SurfaceGenie Data Upload Instructions

Data Format

A .csv file containing a list of proteins (UniProt Accession) and a surrogate value representative of abundance (*e.g.* number of peptide spectrum matches, peak area) identified within a set of samples.

The first column of your data file **must be labeled 'Accession'** with no extra characters (*e.g.* not 'Accession #'). This column should contain the UniProt accession numbers of the proteins in your samples. You may include isoforms. To convert from a different protein ID type to UniProt, bulk conversion is available [here](https://www.uniprot.org/uploadlists/) . Under 'Select options', select your ID type in the 'From' field and then 'UniProt KB' in the 'To' field.

Additionally, data files must be in **csv format**. If you are working in Excel, click 'File --> Save As' and select csv in the drop-down menu to convert from .xlsx to .csv.

Example Data

| **Accession** | **d00** | **d10** | **d13** | **d25** | **d31** | **d40** | **d50** | **d80** | **d90** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| A0AVT1-1 | 6 | 4 | 4 | 6 | 1 | 5 | 2 | 3 | 4 |
| A0FGR8-6 | 0 | 0 | 2 | 0 | 4 | 3 | 2 | 3 | 10 |
| A1L0T0 | 1 | 2 | 4 | 6 | 7 | 2 | 2 | 10 | 5 |
| A1X283 | 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| A5A3E0 | 0 | 0 | 56 | 54 | 59 | 52 | 55 | 51 | 55 |

Data Proccessing Options

Surface Protein Concensus (SPC) Score Consideration

If you are interested in finding cell surface markers, you will want to consider SPC score when calculating the Genie Score. This is the default setting. If you wish to ignore the SPC score for your proteins when generating Genie Scores, you may uncheck this option and the SPC score will be set to 1 for all proteins and will not be weighed into the Genie Score. You can confirm this in the 'CSV' tab and then uncheck 'SPC' in the export options to remove this from the download file. This is a feature designed to enable identification of molecules that may differ among cell types but that may be localized inside the cell.

HLA Molecule Exclusion

Human leukocyte antigen (HLA) molecules are typically found on the cell surface of most cell types and due to high sequence similarity among these proteins (*e.g.* HLA-A3 vs. HLA-A30), it is often challenging to be certain of the specific gene product based solely on peptide-level evidence. As a result, it may be useful to exclude these from consideration when attempting to identify cell surface makers for a specific cell type.

Find Markers for a Specific Sample

If you are interested in identifying markers that are present in a specific sample (*e.g.* positive selection marker for a cell type or experimental condition), SurfaceGenie can exclude proteins that are not observed in that sample. To do this, select the option “Find markers for specific sample”. A text box will then appear. In the text box, enter sample name of interest and make sure it exactly matches what is contained in the file header (i.e. 'd00' for the example dataset). If you have also selected to have SurfaceGenie group your samples (see 'Sample Grouping below') then you may also indicate a group (i.e. 'Group 1').

Sample Grouping

Ideally, similar samples such as technical replicates or biological replicates will have values averaged or summed together into a single column. However, SurfaceGenie will carry out this step for you if you select 'Group samples'. If this box is checked, you will need to provide the grouping method as well as the column numbers for each group. For example, If columns 2, 3, and 5 of your dataset should be grouped together and columns 4 & 6 comprise another group, you should indicate the presence of 2 groups using the slider and then enter the corresponding column numbers below separated by commas: Group 1: '2, 3, 5', Group 2: '4, 6'. Remember that column 1 will contain accession numbers and cannot be grouped with other columns.

Data Export Options

Plots

Several visualizations are made available by SurfaceGenie:

* SurfaceGenie Plot: SurfaceGenie scores plotted in order of priority for all proteins in a dataset.
* SPC Histogram: Shows the distribution of SPC scores
* Clustered Heatmap: Visualize the relationship among samples within a dataset based on the relative abundance measurement contained in the .csv file.
* Distribution Score: Shows the distribution of Genie Scores.

CSV File

You may select data to export as columns appended to the right of your original data. The following variables are available for export:

* Surface Protein Concensus score (SPC): A predictive measure of the likelihood that a particular protein can be present at the cell surface.
* Distribution Score: A measure of the distribution of the protein amongst samples. A higher value corresponds to a more localized distribution. [Wikipedia - Gini coefficient](https://en.wikipedia.org/wiki/Gini_coefficient)
* Signal strength (SS): A weighted value of the maximum value reported among samples for the protein
* Genie Score (GS): SurfaceGenie's measure for the value of a protein as a potential marker of interest.
* CD molecules (CD): Cluster of differentiation (CD) molecules.
* Number of CSPA experiments (CSPA-NE): ---
* UniProt Linkout: Link to the UniProt for information on the protein.