Biological Sciences

SurfaceGenie: A web-based application for prioritizing cell-type specific marker candidates

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: cell surface markers, candidate prioritization, drug targets, web-application, immunophenotyping

Abstract (250 words)

Cell surface proteins play critical roles in a wide range of biological functions and disease processes through mediation of contacts and signals between a cell and its environment. Owing to their biological significance and accessibility, cell surface proteins are attractive targets for developing tools and strategies to identify, study, and manipulate specific cell types of interest. Applications ranging from immunophenotyping and immunotherapy to targeted drug delivery and *in vivo* imaging are enabled by exploitation of cell-type specific surface proteins. Despite their utility and 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 surface proteins, 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 *GenieScore*, a prioritization metric that integrates a consensus-based prediction of cell surface localization with user-input data to rank-order candidate cell-type specific surface markers. In this report, we outline the development of this prioritization strategy and demonstrate its utility for analyzing human and rodent data from proteomic and transcriptomic experiments in the areas of cancer, stem cell, and islet biology. The calculation of *GenieScores*, as well as additional scoring algorithm permutations that enable prioritization of co-expressed and intracellular cell-type specific candidate markers, is made accessible via the freely available SurfaceGenie web-application at www.cellsurfer.net/surfacegenie.

Introduction

Cell surface proteins play key roles in diverse biological processes and disease pathogenesis through mediation of adhesion and signaling between the extracellular and intracellular space. Owing to their accessible location, cell surface proteins can be exploited as valuable markers for a range of research and clinical applications including immunophenotyping live cells, targeted drug delivery, and *in vivo* imaging. As such, 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 innovation regarding surface protein discovery and annotation efforts are needed.

Specialized chemoproteomic approaches which specifically label and enrich cell surface proteins can provide direct evidence of cell surface localization resulting in empirically-derived snapshot views of the cell surface proteome (4-6). However, the large sample requirements and technical sophistication required for these experiments preclude their widespread use and application to sample-limited cell types. For these reasons, whole-cell proteomic and transcriptomic-based approaches that can be applied to identify and quantify thousands of molecules from fewer cells will continue to be useful in the search for cell surface proteins that are informative for particular cell types or disease stages.

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 (7-10). Though transcriptomic and proteomic approaches offer significant advantages over antibody screening with regards to throughput and depth of coverage, candidate markers identified from discovery approaches must be subsequently validated as viable immunodetection or payload-delivery

targets. Considering the significant cost and time required for the development of *de novo* affinity reagents, it is prudent to select candidates in a manner that considers whether a marker is likely to be both accessible and detectable by affinity reagents in a manner that allows cell types of interest (*i.e.* target cells) to be discriminated from non-target cells. Moreover, these assessments should be objective and suited to the analysis of large datasets such as those generated by proteomic and transcriptomic studies. To address these outstanding needs, we developed *GenieScore*, a metric to rank and prioritize candidate cell type specific surface markers - calculated by integrating a consensus-based prediction of cell surface localization with user-input quantitative data. Here, we describe the development of *GenieScore* and demonstrate its utility for prioritizing candidate cell surface markers using data obtained from proteomic workflows that specifically identify cell surface proteins (*e.g.* Cell Surface Capture (CSC)) and more general strategies (*e.g.* whole-cell lysate proteomics and transcriptomics). We also demonstrate that permutations of *GenieScore* can efficiently prioritize co-expressed and intracellular cell-type specific markers. To facilitate its implementation among users, we developed SurfaceGenie, an easy-to-use web application that calculates *GenieScores* for user-input data and annotates the data with ontology information particularly relevant for cell surface proteins. SurfaceGenie is freely available at www.cellsurfer.net/surfacegenie.

Results

Generation of a surface prediction consensus dataset for predictive localization

Four previous bioinformatic-based constructions of the human cell surface proteome were compiled into a single, surface prediction consensus (SPC) dataset containing 5,407 protein accession numbers (Dataset S1.1). The strategies used to generate these predicted human surface protein datasets varied markedly, from manual curation to machine learning, and resulted in dataset sizes ranging from 1090 to 4393 surface proteins (Figure 1A). 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. The number of proteins exclusive to a prediction strategy is positively correlated to the original dataset size, albeit not linearly, comprising 1.7%, 4.4%, 9.6%, and 26.5% for the Diaz-Ramos, Bausch-Fluck, Town, and Cunha datasets, respectively (Figure 1B). To reflect the difference in the consensus of surface localization, each protein was assigned one point for each of the individual predicted datasets in which that protein appeared, termed *Surface Prediction Consensus* (*SPC) score* (Figure 1B, Dataset S1.1). The distribution of *SPC score*s is shown in Figure 1B where 1671, 1507, 1497, and 732 proteins are assigned a score of 1, 2, 3, and 4, respectively. To enable more widespread applicability, mouse and rat *SPC score* databases were generated by mapping the human proteins to mouse and rat homologs using the Mouse Genome Informatics database (http://www.informatics.jax.org, Dataset S1.2-3).

Benchmarking the SPC dataset against other annotations

The SPC dataset was compared to three established resources for obtaining cell surface localization annotations: 1) Gene Ontology Cellular Component Ontology (GO-CCO) (11, 12), 2) annotations within the Cell Surface Protein Atlas (CSPA) (1), and 3) annotations generated through application of HyperLOPIT (13). Comparisons to GO-CCO were consistent with expectations as 'nucleus' and 'cytoplasm' were the two most common terms for proteins with *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 1C). The 'confidence' assignment to proteins in the CSPA agreed with *SPC scores* for both human and mouse, with the notable exception of ~17% of proteins assigned 'high confidence' having an *SPC score* of 0 (Figure S1A). However, upon closer inspection, 95% of these proteins are predicted to be secreted or extracellular matrix proteins (Secretome P, (14)), which can be captured in CSC experiments but are not integral membrane proteins. The most common HyperLOPIT annotation in proteins with *SPC scores* of 3 or 4 was 'plasma membrane'; however, 'ER/Golgi apparatus'

was the most common annotation in proteins with *SPC scores* of 1 or 2 (Figure S1B). 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. Furthermore, 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. Moreover, the calculation of *SPC score* is straightforward and flexible to allow easy integration of results from future efforts of cell surface localization prediction.

Defining features of a cell surface protein marker based on first principles

By defining the term 'marker' to designate a cell surface protein which is capable of distinguishing between cell types of interest on the basis of signal obtained by immunodetection, there are three features that can be used to evaluate the capacity of a protein to serve as a marker (Figure 2A). These include (1) *SPC score* - presence at the cell surface, (2) *signal dispersion* - difference in abundance among cell types, and (3) *signal strength* - sufficient abundance for specific antibody-based detection. The product of these three terms, which we define as *GenieScore*, is a metric that can be used to rank proteins from experimental data for their capacity to serve as a marker. Importantly, 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. Hence, although a consensus-based predictive approach can be adopted to represent whether a protein is capable of being present at the cell surface (*SPC score*), the *signal dispersion* and *signal strength* must be determined empirically, as these will differ among cell types. As the two mathematical terms chosen to represent *signal dispersion* and *signal strength* are agnostic to the data type, we investigated the effects of different sources of input-data to these terms with respect to the calculated *GenieScores*.

Testing GenieScore by comparing two proteomic approaches for marker discovery

We previously demonstrated that CSC applied to four human lymphocyte cell lines resulted in sufficient depth of surface protein identification to allow discrimination among the lines (15). Here, we performed whole-cell lysate (WCL) digestion of these same cell lines to determine whether a generic proteomic approach would be sufficient to detect divergent cell surface proteins to do the same. Notably, the WCL approach only required a peptide amount equivalent to ~1000 cells compared to CSC which used ~100 million cells/experiment. While the majority, 75% (325), of the CSC-identified proteins are predicted to be cell surface localized (*i.e. SPC scores* of 1-4), only 13% (485) of the WCL proteins met this criterion. Although these datasets were collected on the same cell lines, only 91 proteins with *SPC score*s 1-4 were observed in both datasets, which represent 28% and 19% of the CSC and WCL predicted surface proteins, respectively. Despite these differences, applying hierarchical clustering to the subset of proteins in each dataset with *SPC scores* of 1-4 recapitulated the clustering predicted based on the entire dataset for both proteomic approaches (Figure S2). These data highlight the utility of applying the SPC metric as a strategy to filter and compare between datasets and demonstrate that a generic proteomic strategy can provide sufficient surface protein detection to differentiate among cell types using 0.1% of the cellular material required for CSC.

GenieScores were calculated for each protein in the CSC and WCL datasets using peptide-spectrum matches (PSMs) as input for *signal dispersion* and *signal strength* calculations (Dataset S.1-2). Though predicted surface proteins were identified by both proteomic approaches, the distributions of *SPC scores, signal dispersion,* and *signal strength* were markedly different between CSC and WCL (Figure 2B-D). These observations are expected due to the highly-selective nature of CSC which primarily captures *N*-glycosylated peptides resulting in higher specificity for *bona fide* surface proteins and fewer peptides identified per protein (4, 15-17). *GenieScores* were plotted against the rank 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 lower-scoring proteins (Figure 2E-F) with a similar range (6.59 and 6.16 for CSC and WCL, respectively) but significant difference in the distribution. *GenieScores* for the 91 proteins identified in both proteomic approaches were strongly correlated (r_s = 0.63) (Dataset S2.3, Figure 2G).

The top-scoring candidate markers in both the CSC and WCL data sets are proteins for which the majority (if not the totality) of PSMs are in a single cell line. The numbers of PSMs per cell line for selected proteins are shown for CSC and WCL in Figure 3 along with the ranks determined by application of *GenieScore* (plotted in Figure 2E-F). Many of the high ranking candidates have previously been reported as markers for cancer types modeled by the cell lines used here - including ATP1B1, CD39, and HLA-DR for chronic lymphocytic leukemia (CLL) (18-20) (Hg-3 cell line); CD10 and CD79b for Burkitt Lymphoma (21, 22) (Ramos cell line) (Figure 3). Proteins with moderate ranks often had PSMs spread evenly among two or more of the cell lines. The examples here include CD5 - a known T cell marker that is often expressed in CLL (18) (accounting for its observation in Jurkat and Hg-3 cell lines, respectively), and CD47 – a protein reported to be upregulated in many cancer subtypes (23). Proteins with low ranks are equally spread among all the cell lines, often 'housekeeping' type proteins such as transferrin receptor 1 and mannose-6-phosphate receptor (Figure 3).

As the calculation of *GenieScore* relies on averages (as opposed to individual replicate measurements) the relationship between the product of the experimental terms (*signal dispersion* and *signal strength*) used to calculate *GenieScore* and the statistical difference (which considers variability in measurement) between cell lines was investigated. A positive relationship was observed, with Spearman's correlations (r_s) of 0.66 and 0.64 for WCL and CSC, respectively, suggesting that the equation for *GenieScore* is likely to be prioritizing proteins for which there is a statistical difference (Figure S3A). Finally, recognizing the limitations of relying on PSMs for

quantitative comparisons, *GenieScores* were calculated using MS1 peak areas for selected proteins. Results from this strategy had a very strong correlation to *GenieScores* using PSMs; r_s = 0.88 and 0.80 for WCL and CSC, respectively (Figure S3B). Altogether, application of *GenieScore* to the data collected from diverse lymphocyte cell lines produced similar ranking of candidate markers independent of the type of input data, including several surface proteins previously linked to relevant cancer subtypes, which indicates *GenieScore* is a robust and valid prioritization metric.

Benchmarking GenieScore against two published surface protein marker studies

A major application of *GenieScore* is to prioritize candidate markers for immunophenotyping. Hence, we sought to benchmark the performance of *GenieScore* ranking against two published studies that performed flow cytometry analyses to orthogonally validate putative markers for cell types of interest which were originally identified from proteomics and/or transcriptomic data. In the first study, Martinko *et al.* performed CSC and RNA-Seq on MCF10A $KRAS^{G12V}$ cells (comparing the results to empty vector control MCF10A cells) to identify surface proteins indicative of RAS-driven cancer phenotype (24). Antibodies were subsequently developed against seven candidate markers, all of which demonstrated positive signal on the MCF10A $KRAS^{G12V}$ cells. Using *GenieScore* as a prioritization metric, we investigated the relative ranks of these validated markers among the CSC and RNA-Seq datasets. As the goal of the original study was to identify surface proteins which were upregulated in cancer, *GenieScores* were only calculated for predicted surface proteins (*SPC score* >0) which met this additional criterion – resulting in 122 candidates from CSC and 330 candidates from RNA-Seq (Figure 4A, Dataset S2.1-2). The validated proteins were among the highest scoring candidates in both the CSC and RNA-Seq data sets (Figure 4A). *GenieScores* calculated from the CSC and RNA-Seq data had a moderate correlation (r_s = 0.41) with most of the validated markers scoring highly for both CSC and RNA-Seq (Figure 4B). The rank of the putative markers by *GenieScore* was

compared to the rank by log_2 fold changes (a metric denoted as selection criteria in the original manuscript) (Figure 4C). In all but one case, the candidates rank higher by *GenieScore* than by log_2fold ratio. These results support *GenieScore* as a useful, single metric that enables selection of cell surface proteins which can serve as markers for immunodetection. Though the validated markers were among the top-ranking candidates, other proteins with high *GenieScores* emerge as potential targets, highlighting the utility of *GenieScore* to reveal new biological insights or targets from previously published data. Two such targets that scored well by both CSC and RNA-Seq are THBD and NRP1, which have been previously implicated in a KRAS-driven myeloid malignancy and KRAS-driven tumorigenesis (25, 26). Additionally, several proteins score well in one dataset (*i.e.* CSC or RNA-Seq) but are completely absent from the other. SAT-1, ranked 5 within the RNA-Seq data but not observed by CSC, plays a role in a polyamine synthesis pathway that is upregulated by KRAS-driven cancers (27, 28). Inspection of the primary sequence of SAT-1 reveals that the *N*-glycosite is located within a peptide that would make it unlikely to be detected by mass spectrometry. Conversely, there are proteins within the CSC dataset for which there were no matching transcripts such as LRP1, a protein associated with tumorigenesis and tumor progression (29, 30). Altogether, these results present the complementary nature of CSC and RNA-Seq as discovery techniques and demonstrate that *GenieScore*-based analyses of these data, either independently or together, provide a rapid strategy for prioritizing candidates for immunophenotyping.

In the second study, Boheler *et al*. performed CSC on human fibroblasts, embryonic stem cells, and induced pluripotent stem cells to identify surface markers for stem cells (16). Candidate pluripotency markers were selected by comparing the set of proteins observed on stem cells to CSC data from the CSPA, specifically, requiring that a protein was not detected in fibroblasts and detected in fewer than four other somatic cell types (excluding cancer cell types). Negative markers of pluripotency were selected in a similar manner, specifically, not detected in stem cells

and detected in six or more non-diseased cell types in the CSPA. Flow cytometry analysis of human fibroblasts and stem cells was used to orthogonally validate seventeen putative positive and three putative negative pluripotency markers and included three previously reported positive pluripotency markers as controls. The *GenieScores* for proteins observed in the human fibroblast and stem cell CSC experiments were plotted against the $\log_2$fold ratio of PSMs between the cell types providing a visual depiction of capacity to serve as a marker segregated by cell type wherein the reference, putative negative, and putative positive pluripotency markers are denoted in individual plots (Figure 4D, Dataset S3.1). The reference stem cell markers are among the top scoring (with ranks of 2 and 10) candidate pluripotency markers from the CSC dataset, except for Thy1, a protein for which both CSC and flow cytometry results provided evidence for its presence in fibroblasts and stem cells. The *GenieScores* for the putative negative and positive markers were spread over a greater range in this dataset compared to the distribution observed for the Martinko *et al.* study. This difference is likely attributable to the notably divergent strategies employed for candidate selection. Specifically, the Boheler study relied on qualitative (presence/absence) rather than quantitative comparisons, considered data from cell types outside those included in the study, and restricted validation to candidates for which commercially-available monoclonal antibodies were available. Notably, several of the validated markers were identified by relatively few PSMs in the original dataset (IL27RA, EFNA3). While the number of PSMs is sometimes used as a filter to eliminate proteins from consideration, in this case, comparisons to 50 other cell types suggested these candidates are putatively restricted to stem cells. Thus, despite being identified by relatively few PSMs in CSC analyses, proteins that are uniquely observed in a single cell type can be valuable immunophenotyping markers provided there are data of a similar type and quality on other cell types for comparison. Altogether, these data highlight the importance of context during marker selection and the value of considering additional datasets. Specifically, if additional datasets are integrated prior to calculation of *GenieScore,* candidates with a lower *signal strength* (few PSMs) would rank more highly because

they would have a higher *signal dispersion* (all PSMs coming from a single cell type). Overall, these evaluations of previously validated datasets illustrate how *GenieScore* is a useful strategy to prioritize candidate cell surface markers using both proteomic and transcriptomic datasets.

Integrating GenieScores of proteomic and transcriptomic data to reveal candidate markers for mouse islet cell types

As *GenieScore* provided a useful rank-ordering of potential protein markers from both RNA-Seq and CSC data that was consistent with published results, we sought to evaluate its utility for integrating data from disparate studies for marker discovery. To this end, we performed CSC on mouse α and β cell lines and compared the results to published RNA-Seq data acquired on primary α and β cells from dissociated mouse islets (31). The datasets shared 321 predicted surface proteins (Figure 5A, Dataset S4.1), and *GenieScores* from the CSC data were plotted against *GenieScores* from the RNA-Seq data (Figure 5B). A possible explanation for the weak correlation (r_s = 0.26) between *GenieScores* is that the CSC dataset was acquired on cell lines and the RNA-Seq dataset was acquired on primary cells. However, in the context of marker discovery, each of these approaches offers advantages, namely, the CSC data provides experimental evidence regarding protein abundance at the cell surface and the RNA-Seq analysis of primary cells avoids possible artifacts introduced by culturing cells *ex vivo*. Recognizing the complementary benefits of these approaches, the data were combined in a manner that weights them equally, namely, the *signal dispersion* was calculated using the average of the normalized CSC and normalized RNA-Seq data. The combined *GenieScores* were distributed similarly to scores calculated using CSC or RNA-Seq individually and when plotted against the $\log_2$fold ratio between α and β cells allow for visual discrimination of the candidate markers for each cell type (Figure 5C, Dataset S4.2). Among the top candidate markers for α and β cells revealed by this combined approach are proteins with well-established roles in islet biology including GLP1, GABBR2, GALR1, KCNK3, SLC7A2 - proteins highlighted in a recent review of the β cell literature

(32). ALCAM (CD166), CHR1, and CEACAM1 are proteins which have been studied in the context of the islet biology, though have less defined roles (33-35). Altogether, *GenieScore* provided a useful framework for integrating proteomic and transcriptomic data for surface marker prioritization.

To extend the analysis beyond the identification of proteins which might be capable of distinguishing α and β cells to finding cell-type specific markers within the context of the islet, we applied *GenieScore* to a single-cell RNA-Seq dataset that was collected on cells from dissociated human islets (36) (Dataset S5.1). Lawlor *et al* partitioned the data on single cells into seven different cell types – α, β, γ, δ, acinar, ductal, stellate - based on a subset of genes that were determined to be representative of each of the clusters. Top ranking markers for each of the seven cell types are listed in Figure 5D. Many of the proteins identified as capable of distinguishing between α and β cells in the analysis of CSC and RNA-Seq data were not cell-type specific when data from other cell types found within the islet were considered. For example, NRCAM and SLC4A10 are proteins more abundant in β than α cells, but the levels expressed in β cells are equivalent to γ or δ cells, respectively. PTPRK is expressed at a higher level in α than β cells in all studies, but the level of expression is 26-fold lower than acinar cells and 35-fold lower than ductal cells. Altogether, while the cell-type specificity that is ultimately required will depend on the desired downstream application, these observations highlight that consideration of a larger cellular context is important for the identification of cell-type specific markers.

Recognizing the utility of the *GenieScore* approach for prioritizing cell-type specific surface proteins, the equation was further adapted to enable prioritization of other classes of proteins using the same input data. First, removal of the *SPC score* component from *GenieScore*, a permutation termed *OmniGenieScore,* allowed for the identification of proteins which can be used as cell-type specific markers without considering their surface localization. Application of *OmniGenieScore* to the islet cell single-cell RNA-seq data revealed many known cell-type specific

markers such as glucagon (GCG) for α cells, insulin (INS) for β cells, pancreatic polypeptide (PPY) for γ cells, and somatostatin (SST) for δ cells (Figure 5D). By inverting the *signal dispersion* term (*i.e.* $1 - (\frac{G}{G_{max}})^2$), a permutation termed *IsoGenieScore*, the set of cell surface proteins which are relatively abundant and similar in signal among all cell types in the analysis were prioritized. The classes of proteins (*e.g.* adhesion, cell growth, insulin signaling) which were at the top of this ranking system were largely involved in generic processes that are not specific to any cell type (Figure 5D). Reversing the *signal dispersion* and ignoring the *SPC score*, termed *IsoOmniGenieScore*, resulted in prioritization of proteins typically selected to be loading controls for Western blotting or reference genes for PCR (*e.g.* GAPDH, ACTB, B2M) in addition to many of the proteins involved in mitochondrial oxidation (Figure 5D). Altogether, these four permutations of the *GenieScore* enabled the prioritization of candidate markers for a broad range of applications, including cell surface and intracellular markers that distinguish cell types as well as those that are co-expressed among cell types.

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

To enable calculation of *GenieScores* for user input data, a shinyApp, SurfaceGenie, was developed in R. In this interface, users upload data from proteomic or transcriptomic experiments as a .csv file and can view the distribution of *GenieScores* and *SPC scores* for the proteins contained in their data (Figure 6A). SurfaceGenie is compatible with human, mouse, and rat data. As part of the analysis, input proteins are annotated with ontological information including CD and HLA molecule designations. In addition, proteins are annotated with the number of cell types within the CSPA in which the protein has been observed – a factor found to be relevant for marker prioritization in the Boheler *et al* data. The plots and data generated are available for download, including the results for individual terms used to calculate *GenieScore*. The permutations of *GenieScore* applied in Figure 5D are also available. 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 (screen captures shown in Figure 6B).

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 as a baseline for comparison to experimentally-perturbed or diseased phenotypes. 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. However, with technologies that allow for 'omic' evaluation of individual cell types, there is a need to develop methodologies to prioritize the value contained within these studies in order to extract useful knowledge from acquired data.

Here, we describe the development of *GenieScore,* a prioritization approach that integrates a predictive metric regarding surface localization with experimental data to rank-order proteins which may be useful as cell surface markers. We demonstrate that *GenieScore* is compatible with quantitative data from CSC, WCL, and RNA-Seq experiments and is a useful strategy by which to integrate multiple sources of data for candidate marker prioritization. SurfaceGenie, a web-based application, was developed to enable the calculation of *SPC scores* and *GenieScores*, and the various permutations thereof, from user-input data. SurfaceGenie also supplements the data with annotations relevant for marker selection.

Beyond immunophenotyping, SurfaceGenie is expected to facilitate the identification of valuable drug targets as the features of cell surface markers (*e.g.* surface localization and cell-type specificity) are also advantageous when designing efficient and specific therapies. Independent of *GenieScore*, the ability to query *SPC scores* within SurfaceGenie can deliver value in-and-of-itself, providing users with an additional resource to interrogate surface localization for proteins which are not yet characterized experimentally. 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 1) lack traditional sequence motifs (*e.g.* signal peptides), 2) are only trafficked to the cell surface upon ligand binding (*e.g.* glucose transporter 3, GLUT3), or 3) have proteoforms that exhibit different subcellular localization than the canonical version of a protein for which predictions are typically based upon. 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.

In conclusion, we anticipate that SurfaceGenie will enable effective prioritization of informative candidate cell surface markers to support a broad range of research questions, from mechanistic to disease-related studies. The candidates prioritized using SurfaceGenie are expected to be of use to a range of applications including immunophenotyping, immunotherapy, and drug targeting.

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 (15). α TC1 clone 6 (ATCC, CRL-2934) and β-TC-6 (ATCC, CRL-11506) cells were maintained at 37˚C and 5% CO_2, cultured in Dulbecco's Modified Eagle's Medium (Gibco) 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 100 mM ammonium bicarbonate containing 20% acetonitrile and 40% Invitrosol (Thermo Fisher Scientific), digested with trypsin (Promega, Madison, WI) overnight, and cleaned by SP2 (37). Peptides were quantified using Pierce Quantitative Fluorometric Peptide Assay (Thermo Fisher Scientific) according to manufacturer's instructions on a Varioskan LUX Multimode Microplate Reader and SkanIt 5.0 software (Thermo Fisher Scientific). For CSC analysis of mouse islet cell lines, samples were prepared as previously described (15-17).

Mass Spectrometry Acquisition and Analysis

Lymphocyte peptides and CSC samples of mouse islet cell types were analyzed by LC-MS/MS using a Dionex UltiMate 3000 RSLCnano system (Thermo Fisher Scientific) in line with a Q Exactive (Thermo Fisher 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 and queued in blocked and randomized order with two technical replicates analyzed per sample. CSC samples of mouse islet cell types were analyzed as described (38, 39). MS data were analyzed using Proteome Discoverer 2.2 (Thermo Fisher Scientific) and SkylineDaily (v4.2.1.19095) (40).

Construction of a consensus dataset of predicted surface proteins

Four published surfaceome datasets (7-10), 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. The 'retrieve/mapping ID' function within UniProt (www.uniprot.org) was used to convert the gene names provided in the published datasets to UniProt Accession numbers. Ambiguous matches were clarified by any supplementary information provided in the datasets in addition to gene name (*e.g.* alternate name, molecule name, chromosome).

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 qualify as a cell surface marker for distinguishing between cell types or experimental groups. The equation, which returns a metric termed *GenieScore*, is the product of 1) the *SPC scores* (described above); 2) *signal dispersion*, a measure of the disparity in observations among investigated samples that is mathematically equivalent to the square of the normalized Gini coefficient (41); and 3) *signal strength*, a logarithmic transformation of the experimental data (*e.g.* number of PSMs, MS1 peak area, FKPM, or RKPM). A thorough definition and rationalization of the individual equation terms is provided in Supporting Information.

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

Application of GenieScore

Details for the strategies applied to calculate *GenieScores* for each study are provided in Supporting Information.

SurfaceGenie Web application

A web application for accessing SurfaceGenie was developed as an interactive Shiny app written in R and is available for use at www.cellsurfer.net/surfacegenie. Source code is available at www.github.com\GundryLab\SurfaceGenie.

Supporting Information

1. Supplemental Methods
2. Figure S1 – Benchmarking of *Surface Protein Consensus* (*SPC*) database against CSPA and HyperLOPIT annoations
3. Figure S2 – Hierarchical clustering applied to all and predicted surface proteins for Cell Surface Capture and whole-cell lysate data
4. Figure S3 – Correlation of *GenieScore* experimental terms with statistical significance and correlation of *GenieScores* calculated using PSMs or MS1-based peak area
5. Dataset S1 – (1) Human SPC dataset, (2) Mouse SPC dataset, (3) Rat SPC dataset,
6. Dataset S2 - (1) Lymphocyte WCL data with *GenieScores*, (2) Lymphocyte CSC data with *GenieScores*, (3) *GenieScores* for proteins common to CSC and WCL
7. Dataset S3 – (1) CSC data on MCF10A KRAS^{G12V} and empty vector controls with *GenieScores*, (2) RNA-Seq data on MCF10A KRAS^{G12V} and empty vector controls with *GenieScores*
8. Dataset S4 – (1) Human dermal fibroblast and stem cell CSC data with *GenieScores*
9. Dataset S5 – (1) CSC data on mouse α and β cells with *GenieScores*, (2) RNA-Seq data on mouse α and β cells with *GenieScores*
10. Dataset S6 – (1) Single-cell RNA-Seq on human islet cells with *GenieScores* and modified *GenieScores*

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 R 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.; DK-052194 and AI-44458 to J.A.C.]; S.S. is a member of the MCW-MSTP which is partially supported by a T32 grant from NIGMS, GM080202. Special thanks to Dr. Christopher Ashwood and Linda Berg Luecke for critical review of the manuscript and insightful discussions. Funding sources were not involved in study design, data collection, interpretation, analysis or publication.

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. Wollscheid B*, et al.* (2009) Mass-spectrometric identification and relative quantification of N-linked cell surface glycoproteins. *Nature biotechnology* 27(4):378-386.
5. 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.
6. 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.
7. Bausch-Fluck D*, et al.* (2018) The in silico human surfaceome. *Proc Natl Acad Sci U S A* 115(46):E10988-E10997.
8. da Cunha JP*, et al.* (2009) Bioinformatics construction of the human cell surfaceome. *Proc Natl Acad Sci U S A* 106(39):16752-16757.
9. 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.
10. Diaz-Ramos MC, Engel P, & Bastos R (2011) Towards a comprehensive human cell-surface immunome database. *Immunol Lett* 134(2):183-187.
11. Ashburner M*, et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* 25(1):25-29.
12. The Gene Ontology C (2019) The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Res* 47(D1):D330-D338.
13. Christoforou A*, et al.* (2016) A draft map of the mouse pluripotent stem cell spatial proteome. *Nat Commun* 7:8992.
14. Bendtsen JD, Kiemer L, Fausboll A, & Brunak S (2005) Non-classical protein secretion in bacteria. *BMC Microbiol* 5:58.
15. Haverland NA*, et al.* (2017) Cell Surface Proteomics of N-Linked Glycoproteins for Typing of Human Lymphocytes. *Proteomics* 17(19).
16. 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.
17. Gundry RL*, et al.* (2012) A cell surfaceome map for immunophenotyping and sorting pluripotent stem cells. *Molecular & cellular proteomics : MCP* 11(8):303-316.
18. Damle RN*, et al.* (2002) B-cell chronic lymphocytic leukemia cells express a surface membrane phenotype of activated, antigen-experienced B lymphocytes. *Blood* 99(11):4087-4093.
19. Pulte D*, et al.* (2007) CD39 activity correlates with stage and inhibits platelet reactivity in chronic lymphocytic leukemia. *J Transl Med* 5:23.
20. Johnston HE*, et al.* (2018) Proteomics Profiling of CLL Versus Healthy B-cells Identifies Putative Therapeutic Targets and a Subtype-independent Signature of Spliceosome Dysregulation. *Molecular & cellular proteomics : MCP* 17(4):776-791.
21. He X*, et al.* (2018) Continuous signaling of CD79b and CD19 is required for the fitness of Burkitt lymphoma B cells. *EMBO J* 37(11).
22. Anonymous (2016) WHO Classification: Tumours of the Haematopoietic and Lymphoid Tissues (2008). *Postgraduate Haematology, 7th Edition*:885-887.

23. Takimoto CH*, et al.* (2019) The Macrophage 'Do not eat me' signal, CD47, is a clinically validated cancer immunotherapy target. *Ann Oncol* 30(3):486-489.
24. Martinko AJ*, et al.* (2018) Targeting RAS-driven human cancer cells with antibodies to upregulated and essential cell-surface proteins. *Elife* 7.
25. Vivekanandhan S*, et al.* (2017) Genetic status of KRAS modulates the role of Neuropilin-1 in tumorigenesis. *Sci Rep* 7(1):12877.
26. Meyerson H*, et al.* (2017) Juvenile myelomonocytic leukemia with prominent CD141+ myeloid dendritic cell differentiation. *Hum Pathol* 68:147-153.
27. Linsalata M*, et al.* (2004) Polyamine biosynthesis in relation to K-ras and p-53 mutations in colorectal carcinoma. *Scand J Gastroenterol* 39(5):470-477.
28. Arruabarrena-Aristorena A, Zabala-Letona A, & Carracedo A (2018) Oil for the cancer engine: The cross-talk between oncogenic signaling and polyamine metabolism. *Sci Adv* 4(1):eaar2606.
29. Xing P*, et al.* (2016) Roles of low-density lipoprotein receptor-related protein 1 in tumors. *Chin J Cancer* 35:6.
30. Chen JS*, et al.* (2010) Secreted heat shock protein 90alpha induces colorectal cancer cell invasion through CD91/LRP-1 and NF-kappaB-mediated integrin alphaV expression. *J Biol Chem* 285(33):25458-25466.
31. 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.
32. Rorsman P & Ashcroft FM (2018) Pancreatic beta-Cell Electrical Activity and Insulin Secretion: Of Mice and Men. *Physiol Rev* 98(1):117-214.
33. Fujiwara K*, et al.* (2014) CD166/ALCAM expression is characteristic of tumorigenicity and invasive and migratory activities of pancreatic cancer cells. *PloS one* 9(9):e107247.
34. Schmid J*, et al.* (2011) Modulation of pancreatic islets-stress axis by hypothalamic releasing hormones and 11beta-hydroxysteroid dehydrogenase. *Proc Natl Acad Sci U S A* 108(33):13722-13727.
35. DeAngelis AM*, et al.* (2008) Carcinoembryonic antigen-related cell adhesion molecule 1: a link between insulin and lipid metabolism. *Diabetes* 57(9):2296-2303.
36. Lawlor N*, et al.* (2017) Single-cell transcriptomes identify human islet cell signatures and reveal cell-type-specific expression changes in type 2 diabetes. *Genome Res* 27(2):208-222.
37. 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*.
38. 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.
39. Mallanna SK, Waas M, Duncan SA, & Gundry RL (2016) N-glycoprotein surfaceome of human induced pluripotent stem cell derived hepatic endoderm. *Proteomics*.
40. Schilling B*, et al.* (2012) Platform-independent and label-free quantitation of proteomic data using MS1 extracted ion chromatograms in skyline: application to protein acetylation and phosphorylation. *Mol Cell Proteomics* 11(5):202-214.
41. Giugni C (1912) *Variabilità e mutabilità: Contributo allo studio delle distribuzioni e delle relazioni statistiche* (Cuppini).
42. Conway JR, Lex A, & Gehlenborg N (2017) UpSetR: an R package for the visualization of intersecting sets and their properties. *Bioinformatics* 33(18):2938-2940.
43. Lex A, Gehlenborg N, Strobelt H, Vuillemot R, & Pfister H (2014) UpSet: Visualization of Intersecting Sets. *IEEE Trans Vis Comput Graph* 20(12):1983-1992.

Figure Legends

Figure 1: Generation and benchmarking of a *Surface Prediction Consensus* (*SPC*) score. (A) The four previously published human surfaceome databases used, designated by first author of the corresponding publication, with details about how the databases were generated and the number of Uniprot Accessions within each database. (B) An UpSet plot (42, 43) depicting the intersections between the individual surfaceome databases. The proteins were stratified by the number of individual datasets they appeared in, termed *Surface Prediction Consensus* (*SPC*). The number of proteins with each SPC score is shown. The full dataset is provided in the Supporting Information (Dataset S1, 4.1) (C) The distribution of Gene Ontology Cellular Component Ontology (GO-CCO) annotations across different *SPC scores* depicted as a bubble chart, where the size of the bubble represents the number of proteins in the intersection between the particular *SPC score* and GO-CCO annotations.

Figure 2. GenieScore components and application to two proteomic analyses of four lymphocyte lines. (A) The features of a protein that were hypothesized to predominate the capacity of a protein to serve as a cell surface marker are shown with the names of the mathematical terms derived to represent them. The marker potential features are annotated by the applied approach (i.e. predictive or experimental) to answer the relevant questions. The remaining panels depict the distribution of the individual components and *GenieScores* calculated from the data acquired from application of whole-cell lysate (WCL) or Cell Surface Capture (CSC) to four lymphocyte cell line (n = 3 per cell line, N = 485 data points for WCL, N = 325 data points for CSC). (B) A histogram depicting the distribution of *SPC scores* within predicted surface proteins (SPC score >0) identified by application of WCL and CSC. (C) A violin plot depicting the distribution of *signal dispersion* for the predicted surface proteins identified by WCL and CSC. (D) A violin plot depicting the distribution of *signal strength* for the predicted surface proteins identified by WCL and CSC. (E) Plot of *GenieScore* against rank-order of candidate cell surface markers for predicted surface proteins identified by WCL. (F) Plot of *GenieScore* against rank-order of candidate cell surface markers for predicted surface proteins identified by CSC. (G) *GenieScores* calculated using either WCL or CSC data are plotted against each other the 91 proteins identified by both approaches along with the Spearman's Correlation for those scores.

Figure 3. Distributions of observed abundance for selected proteins in the lymphocyte data with a range of *GenieScores*. The number of peptide-spectrum matches (PSMs) assigned to selected proteins for both Cell Surface Capture (CSC) and whole-cell lysate (WCL) experiments. Biological replicates (n = 3) are shown as data points and averages are shown as columns. The ranks assigned to each protein, according to the set of calculated *GenieScores*, are shown for both CSC and WCL datasets.

Figure 4. Benchmarking *GenieScore* against two published cell surface marker studies which validated candidate markers by flow cytometry. Panels A-C depict data from application of *GenieScore* to data from Martinko *et al*. Panel D depicts data from application of *GenieScore* to data from Boheler *et al.* (A) The subset of proteins for which *GenieScores* were calculated is the intersection of the set of proteins with SPC scores >0 with the set of proteins that were increased in the KRAS mutant - shown by the shaded overlap. Plots of *GenieScores* against

candidate rank are shown for the Cell Surface Capture (CSC) and RNA-Seq datasets. Proteins selected in the original manuscript for antibody development and subsequently validated as surface markers by flow cytometry are shown as black diamonds and labeled with gene names. (B) The *GenieScores* calculated using either CSC or RNA-Seq data are plotted against each other for the 211 surface proteins identified by both approaches along with the Spearman's Correlation of those scores. The flow cytometry-validated markers are shown as black diamonds. (C) A table containing the ranks assigned, according to either *GenieScores* or $\log_2$fold, for each protein. The change in rank, calculated as *GenieScore* rank minus $\log_2$fold rank, is shown for each flow cytometry-validated marker. (D) *GenieScores* for the 495 proteins identified by CSC in human fibroblast and stem cells are plotted against the $\log_2$fold ratio. Refence stem cell markers, as well as the negative and positive markers for pluripotency selected for validation by flow cytometry are highlighted in their own plots.

Figure 5. Application of *GenieScore* and its permutations to islet cell types*.* Panels A-C depict data from application of *GenieScore* to Cell Surface Capture (CSC) data from mouse α and β cell lines collected as part of this study integrated with RNA-Seq on mouse primary α and β cells from Benner *et al*. Panel D depicts data from application of *GenieScore* and its permutations to human islet single-cell RNA-Seq data from Lawlor *et al.* (A) The subset of proteins for which *GenieScores* were calculated is the set of proteins with *SPC scores* >0 that were identified by both CSC and RNA-Seq, shown as the shaded overlap. (B) The *GenieScores* calculated using either CSC or RNA-Seq data are plotted against each other for the 321 proteins identified by both approaches along with the Spearman's Correlation of those scores. (C) *GenieScores* calculated using the combined CSC and RNA-Seq data are plotted against candidate rank and against the $\log_2$fold ratio (N = 321 proteins). Selected candidate markers which have previously been associated with islet cell biology are labeled with gene names. (D) The top scoring proteins from application of the different permutations of *GenieScore* are shown grouped either by cell type or by biological function.

Figure 6. Overview of the utility of SurfaceGenie and screen captures from the web application*.* (A) A schematic depicting the tested inputs and potential applications of SurfaceGenie, a web-based application which calculates *GenieScore* permutations from user-input data. (B) Screen captures of the different modes of use for the SurfaceGenie web application.

Figures:

Figure 1

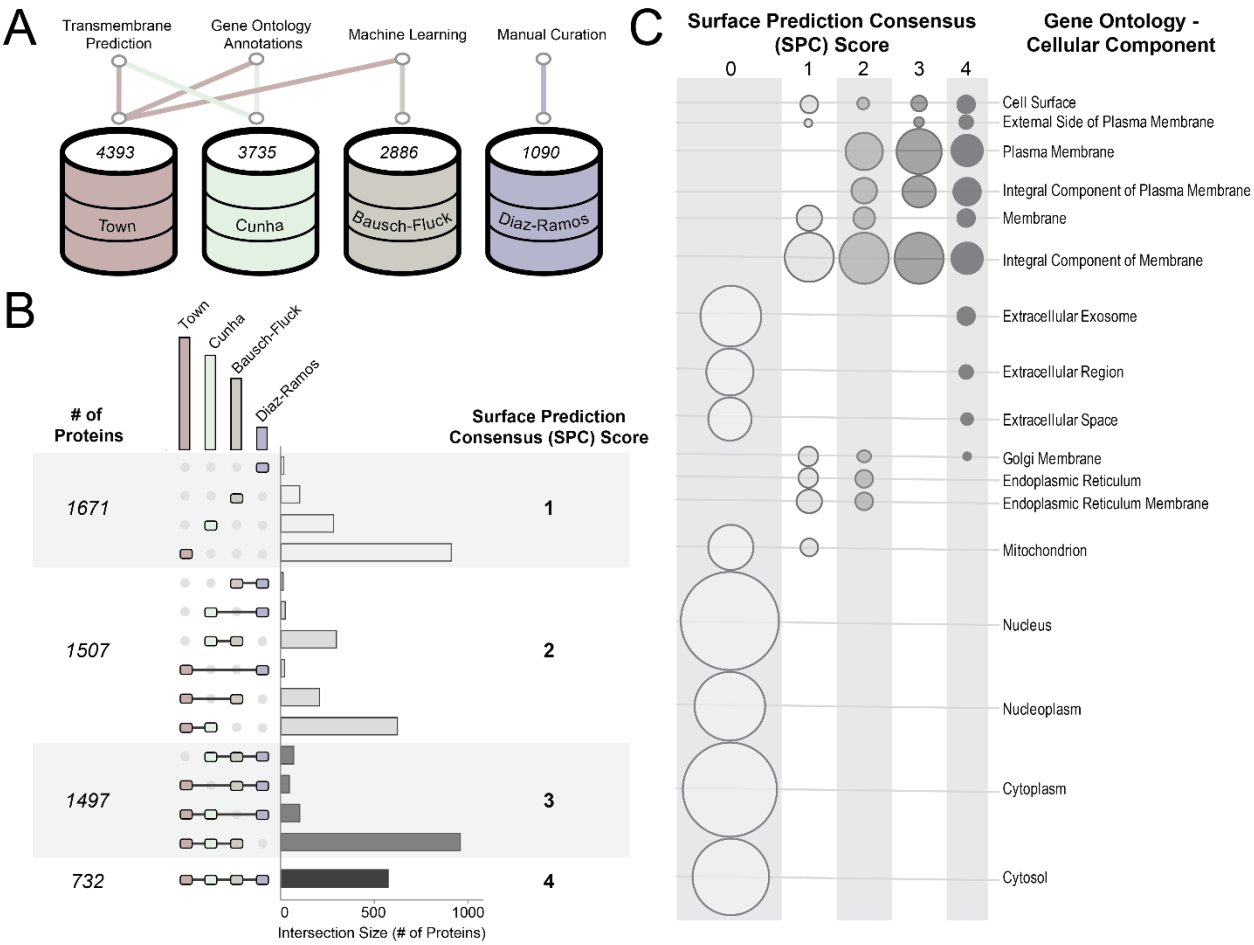


Figure 2

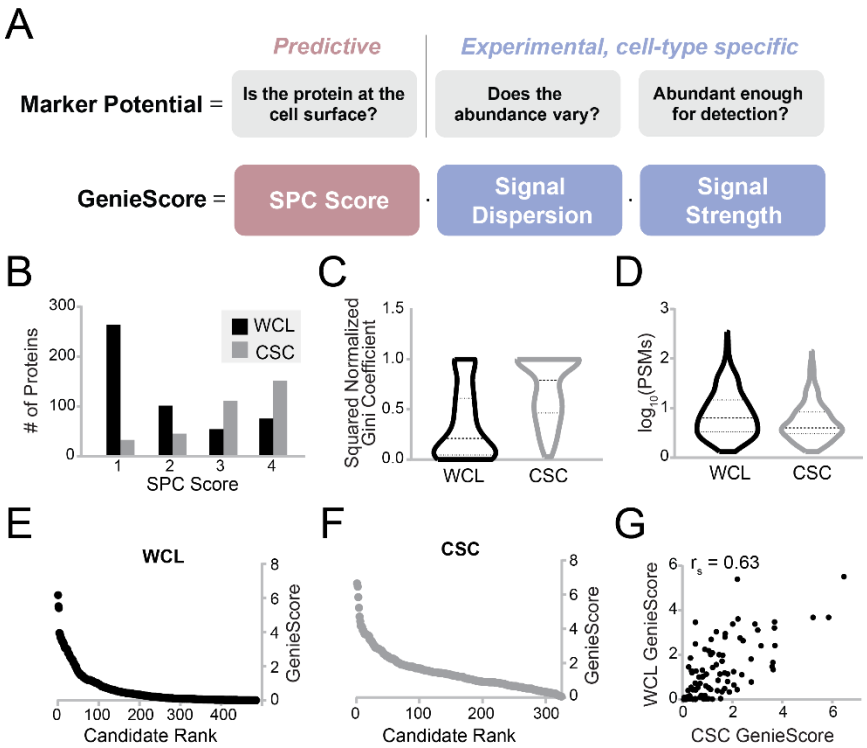

A
Predictive
Experimental, cell-type specific
Marker Potential =
Is the protein at the cell surface?
Does the abundance vary?
Abundant enough for detection?
GenieScore =
SPC Score
Signal Dispersion
Signal Strength
B
# of Proteins
WCL
CSC
SPC Score
C
Squared Normalized Gini Coefficient
WCL
CSC
D
log10(PSMs)
WCL
CSC
E
WCL
GenieScore
Candidate Rank
F
CSC
GenieScore
Candidate Rank
G
rs = 0.63
WCL GenieScore
CSC GenieScore

Figure 3

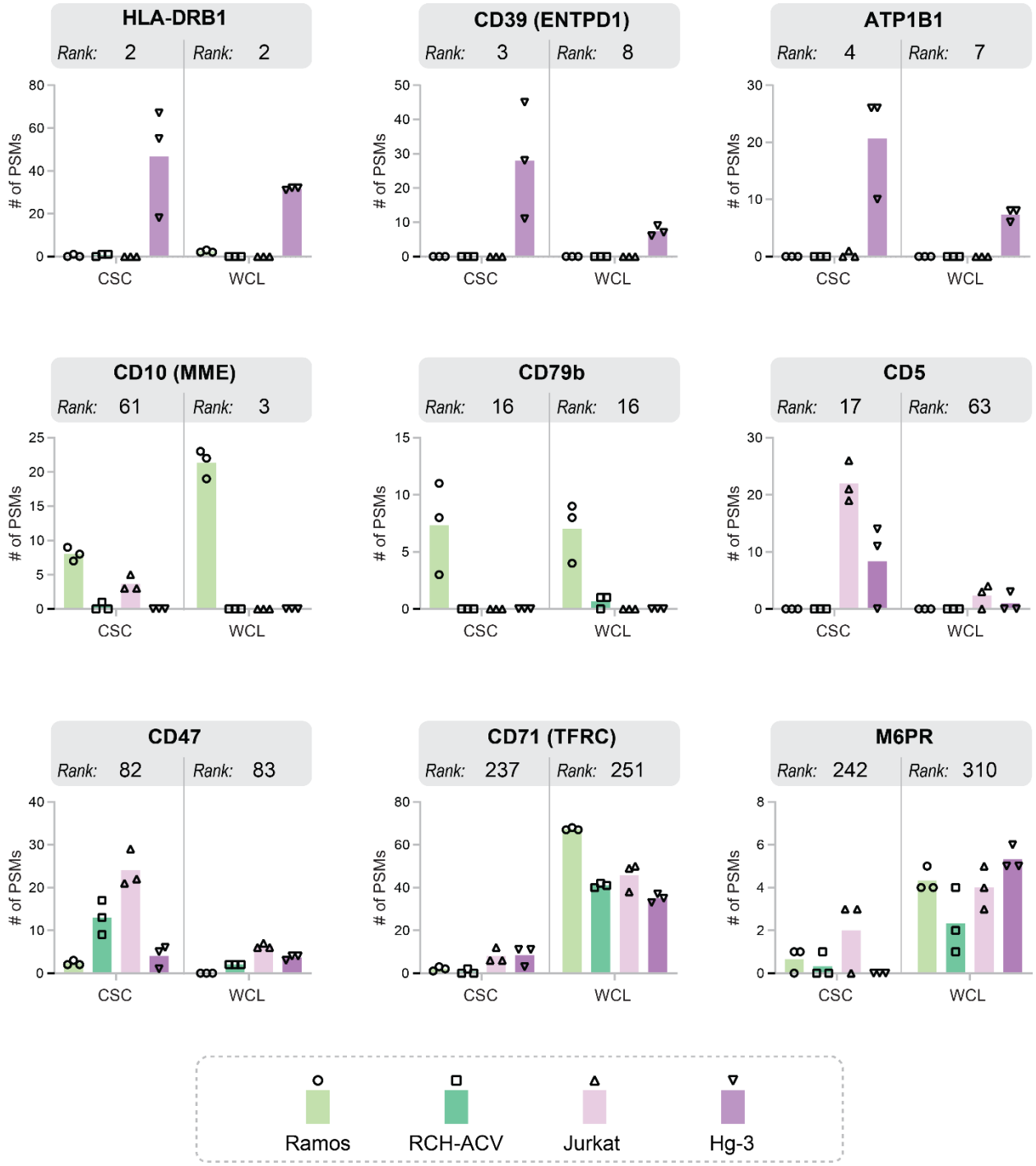
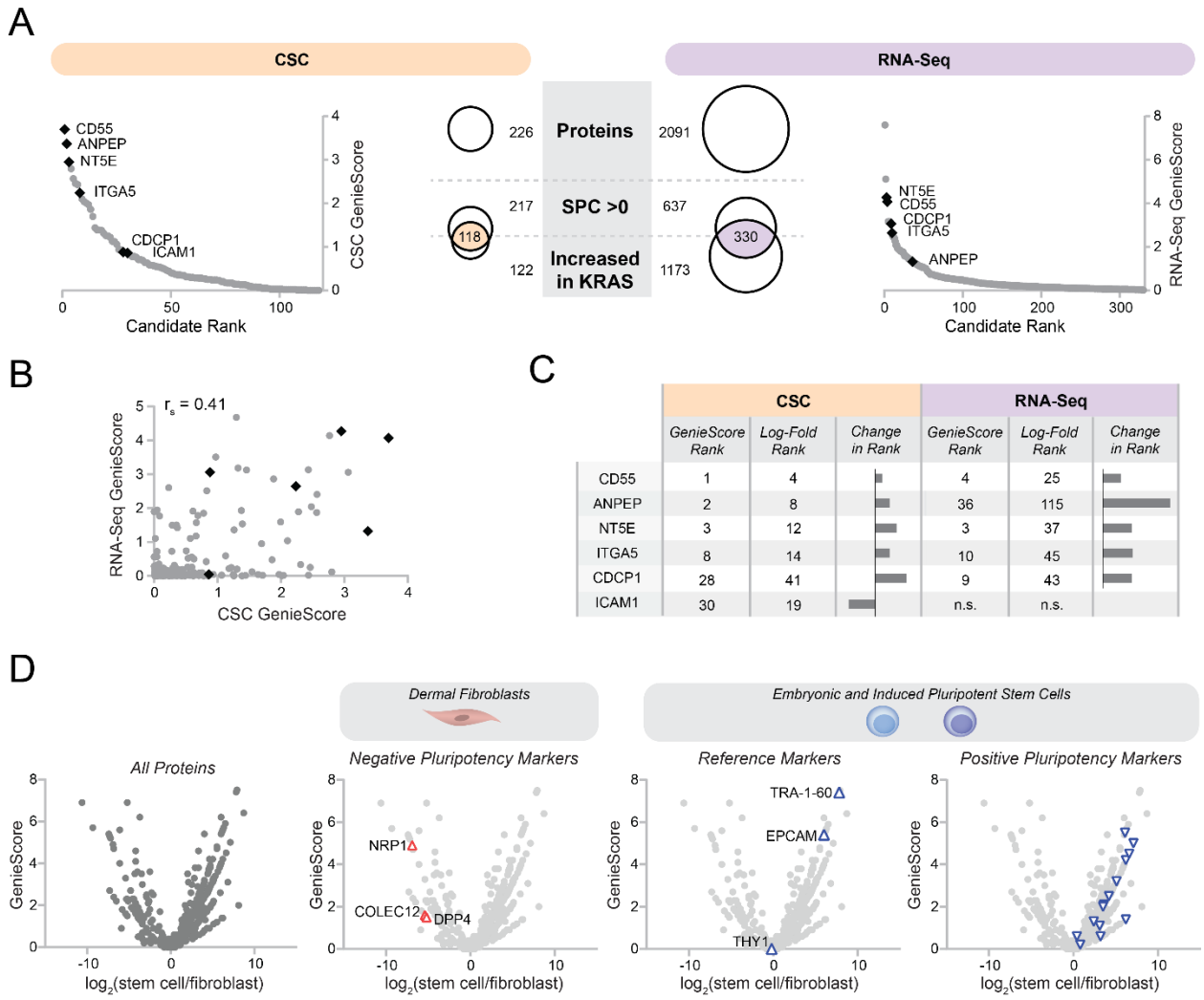

HLA-DRB1
Rank: 2
Rank: 2
CD39 (ENTPD1)
Rank: 3
Rank: 8
ATP1B1
Rank: 4
Rank: 7
CD10 (MME)
Rank: 61
Rank: 3
CD79b
Rank: 16
Rank: 16
CD5
Rank: 17
Rank: 63
CD47
Rank: 82
Rank: 83
CD71 (TFRC)
Rank: 237
Rank: 251
M6PR
Rank: 242
Rank: 310
# of PSMs
CSC
WCL
Ramos
RCH-ACV
Jurkat
Hg-3

Figure 4



	CSC			RNA-Seq		
	GenieScore Rank	*Log-Fold Rank*	*Change in Rank*	*GenieScore Rank*	*Log-Fold Rank*	*Change in Rank*
CD55	1	4		4	25	
ANPEP	2	8		36	115	
NT5E	3	12		3	37	
ITGA5	8	14		10	45	
CDCP1	28	41		9	43	
ICAM1	30	19		n.s.	n.s.	

Figure 5

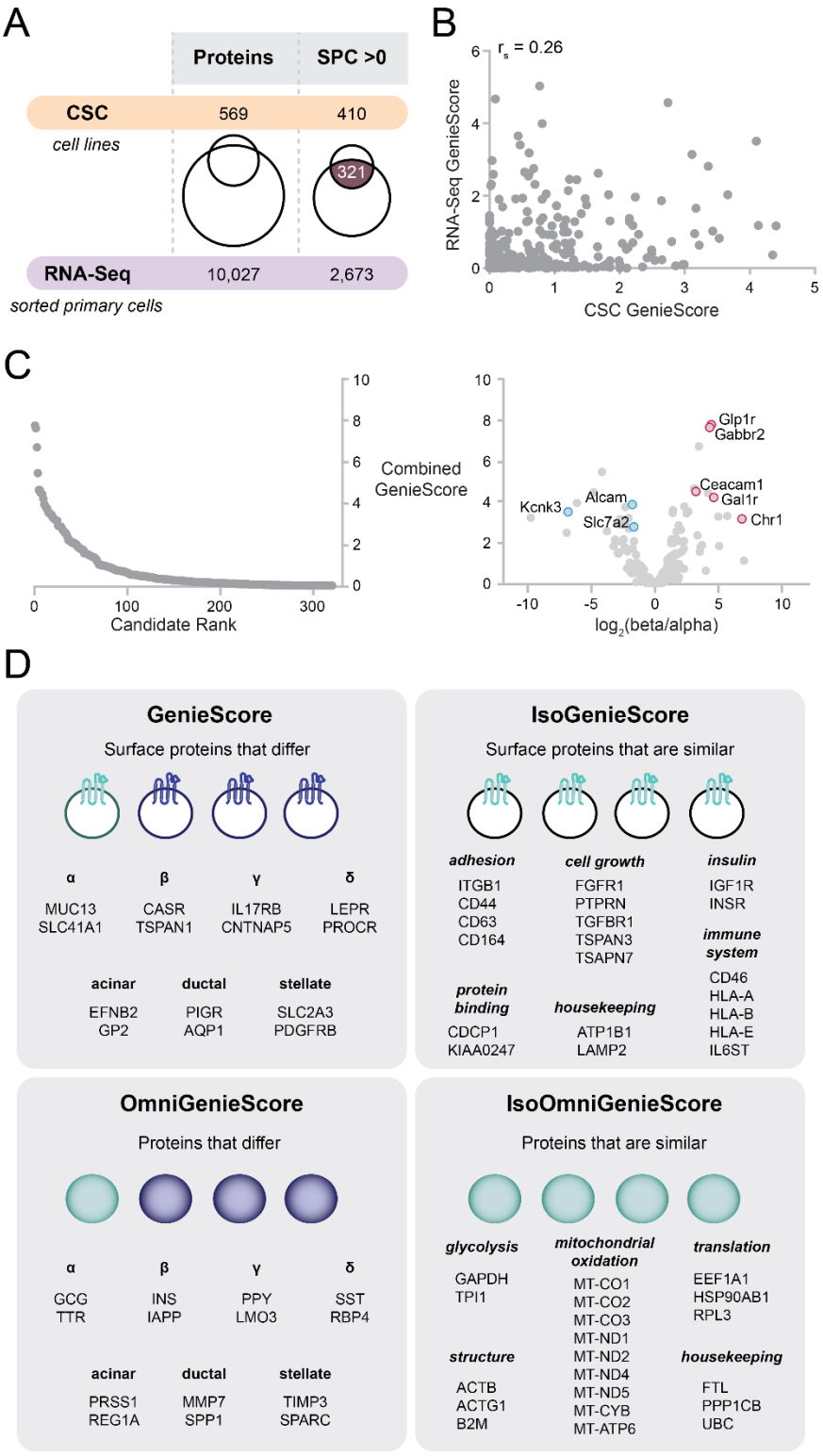
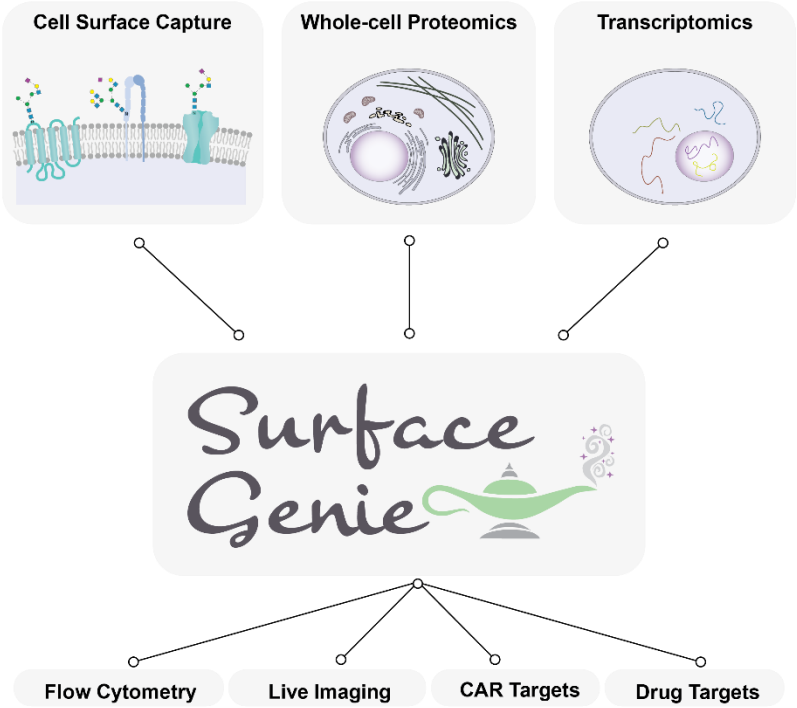
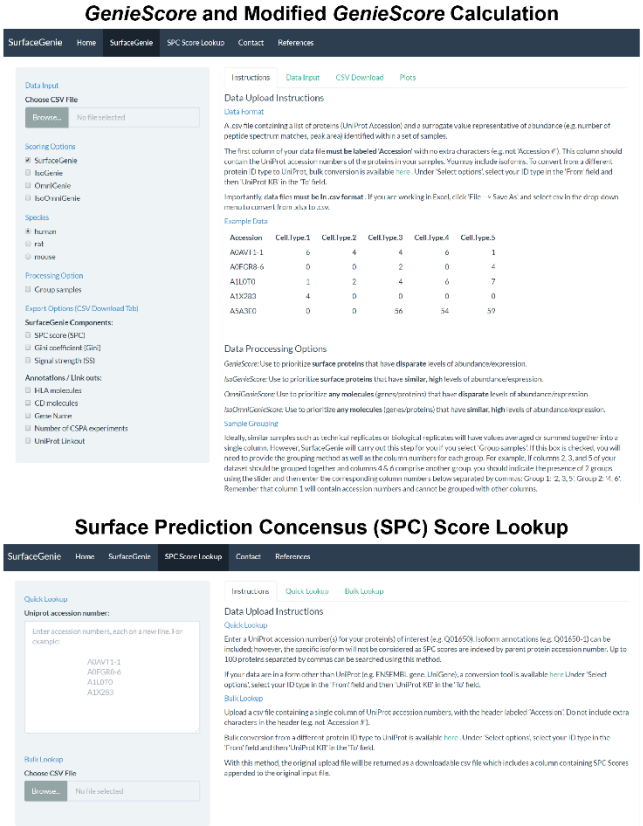

A
Proteins
SPC >0
CSC
569
410
cell lines
321
RNA-Seq
10,027
2,673
sorted primary cells
B
r_s = 0.26
RNA-Seq GenieScore
CSC GenieScore
C
Candidate Rank
Combined GenieScore
Glp1r
Gabbr2
Ceacam1
Gal1r
Chr1
Kcnk3
Alcam
Slc7a2
log2(beta/alpha)
D
GenieScore
Surface proteins that differ
α
MUC13
SLC41A1
β
CASR
TSPAN1
γ
IL17RB
CNTNAP5
δ
LEPR
PROCR
acinar
EFNB2
GP2
ductal
PIGR
AQP1
stellate
SLC2A3
PDGFRB
IsoGenieScore
Surface proteins that are similar
adhesion
ITGB1
CD44
CD63
CD164
cell growth
FGFR1
PTPRN
TGFBR1
TSPAN3
TSAPN7
insulin
IGF1R
INSR
immune system
CD46
HLA-A
HLA-B
HLA-E
IL6ST
protein binding
CDCP1
KIAA0247
housekeeping
ATP1B1
LAMP2
OmniGenieScore
Proteins that differ
α
GCG
TTR
β
INS
IAPP
γ
PPY
LMO3
δ
SST
RBP4
acinar
PRSS1
REG1A
ductal
MMP7
SPP1
stellate
TIMP3
SPARC
IsoOmniGenieScore
Proteins that are similar
glycolysis
GAPDH
TPI1
mitochondrial oxidation
MT-CO1
MT-CO2
MT-CO3
MT-ND1
MT-ND2
MT-ND4
MT-ND5
MT-CYB
MT-ATP6
translation
EEF1A1
HSP90AB1
RPL3
structure
ACTB
ACTG1
B2M
housekeeping
FTL
PPP1CB
UBC

Figure 6

A


Cell Surface Capture
Whole-cell Proteomics
Transcriptomics
Surface Genie
Flow Cytometry
Live Imaging
CAR Targets
Drug Targets

B


GenieScore and Modified *GenieScore* Calculation
SurfaceGenie
Home
SurfaceGenie
SPC Score Lookup
Contact
References
Data Input
Choose CSV File
Browse...
No file selected
Scoring Options
SurfaceGenie
IsoGenie
OmniGenie
IsoOmniGenie
Species
human
rat
mouse
Processing Option
Group samples
Export Options (CSV Download Tab)
SurfaceGenie Components:
SPC score (SPC)
Gini coefficient (Gini)
Signal strength (SS)
Annotations / Link outs:
HLA molecules
CD molecules
Gene Name
Number of CSPA experiments
UniProt Linkout
Instructions
Data Input
CSV Download
Plots
Data Upload Instructions
Data Format
A .csv file containing a list of proteins (UniProt Accession) and a surrogate value representative of abundance (e.g. number of peptide spectrum matches, seak area identified with a set of samples.
The first column of your data file must be labeled 'Accession' with no extra characters (e.g. not 'Accession #'). This column should contain the UniProt accession numbers of the proteins in your samples. You may include isoforms. To convert from a different protein ID type to UniProt, bulk conversion is available here. Under 'Select options', select your ID type in the 'From' field and then 'UniProt KB' in the 'To' field.
Importantly, data files must be in .csv format. If you are working in Excel, click 'File' > Save As' and select csv in the drop down menu to convert from .xlsx to .csv.
Example Data
Accession Cell.type.1 Cell.type.2 Cell.type.3 Cell.type.4 Cell.type.5
A0AVT1-1 6 4 4 6 1
A0FGR8-6 0 0 2 0 4
A1L0T0 1 2 4 6 7
A1X283 4 0 0 0 0
A5A3E0 0 0 56 54 59
Data Processing Options
GenieScore: Use to prioritize surface proteins that have disparate levels of abundance/expression.
IsoGenieScore: Use to prioritize surface proteins that have similar, high levels of abundance/expression.
OmniGenieScore: Use to prioritize any molecules (genes/proteins) that have disparate levels of abundance/expression.
IsoOmniGenieScore: Use to prioritize any molecules (genes/proteins) that have similar, high levels of abundance/expression.
Sample Grouping
Ideally, similar samples such as technical replicates or biological replicates will have values averaged or summed together into a single column. However, SurfaceGenie will carry out this step for you if you select 'Group samples'. If this box is checked, you will need to provide the grouping method as well as the column numbers for each group. For example, if columns 2, 3, and 5 of your dataset should be grouped together and columns 4 & 6 comprise another group, you should indicate the presence of 2 groups using the slider and then enter the corresponding column numbers below separated by commas: Group 1: '2, 3, 5'; Group 2: '4, 6'. Remember that column 1 will contain accession numbers and cannot be grouped with other columns.
Surface Prediction Concensus (SPC) Score Lookup
SurfaceGenie
Home
SurfaceGenie
SPC Score Lookup
Contact
References
Quick Lookup
Uniprot accession number:
Enter accession numbers, each on a new line. For example:
A0AVT1-1
A0FGR8-6
A1L0T0
A1X283
Bulk Lookup
Choose CSV File
Browse...
No file selected
Instructions
Quick Lookup
Bulk Lookup
Data Upload Instructions
Quick Lookup
Enter a UniProt accession number(s) for your protein(s) of interest (e.g. Q01650). Isoform annotations (e.g. Q01650-1) can be included; however, the specific isoform will not be considered as SPC scores are indexed by parent protein accession number. Up to 100 proteins separated by commas can be searched using this method.
If your data are in a form other than UniProt (e.g. ENSEMBL gene, UniGene), a conversion tool is available here. Under 'Select options', select your ID type in the 'From' field and then 'UniProt KB' in the 'To' field.
Bulk Lookup
Upload a csv file containing a single column of UniProt accession numbers, with the header labeled 'Accession'. Do not include extra characters in the header (e.g. not 'Accession #').
Bulk conversion from a different protein ID type to UniProt is available here. Under 'Select options', select your ID type in the 'From' field and then 'UniProt KB' in the 'To' field.
With this method, the original upload file will be returned as a downloadable csv file which includes a column containing SPC Scores appended to the original input file.

Supplementary Information Text

Rationalization and description of GenieScore equation components.

Surface Protein Consensus (SPC) score was generated from concatenating four individual human surfaceome databases and assigning a point for each of the individual datasets in which the protein was predicted to be localized to the cell surface. *SPC scores* range 0-4 such that proteins with more consensus of surface localization are prioritized over proteins with less consensus. Human, mouse and Rat SPC scores are in Dataset S1.

Signal dispersion is calculated for each protein based on the quantitative measurements from each cell type. First, the Gini coefficient, a measure of disparity, is calculated on the array of measurements. Next, this value is normalized by dividing by the maximum Gini coefficient possible, (1 – 1/N), where N is equal to the number of cell types. Finally, this value is squared to increase the weight assigned to this term. The values for this term range 0-1. Proteins with exactly equal measurements across cell types will score 0, proteins only observed in a single cell type will score 1. This measurement does not assume the normal distribution of data and requires no imputation of zero-values making it amenable to many types of quantitative measurements.

Signal strength is calculated for each protein based on the quantitative measurements from each cell type. First, the maximum measurement is calculated for each protein. Next, the $\log_{10}$ is calculated for 1 plus this value, in order to force all the values to be returned as positive numbers. This results in proteins at the lower limit of detection being of lower priority than those with a stronger signal, because it is expected that those of higher abundance will practically serve as more accessible markers for downstream technologies. *Signal strength* is not a bounded term and the range highly depends on the type of quantitative measurement.

Modifications to the GenieScore equation.

IsoGenieScore utilizes the same three calculations as *GenieScore* (see above), however, it uses (1 - *signal dispersion*). This prioritizes proteins with equal and intense measurements as opposed to those with disparate measurements.

OmniGenieScore is equal to the product of *signal dispersion* and *signal strength*. This prioritizes molecules with disparate measurements without considering the surface localization. As this score doesn't apply any protein-specific information, it can be calculated on any type on quantitative data.

IsoOmniGenieScore is equal to the product of (1 - *signal dispersion*) and *signal strength*. This prioritizes molecules with equal and intense measurements without considering the surface localization. As this score doesn't apply any protein-specific information, it can be calculated on any type on quantitative data.

Methods

Cell lysis, protein digestion, and peptide cleanup. For whole-cell lysate analysis of lymphocyte cell lines, pellets of 5 million cells were lysed in 500 µL of 2x Invitrosol (40% v/v; Thermo Fisher Scientific), 20% acetonitrile in 50 mM ammonium bicarbonate. Sample was sonicated (VialTweeter; Hielscher Ultrasonics, Teltow, Germany) by three ten-second pulses, set on ice for one minute, and then sonicated by three ten-second pulses. Samples were brought to 5mM TCEP and reduced for 30 min at 37°C on a Thermomixer at 1200 rpm. Samples were brought to 10 mM IAA and alkylated for 30 min at 37°C on a Thermomixer at 1200 rpm in the dark. 20 µg trypsin was added to each sample and was digested at 37°C overnight on a Thermomixer at 1200 rpm.

Hierarchical clustering of lymphocyte whole-cell lysate (WCL) and Cell Surface Capture (CSC) data. Data containing the number of peptide-spectrum matches were uploaded into SPSS

(v. 22). Hierarchical clustering was performed using Phi-square measure of distance (appropriate for count data) and furthest neighbor (complete) linkage. Clustering was performed on entire dataset and then repeated for predicted surface proteins.

MS1 Peak Area Quantification. RAW and searched MS data for CSC and WCL were imported into SkylineDaily (v4.2.1.19095) (1). For both CSC and WCL, peptide inclusion criteria were (1) fully tryptic, (2) no missed cleavages, (3) length 6-30, (4) exclude 25 N-terminal amino acids, and (5) no methionine residues. For WCL samples, all default, sequence-based exclusion criteria in Skyline except cysteine were further applied. For CSC, proteins with ≥ 3 peptides were selected for MS1-based quantification. For WCL, proteins with ≥ 5 peptides were candidates for MS1-based quantification. From among these candidates, the proteins with the top 15 and bottom 15 *GenieScores* were selected for MS1-based quantification.

Implementation of GenieScore for each dataset.

Lymphocyte whole-cell lysate (WCL) and Cell Surface Capture (CSC) data: WCL data were acquired and searched as part of this study using parameters in Tables S1 and S2. Searched WCL data were filtered to include only proteins with ≥2 unique peptides. RAW files were obtained from MassIVE (massive.ucsd.edu; accession number MSV000080532) for CSC experiments performed by Haverland *et al.* (2) and re-searched using parameters in Table S2 (CSC Hi-Hi). Searched CSC data were filtered to include only proteins with ≥ 2 peptide-spectrum matches (PSMs) among all samples. All data are in Dataset S2.

CSC and RNA-Seq data on MCF10A KRAS^{G12V} and empty vector controls: CSC and RNA-Seq data were obtained from Supplemental Files 1 and 5, respectively, from Martinko *et al*. (3). Only transcripts marked as "significantly different" were included in RNA-Seq analysis. As only $\log_2$fold changes were provided for CSC data, these data were transformed to allow calculation of *signal dispersion*. The *signal strength* component calculated from FPKM values were used for both CSC and RNA-Seq analyses. All data are in Dataset S3.

Human stem cell and dermal fibroblast CSC data: RAW files were obtained from MassIVE (massive.ucsd.edu; accession number MSV000083846) for CSC experiments performed by Boheler *et al.* (4). RAW files for embryonic stem cells (DR-11, DR-17, DR-27, DR-29), induced pluripotent stem cells (DR-28, DR-30, DR-31), and dermal fibroblasts (DR-12, RG-107, RG-108) were re-searched using parameters in Table S2 (CSC Hi-Lo). Searched CSC data were filtered to include only proteins with ≥ 2 PSMs among all samples. Embryonic and induced pluripotent stem cells were treated as a single group, averaging the number of PSMs for each protein. All data are in Dataset S4.

α and β cell CSC and RNA-Seq data: CSC data were acquired and RAW files were searched as part of this study according to parameters in Table S1 and S2 (CSC Hi-Hi). Searched CSC data were filtered to include only proteins occurring in ≥ 2 biological replicates. RNA-Seq data were obtained from Supplemental File 12 from Benner *et al.*(5). The combined *GenieScore* calculations were performed by first normalizing the PSMs and RPKM measurements individually to the maximum value for each protein. Next, the average of the normalized values was calculated for both α and β cells. Finally, the *signal dispersion* was calculated using these averaged, normalized measurements. The sum of the CSC and RNA-Seq *signal strength* values was used for the calculation of the combined *GenieScores*. All data are in Dataset S5.

Islet cell single-cell RNA-Seq: RNA-Seq data were obtained from Supplemental File 6 from Lawlor *et al.* (6). All data are in Dataset S6.

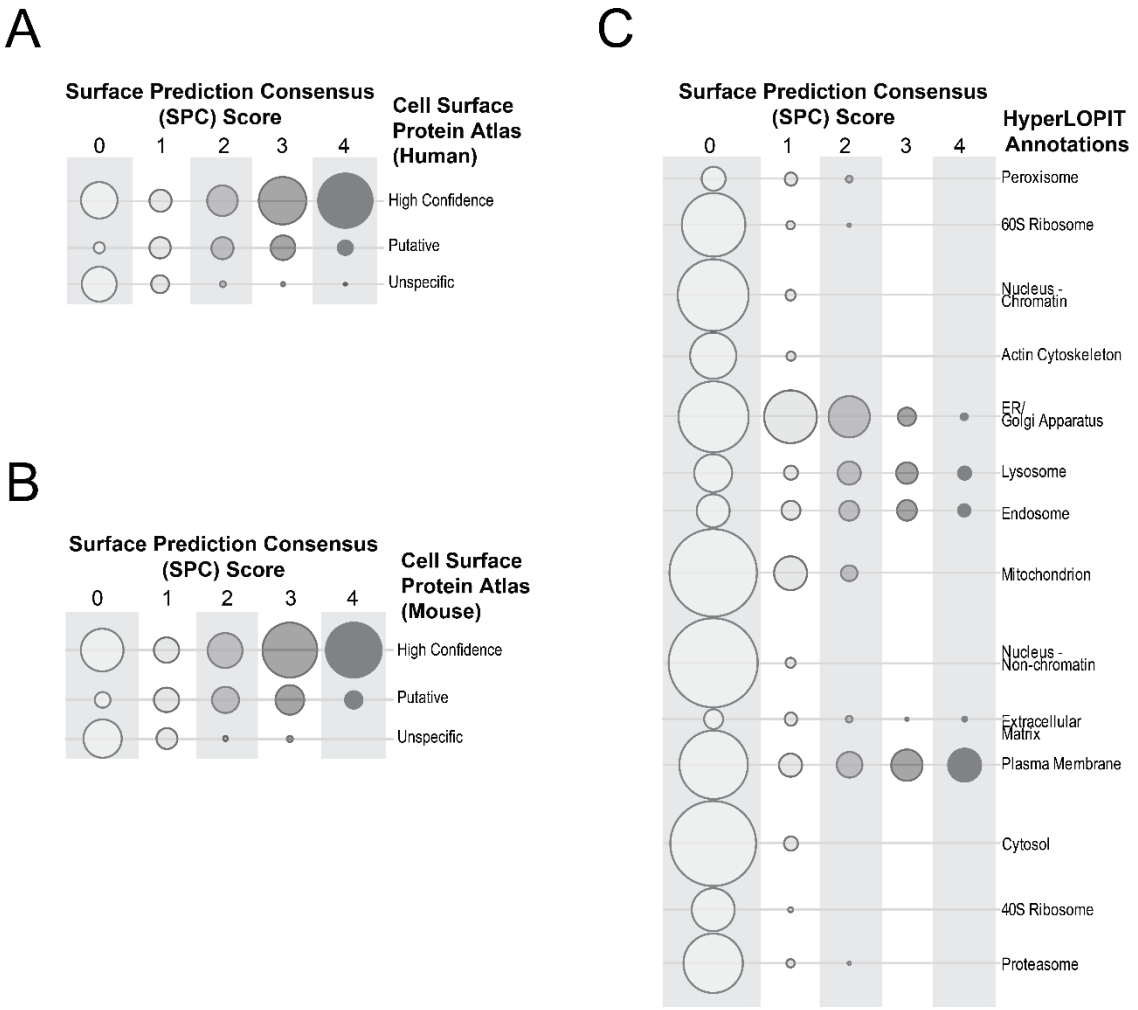


Fig. S1. Benchmarking Surface Protein Consensus (SPC) scores. (A-B) The distribution of confidence assignments within the Cell Surface Protein Atlas (CSPA) (7) across different *SPC scores* for human and mouse datasets depicted as bubble charts, where the size of the bubble represents the number of proteins in the intersection between the particular *SPC score* and CSPA annotations. (C) The distribution of annotations assigned by application of HyperLOPIT (8) to mouse stem cells across different *SPC scores* depicted as a bubble chart, where the size of the bubble represents the number of proteins in the intersection between the particular *SPC score* and HyperLOPIT annotation.

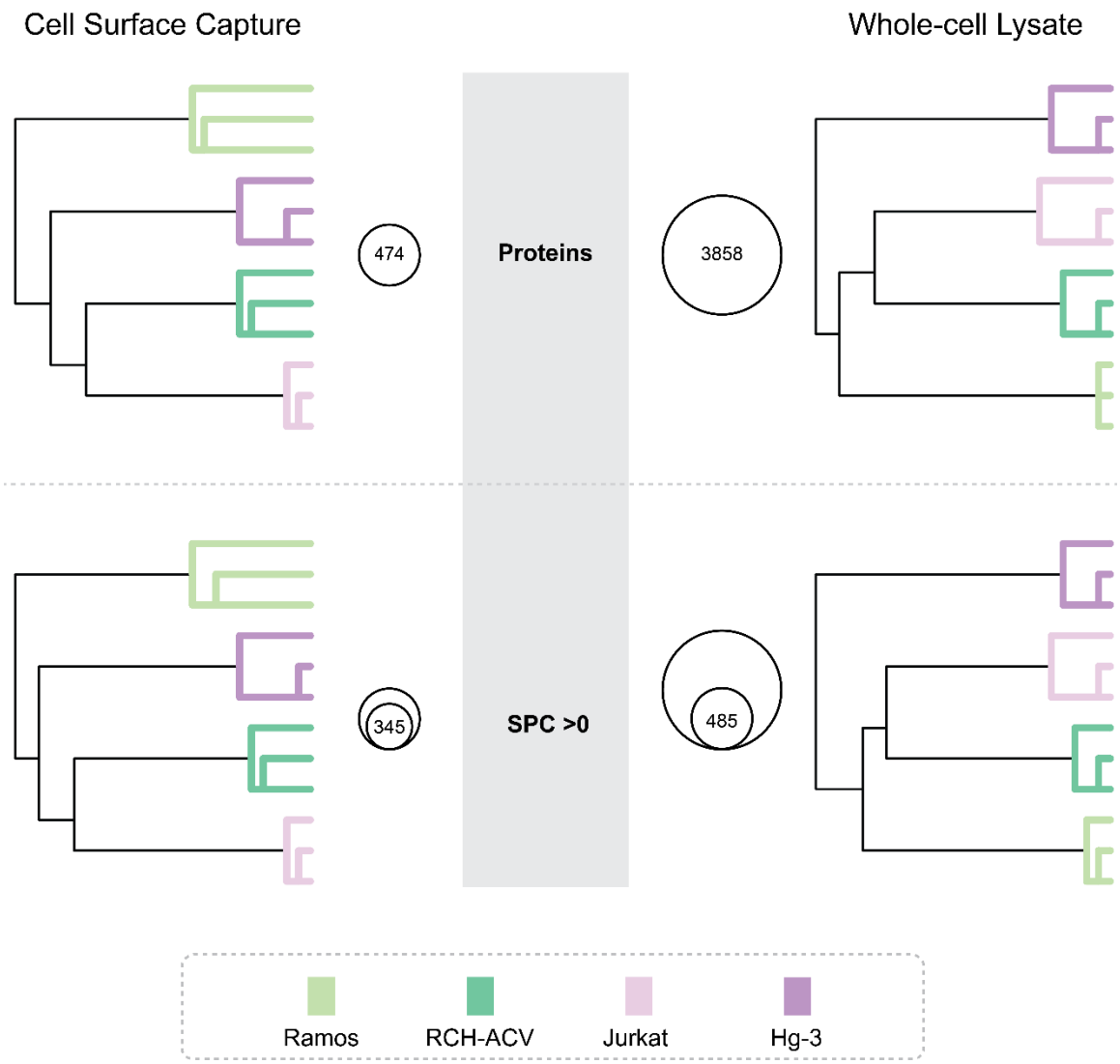


Fig. S2. Hierarchical clustering of lymphocyte Cell Surface Capture and Whole-cell Lysate data. Dendrograms depicting the relationships inferred by hierarchical clustering. All three biological replicates cluster for each of the four lymphocyte cells lines whether using all identified proteins or the subset of proteins predicted to be surface-localized by *SPC scores*.

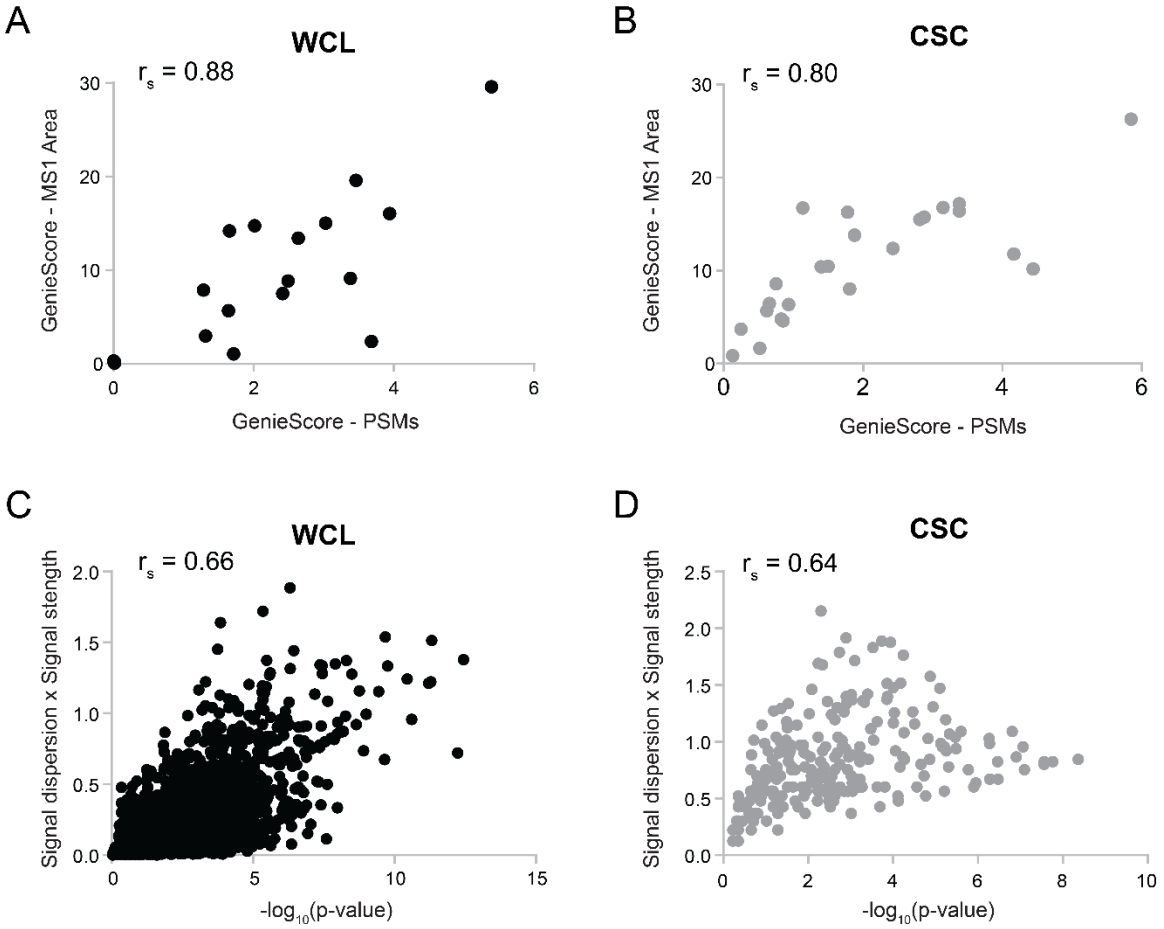


Fig. S3. Correlations of *GenieScores* and GenieScore components. *GenieScores* calculated MS1-based peak area plotted against *GenieScores* for the same proteins calculated using peptide-spectrum matches shown with calculated Spearman's correlation for whole-cell lysate (WCL) and Cell Surface Capture (CSC) data. The product of signal distribution and signal strength plotted against the statistical significance calculated using a one-way ANOVA shown with calculated Spearman's correlation for CSC and WCL data.

Table S1. Mass spectrometry acquisition settings

	Whole-cell Lysate	Cell Surface Capture
Injection Mode	Full Loop	uL PickUp
Sample Loop	20 µL	20 µL
Stationary Phase	Acclaim PepMap C 18 100 Å, 75 µm, 2 µm, 25 cm	Michrom Bioresources Magic C18AQ 200 Å, 3 µm, 10 cm
LC Solvent A	100% H_2O, 0.1% formic acid	100% H_2O, 0.1% formic acid
LC Solvent B	80% MeCN, 0.1% formic acid	80% MeCN, 0.1% formic acid
LC Gradient	7-7% B in 5 min 7-28% B in 123 min 28-40% B in 25 min 40-98% in 3 min	2-2% B in 10 min 2-35% B in 40 min 35-98% B in 10 min
LC Flow Rate	300 nL/min	300 nL/min
Mass Spectrometer	Thermo Orbitrap Q Exactive	Thermo Orbitrap Q Exactive
Method Type	Data dependent MS2, Top15	Data dependent MS2, Top15
Spray Voltage	2 kV	3.8 kV
MS^1 Detector	Orbitrap	Orbitrap
MS^1 scan range	350-1600 m/z	300-1600 m/z
MS^1 resolution	70,000 @ 200 m/z	70,000 @ 200 *m/z*
MS^1 AGC Target	1e6	1e6
MS^1 Maximum IT	50 ms	50 ms
MS^2 Detector	Orbitrap	Orbitrap
MS^2 resolution	17,500 @ 200 m/z	17,500 @ 200 m/z
Isolation Window	2.0 m/z	2.0 m/z
MS^2 AGC Target	5e4	1e5
MS^2 Maximum IT	110 ms	110 ms
Activation Type / Collision Energy	HCD 27%	HCD 27%
Minimum AGC Target.	5.0e2	1.0e3
Intensity Threshold	4.5e3	9.1e3
Dynamic Exclusion	30 s	60 s

Table S2. Peptide search and post-search validation parameters

Sample	**Whole-cell lysate**	**Cell Surface Capture (Hi-Hi)**	**Cell Surface Capture (Hi-Lo)**
Platform	ProteomeDiscoverer 2.2	ProteomeDiscoverer 2.2	ProteomeDiscoverer 2.2
Search Algorithm	SequestHT	SequestHT	SequestHT
Validation	Percolator Peptide Validator Protein FDR Validator	Percolator Peptide Validator Protein FDR Validator	Percolator Peptide Validator Protein FDR Validator
Database	SwissProt; Human; created 6/7/2017	SwissProt; Human; created 6/7/2017 or SwissProt; Mouse; created 11/30/2018	SwissProt; Human; created 6/7/2017
Enzyme (semi/full)	Trypsin (full)	Trypsin (semi)	Trypsin (semi)
Missed Cleavages	2	2	2
Precursor mass tolerance	10 ppm	10 ppm	10 ppm
Fragment mass tolerance	0.02 Da	0.02 Da	0.6 Da
Static Modifications	Carbamidomethyl (C)	Carbamidomethyl (C)	Carbamidomethyl (C)
Dynamic Modifications	Oxidation (M), Acetylation (N-term)	Oxidation (M), Acetylation (N-term) Deamidation (N)	Oxidation (M), Acetylation (N-term) Deamidation (N)
Target FDR (Strict):	0.01	0.01	0.01
Target FDR (Relaxed):	0.05	0.05	0.05
Validation basis	q-Value	q-Value	q-Value

Additional dataset S1 (separate file – descriptions of separate tabs are below)

(1) Human SPC dataset
(2) Mouse SPC dataset
(3) Rat SPC dataset

Additional dataset S2 (separate file – descriptions of separate tabs are below)

(1) Lymphocyte WCL data with *GenieScores*
(2) Lymphocyte CSC data with *GenieScores*
(3) *GenieScores* for proteins common to CSC and WCL

Additional dataset S3 (separate file – descriptions of separate tabs are below)

(1) CSC data on MCF10A $KRAS^{G12V}$ and empty vector controls with *GenieScores*
(2) RNA-Seq data on MCF10A $KRAS^{G12V}$ and empty vector controls with *GenieScores*

Additional dataset S4 (separate file – descriptions of separate tabs are below)

(1) Human dermal fibroblast and stem cell CSC data with *GenieScores*

Additional dataset S5 (separate file – descriptions of separate tabs are below)

(1) CSC data on mouse α and β cells with *GenieScores*
(2) RNA-Seq data on mouse α and β cells with *GenieScores*

Additional dataset S6 (separate file – descriptions of separate tabs are below)

(1) Single-cell RNA-Seq on human islet cells with *GenieScores* and modified *GenieScores*

References

1. Schilling B, *et al.* (2012) Platform-independent and label-free quantitation of proteomic data using MS1 extracted ion chromatograms in skyline: application to protein acetylation and phosphorylation. *Molecular & cellular proteomics : MCP* 11(5):202-214.
2. Haverland NA, *et al.* (2017) Cell Surface Proteomics of N-Linked Glycoproteins for Typing of Human Lymphocytes. *Proteomics* 17(19).
3. Martinko AJ, *et al.* (2018) Targeting RAS-driven human cancer cells with antibodies to upregulated and essential cell-surface proteins. *Elife* 7.
4. 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.
5. 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.
6. Lawlor N, *et al.* (2017) Single-cell transcriptomes identify human islet cell signatures and reveal cell-type-specific expression changes in type 2 diabetes. *Genome Res* 27(2):208-222.
7. Bausch-Fluck D, *et al.* (2015) A mass spectrometric-derived cell surface protein atlas. *PloS one* 10(3):e0121314.
8. Christoforou A, *et al.* (2016) A draft map of the mouse pluripotent stem cell spatial proteome. *Nat Commun* 7:8992.