# Overview

This User Guide will cover all of the features available in the SurfaceGenie web application including some background on the theory and calculations. 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.

In developing SurfaceGenie we aimed to create an accessible tool for calculation of GenieScore and GenieScore components from input data. For all calculations, users will be 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](https://github.com/GundryLab/SurfaceGenie).

#### Before you begin

Currently, the functions on 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](https://www.uniprot.org/uploadlists/). Note that conversion between IDs is not always one-to-one. Manual curation of the results from the ID mapping is advisable.

Contents

[Overview 1](#_Toc24364049)

[GenieScore Calculation 3](#_Toc24364050)

[What is a GenieScore? 3](#_Toc24364051)

[Assumptions/Caveats 3](#_Toc24364052)

[Overview 4](#_Toc24364053)

[Input: 4](#_Toc24364054)

[Data Processing Summary: 4](#_Toc24364055)

[Output: 4](#_Toc24364056)

[GenieScore Permutations 5](#_Toc24364057)

[GenieScore Calculator Tutorial 6](#_Toc24364058)

[SPC Score Lookup 7](#_Toc24364059)

[What is SPC Score? 7](#_Toc24364060)

[Assumptions/Caveats 7](#_Toc24364061)

[Overview 7](#_Toc24364062)

[Input 7](#_Toc24364063)

[Output 7](#_Toc24364064)

[SPC Score Lookup Tutorial 8](#_Toc24364065)

[Additional Information 9](#_Toc24364066)

# GenieScore Calculation

## What is a GenieScore?

GenieScore is a metric that was designed to provide a single metric that could 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

## 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 are XX, meaning that 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.

## Overview

### Input:

SurfaceGenie accepts files (tsv, txt, csv, xlsx) containing a list of proteins (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. SurfaceGenie has SPC datasets for human, mouse, and rat.

### Calculations:

For each protein in the dataset, SurfaceGenie calculates the product of three independent scores:

#### Surface Protein Consensus (SPC) score

A predictive measure of the likelihood that a particular protein can be 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. For more details on the predictive datasets used, see the References tab.

#### Distribution Score

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 for calculating statistical dispersion of values. Scores range 0 – 1.

#### Signal Strength

An approximate 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 .

### Output:

* CSV Download - Columns of selected data types (e.g. SPC score, CD molecule annotation, etc) are appended to each entry in the original input file. More information regarding annotation can be found XX.
* Plots - Scores from each of the 4 permutations are plotted in order of priority for all proteins within a dataset
* SPC Histogram - Displays the distribution of SPC scores

## GenieScore Permutations

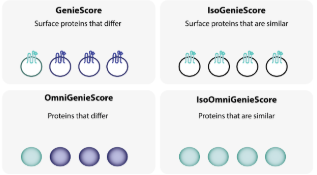
While a major benefit of GenieScore is the ability to prioritize proteins that are localized to the cell surface, it is also possible to analyze data without this parameter to find proteins of interest that reside in other subcellular localizations. See the descriptions for each of the four permutations of the scoring algorithm. GenieScore has been tested with semi-quantitative proteomic and transcriptomic datasets, but the underlying calculation may be useful in other contexts.

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.

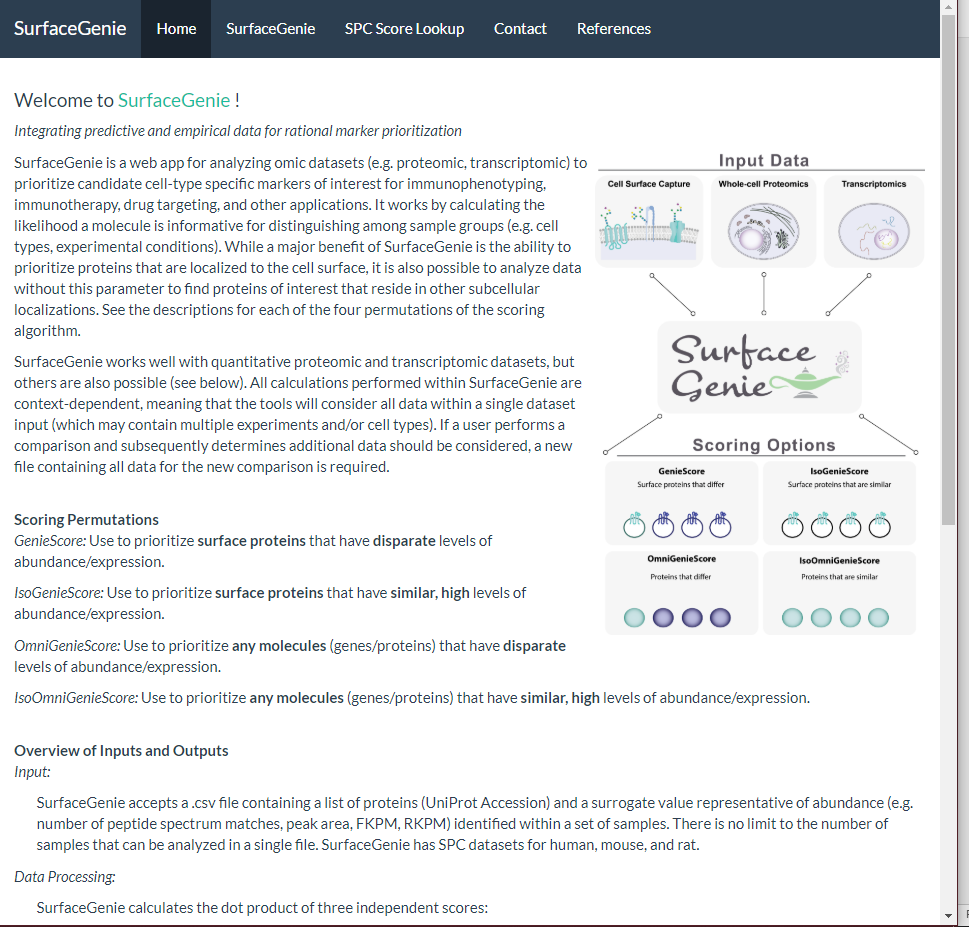
IsoOmniGenieScore: Use to prioritize **any molecules** (genes/proteins) that have **similar, high levels** of abundance/expression.



## GenieScore Calculator Tutorial

This tutorial assumes you have read the preceding information and therefore have some familiarity with the SurfaceGenie vocabulary.

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



# SPC Score Lookup

## What is SPC Score?

## Assumptions/Caveats

## Overview

### Input

### Output

## SPC Score Lookup Tutorial

# Additional Information

## Rationale and Calculation of GenieScore Components

## Optional Annotations for Data Export

CD molecules (CD) : Cluster of differentiation (CD) molecules are annotated with CD nomenclature. CD molecules have validated antibodies against them and therefore are attractive candidate markers for immunodetection -based applications.

HLA molecules (HLA) 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.

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.

UniProt Linkout : Link to the UniProt entry for input proteins providing effortless access to additional information about candidate markers.