-surface proteins. *Elife* 7.
27. Benner C*, et al.* (2014) The transcriptional landscape of mouse beta cells compared to human beta cells reveals notable species differences in long non-coding RNA and protein-coding gene expression. *BMC Genomics* 15:620.
28. Thorens B (1992) Expression cloning of the pancreatic beta cell receptor for the gluco-incretin hormone glucagon-like peptide 1. *Proc Natl Acad Sci U S A* 89(18):8641-8645.
29. Ye R*, et al.* (2018) Intracellular lipid metabolism impairs beta cell compensation during diet-induced obesity. *J Clin Invest* 128(3):1178-1189.
30. Danzer C*, et al.* (2012) Comprehensive description of the N-glycoproteome of mouse pancreatic beta-cells and human islets. *J Proteome Res* 11(3):1598-1608.
31. Waas M, Pereckas M, Jones Lipinski RA, Ashwood C, & Gundry RL (2019) SP2: Rapid and Automatable Contaminant Removal from Peptide Samples for Proteomic Analyses. *J Proteome Res*.
32. Mallanna SK, Cayo MA, Twaroski K, Gundry RL, & Duncan SA (2016) Mapping the Cell-Surface N-Glycoproteome of Human Hepatocytes Reveals Markers for Selecting a Homogeneous Population of iPSC-Derived Hepatocytes. *Stem cell reports* 7(3):543-556.
33. Mallanna SK, Waas M, Duncan SA, & Gundry RL (2016) N-glycoprotein surfaceome of human induced pluripotent stem cell derived hepatic endoderm. *Proteomics*.