The in silico human surfaceome

Damaris Bausch-Fluck[a,b,1], Ulrich Goldmann[a,1,2], Sebastian Müller[a,3], Marc van Oostrum[a,b], Maik Müller[a,b], Olga T. Schubert[a,4], and Bernd Wollscheid[a,b,5]

[a]Institute of Molecular Systems Biology at the Department of Biology, ETH Zurich, 8093 Zurich, Switzerland; and [b]Biomedical Proteomics Platform, Department of Health Sciences and Technology, ETH Zurich, 8093 Zurich, Switzerland

Edited by James A. Wells, University of California, San Francisco, CA, and approved October 2, 2018 (received for review May 24, 2018)

Cell-surface proteins are of great biomedical importance, as demonstrated by the fact that 66% of approved human drugs listed in the DrugBank database target a cell-surface protein. Despite this biomedical relevance, there has been no comprehensive assessment of the human surfaceome, and only a fraction of the predicted 5,000 human transmembrane proteins have been shown to be located at the plasma membrane. To enable analysis of the human surfaceome, we developed the surfaceome predictor SURFY, based on machine learning. As a training set, we used experimentally verified high-confidence cell-surface proteins from the Cell Surface Protein Atlas (CSPA) and trained a random forest classifier on 131 features per protein and, specifically, per topological domain. SURFY was used to predict a human surfaceome of 2,886 proteins with an accuracy of 93.5%, which shows excellent overlap with known cell-surface protein classes (i.e., receptors). In deposited mRNA data, we found that between 543 and 1,100 surfaceome genes were expressed in cancer cell lines and maximally 1,700 surfaceome genes were expressed in embryonic stem cells and derivative lines. Thus, the surfaceome diversity depends on cell type and appears to be more dynamic than the nonsurface proteome. To make the predicted surfaceome readily accessible to the research community, we provide visualization tools for intuitive interrogation (wlab.ethz.ch/surfaceome). The in silico surfaceome enables the filtering of data generated by multiomics screens and supports the elucidation of the surfaceome nanoscale organization.

surfaceome | SURFY | machine learning | cell surface protein | multiomics

The cell surface is the gateway that regulates information transfer from and to the outside world. Proteins at the cell surface connect intracellular and extracellular signaling networks and largely determine a cell's capacity to communicate and interact with its environment. The entirety of all possible cell-surface proteins, the surface proteome or surfaceome, consists of receptors, transporters, channels, cell-adhesion proteins, and enzymes and is a source of potential diagnostic biomarkers of disease and therapeutic targets (1). We define the surfaceome as all plasma membrane proteins that have at least one amino acid residue exposed to the extracellular space. As such, the surfaceome is a subset of the plasma membrane proteome, which is a subset of the membrane proteome, the entirety of all membrane proteins. Integral monotopic membrane proteins that are attached to the extracellular lipid leaflet [e.g., via a glycosylphosphatidylinositol (GPI) anchor] are part of the human surfaceome, but most of the surfaceome consists of α-helical transmembrane (TM) proteins (Fig. 1*A*). The bioinformatic differentiation between proteins residing in intracellular membranes (i.e., Golgi or endoplasmic reticulum), in the plasma membrane, and on the cell surface is not straightforward, and current classifications are mainly based on experimental or functional evidence (Fig. 1*B*). The lack of an accurate and comprehensive classification of all existing cell-surface proteins impedes their measurement and thereby negatively impacts research to better understand their role in biological processes and their clinical potential.

Predicting that a protein resides at the cell surface requires (*i*) the detection of a TM domain or a lipid anchor; (*ii*) the definition of the orientation of a protein within the membrane, including the identification of an extracellular exposed domain; and (*iii*) subcellular location prediction. (*i*) Bioinformatic tools for predicting TM domains, signal peptides, and GPI-linked proteins are available (2–9). (*ii*) Prediction of the correct orientation of the protein within the membrane is computationally more challenging, and experimental evidence for TM topologies is scarce, as are atomic-resolution structures of human TM proteins. The Membrane Protein Data Bank (10) lists only 124 human membrane protein structures; this is only 2% of proteins for which structures are known (1). (*iii*) Subcellular localization of a protein can to some extent be predicted by computational methods (11). Although multiple tools have been published for generalized subcellular localization prediction (12, 13), only two of these are specific for prediction of the subcellular localization of membrane proteins (14, 15). Both apply machine-learning algorithms, which rely on large training sets. Because of limited experimental data, the training sets used are based on annotations from the Gene Ontology Consortium, which are themselves frequently inferred or predicted (16). Previous attempts to compile a human surfaceome resource (17–19) integrated existing annotation and prediction of

Significance

Despite the fundamental importance of the surfaceome as a signaling gateway to the cellular microenvironment, it remains difficult to determine which proteoforms reside in the plasma membrane and how they interact to enable context-dependent signaling functions. We applied a machine-learning approach utilizing domain-specific features to develop the accurate surfaceome predictor SURFY and used it to define the human in silico surfaceome of 2,886 proteins. The in silico surfaceome is a public resource which can be used to filter multiomics data to uncover cellular phenotypes and surfaceome markers. By our domain-specific feature machine-learning approach, we show indirectly that the environment (extracellular, cytoplasm, or vesicle) is reflected in the biochemical properties of protein domains reaching into that environment.

Author contributions: D.B.-F., U.G., and B.W. designed research; D.B.-F., U.G., S.M., M.v.O., M.M., O.T.S., and B.W. performed research; D.B.-F., U.G., and B.W. contributed new reagents/analytic tools; D.B.-F., U.G., S.M., M.v.O., M.M., O.T.S., and B.W. analyzed data; and D.B.-F., U.G., and B.W. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Data deposition: Data are available at wlab.ethz.ch/surfaceome.

[1]D.B.-F. and U.G. contributed equally to this work.

[2]Present address: CeMM Research Center for Molecular Medicine of the Austrian Academy of Sciences, A-1090 Vienna, Austria.

[3]Present address: Biognosys, 8952 Schlieren, Switzerland.

[4]Present address: Department of Human Genetics, University of California, Los Angeles, CA 90095.

[5]To whom correspondence should be addressed. Email: wbernd@ethz.ch.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1808790115/-/DCSupplemental.

Published online October 29, 2018.

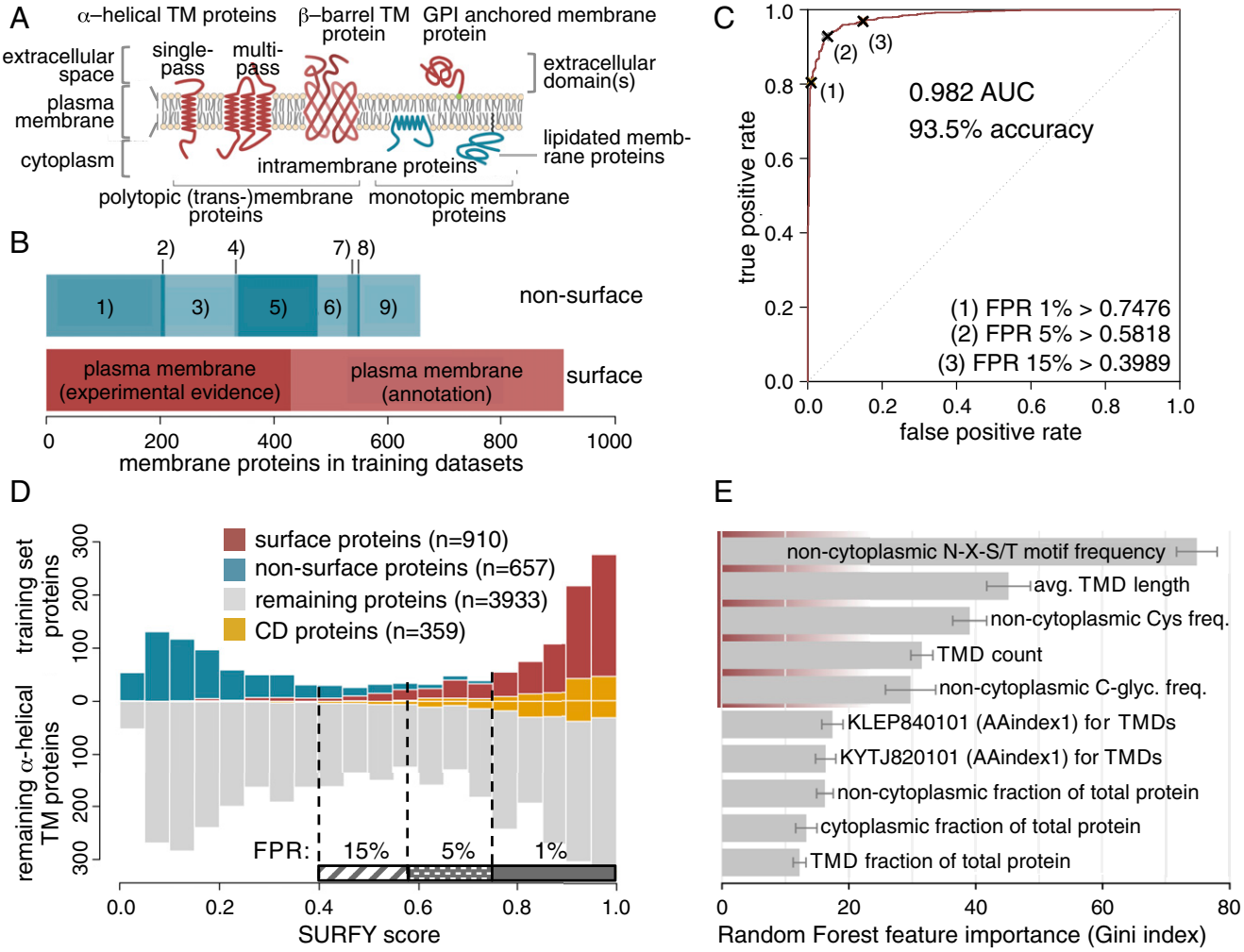

Fig. 1. Surfaceome definition and construction. (*A*) Visual representation of surfaceome definition. Proteins shown in red are regarded as surfaceome members; those in blue are not. (*B*) Compositions of nonsurface (negative) and surface (positive) training sets used for the machine-learning model. Subcellular location of the nonsurface training set are labeled as follows: 1, endoplasmic reticulum; 2, endosome; 3, Golgi apparatus; 4, lysosome; 5, mitochondrion; 6, nucleus; 7, peroxisome; 8, cytosol; and 9, multiple locations. (*C*) Receiver operating characteristics for the full model derived from out-of-bag error estimates (red line) (Dataset S1, 11.6). Gray line indicates the performance of random guessing. The three SURFY score cutoffs at 1%, 5%, and 15% FPRs are indicated. (*D*) Distribution of the predicted scores for the training sets (*Upper*) and for the remaining α-helical TM proteins (*Lower*). The bars of the nonsurface training set are stacked on top of the bars of the surface training set. Score cutoffs for estimated 1%, 5%, and 15% FPRs are indicated at the bottom. The predicted score distribution for CD antigens in and outside the training set are highlighted in yellow. (*E*) Gini index scores (41) for the 10 most important features used in building the predictive random forest model used by SURFY. Scores are plotted as means ± SDs. Features used for calculating SURFY scores are highlighted in red. AUC, area under the curve; Avg., average; C-glyc., *C*-glycosylation; TMD, TM domain.

TM domains and signal peptides, but did not make use of the latest experimental techniques or computational modeling.

The experimental assessment of the surfaceome is complicated by the hydrophobic nature of TM domains and the low abundance of cell-surface proteins compared with intracellular proteins. To overcome these hurdles, we previously developed the Cell Surface Capture (CSC) technology (20) that allows large-scale and highly specific identification and quantification of glycosylated cell-surface-residing proteins. The steps of the chemoproteomic CSC strategy are highly specific chemical labeling of cell-surface proteins on living cells, subsequent purification of *N*-glycosylated peptides, and mass-spectrometric identification of extracellularly exposed glycopeptides and their parent cell-surface proteins (21–33). By using this CSC strategy, an extensive experimental resource on cell-surface-residing and extracellularly exposed proteins was constructed. The Cell Surface Protein Atlas (CSPA; wlab.ethz.ch/cspa/) is a public resource which contains experimental evidence for cell-surface proteins identified in 41 human cell types (34).

Although the CSPA is an extensive biomedical resource, it only spans a limited number of cell types and cellular-activation stages and is therefore unlikely to encompass all possible human surfaceome members. To obtain a comprehensive picture of the human surfaceome, we set out to complement the experimental resource using computational prediction. Here, we present the machine-learning-based predictor SURFY, which leverages the CSPA as the necessary basis for a high-quality training dataset. Machine-learning algorithms enable pattern recognition, classification, and prediction based on models derived from existing data (35). Applied to the human proteome, SURFY predicted that the human surfaceome includes 2,886 proteins across all human cell types and developmental stages. The predicted set has excellent overlap with existing annotation of bona fide cell-surface proteins, with 99% of Cluster of Differentiation (CD) proteins included.

To understand the surfaceome variability between different cell types, gene-expression profiles from two independent datasets were analyzed according to our in silico surfaceome. From

these analyses, we identified a set of housekeeping surfaceome proteins. We also demonstrated that the surfaceome can be applied to narrow down large datasets for defining candidate sets for further validation, which ultimately leads to relevant therapeutic targets.

In summary, our surfaceome predictor SURFY enabled de novo prediction of surfaceome proteins. The in silico surfaceome is a unique public resource that, when combined with cell-type-specific expression data, can be used for phenotyping of cells guiding subsequent antibody development against prequalified targets and that will serve as a hypothesis generator for more targeted analyses by single-cell proteotype analysis by mass or flow cytometry or by genetically barcoded antibodies called phage-antibody next-generation sequencing (PhaNGS; ref. 36). The in silico surfaceome will allow unprecedented exploration of the surfaceome landscape within and across human cell-type populations.

Results

Training Sets and Protein Topology Assessment for Surfaceome Prediction. Since the milieu in the extracellular space and in the cytoplasm differ in pH, redox state, and interaction partners, we hypothesized that this should be reflected in different biochemical properties of extracellular, TM, and intracellular domains of membrane proteins, which could inform a machine-learning approach to discriminate between intracellular membrane and cell-surface-residing proteins. To calculate domain-specific features for the machine-learning algorithm, we first needed to define these domains. The TM topology is defined by a protein's TM domains, its orientation within the membrane, and locations of additional membrane attachment sites, if present (Fig. 1*A*). In a hierarchical approach, we combined topology information provided by the CSPA in the form of noncytoplasmic identification of *N*-glycopeptides for 1,309 proteins, by UniProt in the form of TM domain annotations for 4,471 proteins, and by prediction using Phobius (7) for 2,123 proteins. This yielded complete topologies for 7,903 human proteins, of which there were 5,500 α-helical TM proteins, two β-barrel proteins, 12 proteins with intramembrane loops, and 2,389 proteins with a signal peptide but no TM domain. As only a small fraction of the proteins with just a signal peptide are expected to be membrane-anchored by some other means (e.g., lipidation), we focused our machine-learning approach on the α-helical TM proteins. The exact borders of all α-helical TM regions were reassessed by using Phobius (7). The α-helical TM domains of our final topologies were on average 21.9 ± 2.3 residues (mean ± SD) in length.

Development of the Surfaceome Predictor SURFY. Machine-learning classification requires a numerical representation of the objects to be classified and negative and positive training sets. In contrast to previous approaches for subcellular localization prediction, we calculated most protein features individually for each topological domain (noncytoplasmic, cytoplasmic, signal peptide, and TM domain). We selected the following sequence-, annotation-, and prediction-based features to represent each protein: number, average length, and relative fraction of α-helical TM regions and signal peptides; frequencies of each of the 20 amino acids; four carefully selected physicochemical properties from AAindex1 (37); frequency of *N*-glycosylation consensus sequence motifs (*N*-X-S/T); predicted numbers of *C*-, *N*-, and *O*-linked glycosylation sites [GlycoMine (38)]; and 12 selected UniProt features annotating functional domains such as EGF-like domains or protein kinase domains. Details of the selected features from AAindex1 and UniProt are discussed in *Methods*. The final feature matrix consisted of 131 numerical features for each of the 7,903 proteins. The positive and negative training sets (910 and 657 proteins, respectively; Dataset S1, 11.1 and 11.2) were composed of data from the CSPA and other annotation resources (Fig. 1*B* and *SI Appendix*, Fig. S1) and were generated as described in *Methods*. The training sets were used to build a random forest model that showed a predictive accuracy of 93.5% (Fig. 1*C*).

Once trained, we used SURFY to calculate surface scores for the 5,500 α-helical TM proteins (Dataset S1, 11.3). The surface-score distribution showed a clear bimodal appearance, suggesting that the model successfully captured features that distinguish surface from nonsurface proteins (Fig. 1*D*). Feature importance was assessed by using the Gini index (39) and revealed five features to be most important (Fig. 1*E* and Dataset S1, 11.4 and 11.5). The frequency of noncytoplasmic *N*-X-S/T motifs was by far the most discriminative feature. Furthermore, the length and number of TM domains (Figs. 1*E* and 2 *A* and *B*), the frequency of noncytoplasmic cysteines, and the frequency of predicted noncytoplasmic *C*-glycosylation sites substantially influenced the surfaceome score calculated by SURFY.

Human Surfaceome Predicted by SURFY. To define the proteins belonging to the surfaceome, we compared datasets at different estimated false positive rates (FPRs) of 1%, 5%, and 15% (Fig. 1*C*); sensitivities of 96.9%, 92.7%, and 80.7% were obtained for the 2,242; 2,756; and 3,284 proteins, respectively. We choose the surfaceome with a 5% FPR as the representative human surfaceome (SURFY scores > 0.5818), as this dataset balanced sensitivity and specificity. In addition, 130 annotated GPI-linked human proteins (UniProt keyword: GPI-anchor) that do not have TM domains were included, resulting in a total surfaceome set of 2,886 human proteins (Dataset S1, 11.7). The nonsurfaceome set consisted of the 689 proteins from the negative training set and another 1,527 proteins with low SURFY scores (scores < 0.3989; 2,216 proteins in total). The 528 proteins with midrange scores (0.3989 < score < 0.5818) were considered borderline cases and were not included in further analyses. The surfaceome set contained 99% of all CD proteins (Fig. 1*D*), supporting the validity of the surfaceome prediction. Distributions of the SURFY scores assigned to other protein classes such as receptors and enzymes are plotted in *SI Appendix*, Fig. S2. The model was also applied to analyze the 2,403 proteins without α-helical TM domains, including the 130 proteins with GPI anchors (*SI Appendix*, Fig. S3), but predictions for these additional classes of proteins, most of them presumably secreted, should be considered speculative.

Characterization of the Predicted Human Surfaceome. To uncover the structural and biochemical differences in surface proteins and proteins that are localized to intracellular membranes such as the Golgi and endoplasmic reticulum, we compared the set of 2,886 surfaceome proteins to the set of 2,216 nonsurfaceome membrane proteins based on the five most discriminant SURFY score features (Fig. 1*E*). The frequency of *N*-glycosylation sequence motifs (*N*-X-S/T) in the noncytoplasmic regions of a TM protein was the most discriminative feature for the model. Nonsurface membrane proteins had an average of 0.42 such motifs per 100 noncytoplasmic amino acids, whereas surfaceome proteins showed a 3.5-fold enrichment to an average of 1.47 of these motifs per 100 extracellular amino acids (Fig. 2 *A*, *Left*). Only 137 surface proteins (4.7%) had no extracellular *N*-glycosylation motif at all, but 1,298 of the nonsurface membrane proteins (58.6%) lacked a noncytoplasmic *N*-glycosylation motif.

A clear enrichment of surface membrane proteins among all membrane proteins was found for proteins with seven TM domains, a class consisting of >95% G-protein-coupled receptors (GPCRs). In contrast, intracellular membrane proteins were enriched in double-pass TM domain proteins (Fig. 2*B*). The α-helical TM domains were slightly longer in surface proteins, with a median length of 22 amino acids, than were these domains in nonsurface proteins, which had a median length of 21 residues (Fig. 2*C*). This likely reflects differences in lipid bilayer composition.

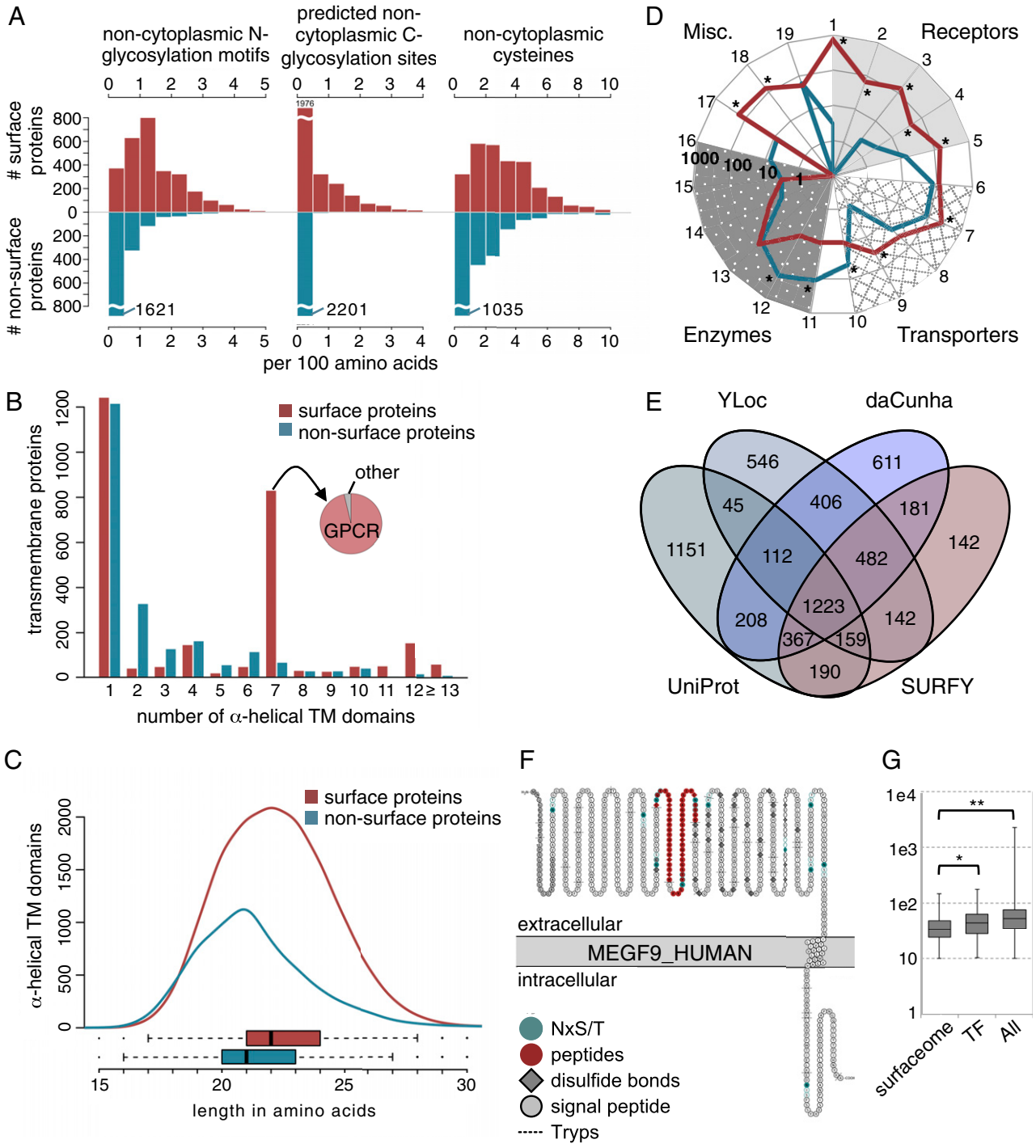

Fig. 2. Characterization of the predicted surfaceome. (*A*–*C*) Comparison of 2,886 surfaceome proteins in red, with 2,216 nonsurfaceome membrane proteins in blue. (*A*) Distributions of sequence features in noncytoplasmic domains of surfaceome (upper graphs) and nonsurfaceome membrane (lower graphs) proteins, calculated as frequency per 100 amino acids. *A*, *Left* shows the distribution of numbers of *N*-glycosylation sequence motifs (*N*-X-S/T) per 100 amino acids. Proteins with more than five motifs were excluded from this graph. *A*, *Center* shows the distribution of numbers of *C*-glycosylation sites per 100 amino acids predicted by using GlycoMine. Proteins with more than four predicted sites were excluded. *A*, *Right* shows the distribution of numbers of cysteine residues per 100 amino acids. Proteins with >10 cysteines were excluded. (*B*) Distribution of the number of α-helical TM domains per protein. The pie chart shows the proportion of GPCRs within the set of surface proteins with seven TM domains (red bar). (*C*) Distribution of the length of α-helical TM domains. (*D*) Classification of surface and nonsurface proteins into functional classes. $*P < 10^{-2}$. Functional classes are numbered as follows: 1, GPCRs; 2, receptor-type tyrosine kinases; 3, receptors of the Ig superfamily; 4, scavenger receptors; 5, other receptors; 6, channels; 7, solute carrier superfamily; 8, active transporters; 9, auxiliary transport proteins; 10, other transporters; 11, oxidoreductases; 12, transferases; 13, hydrolases; 14, lyases; 15, isomerases; 16, ligases; 17, structure/adhesion proteins; 18, ligand proteins; and 19, proteins of unknown function. (*E*) Overlap of proteins of the human surfaceome annotated in UniProt, predicted by YLoc (13), predicted by da Cunha et al. (17), and predicted by SURFY. (*F*) Protter image of human MEGF9. *N*-X-S/T motifs are marked in light blue, with the corresponding asparagine (N) in dark blue. CSC identified peptides are marked in purple. (*G*) Half-life distributions of surfaceome proteins, transcription factors (TF), and all quantified proteins from Mathieson et al. (45). Misc., miscellaneous.

C-linked mannosylation sites predicted by GlycoMine appeared almost exclusively in the noncytoplasmic regions of surfaceome proteins (Fig. 2*A*, *Center*). Surface proteins also showed a twofold enrichment in cysteine, with 3.38% cysteines in their extracellular domains compared with 1.72% cysteines in nonsurface proteins (Fig. 2 *A*, *Right*). The compositions of intracellular domains of surface proteins resembled those of the intracellular domains of nonsurface proteins with respect to the features analyzed, especially in the frequency of *N*-glycosylation sequence motifs, predicted *C*-glycosylation sites, and cysteine frequency. This observation could help increase performance of TM topology prediction tools in prediction of the orientation of plasma membrane proteins (40).

The functional classification showed that receptors and ligand proteins were assigned nearly exclusively by SURFY to the surfaceome set (Fig. 2*D* and Dataset S1, 11.7). This demonstrates the accuracy of the SURFY-predicted surfaceome. Other functional classes are evenly distributed among different cellular membrane structures (e.g., transporters). Some are more prominent among intracellular membrane proteins (e.g., enzymes).

Characteristics of Previously Unclassified Surface Proteins. The predicted human surfaceome showed good overlap with current subcellular localization annotation resources [UniProt keyword "cell membrane," YLoc subcellular localization prediction (13), and the bioinformatic surfaceome of da Cunha et al. (17)], with 1,223 proteins appearing in all four surfaceome lists (Fig. 2*E*). Proteins with a cell membrane annotation in UniProt, but not present in our surfaceome (1,480 proteins), are mostly proteins that are attached or integrated into the plasma membrane from the intracellular side that do not have an extracellularly exposed amino acid. By our definition, these proteins do not belong to the surfaceome and contain, for example, Ras/Rho-related proteins or phosphatidylinositol phosphate kinases. Conversely, a large majority of the surfaceome proteins identified by SURFY that have no cell membrane annotation are simply annotated in UniProt as "membrane" without further specification. YLoc is a machine-learning approach for the prediction of subcellular location (13) that does not consider features in a topological domain-specific manner, which we found was critical. The bioinformatic surfaceome of da Cunha et al. (17) solely relies on Gene Ontology annotations and shows limited accuracy for cell-surface proteins.

Among the 142 proteins exclusively present in our surfaceome, 49 proteins belong to the family of HLA class I histocompatibility antigens, and 26 proteins were identified in the CSPA. For 67 proteins predicted by SURFY to be cell-surface proteins, there is no evidence for cell-surface localization in UniProt, YLoc, or the da Cunha surfaceome. As a proof-of-concept study, we validated proteins, for which antibodies were available, by immunofluorescence, flow cytometry (*SI Appendix*, Fig. S4), and also a deep-coverage CSC experiment (Fig. 2*F* and Dataset S1, 11.8). Most of these proteins are not well characterized; however, we found some additional proteins, to be present in the Cell Atlas of the Human Protein Atlas data and in recent publications, confirming cell-surface localization (41–43) (Dataset S1, 11.9). These lines of evidence underscore the value of our computational approach to identify previously uncharacterized cell-surface proteins.

Surfaceome Proteostasis. Protein stability varies widely among proteins in a cell, and a global assessment of gene-expression control revealed that cell-surface proteins have rather stable mRNAs but high protein turnover rates (44). In a recently published dataset with protein turnover rates for surfaceome proteins (45), we found half-lives for ∼300 surfaceome members. A comparison of the half-lives of surfaceome proteins to those of transcription factors or to all proteins measured revealed significantly lower half-lives for the surfaceome proteins (Fig. 2*G*). The set of 300 surface proteins for which data were available represents only 10% of the surfaceome, but their higher turnover rates are in line with the findings of Schwanhäusser et al. (44) and support the hypothesis that high surfaceome turnover rates reflect the cellular ability to react quickly to extracellular stimuli.

Cell-Surface Phenotyping Using Transcriptome Data. The availability of our predicted surfaceome set enables analyses of surfaceomes across cell types and states. Even though quantitative proteomic data can provide information about the actual abundance of a protein of interest in a specific subcellular location, the high dynamic range of protein abundances typically hinders the identification of low-abundance proteins (46). Transcriptome data are more likely to include genes expressed at low levels due to the PCR amplification step and could therefore provide a blueprint of surfaceome candidates present in a specific cell type that could then be investigated further by more sensitive targeted proteomics methods such as selected reaction monitoring/parallel reaction monitoring (PRM) (47) and/or untargeted data-independent acquisition (DIA) of all theoretical fragment ion spectra mass spectrometry analyses (48) or biochemical methods.

To assess the overall variability of surfaceome across different cell types, we matched the surfaceome with transcriptome data from 610 cancer cell lines (49). We found that 2,331 surfaceome genes (of 2,704 matched IDs) were expressed in at least one cell line. Of the 373 surfaceome genes (13%) that were not expressed in any of these cell lines, the majority (297) encode olfactory receptors. The number of surfaceome genes expressed in individual cell lines ranged from 543 to 1,100 (Fig. 3*A*). The number of CD genes expressed varied from 74 to 196. Interestingly, lymphoid cells had a substantially less diverse surfaceome expression profile than cells from other tissues (Fig. 3 *A* and *C*). Lung and neuronal cells were among the cell types with the most diverse surfaceome profiles. The size of the surfaceome was correlated with the sum of expression levels of all surfaceome genes (Fig. 3*A*), even though the sum of all expressed genes was rather constant across cell lines (*SI Appendix*, Fig. S5). There was not a significant correlation between the size of the surfaceome and the physical size of the cell (Fig. 3*D*). Notably, the distribution of expression levels of surfaceome genes per cell line was relatively constant (Fig. 3*B*). This means that, independent of the size of the surfaceome (i.e., the number of different surfaceome genes expressed) (Fig. 3*A*), the median amount of mRNA molecules encoding for surfaceome genes stayed constant (Fig. 3*B*). It should be noted that gene expression numbers do not necessarily translate into quantity of protein at the cell surface due to coexisting intracellular protein pools. Nevertheless, single genes can vary substantially in their gene expression level from cell line to cell line, as exemplified by *PD-L1* (Fig. 3*E*).

To investigate how specific the expression of cell-surface proteins is to individual cell types, we counted for each surfaceome member in how many different cell lines it was expressed. The occurrence of expressed surfaceome genes over the 610 cell lines showed a sigmoidal shape, with many genes only expressed in one cell line (23%; 500 genes) and only 10% of genes expressed in every cell line (240 genes) (Fig. 3*F* and Dataset S1, 11.10). This distribution pattern shows that the surfaceome is highly cell-line-specific, even compared with other gene sets that are thought to be cell-type specific. For example, 44% of transcription factors are expressed in each of the 610 cell lines (*SI Appendix*, Fig. S6). To investigate functional differences between ubiquitously expressed and cell-type-specific expressed genes, we grouped them based on frequency of expression across the cell lines evaluated (Fig. 3*F* and Dataset S1, 11.10). Group 1 has relatively low median gene expression levels (0.77 log2 RPKME; *SI Appendix*, Fig. S7), but, if proteins are detectable at the cell surface, these proteins could serve as cellular markers. Group 5 genes are expressed in most cell lines (Fig. 3*F*) and are generally more strongly expressed (4.14 log2 RPKME; *SI Appendix*,

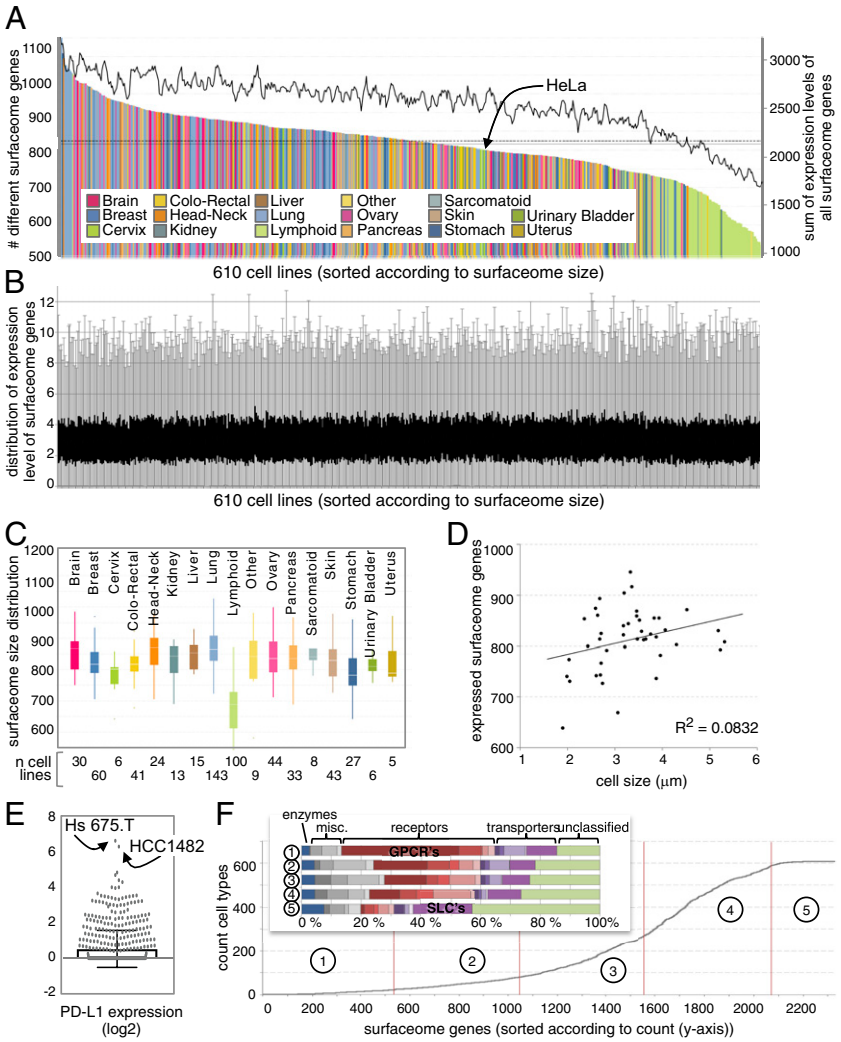

Fig. 3. Surfaceome expression in 610 cancer cell lines. (*A*) Distribution of cell-specific surfaceome diversity (count of expressed surfaceome genes; left axis), sorted from large to small. Cell lines are colored based on their tissue type, as indicated by the color code. The straight gray line marks the average, and the dashed line marks the median surfaceome diversity. The sum of the expressed surfaceome genes is indicated by the black line corresponding to the right axis. (*B*) Distribution of surfaceome diversities (count of expressed surfaceome genes) based on tissue type. Tissues are color-coded as in *A*. The number of cell lines belonging to each tissue is indicated on the horizontal axis. (*C*) Scatter plot of count of expressed surfaceome genes vs. physical cell size. Squared Pearson correlation coefficient is indicated. (*D*) Box plots of the surfaceome gene expression level distribution for each cell line, sorted based on surfaceome diversity as in *A* from large to small. The black range represents the interquartile range; whiskers are depicted in gray. (*E*) Distribution of log2 expression level of *PD-L1*. Cell lines with the highest expression are indicated. (*F*) Surfaceome genes sorted by number of cell lines in which each gene is expressed enabled categorization into five groups. Functional classification for each group of genes based on Almén et al. (1) is shown in the bar chart in *F*, *Inset*. Misc., miscellaneous.

Fig. S7). We refer to this group of genes as surfaceome "housekeeping" genes (Dataset S1, 11.10). Functional annotation revealed many receptors, especially GPCRs, in group 1, whereas in group 5, many proteins are unclassified. Surfaceome genes between groups 1 and 5 were evenly divided in three groups (groups 2, 3, and 4; Fig. 3*F*), but did not differ drastically in terms of their functional annotations. In summary, we show that integrating the surfaceome set with complementary data results in cell-type-specific surfaceome landscapes and allows for the identification of a small testing set of candidates for cell-line-/type-/state-specific cell-surface markers. However, selected candidates based on mRNA expression levels still need to be validated on the protein level, since correlation between mRNA expression and protein expression can be rather poor, as discussed later.

Surfaceome Genes Allow for Cell-Line Classification. To test the hypothesis that the surfaceome contains information about cellular identity, we performed a principal component analysis with the expressed surfaceome genes of the 610 cell lines (*SI Appendix*, Fig. S8). This analysis revealed that the lymphoid cells (light green) and the lung cells (light blue) are clearly separated from the other tissues. The remaining tissue types are less clearly separated from each other; however, many of them group together [i.e., skin cells (light brown) and head–neck cells (orange)]. To more specifically compare and visualize the functional differences of cell-type-specific surfaceomes, we assembled them in a Voronoi tree map. Proteins were hierarchically grouped by functional classification and colored according to their expression levels. This allowed identification of cell-specific gene modules, as demonstrated in Fig. 4. In the Voronoi tree map of the B-cell line RAMOS, the concerted expression of MHC class I proteins was immediately detectable (Fig. 4*A*), suggesting the identity of a B-cell line without consultation of orthogonal data. In the colorectal carcinoma cell line HT-29, the cluster of expressed Integrins was apparent (Fig. 4*B*), which correlates well with the adhesive behavior of these cells (50). In the neuroblastoma

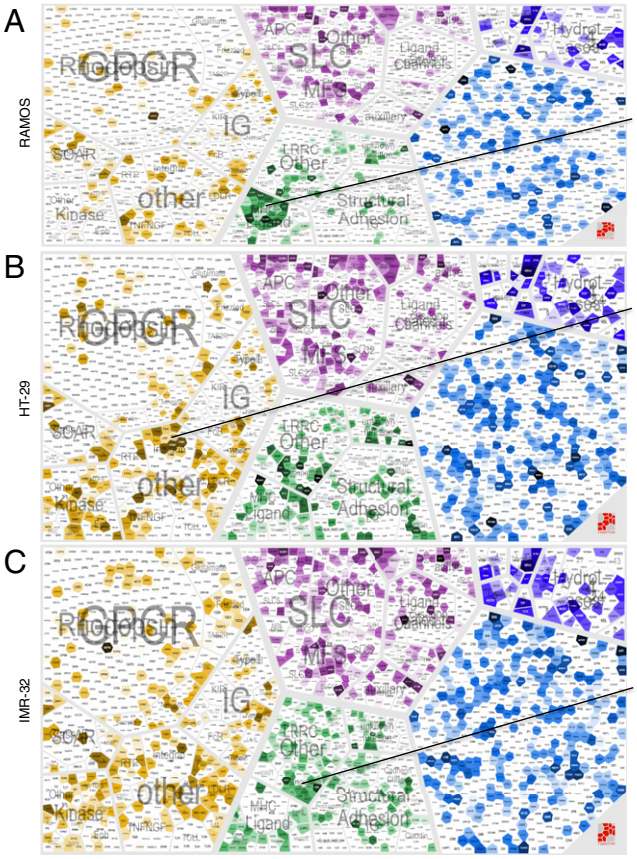

Fig. 4. Voronoi tree maps generated on wlab.ethz.ch/surfaceome. Maps for RAMOS (*A*), HT-29 (*B*), and IMR-32 (*C*) are shown. RPKME values of each cell line were scaled from 0 to 1 and mapped onto the whole in silico surfaceome. Light color indicates low expression; dark color indicates strong expression. White genes are not expressed. Characteristic functional protein groups of these cell lines are highlighted on the right.

cell line IMR-32, the majority of tetraspanin was expressed, which has been proposed to influence cancer invasion and metastasis (51).

Our online tool allows creation of Voronoi tree maps for user-defined surfaceome interrogation (wlab.ethz.ch/surfaceome). Input data are mapped onto the functionally annotated in silico surfaceome for immediate visualization of coregulated proteins.

Surfaceome Adaptation During Neurogenesis. The quest to find accurate cell-surface markers for the identification and classification of stem and progenitor cells at various differentiation stages remains a high priority in the field of stem-cell research. To determine whether surfaceome analysis could be used to aid in classification, we analyzed surfaceome gene expression in a transcriptional dataset of human embryonic stem cells (hESCs) and neural progenitors at various stages of differentiation (52). The number of surfaceome genes expressed varied from 1,610 to 1,700 (Fig. 5*A*). Interestingly, this was substantially higher (by ∼50%) than in any of the 610 cancer cell line transcriptomes analyzed. This finding is in line with a previous study showing that cancer cells harbor a reduced protein repertoire compared with primary cells (53).

We hypothesized that the onset of neurogenesis should be reflected in the surfaceome. The number of different surfaceome genes expressed remained relatively constant over time, but the amount of total surfaceome-coding mRNA synthesized increased between day 12 and 16, which possibly reflects a boost in proliferation (Fig. 5*A*). Li et al. termed this stage the neural progenitor cell (NPC) proliferation state (52). Furthermore, we soft-clustered the expression profiles of surfaceome genes and found five clusters (Fig. 5*C* and Dataset S1, 11.11). Functional annotation of the clustered genes revealed that cluster 1 contains an enrichment in genes coding for axon guidance and synapse formation. Cluster 1 surfaceome genes are most strongly expressed on day 0, which suggests that already hESCs are expressing genes for neurogenesis. Cluster 2 harbors gene-expression profiles for proteins involved in migration. An increased number of genes involved in neurotransmitter channel (namely, the GABA receptors) and ion transport was found in cluster 3, and in cluster 4, expression profiles of genes encoding for proteins involved in neurogenesis and neural migration were gathered. Surfaceome genes grouped in cluster 5 showed an enrichment in channels, transporter, and adhesion functionality, possibly reflecting the outgrowth of axons and increased attachment to the extracellular matrix (Fig. 5*C*).

To compare the surfaceome of undifferentiated cells (day 0) to a surfaceome of a progeny (day 22), we created a Voronoi tree map of expression differences (Fig. 5*D*). Since gene entries in the Voronoi tree map are grouped based on their assigned functions, strong up- or down-regulation of functional groups are revealed and suggested to be coregulated. For example, several gap junction proteins (GJA1, GJB2, and GJB3), known to be major mediators of cell–cell communication during embryogenesis (54), were strongly expressed at day 0 and decreased with increased differentiation (Fig. 5 *B* and *D*).

The in silico surfaceome advances the quest for stem cell markers, since it identifies protein candidates that are directly accessible from the extracellular space and could be used as stem-cell-specific markers. In addition, surfaceome genes that display similar expression profiles could occur in a functionally relevant signaling synapse and directly reveal cell-surface interactions (49). Such nanoscale information about the surfaceome will allow informed development of multivalent affinity binders for highly specific targeting and enrichment of stem cells and their derivatives.

Discussion

SURFY Accurately Identifies Cell-Surface Proteins. By integrating annotation, experimental evidence, and computational methods, we generated a surfaceome-specific protein classifier termed SURFY. SURFY, with an accuracy of 93.5% based on a random forest model analysis of training sets, clearly outperformed other location prediction tools (YLoc: 62.2%; MemLoci: 77.8%) (13, 15) as well as TMHMM (3) and SignalP (55) used in combination with subcellular localization databases (19, 56). SURFY is an approach for subcellular localization prediction based on a number of biologically relevant features and consideration of TM topologies. Within the human proteome of 20,193 proteins (UniProt version 2015_01), SURFY classified 2,756 α-helical TM proteins as located and exposed at the cell surface. The addition of 130 currently annotated GPI-anchored proteins resulted in a 2,886-protein surfaceome, which corresponds to 14.3% of the entire human proteome.

Distinguishing Features of Cell-Surface Proteins. The predictive accuracy of SURFY is achieved by the combination of five discriminative features. The individual sequence- and topology-based features alone are insufficient for proper classification of surface proteins. Nevertheless, these five features harbor notable biological relevance for cell-surface proteins. The first distinguishing feature is the *N*-X-S/T sequence motif. *N*-glycosylation within this motif plays an important role in functions of many surface proteins (57), but not every sequence motif is *N*-glycosylated (58). We show that this *N*-glycosylation sequence motif is 3.5-fold enriched in extracellular domain sequences compared with sequences of intracellular proteins or intracellular domains of surface proteins. In

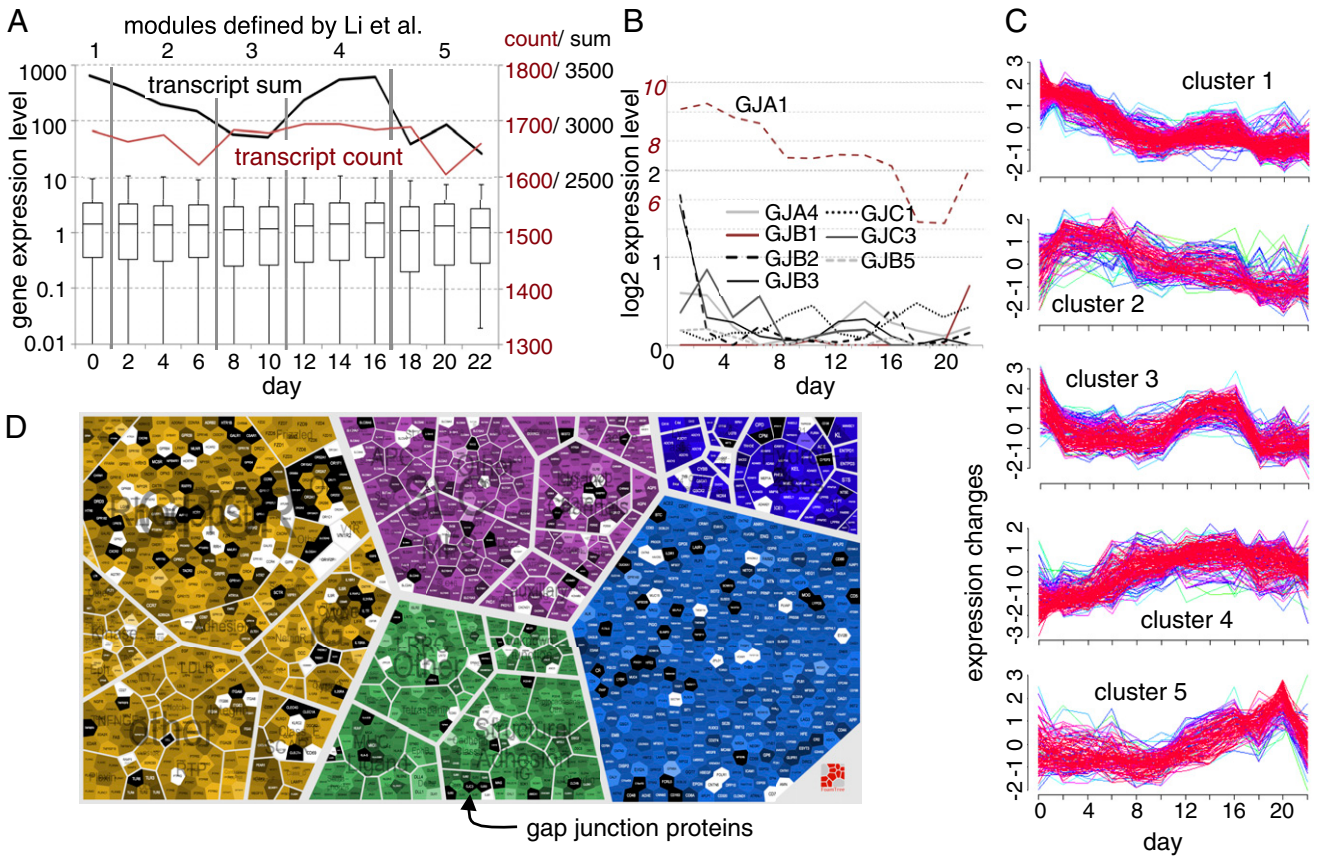

Fig. 5. Surfaceome changes during neurogenesis. (*A*) *Left* axis: Surfaceome gene level distribution from day 0 to day 22. *Right* axis: The red line shows the total number of expressed surfaceome genes, and the brown line shows the sum of expression levels over all expressed surfaceome genes. Transcriptomic data and definition of developmental stages (1, pluripotency stage; 2, differentiation initiation stage; 3, neural commitment stage; 4, NPC proliferation stage; 5, neuronal differentiation stage) were obtained from Li et al. (52). (*B*) Expression of selected gap junction genes from day 0 to day 22. (*C*) Identified clusters among surfaceome gene expression profiles based on c-means soft clustering. Red, higher correlation with cluster; light blue, lower correlation with cluster. (*D*) Voronoi tree map of log2 expression ratios between day 0 and day 22; the darker means more expressed at day 0, and the brighter color means more expressed at day 22. Surfaceome genes are hierarchically grouped by functional classification [receptors (orange), transporters (purple), hydrolases (dark blue), unclassified (blue), and miscellaneous (green)].

absolute terms, there are on average an additional 1.05 *N*-X-S/T motifs per 100 amino acids in extracellular domains than in intracellular proteins. Under the assumption that the 0.42 motifs per 100 amino acids in nonsurface proteins represents the frequency of unoccupied motifs and the increase by 1.05 motifs in the extracellular domains represent the frequency of occupied motifs, we can conclude an *N*-glycosylation site occupancy of 71%, which is in good agreement with a previous analysis on a smaller dataset, reporting 65% site occupancy (59).

The second distinguishing feature of surface-protein extracellular domains is the increased frequency of cysteine residues (3.4% compared with 2.3% proteome-wide). This presumably reflects the fact that disulfide cross-links are vital for folding surface proteins into their native conformation and for their structural stability (60).

The third feature is the presence of a possible *C*-glycosylation site. The biological function of *C*-glycosylation is unknown, and only 17 human proteins are annotated with this modification so far (UniProtKB version 2015_01). That 1,070 surface proteins and only 35 nonsurface TM proteins had at least one predicted noncytoplasmic *C*-glycosylation site clearly indicates an extracellular function of this modification or the underlying predicted motif.

The final two distinguishing features of surface proteins are length and number of TM domains. We found that α-helical TM domains of surface proteins are on average 1 amino acid longer than nonsurface proteins. It has been reported that α-helical TM domains of human plasma membrane proteins are on average ∼4 amino acids longer and also less bulky than proteins localized to the endoplasmic reticulum or Golgi (61). Sharpe et al. (61) restricted their analysis to single-pass TM proteins, where the differences in TM domain structure are expected to be more pronounced than those we observed. Cholesterol is known to increase the thickness of lipid bilayers and is the reason that the plasma membrane is thicker than endoplasmic reticulum or Golgi membranes. A difference in thickness is also observed between cholesterol-enriched domains (e.g., lipid rafts) and disordered regions within the plasma membrane (62). Our random forest approach could be expanded to explore potential correlations with TM domain size and lipid raft localization. The number of TM domains of surface proteins displays a clear bias toward 7- and >10-TM domains, reflecting cell-surface-related functions from GPCRs, as well as transporters and channels, maintaining cellular metabolism.

Proteins Not Included in the in Silico Surfaceome. The surfaceome predicted by SURFY shows all known cell-surface protein groups. It encompasses virtually the complete set of receptor proteins; the only exceptions are some small groups of intracellular receptor proteins. All but 3 of the 804 GPCRs, a class of proteins which is a major target for drugs (63), are found in

the in silico surfaceome. Inherent to every prediction is that a certain cutoff has to be selected as a tradeoff between sensitivity and specificity. The presented in silico surfaceome was defined with an estimated 5% FPR. To better understand the FPR, we took a closer look at the 33 proteins within the negative training set that have high SURFY scores, qualifying them as surface proteins and presumably false-positive hits. However, current versions of UniProt and COMPARTMENTS annotations indicate that some of them (namely, LRRT1, SE6L1, TM130, TMM59, and TPC2) are expressed at the cell surface, indicating that our FPR estimate of 5% is rather conservative. We also found that a few proteins known to localize to the cell-surface proteins scored below the threshold and are thus not predicted to be at the cell surface. An example is TFR1 (score 0.3272). TFR1 has a large extracellular domain and also *N*-glycosylation sites, but the frequency of *N*-glycosylation sites (0.69) and of cysteine (1.1) per 100 noncytoplasmic amino acids is low, and its TM domain is rather short (18 amino acids). The combination of these features resulted in a low SURFY score. The frequency of noncytoplasmic *N*-glycosylation sites is the most discriminant feature of the SURFY algorithm since the majority of surfaceome proteins are indeed glycosylated (57). Proteins with no noncytoplasmic *N*-X-S/T motif are not likely to score as surfaceome proteins (as was the case for CD3e); however, the in silico surfaceome contains 59 proteins without the *N*-X-S/T motif. Based on the different assessments with known protein groups (i.e., CD proteins and receptors) and the characteristics of the predicted surface proteins, we are confident that the surfaceome encompasses to the most complete and most correct list of cell-surface proteins reported so far.

The Expressed Surfaceome. By matching the surfaceome to the transcriptional profile of 610 cancer cell lines, we showed that 2,331 predicted cell-surface proteins are expressed in at least one cell line, and the majority of surfaceome genes that are not expressed in the cell lines analyzed encode olfactory receptors. Deep surfaceome profiling using CSC technology identified 507 surfaceome proteins on HeLa cells (Dataset S1, 11.8), whereas RNA-sequencing data indicated that 801 surfaceome genes are expressed (49). Surface proteins identified by CSC displayed generally a stronger expression at the mRNA level, indicating that surfaceome genes with low expression are below the current detection limit of mass-spectrometry-based methods (*SI Appendix*, Fig. S9). Surfaceome gene expression analysis therefore has the possibility to identify low-abundance cell-surface proteins to target them by more sensitive approaches (like PRM or DIA).

Care must be taken when translating gene expression levels into cell-surface protein abundance. Although it has been shown that this derivation is possible with a gene-specific conversion factor (64), it was also demonstrated that the mRNA–protein correlation is less accurate for cell surface than for cytoplasmic proteins (44). The comparison of abundance of surfaceome proteins from the quantitative proteome of Beck et al. (65) with gene expression data from Klijn et al. (49) showed only a low correlation (*SI Appendix*, Fig. S10). In addition, the total protein pool of a cell-surface protein can differ from its actual abundance at the cell surface. For example, the glucose transporter GTR4 is stored in vesicles and only transported to the cell surface after insulin stimulation (66). Based on mRNA expression levels from a large screen, SURFY allows for the selection of a small set of cell-specific precandidates, which are feasible to validate on the protein level.

Interestingly, certain surfaceome proteins are expressed on literally every cell line (231 proteins) or only on a single or very few cell lines (529 proteins) (Dataset S1, 11.10). We hypothesize that the group of constitutively expressed proteins are encoded by surfaceome housekeeping genes. These genes are expressed at substantially higher median expression over all cell lines than are those genes that encode surface proteins expressed by very few lines (*SI Appendix*, Fig. S7). If the housekeeping surfaceome proteins show stable abundance at the cell surface, they could serve as a basis for normalization in absolute quantitative surfaceome proteotype screens.

In summary, we have presented a comprehensive description of the in silico human surfaceome. This public resource of 2,886 proteins can be used to query proteotype and transcriptomic datasets for context-dependent biomarkers and drug-target candidates (19, 56). Furthermore, the in silico surfaceome will provide a guiding role in future studies assessing the plasticity and response of the cell-surface proteome to differentiation, perturbations, or other cellular processes. The recently published GeneGini coefficient to compare expression levels will be a helpful tool for such surfaceome studies (67). To interrogate the in silico surfaceome, we developed a publicly accessible online tool enabling the interactive visualization of quantitative surfaceomes based on a preconfigured Voronoi tree map. The surfaceome website allows for quantitative visual comparison of expressed surfaceomes on static background Voronoi tree maps and will provide insights into the systemic and dynamic response of a cell's surfaceome to perturbations or drugs (wlab.ethz.ch/surfaceome). The human in silico surfaceome will further provide the basis for dynamic interactome studies to decipher cell-surface-specific proximal interaction networks and cellular nanoscale organization, deepening our understanding of how cells sense and communicate with their microenvironment. The computational machine-learning strategy shown here could also be developed further to enable new insights into model organism surfaceomes.

Methods

Consolidating TM Topologies for the Human Membrane Proteome. The basis and reference for the presented human surface proteome analysis was the human proteome in UniProtKB/Swiss-Prot (Version 2015_01) (68). We first matched the 8,010 human peptides from the CSPA to the reference proteome, generating a list of constraints for the topology of a total of 1,387 proteins by enforcing identified *N*-glycosylation sites to be noncytoplasmic. Additional details can be found *SI Appendix, SI Methods*.

Defining Training Sets for Predictive Modeling. The positive training set was composed of human α-helical TM domain-containing proteins appearing in at least two of the following three datasets: (*i*) the "high confidence" subset of the CSPA containing 735 proteins, (*ii*) the UniProtKB/Swiss-Prot (Version 2015_01) containing 2,043 proteins attributed with the "cell membrane" keyword, and (*iii*) the subcellular localization database COMPARTMENTS (69) containing 826 high-confidence plasma membrane proteins (five stars), which belong to the COMPARTMENTS inherent "plasma membrane" positive benchmark set and also belong to the COMPARTMENTS inherent negative benchmark sets for each of the remaining subcellular locations (all but "extracellular space"). Additional details can be found in *SI Appendix, SI Methods*.

Derivation of a Numerical Feature Vector for Each Protein. Numerical features were defined as follows for each topological domain (cytoplasmic, noncytoplasmic, TM, and signal peptide, if applicable): amino acid frequencies; the number, average length, and relative fraction of α-helical TM regions; the presence of and average length of a signal peptide; and the relative fraction of cytoplasmic and noncytoplasmic regions. Additional details can be found in *SI Appendix, SI Methods*.

Supervised Learning and Predictive Modeling of Human Cell-Surface Proteins. We chose the random forest algorithm (70) as implemented in the "randomForest" package (71) of the statistical computing environment R (Version 3.1.0) (72) for supervised, binary classification. Additional details can be found *SI Appendix, SI Methods*.

ACKNOWLEDGMENTS. We thank the whole B.W. laboratory for suggestions and support at all stages of the project. This work was supported by Swiss National Science Foundation Grant 31003A_160259 (to B.W.), the Swiss Initiative in Systems Biology SystemsX.ch InfectX project (B.W.), and the Commission of Technology and Innovation (B.W.).

1. Almén MS, Nordström KJV, Fredriksson R, Schiöth HB (2009) Mapping the human membrane proteome: A majority of the human membrane proteins can be classified according to function and evolutionary origin. *BMC Biol* 7:50.
2. Reeb J, Kloppmann E, Bernhofer M, Rost B (2015) Evaluation of transmembrane helix predictions in 2014. *Proteins* 83:473–484.
3. Krogh A, Larsson B, von Heijne G, Sonnhammer ELL (2001) Predicting transmembrane protein topology with a hidden Markov model: Application to complete genomes. *J Mol Biol* 305:567–580.
4. Jones DT (2007) Improving the accuracy of transmembrane protein topology prediction using evolutionary information. *Bioinformatics* 23:538–544.
5. Viklund H, Elofsson A (2008) OCTOPUS: Improving topology prediction by two-track ANN-based preference scores and an extended topological grammar. *Bioinformatics* 24:1662–1668.
6. Nielsen H, Engelbrecht J, Brunak S, von Heijne G (1997) Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Eng* 10: 1–6.
7. Käll L, Krogh A, Sonnhammer ELL (2004) A combined transmembrane topology and signal peptide prediction method. *J Mol Biol* 338:1027–1036.
8. Eisenhaber B, Bork P, Eisenhaber F (1999) Prediction of potential GPI-modification sites in proprotein sequences. *J Mol Biol* 292:741–758.
9. Fankhauser N, Mäser P (2005) Identification of GPI anchor attachment signals by a Kohonen self-organizing map. *Bioinformatics* 21:1846–1852.
10. Raman P, Cherezov V, Caffrey M (2006) The membrane protein data bank. *Cell Mol Life Sci* 63:36–51.
11. Chou K-C, Shen H-B (2007) MemType-2L: A web server for predicting membrane proteins and their types by incorporating evolution information through Pse-PSSM. *Biochem Biophys Res Commun* 360:339–345.
12. Emanuelsson O, Nielsen H, Brunak S, von Heijne G (2000) Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *J Mol Biol* 300: 1005–1016.
13. Briesemeister S, Rahnenführer J, Kohlbacher O (2010) YLoc—An interpretable web server for predicting subcellular localization. *Nucleic Acids Res* 38(Suppl 2):W497–W502.
14. Du P, Tian Y, Yan Y (2012) Subcellular localization prediction for human internal and organelle membrane proteins with projected gene ontology scores. *J Theor Biol* 313: 61–67.
15. Pierleoni A, Martelli PL, Casadio R (2011) MemLoci: Predicting subcellular localization of membrane proteins in eukaryotes. *Bioinformatics* 27:1224–1230.
16. Rhee SY, Wood V, Dolinski K, Draghici S (2008) Use and misuse of the gene ontology annotations. *Nat Rev Genet* 9:509–515.
17. da Cunha JP, et al. (2009) Bioinformatics construction of the human cell surfaceome. *Proc Natl Acad Sci USA* 106:16752–16757.
18. Díaz-Ramos MC, Engel P, Bastos R (2011) Towards a comprehensive human cell-surface immunome database. *Immunol Lett* 134:183–187.
19. Town J, et al. (2016) Exploring the surfaceome of Ewing sarcoma identifies a new and unique therapeutic target. *Proc Natl Acad Sci USA* 113:3603–3608.
20. Wollscheid B, et al. (2009) Mass-spectrometric identification and relative quantification of N-linked cell surface glycoproteins. *Nat Biotechnol* 27:378–386.
21. Gundry RL, Boheler KR, Van Eyk JE, Wollscheid B (2008) A novel role for proteomics in the discovery of cell-surface markers on stem cells: Scratching the surface. *Proteomics Clin Appl* 2:892–903.
22. Schiess R, Wollscheid B, Aebersold R (2009) Targeted proteomic strategy for clinical biomarker discovery. *Mol Oncol* 3:33–44.
23. Gundry RL, et al. (2009) The mouse C2C12 myoblast cell surface N-linked glycoproteome: Identification, glycosite occupancy, and membrane orientation. *Mol Cell Proteomics* 8:2555–2569.
24. Hofmann A, et al. (2010) Proteomic cell surface phenotyping of differentiating acute myeloid leukemia cells. *Blood* 116:e26–e34.
25. Bock T, Bausch-Fluck D, Hofmann A, Wollscheid B (2012) CD proteome and beyond-technologies for targeting the immune cell surfaceome. *Front Biosci* 17:1599–1612.
26. Ziegler A, et al. (2012) Proteomic surfaceome analysis of mesothelioma. *Lung Cancer* 75:189–196.
27. Boysen G, et al. (2012) Identification and functional characterization of pVHL-dependent cell surface proteins in renal cell carcinoma. *Neoplasia* 14:535–546.
28. Cerciello F, et al. (2013) Identification of a seven glycopeptide signature for malignant pleural mesothelioma in human serum by selected reaction monitoring. *Clin Proteomics* 10:16.
29. Mirkowska P, et al. (2013) Leukemia surfaceome analysis reveals new disease-associated features. *Blood* 121:e149–e159.
30. Moest H, et al. (2013) Malfunctioning of adipocytes in obesity is linked to quantitative surfaceome changes. *Biochim Biophys Acta* 1831:1208–1216.
31. Hofmann A, Bausch-Fluck D, Wollscheid B (2013) CSC technology: Selective labeling of glycoproteins by mild oxidation to phenotype cells. *Methods Mol Biol* 951:33–43.
32. DeVeale B, et al. (2014) Surfaceome profiling reveals regulators of neural stem cell function. *Stem Cells* 32:258–268.
33. Kropp EM, et al. (2014) N-glycoprotein surfaceomes of four developmentally distinct mouse cell types. *Proteomics Clin Appl* 8:603–609.
34. Bausch-Fluck D, et al. (2015) A mass spectrometric-derived cell surface protein atlas. *PLoS One* 10:e0121314.
35. Tarca AL, Carey VJ, Chen X-W, Romero R, Drăghici S (2007) Machine learning and its applications to biology. *PLoS Comput Biol* 3:e116.
36. Pollock SB, et al. (2018) Highly multiplexed and quantitative cell-surface protein profiling using genetically barcoded antibodies. *Proc Natl Acad Sci USA* 115: 2836–2841.
37. Kawashima S, et al. (2008) AAindex: Amino acid index database, progress report 2008. *Nucleic Acids Res* 36:D202–D205.
38. Li F, et al. (2015) GlycoMine: A machine learning-based approach for predicting N-, C- and O-linked glycosylation in the human proteome. *Bioinformatics* 31:1411–1419.
39. Strobl C, Boulesteix A-L, Augustin T (2007) Unbiased split selection for classification trees based on the Gini index. *Comput Stat Data Anal* 52:483–501.
40. Tsirigos KD, Hennerdal A, Käll L, Elofsson A (2012) A guideline to proteome-wide α-helical membrane protein topology predictions. *Proteomics* 12:2282–2294.
41. Lu YC, et al. (2015) Structural basis of latrophilin-FLRT-UNC5 interaction in cell adhesion. *Structure* 23:1678–1691.
42. Heidmann O, et al. (2017) HEMO, an ancestral endogenous retroviral envelope protein shed in the blood of pregnant women and expressed in pluripotent stem cells and tumors. *Proc Natl Acad Sci USA* 114:E6642–E6651.
43. Thul PJ, et al. (2017) A subcellular map of the human proteome. *Science* 356:eaal3321.
44. Schwanhäusser B, et al. (2011) Global quantification of mammalian gene expression control. *Nature* 473:337–342.
45. Mathieson T, et al. (2018) Systematic analysis of protein turnover in primary cells. *Nat Commun* 9:689.
46. Surinova S, et al. (2011) On the development of plasma protein biomarkers. *J Proteome Res* 10:5–16.
47. Peterson AC, Russell JD, Bailey DJ, Westphall MS, Coon JJ (2012) Parallel reaction monitoring for high resolution and high mass accuracy quantitative, targeted proteomics. *Mol Cell Proteomics* 11:1475–1488.
48. Gillet LC, et al. (2012) Targeted data extraction of the MS/MS spectra generated by data-independent acquisition: A new concept for consistent and accurate proteome analysis. *Mol Cell Proteomics* 11:O111.016717.
49. Klijn C, et al. (2015) A comprehensive transcriptional portrait of human cancer cell lines. *Nat Biotechnol* 33:306–312.
50. Schreiner C, Bauer J, Margolis M, Juliano RL (1991) Expression and role of integrins in adhesion of human colonic carcinoma cells to extracellular matrix components. *Clin Exp Metastasis* 9:163–178.
51. Detchokul S, Williams ED, Parker MW, Frauman AG (2014) Tetraspanins as regulators of the tumour microenvironment: Implications for metastasis and therapeutic strategies. *Br J Pharmacol* 171:5462–5490.
52. Li Y, et al. (2017) Transcriptome analysis reveals determinant stages controlling human embryonic stem cell commitment to neuronal cells. *J Biol Chem* 292:19590–19604.
53. Pan C, Kumar C, Bohl S, Klingmueller U, Mann M (2009) Comparative proteomic phenotyping of cell lines and primary cells to assess preservation of cell type-specific functions. *Mol Cell Proteomics* 8:443–450.
54. Peiris TH, Oviedo NJ (2013) Gap junction proteins: Master regulators of the planarian stem cell response to tissue maintenance and injury. *Biochim Biophys Acta* 1828: 109–117.
55. Petersen TN, Brunak S, von Heijne G, Nielsen H (2011) SignalP 4.0: Discriminating signal peptides from transmembrane regions. *Nat Methods* 8:785–786.
56. Fonseca AL, et al. (2016) Bioinformatics analysis of the human surfaceome reveals new targets for a variety of tumor types. *Int J Genomics* 2016:8346198.
57. Moremen KW, Tiemeyer M, Nairn AV (2012) Vertebrate protein glycosylation: Diversity, synthesis and function. *Nat Rev Mol Cell Biol* 13:448–462.
58. Mellquist JL, Kasturi L, Spitalnik SL, Shakin-Eshleman SH (1998) The amino acid following an asn-X-Ser/Thr sequon is an important determinant of N-linked core glycosylation efficiency. *Biochemistry* 37:6833–6837.
59. Petrescu A-J, Milac A-L, Petrescu SM, Dwek RA, Wormald MR (2004) Statistical analysis of the protein environment of N-glycosylation sites: Implications for occupancy, structure, and folding. *Glycobiology* 14:103–114.
60. Sevier CS, Kaiser CA (2002) Formation and transfer of disulphide bonds in living cells. *Nat Rev Mol Cell Biol* 3:836–847.
61. Sharpe HJ, Stevens TJ, Munro S (2010) A comprehensive comparison of transmembrane domains reveals organelle-specific properties. *Cell* 142:158–169.
62. Lundbaek JA, Andersen OS, Werge T, Nielsen C (2003) Cholesterol-induced protein sorting: An analysis of energetic feasibility. *Biophys J* 84:2080–2089.
63. Filmore D (2004) It's a GPCR world. *Mod Drug Discovery* 7:24–28.
64. Edfors F, et al. (2016) Gene-specific correlation of RNA and protein levels in human cells and tissues. *Mol Syst Biol* 12:883.
65. Beck M, et al. (2011) The quantitative proteome of a human cell line. *Mol Syst Biol* 7: 549.
66. Rea S, James DE (1997) Moving GLUT4: The biogenesis and trafficking of GLUT4 storage vesicles. *Diabetes* 46:1667–1677.
67. O'Hagan S, Wright Muelas M, Day PJ, Lundberg E, Kell DB (2018) GeneGini: Assessment via the Gini coefficient of reference "housekeeping" genes and diverse human transporter expression profiles. *Cell Syst* 6:230–244.e1.
68. The UniProt Consortium (2014) UniProt: A hub for protein information. *Nucleic Acids Res* 43:D204–D212.
69. Binder JX, et al. (2014) COMPARTMENTS: Unification and visualization of protein subcellular localization evidence. *Database (Oxford)* 2014:bau012.
70. Breiman L (2001) Random forests. *Mach Learn* 45:5–32.
71. Liaw A, Wiener M (2002) Classification and regression by randomForest. *R News* 2: 18–22.
72. R Core Team (2014) R: A Language and Environment for Statistical Computing, Version 3.1.0. Available at www.R-project.org/. Accessed July 1, 2014.