SurfaceGenie User Guide

Contents

SurfaceGenie User Guide

1. SurfaceGenie Overview

SurfaceGenie is a web application for analyzing omic datasets (*e.g.* proteomic, transcriptomic) to prioritize candidate cell-type specific markers of interest for immunophenotyping, immunotherapy, drug targeting, and other applications. This User Guide includes instructions for how to use the features available in the ***SurfaceGenie*** web application and describes the theory and calculations used in the scoring algorithm.

In developing ***SurfaceGenie*** we aimed to create an accessible tool for calculation of *GenieScore* and *GenieScore* components from input data. ***SurfaceGenie*** contains two separate, though related, tools – ***GenieScore Calculator*** and ***SPC Score Lookup***. For both tools, users are able to export the calculated values and generated plots.

SurfaceGenie was written in R and the web application was developed using the Shiny library. Source code and all reference lookup tables are publicly available here.

Before you begin

Currently, the primary functions of ***SurfaceGenie*** are available only for human, mouse, and rat data. ***SurfaceGenie*** operates with Uniprot Accession IDs only. Bulk conversion of alternate IDs to Uniprot IDs can be performed using the 'Retrieve/ID mapping tool' available on the Uniprot website, found here. Note that conversion between IDs is not always one-to-one. Manual curation of the results from the ID mapping is advisable.

2. GenieScore Calculator - Basics and Tutorial

2.1. What is a GenieScore?

GenieScore is a metric designed to provide a single value that can be used to rank order molecules based on their capacity to serve as a surface marker for distinguishing among sample groups (e.g. cell types, experimental conditions). *GenieScores* are calculated for each protein and are experiment-specific, meaning that the *GenieScore* for a single protein can vary depending on the data provided as input for the analysis.

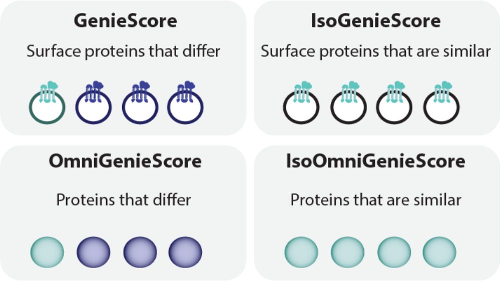
2.2. Assumptions/Caveats

GenieScore was designed to analyze data collected as part of the same batch of studies and therefore does not perform any normalization of datasets prior to analysis. The operating assumption is that the input dataset was either collected in a semi-quantitative manner or curated such that the data from different experimental groups are of the same type and quality. Batch correction of data may enable comparison of disjointed datasets.

All calculations performed within SurfaceGenie consider values from the current input. In other words, the tools will consider all data within a single dataset input (which may contain multiple experiments and/or cell types). If a user performs a comparison and subsequently determines additional data should be considered, a new file containing all data for the new comparison is required.

2.3. GenieScore Permutations

While a major benefit of ***SurfaceGenie*** is the ability to prioritize proteins that are localized to the cell surface, it is also possible to analyze data without this consideration to prioritize potential markers that reside in other subcellular localizations. The descriptions for the four permutations of the scoring algorithm are provided below. *GenieScore* has been tested with semi-quantitative proteomic and transcriptomic datasets, but is expected to be useful for other data types. *OmniGenieScore* and *IsoOmniGenieScore* are independent of Accession ID and therefore suitable for any species or type of data which is not associated with Accession IDs (*e.g.* metabolic data). See Section 4.3 for more info.


GenieScore
Surface proteins that differ
IsoGenieScore
Surface proteins that are similar
OmniGenieScore
Proteins that differ
IsoOmniGenieScore
Proteins that are similar

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.

2.4. Overview of GenieScore Calculator Usage

2.4.1. Input:

GenieScore Calculator accepts text files (tab, tsv, txt, csv, xlsx) containing a list of protein identifiers (UniProt Accession) and a surrogate value representative of abundance (e.g. number of peptide spectrum matches, peak area, FKPM, RKPM) identified within a set of samples. There is no limit to the number of samples that can be analyzed in a single file. The column header of the first column must be labeled with *Accession*. An example file can be downloaded from the Instructions page of ***SurfaceGenie***.

2.4.2. Calculations:

For each protein in the dataset, each of the selected *GenieScore* permutations are calculated utilizing information from three independent scores. For more information regarding their rationale and calculation see Section 4.1.

- *Surface Protein Consensus (SPC) score:* A predictive measure of the likelihood that a particular protein is present at the cell surface. This value is a sum of the number of predictive datasets for which a protein has been predicted to be localized to the cell surface. Scores range 0-4. SurfaceGenie has SPC datasets for human, mouse, and rat.

- *Signal Dispersion*: A measure of how evenly or unevenly distributed a protein is among multiple samples within a comparison dataset. It is based on the Gini coefficient, a measure of statistical dispersion of values. Signal Dispersion scores range 0 – 1.

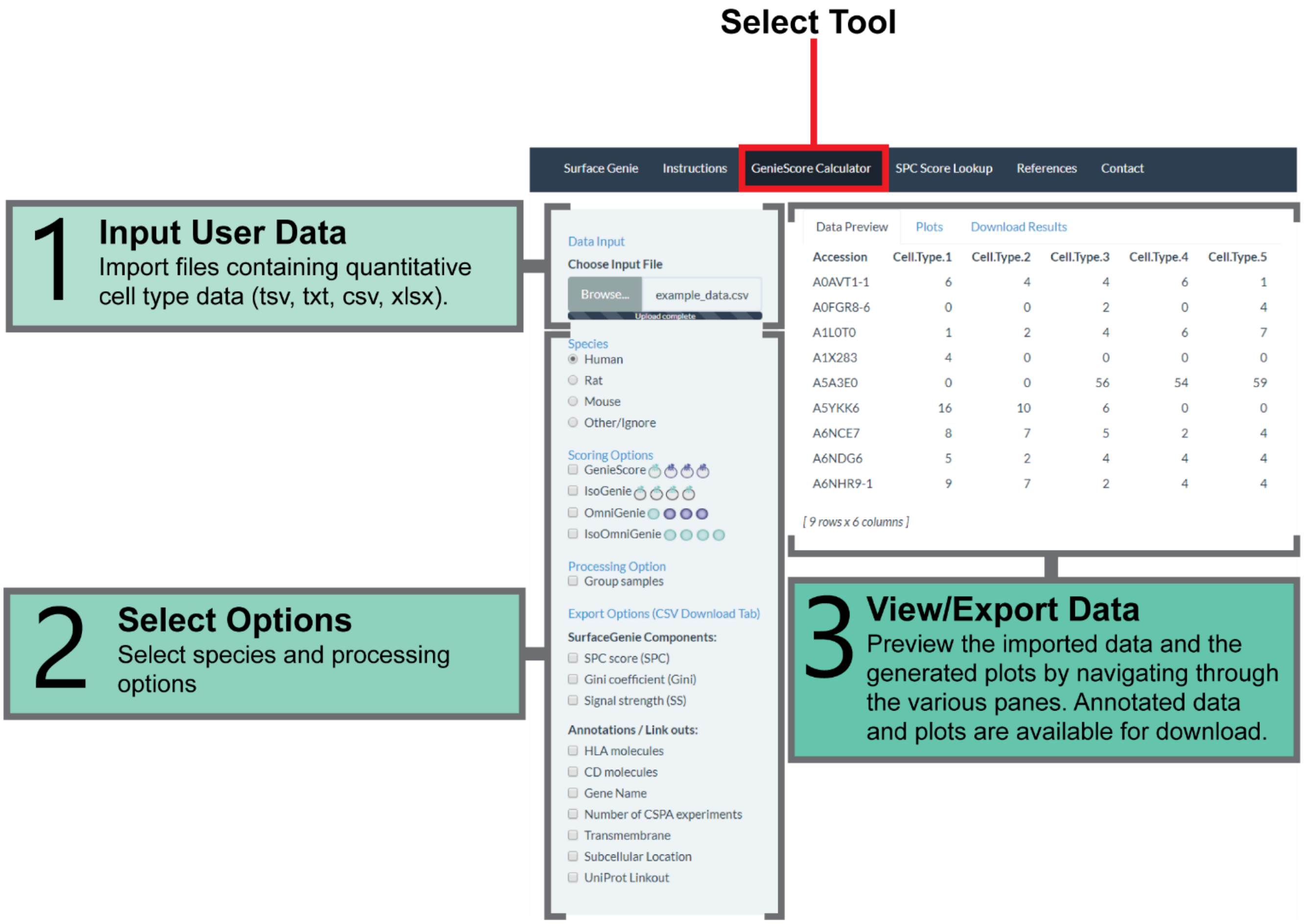
- *Signal Strength:* A measure of protein abundance for cell types in which a protein is observed. Proteins at the lower limit of detection are of lower priority than those with more observations, because it is expected that those of higher abundance will practically serve as more accessible markers for downstream technologies. Scores typically range 0 ~ 4.

2.4.3. Output:

Users are able to export both an appended version of the input data set and the automatically generated plots.

- *Data Download*: Each of the selected *GenieScore* permutations and any additional selected export options (e.g. SPC score, CD molecule annotation, etc) are appended to each entry in the original input file. More information regarding annotation see Section 4.4.
- *GenieScore Plots*: Scores from each of the selected *GenieScore* permutations are plotted in decreasing order.
- *SPC score Histogram*: Displays the distribution of *SPC scores*.

2.5. GenieScore Calculator Quick-start Guide:



1 Input User Data
Import files containing quantitative cell type data (tsv, txt, csv, xlsx).

2 Select Options
Select species and processing options

3 View/Export Data
Preview the imported data and the generated plots by navigating through the various panes. Annotated data and plots are available for download.

Data Input

Choose Input File: example_data.csv

Species
● Human
○ Rat
○ Mouse
○ Other/Ignore

Scoring Options
□ GenieScore
□ IsoGenie
□ OmniGenie
□ IsoOmniGenie

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
□ Transmembrane
□ Subcellular Location
□ UniProt Linkout

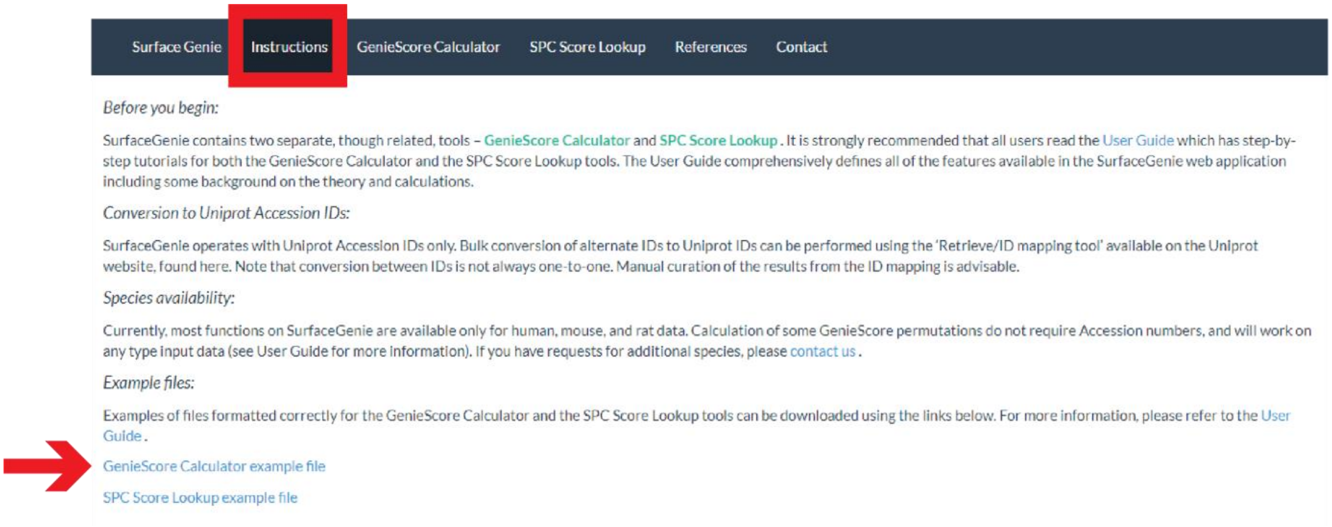
Accession	Cell.Type.1	Cell.Type.2	Cell.Type.3	Cell.Type.4	Cell.Type.5
A0AVT1-1	6	4	4	6	1
A0FGR8-6	0	0	2	0	4
A1L0T0	1	2	4	6	7
A1X283	4	0	0	0	0
A5A3E0	0	0	56	54	59
A5YKK6	16	10	6	0	0
A6NCE7	8	7	5	2	4
A6NDG6	5	2	4	4	4
A6NHR9-1	9	7	2	4	4

[9 rows x 6 columns]

2.6. GenieScore Calculator Tutorial

Before you begin:

This tutorial uses the example data file provided in the ***Instructions*** tab.



Before you begin:

SurfaceGenie contains two separate, though related, tools – GenieScore Calculator and SPC Score Lookup . It is strongly recommended that all users read the User Guide which has step-by-step tutorials for both the GenieScore Calculator and the SPC Score Lookup tools. The User Guide comprehensively defines all of the features available in the SurfaceGenie web application including some background on the theory and calculations.

Conversion to Uniprot Accession IDs:

SurfaceGenie operates with Uniprot Accession IDs only. Bulk conversion of alternate IDs to Uniprot IDs can be performed using the 'Retrieve/ID mapping tool' available on the Uniprot website, found here. Note that conversion between IDs is not always one-to-one. Manual curation of the results from the ID mapping is advisable.

Species availability:

Currently, most functions on SurfaceGenie are available only for human, mouse, and rat data. Calculation of some GenieScore permutations do not require Accession numbers, and will work on any type input data (see User Guide for more information). If you have requests for additional species, please contact us .

Example files:

Examples of files formatted correctly for the GenieScore Calculator and the SPC Score Lookup tools can be downloaded using the links below. For more information, please refer to the User Guide .

GenieScore Calculator example file

SPC Score Lookup example file

Alternatively, users can follow the steps with their own data provided it conforms to the following specifications:

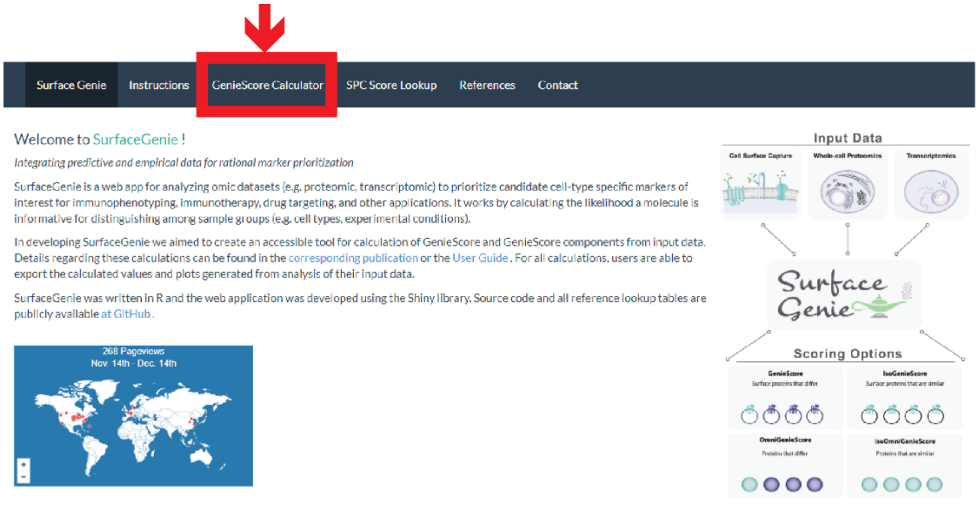
- File type: tab, tsv, txt, csv, xlsx
- Species: Human, Mouse, Rat
- Identifier: UniProt Accession ID
- ** The header of the first column must be *Accession* **

Example of properly formatted file:

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
A5YKK6	16	10	6	0	0
A6NCE7	8	7	5	2	4
A6NDG6	5	2	4	4	4
A6NHR9-1	9	7	2	4	4

SurfaceGenie User Guide

1. From the Home Page of ***SurfaceGenie***, click on the ***GenieScore Calculator*** tab in the header bar.



Welcome to SurfaceGenie !

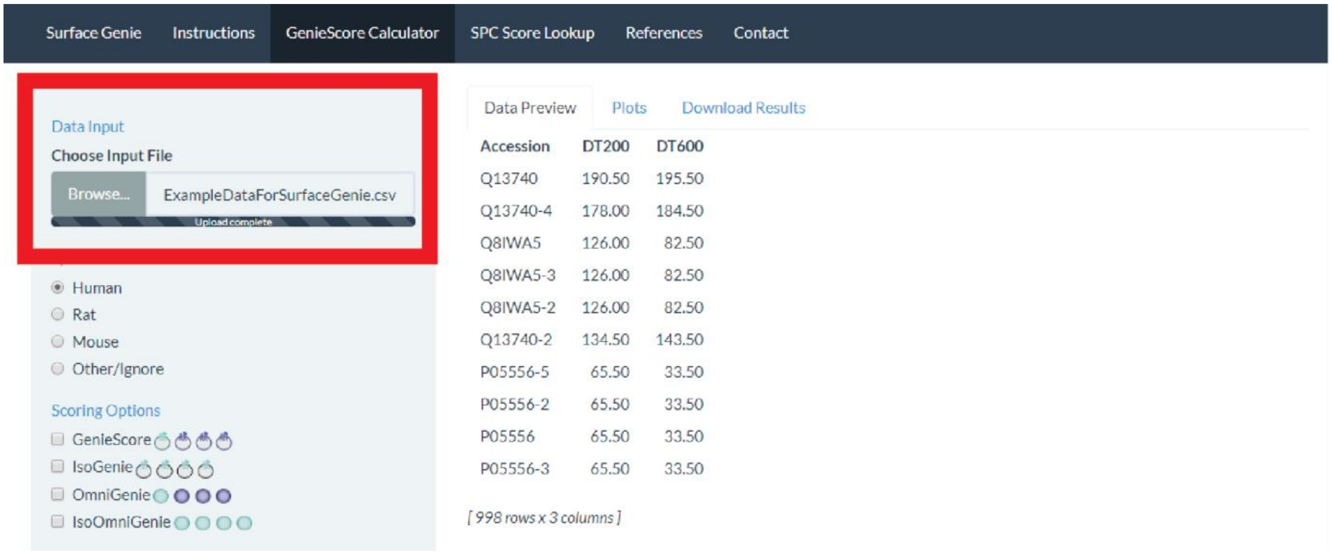
Integrating predictive and empirical data for rational marker prioritization

SurfaceGenie is a web app for analyzing omic datasets (e.g. proteomic, transcriptomic) to prioritize candidate cell-type specific markers of interest for immunophenotyping, immunotherapy, drug targeting, and other applications. It works by calculating the likelihood a molecule is informative for distinguishing among sample groups (e.g. cell types, experimental conditions).

In developing SurfaceGenie we aimed to create an accessible tool for calculation of GenieScore and GenieScore components from input data. Details regarding these calculations can be found in the corresponding publication or the User Guide . For all calculations, users are able to export the calculated values and plots generated from analysis of their input data.

SurfaceGenie was written in R and the web application was developed using the Shiny library. Source code and all reference lookup tables are publicly available at GitHub .

2. Using the ‘Browse’ button, in the ‘Data Input’ section, navigate to and select the data file to be imported. Once the data has been imported, the first ten rows will be visible for manual inspection in the ‘Data Preview’ pane.



Data Input

Choose Input File

ExampleDataForSurfaceGenie.csv

● Human
○ Rat
○ Mouse
○ Other/Ignore

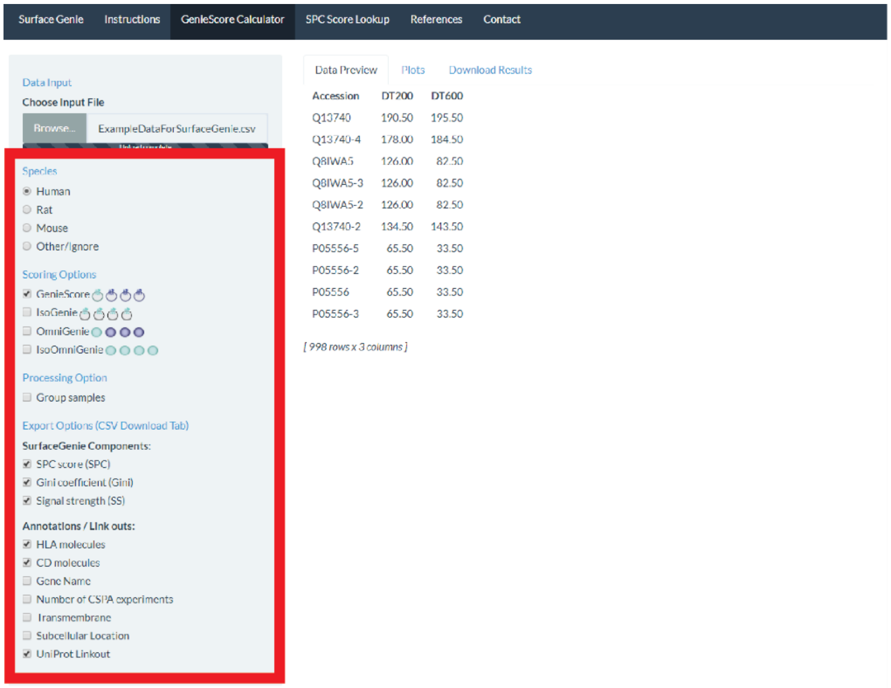
Scoring Options

☐ GenieScore
☐ IsoGenie
☐ OmniGenie
☐ IsoOmniGenie

Accession	DT200	DT600
Q13740	190.50	195.50
Q13740-4	178.00	184.50
Q8IWA5	126.00	82.50
Q8IWA5-3	126.00	82.50
Q8IWA5-2	126.00	82.50
Q13740-2	134.50	143.50
P05556-5	65.50	33.50
P05556-2	65.50	33.50
P05556	65.50	33.50
P05556-3	65.50	33.50

[998 rows x 3 columns]

3. Using the buttons in the left pane, select the correct species (human for the Example File) and choose the desired ‘Scoring Options’ and ‘Export Options’.



Data Input

Choose Input File

ExampleDataForSurfaceGenie.csv

Species

● Human
○ Rat
○ Mouse
○ Other/Ignore

Scoring Options

☑ GenieScore
☐ IsoGenie
☐ OmniGenie
☐ IsoOmniGenie

Processing Option

☐ Group samples

Export Options (CSV Download Tab)

SurfaceGenie Components:

☑ SPC score (SPC)
☑ Gini coefficient (Gini)
☑ Signal strength (SS)

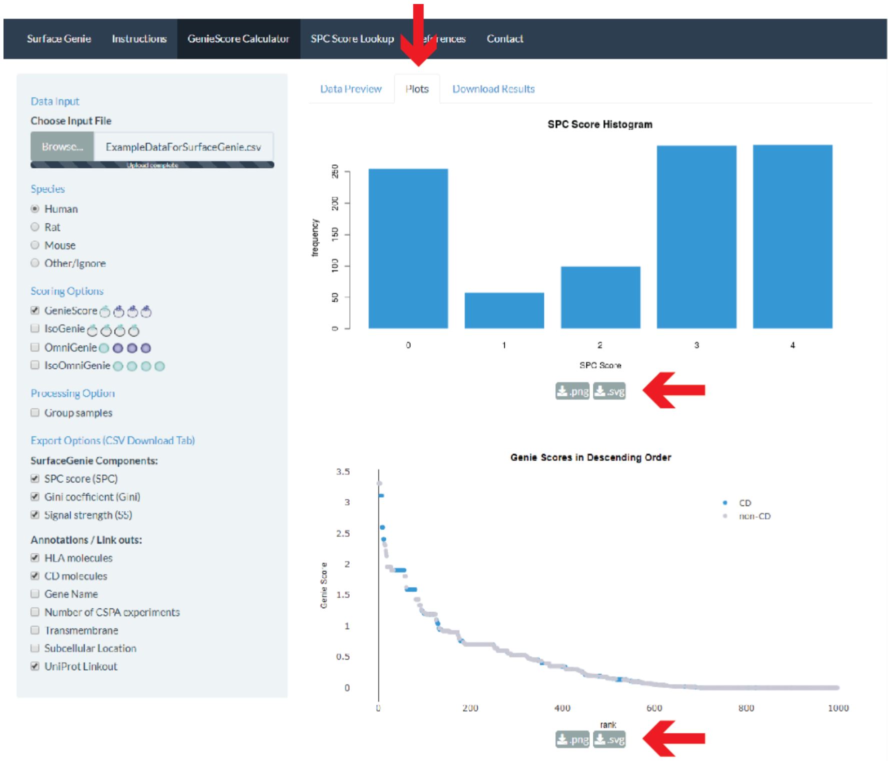
Annotations / Link outs:

☑ HLA molecules
☑ CD molecules
☐ Gene Name
☐ Number of CSPA experiments
☐ Transmembrane
☐ Subcellular Location
☑ UniProt Linkout

Accession	DT200	DT600
Q13740	190.50	195.50
Q13740-4	178.00	184.50
Q8IWA5	126.00	82.50
Q8IWA5-3	126.00	82.50
Q8IWA5-2	126.00	82.50
Q13740-2	134.50	143.50
P05556-5	65.50	33.50
P05556-2	65.50	33.50
P05556	65.50	33.50
P05556-3	65.50	33.50

[998 rows x 3 columns]

4. Navigate to the ‘Plots’ pane to view a histogram of *SPC scores* and a plot of the calculated *GenieScores*. Use the buttons below the plots to download copies.



Data Input

Choose Input File

ExampleDataForSurfaceGenie.csv

Species

● Human
○ Rat
○ Mouse
○ Other/Ignore

Scoring Options

☑ GenieScore
☐ IsoGenie
☐ OmniGenie
☐ IsoOmniGenie

Processing Option

☐ Group samples

Export Options (CSV Download Tab)

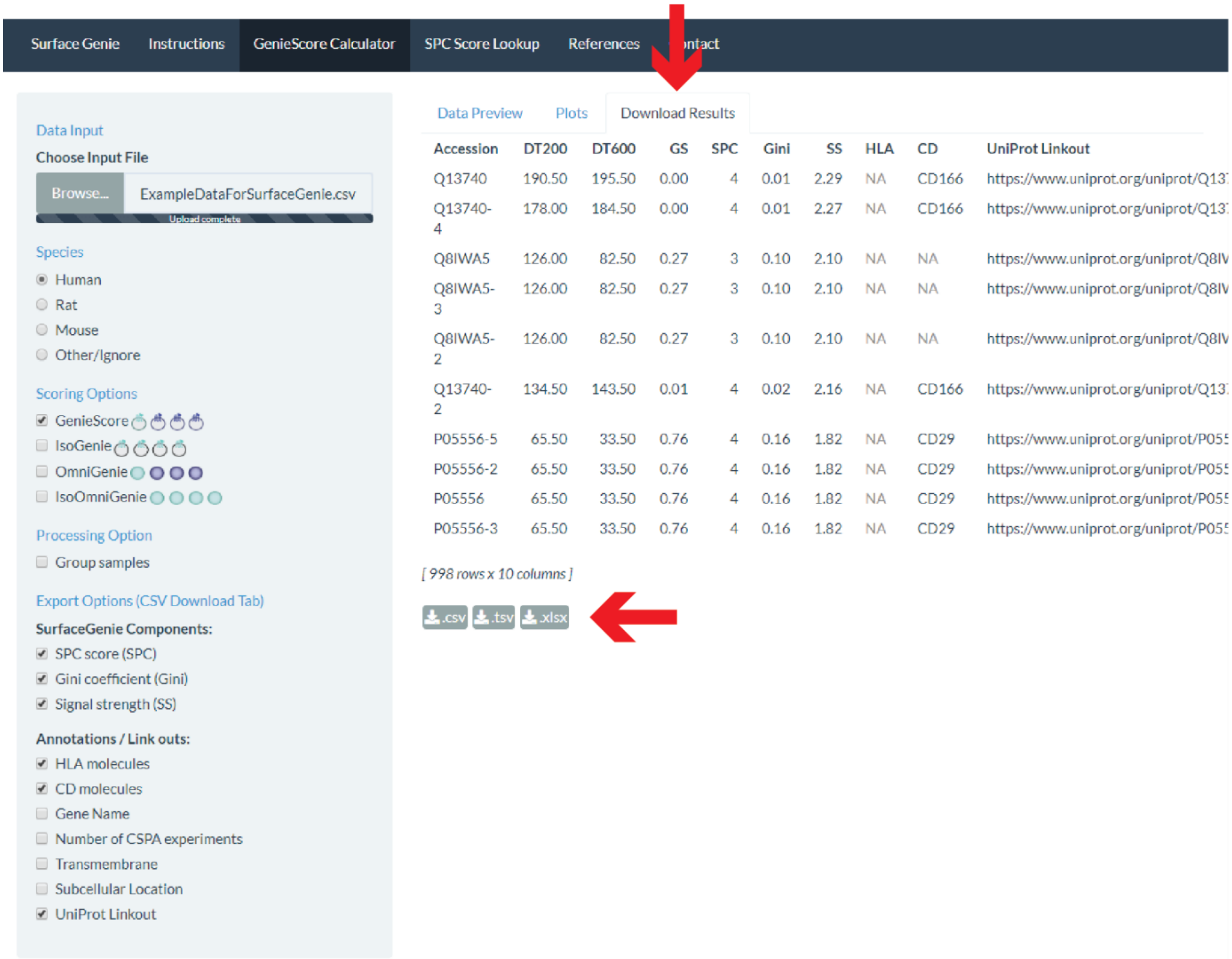
SurfaceGenie Components:

☑ SPC score (SPC)
☑ Gini coefficient (Gini)
☑ Signal strength (SS)

Annotations / Link outs:

☑ HLA molecules
☑ CD molecules
☐ Gene Name
☐ Number of CSPA experiments
☐ Transmembrane
☐ Subcellular Location
☑ UniProt Linkout

5. Navigate to the 'Download Results' pane to visualize the desired export options appended to the original input data. The results can be downloaded in multiple file formats using the buttons under the data.



Data Input

Choose Input File

ExampleDataForSurfaceGenie.csv

Species

● Human
○ Rat
○ Mouse
○ Other/Ignore

Scoring Options

☑ GenieScore
☐ IsoGenie
☐ OmniGenie
☐ IsoOmniGenie

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
☐ Transmembrane
☐ Subcellular Location
☑ UniProt Linkout

Accession	DT200	DT600	GS	SPC	Gini	SS	HLA	CD	UniProt Linkout
Q13740	190.50	195.50	0.00	4	0.01	2.29	NA	CD166	https://www.uniprot.org/uniprot/Q13
Q13740-4	178.00	184.50	0.00	4	0.01	2.27	NA	CD166	https://www.uniprot.org/uniprot/Q13
Q8IWA5	126.00	82.50	0.27	3	0.10	2.10	NA	NA	https://www.uniprot.org/uniprot/Q8IV
Q8IWA5-3	126.00	82.50	0.27	3	0.10	2.10	NA	NA	https://www.uniprot.org/uniprot/Q8IV
Q8IWA5-2	126.00	82.50	0.27	3	0.10	2.10	NA	NA	https://www.uniprot.org/uniprot/Q8IV
Q13740-2	134.50	143.50	0.01	4	0.02	2.16	NA	CD166	https://www.uniprot.org/uniprot/Q13
P05556-5	65.50	33.50	0.76	4	0.16	1.82	NA	CD29	https://www.uniprot.org/uniprot/P05
P05556-2	65.50	33.50	0.76	4	0.16	1.82	NA	CD29	https://www.uniprot.org/uniprot/P05
P05556	65.50	33.50	0.76	4	0.16	1.82	NA	CD29	https://www.uniprot.org/uniprot/P05
P05556-3	65.50	33.50	0.76	4	0.16	1.82	NA	CD29	https://www.uniprot.org/uniprot/P05

[998 rows x 10 columns]

Additional notes:

- The 'Other/Ignore' option for species is a way to process data unrelated to proteins (*i.e.* without Accession IDs) without prompting errors.
- The *GenieScore* plot is interactive and will return information for the data point the pointer is hovering over.
- Separate plots will be generated for each *GenieScore* permutation that is selected.
- Any ID that is not an Accession for the selected species is scored as N/A. Any valid Accession that is not predicted to be surface localized is given a score of 0
- The 'Group samples' functionality is a convenient way to combine columns that represent replicates or otherwise similar cell types from within ***SurfaceGenie***.

3. SPC Score Lookup - Basics and Tutorial

3.1 What is SPC Score?

Surface Prediction Concensus (SPC) score is a predictive measure of the likelihood that a particular protein is present at the cell surface. This value is a sum of the number of predictive datasets for which a protein has been predicted to be localized to the cell surface. Scores range 0-4. ***SurfaceGenie*** has *SPC score* datasets for human, mouse, and rat. For more details on the predictive datasets used, see Section 4.1.1.

3.2 Assumptions/Caveats

By concatenating published 'surfaceome' sets, *SPC score* is a straightforward representation of the proteins that have been predicted to be cell surface localized despite the caveats associated with each of the prediction strategies. By forgoing manual curation, it is likely that the set of proteins predicted to be at the surface by *SPC score* is overly inclusive (*e.g*. includes membrane proteins either not localized or exposed to the surface); however, our approach avoids the complication of introducing further bias by relying on alternative or additional prediction strategies (*e.g.* signal peptide or transmembrane orientation). As the localization of a protein is ultimately cell type- and context dependent (*e.g.* experimental condition, disease and/or stimulus state), every protein candidate must eventually be validated for the application of choice within that system. Our use of an inclusive list is designed with this fact in mind. Ultimately, the score enables prioritization of the marker(s) to pursue in subsequent studies and is not a promise that the top candidate will be a suitable immunophenotyping marker.

Whereas the human *SPC scores* are derived directly from previous constructions of the human 'surfaceome', the mouse and rat scores were assigned by mapping homologous Accession IDs (utilizing Mouse Genome Informatics database (http://www.informatics.jax.org). This introduces the fidelity of homology mapping as an assumption.

3.3 Overview of SPC Score Lookup Usage

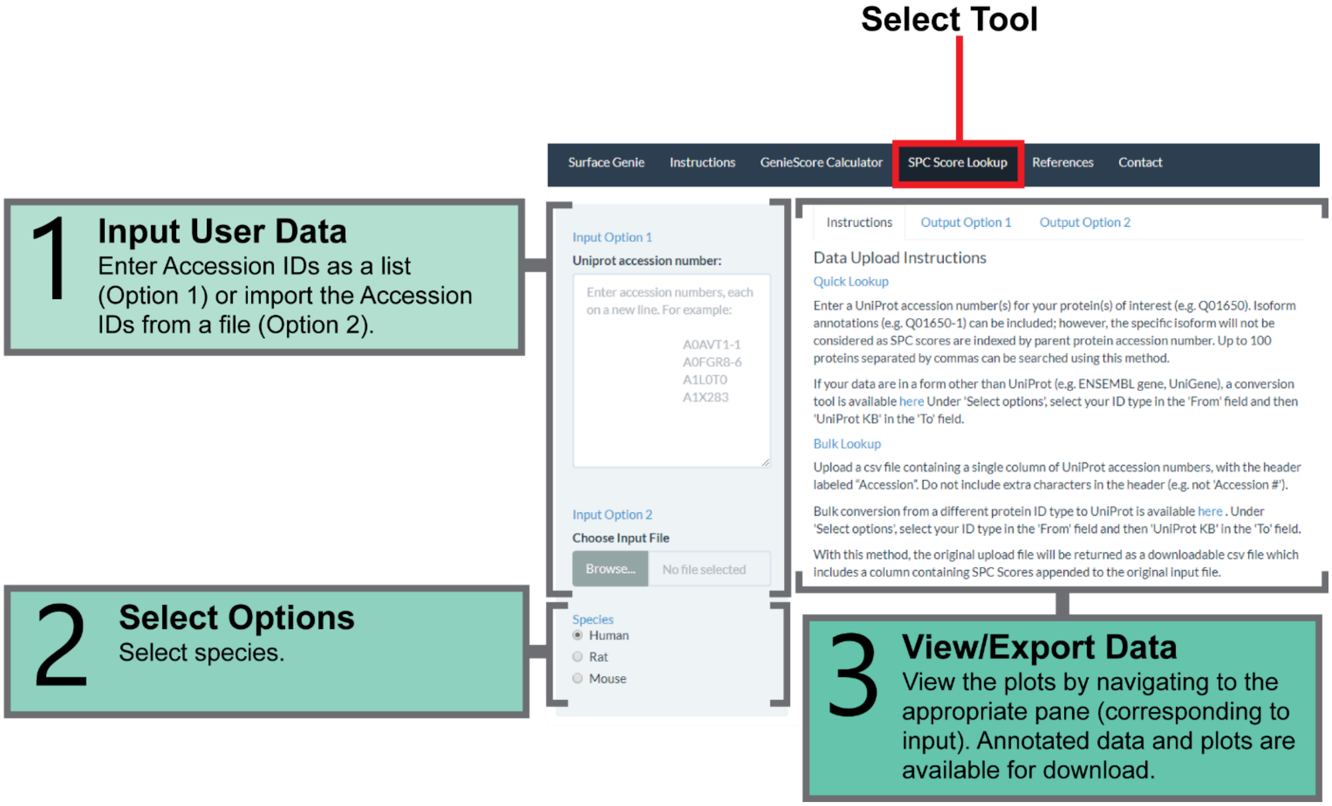
3.3.1 Input

The ***SPC Score Lookup*** tool accepts text files (tab, tsv, txt, csv, xlsx) containing a list of protein identifiers (UniProt Accession) in a column. There is no limit to the number of Accession IDs that can be analyzed in a single file. The column header of the first column must be labeled with *Accession*. An example file can be downloaded from the Instructions page of ***SurfaceGenie***. Alternatively, Accession IDs can be pasted directly into the box labeled 'Input Option 1'.

3.3.2 Output

For human Accession IDs, *SPC scores* are returned along with presence/absence information from each of the four original surfaceome constructions. For mouse and rat Accession IDs, the homologous human Accession ID is returned along with *SPC score*s. For all species, a csv file can also be downloaded containing the previously described information appended to the input list.

3.4 SPC Score Lookup Quick-start Guide



1 Input User Data
Enter Accession IDs as a list (Option 1) or import the Accession IDs from a file (Option 2).

2 Select Options
Select species.

3 View/Export Data
View the plots by navigating to the appropriate pane (corresponding to input). Annotated data and plots are available for download.

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 #').

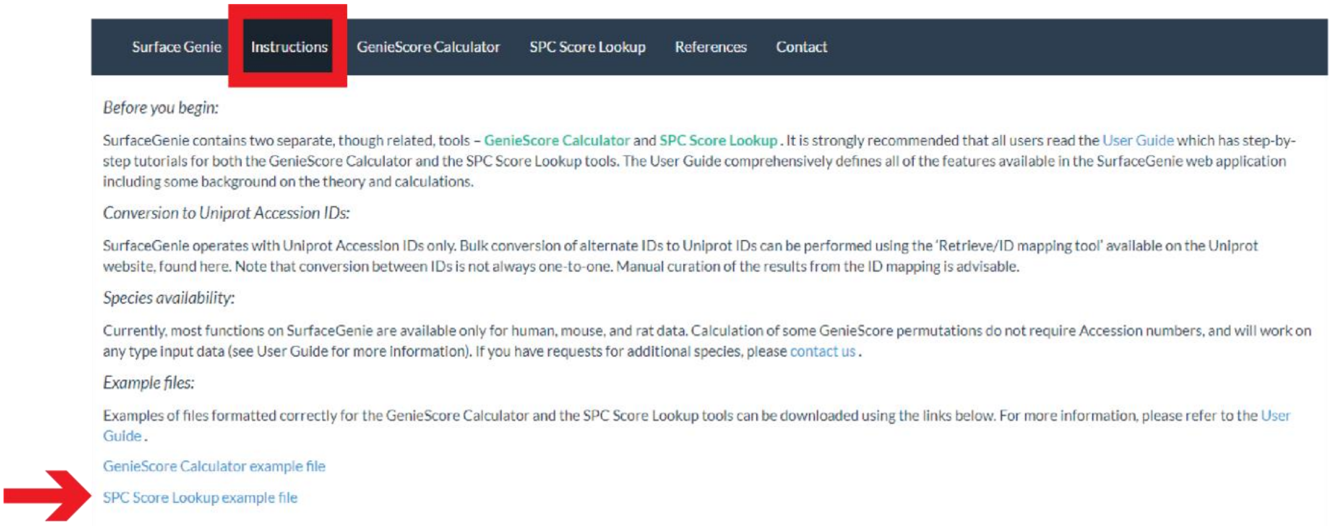
Bulk conversion from a different protein ID type to UniProt is available here . Under 'Select options', select your ID type in the 'From' field and then 'UniProt KB' in the 'To' field.

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.

3.5 SPC Score Lookup Tutorial

Before you begin:

This tutorial uses the example data file provided in the ***Instructions*** tab.



Before you begin:

SurfaceGenie contains two separate, though related, tools – GenieScore Calculator and SPC Score Lookup . It is strongly recommended that all users read the User Guide which has step-by-step tutorials for both the GenieScore Calculator and the SPC Score Lookup tools. The User Guide comprehensively defines all of the features available in the SurfaceGenie web application including some background on the theory and calculations.

Conversion to Uniprot Accession IDs:

SurfaceGenie operates with Uniprot Accession IDs only. Bulk conversion of alternate IDs to Uniprot IDs can be performed using the 'Retrieve/ID mapping tool' available on the Uniprot website, found here. Note that conversion between IDs is not always one-to-one. Manual curation of the results from the ID mapping is advisable.

Species availability:

Currently, most functions on SurfaceGenie are available only for human, mouse, and rat data. Calculation of some GenieScore permutations do not require Accession numbers, and will work on any type input data (see User Guide for more information). If you have requests for additional species, please contact us .

Example files:

Examples of files formatted correctly for the GenieScore Calculator and the SPC Score Lookup tools can be downloaded using the links below. For more information, please refer to the User Guide .

GenieScore Calculator example file

SPC Score Lookup example file

Alternatively, users can follow the steps with their own data provided it conforms to the following specifications:

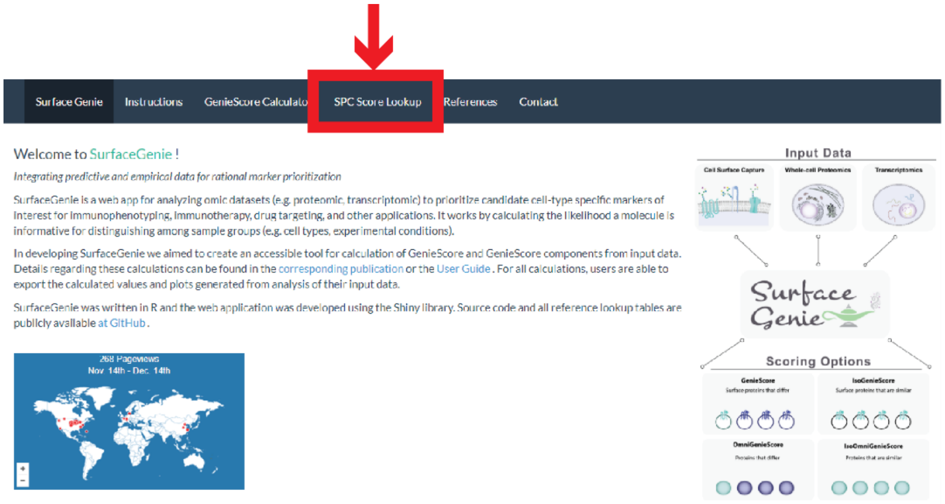
- <u>File type</u>: tab, tsv, txt, csv, xlsx
- <u>Species</u>: Human, Mouse, Rat
- <u>Identifier:</u> UniProt Accession ID
- ** The header of the first column must be *Accession* **

Example of properly formatted file:

Accession
A0AVT1-1
A0FGR8-6
A1L0T0
A1X283
A5A3E0
A5YKK6
A6NCE7
A6NDG6
A6NHR9-1

SurfaceGenie User Guide

1. From the Home Page of ***SurfaceGenie***, click on the ***GenieScore Calculator*** tab in the header bar.



Welcome to SurfaceGenie !

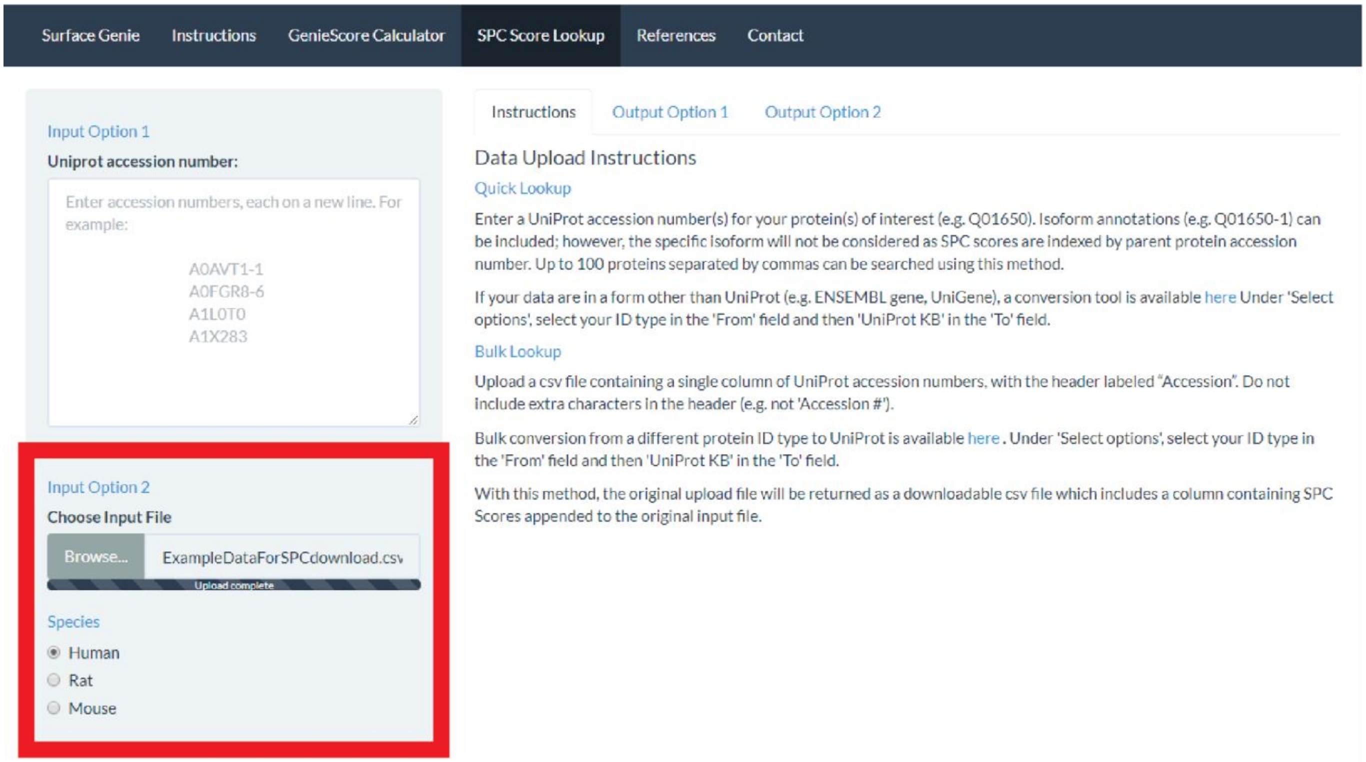
Integrating predictive and empirical data for rational marker prioritization

SurfaceGenie is a web app for analyzing omic datasets (e.g. proteomic, transcriptomic) to prioritize candidate cell-type specific markers of interest for immunophenotyping, immunotherapy, drug targeting, and other applications. It works by calculating the likelihood a molecule is informative for distinguishing among sample groups (e.g. cell types, experimental conditions).

In developing SurfaceGenie we aimed to create an accessible tool for calculation of GenieScore and GenieScore components from input data. Details regarding these calculations can be found in the corresponding publication or the User Guide . For all calculations, users are able to export the calculated values and plots generated from analysis of their input data.

SurfaceGenie was written in R and the web application was developed using the Shiny library. Source code and all reference lookup tables are publicly available at GitHub .

2. Using the 'Browse' button, in the 'Input Option 3' section, navigate to and select the data file to be imported. Make sure to select the correct species (Human for the example file).



Input Option 1

Uniprot accession number:

Enter accession numbers, each on a new line. For example:

A0AVT1-1
A0FGR8-6
A1L0T0
A1X283

Input Option 2

Choose Input File

ExampleDataForSPCdownload.csv

Species

● Human
○ Rat
○ Mouse

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.

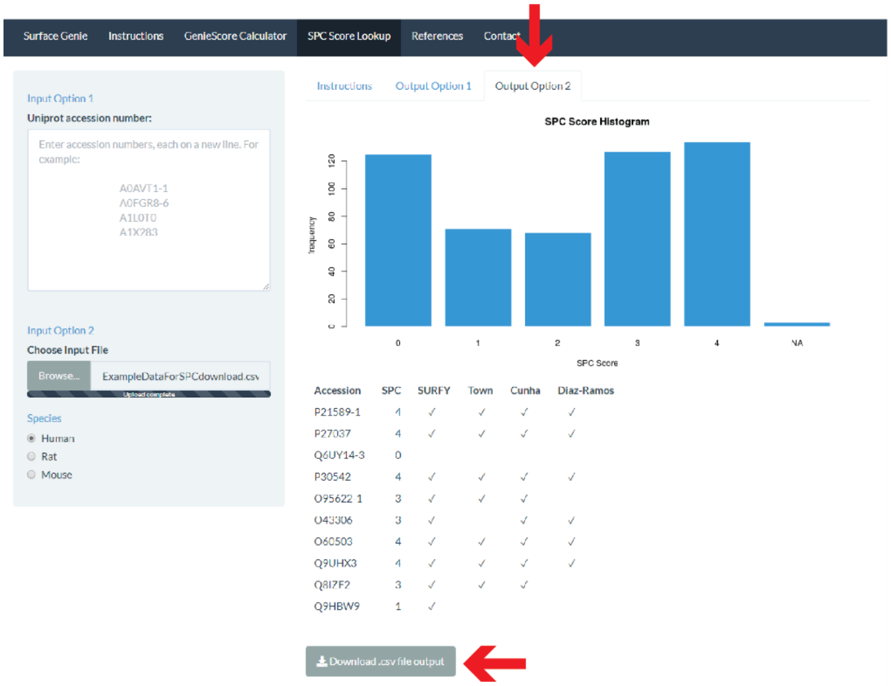
Bulk Lookup

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.

3. Navigate to the 'Output Option 2' pane to view a histogram of *SPC scores*. Below the plot will be a table that previews the first ten rows showing the overall SPC score and in which datasets the protein was predicted as being surface-localized. Use the button below the table to download a table containing the *SPC scores*.



Input Option 1

Uniprot accession number:

Enter accession numbers, each on a new line. For example:

A0AVT1-1
A0FGR8-6
A1L0T0
A1X283

Input Option 2

Choose Input File

ExampleDataForSPCdownload.csv

Species

● Human
○ Rat
○ Mouse

Accession	SPC	SURFY	Town	Cunha	Diaz-Ramos
P21589-1	4	✓	✓	✓	✓
P27037	4	✓	✓	✓	✓
Q6UY14-3	0				
P30542	4	✓	✓	✓	✓
O95622-1	3	✓	✓	✓	
O43306	3	✓		✓	✓
O60503	4	✓	✓	✓	✓
Q9UHX3	4	✓	✓	✓	✓
Q8IZF2	3	✓	✓	✓	
Q9HBW9	1	✓			

Additional notes:

- Any ID that is not an Accession for the selected species is scored as N/A. Any valid Accession that is not predicted to be surface localized is given a score of 0.
- Using 'Input Option 1' Accession IDs can be pasted directly in the provided box.
    - Make sure to select the correct species even if using 'Input Option 1'.
    - The results for this option will appear under the 'Output Option 1' pane.
- For human Accession IDs, *SPC scores* are returned along with presence/absence information from each of the four original surfaceome constructions.
- For mouse and rat Accession IDs, the homologous human Accession ID is returned along with *SPC scores*.

4. Additional Information

4.1 Rationale and Calculation of GenieScore Components

4.1.1 Surface Prediction Consensus (SPC) score

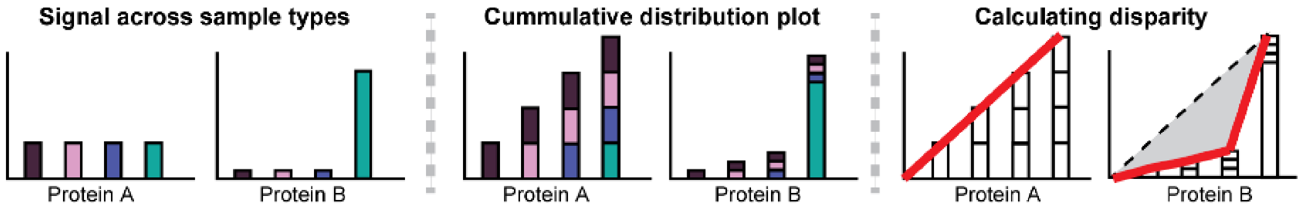
Surface Prediction Consensus (SPC) score was generated from concatenating four individual human surfaceome databases and assigning a point for each of the individual datasets in which the protein was predicted to be localized to the cell surface. For more information about the four individual datasets, please refer to the original publications.

- Bausch-Fluck D, et al. (2018) – machine learning approach trained using experimentally validated surface proteins
- da Cunha JP, et al. (2009) – constructed using ontological annotations and transmembrane prediction
- Town J, et al. (2016) – constructed by combining ontological and machine learning approaches
- Diaz-Ramos MC, Engel P, & Bastos R (2011) – manual curation

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* can be accessed via the Github repository.

4.1.2 Signal dispersion

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. A visual depiction of Gini coefficient calculation is shown in the figure below. 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.



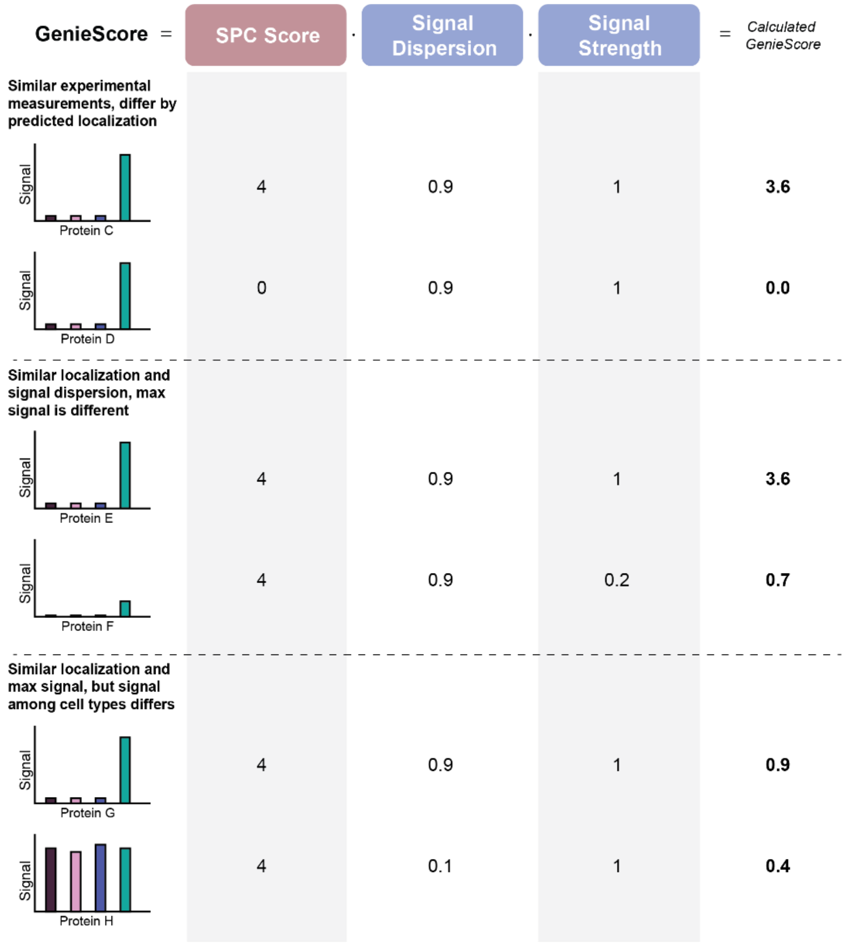
In the figure above, the calculation of Gini coefficient is represented visually for two example proteins, where the gray shaded area represents the calculated disparity between measurements. Protein A has equal measurements in each sample type, resulting in a Gini coefficient of 0 (i.e. tracing the addition to cumulative signal from each sample results in the identity function). The contributions to total Protein B signal are disparate among the cell types. The gray shaded are represents the discrete integral calculated from the identity function to a point-to-point fit of the contributions to cumulative signal. The gray shaded area is used to calculate the Gini coefficient.

4.1.3 Signal strength

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 depends on the type of quantitative measurement.

4.2 Examples of GenieScore Calculations

Examples of *GenieScores* are shown for three pairs of proteins, which differ with respect to one of the individual components of the *GenieScore* equation.



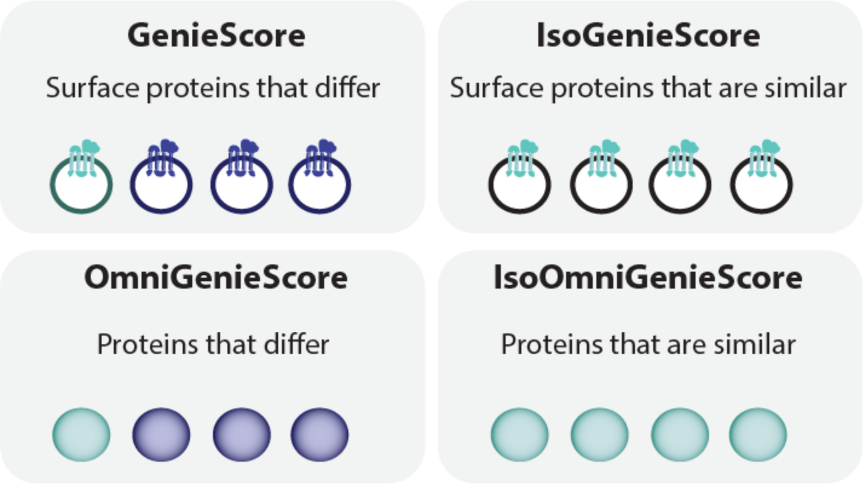
GenieScore =	SPC Score	·	Signal Dispersion	·	Signal Strength	=	Calculated GenieScore
Similar experimental measurements, differ by predicted localization							
Protein C	4		0.9		1		**3.6**
Protein D	0		0.9		1		**0.0**
Similar localization and signal dispersion, max signal is different							
Protein E	4		0.9		1		**3.6**
Protein F	4		0.9		0.2		**0.7**
Similar localization and max signal, but signal among cell types differs							
Protein G	4		0.9		1		**0.9**
Protein H	4		0.1		1		**0.4**

4.3 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.


GenieScore
Surface proteins that differ
IsoGenieScore
Surface proteins that are similar
OmniGenieScore
Proteins that differ
IsoOmniGenieScore
Proteins that are similar

4.4 Optional Annotations for Data Export

Several optional annotations are made available to be appended to user input datasets. Below is a table describing the availability for different species and the source of the information.

4.4.1 Table Summary

Included below is a summary of the source and availability of the various annotations for each species.

Export/Annotation Options	Source	Human	Mouse	Rat
HLA molecules	Manually curated	X		
CD molecules	UniProt	X	X	X
Gene Name	UniProt	X	X	X
Number of CSPA experiments	PMID: 25894527	X	X	
Transmembrane	UniProt	X	X	X
Subcellular Location	UniProt	X	X	X
UniProt Linkout	UniProt	X	X	X

4.4.2 Additional Details

Additional information for annotations, including the potential relevance to marker prioritization.

- HLA molecules: Human leukocyte antigen (HLA) molecules are surface proteins that have high sequence similarity. As such, it is often challenging to be certain of the specific gene product based solely on peptide-level evidence, particularly for Cell Surface Capture experiments. As a result, it may be useful to exclude these from consideration when attempting to identify cell surface makers for a specific cell type.
- CD molecules (CD): Cluster of Differentiation (CD) is a protocol used for the identification and investigation of cell surface molecules providing targets for immunophenotyping of cells. The proposed surface molecule is assigned a CD number once two specific monoclonal antibodies (mAb) are shown to bind to the molecule.
- Gene Name: Gene names are provided for human readability of potential markers.
- Number of CSPA experiments (CSPA): The number of cell types in which this protein was observed in the Cell Surface Protein Atlas. This information can provide context for how specific a protein might be among cell types.
- Transmembrane: Information about predicted transmembrane can help provide context for the localization and *SPC score* assigned to a protein.
- Subcellular Location: Gene Ontology - Cellular Component Annotations can help provide context for the localization and *SPC score* assigned to a protein.
- UniProt Linkout: Link to the UniProt entry for input proteins providing effortless access to additional information about candidate markers.