Contents

1. Veneer Overview

Veneer is a web-based bioinformatic tool to rapidly and automatically assess and curate data from cell surface capture (CSC; *e.g.,* µCSC, autoCSC, classic CSC) and related workflows (*i.e., N*-glycoprotein or *N*-glycopeptide enrichment strategies including ligand receptor capture and methods that incorporate, biocytin hydrazide, aminooxy biotin, and alkoxyamine-PEG_4-biotin) that release captured glycopeptides by peptides-N(4)-(N-acetyl-beta-D-glycosaminyl)asparagine amidase F (PNGase F) treatment. **Veneer** employs a categorization scheme that first assigns levels of evidence (1, 2, 3) and then designates proteins to cell surface protein assignment tiers (high, medium, low, and zero). **Veneer** provides enhanced functional annotations contributing added biological insights with high relevance for cell surface proteins. All inputs and outputs are .xlsx, .xls, .csv, or .tsv file formats and, therefore, are agnostic of vendor or platform used for data acquisition or database searching. **Veneer** enhances consistency in data curation, reduces curation time, calculates key parameters used to assess the quality of experimental output and aid in troubleshooting, and annotates the dataset to facilitate biological interpretation.

2. Curate Data – Basics and Tutorial

2.1. Terminology

2.1.1. Sequon

The term sequon used in this workflow is analogous to "sequence motif" and "consensus motif" found in literature. Sequon refers to the *N*-glycosylation motif (N-X-S/T/C/V where X ≠ P). De-*N*-glycosylated peptides (derived from *N*-glycopeptides that have been treated with PNGase F to release the glycan from the asparagine) are identified by the presence of a deamidated asparagine in the sequon.

2.1.2. Levels of evidence

Level 1 evidence (experimental) is defined as at least one identified peptide containing a deamidation in the sequon. For the sequon, we include S/T/C/V in the third position for completeness and quality control assessment. Level 2 evidence (bioinformatic) is based on evaluating three prediction criteria: 1) *N*-gly-Compiled

Interactive Resource for Extracellular and Surface Studies (CIRFESS) score: indicates whether the protein sequence includes a non-cytoplasmic tryptic peptide of suitable size for MS that contains the NxS/T/C/V motif, 2) the surface protein consensus (SPC) score: indicates whether a protein is predicted to be a surface protein according to four algorithms; and 3) signal peptide prediction: from three algorithms (PrediSi, SignalP, and Phobius). If a protein meets any one of these criteria, level 2 evidence is assigned. Level 3 evidence (bioinformatic) requires a protein to contain at least one predicted extracellular domain in the topological domain annotation field in UniProt.

2.1.3. Surface protein assignment tier

Each protein is assigned a surface protein assignment tier of high, medium, low, or zero based on the levels of evidence present. High tier requires levels 1+2+3 or levels 1+3, medium tier requires level 1 + 2, and low tier are those with only level 1 evidence. Any protein not identified by at least peptide containing a deamidation in the sequon (level 1) is assigned tier zero.

2.2. Overview of Filter and Annotate Usage

2.2.1. Input

Annotate and Assign Evidence accepts .xlsx, .xls, .csv, or .tsv files containing a list of protein identifiers (UniProt Accession) and corresponding peptide spectrum matches (PSM). The column header of the first column must be labeled with *Master Protein Accessions*. The column header for the second column containing PSM must be labeled with *Annotated Sequence*. The PSM must be in the following format [R].TQDEILFSnSTR.[L], where flanking amino acids are in brackets and deamidation (release the glycan from the asparagine) is denoted as small letter n. PSM-level data is required for correct Veneer output. If peptide-level data is used instead of PSM-level data, calculations of PSM-level output as described below in section 3.4. (*e.g.,* total number of PSM with sequon) will be incorrect. Multiple files can be combined into a zip file. An example single and zip file can be downloaded from the Instructions page of Veneer. The input file cannot exceed 150 MB. If files exceed this limit, users are advised to download the application from Github and run it locally.

2.2.2. Output

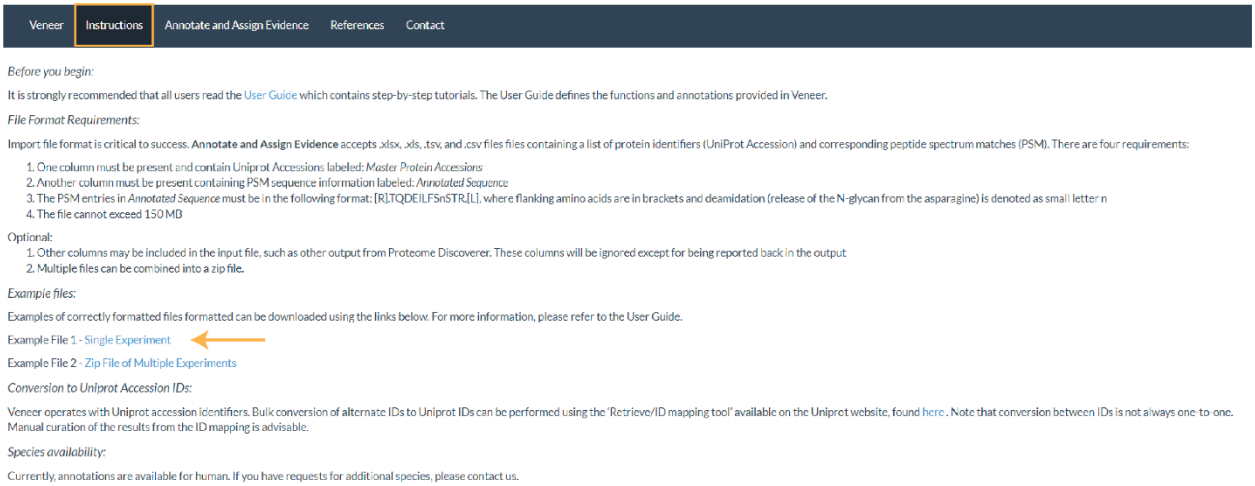
Users can view an overview of the number of identified high, medium, low, and zero proteins. User can export an xlsx file containing curated and annotated data.

2.3. Annotate and Assign Evidence: Quick-start Guide

2.4. Annotate and Assign Evidence: Tutorial

Before you begin:

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



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

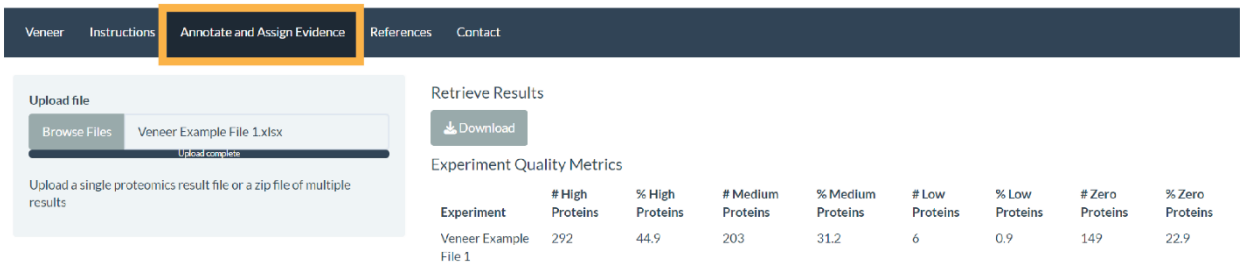
- File type: xlsx, xls, csv
- Species: Human
- Identifier: UniProt Accession ID
- Data: PSM
- ** The header of the first column must be *Master Protein Accessions* **
- ** The header of the second column must be *Annotated Sequence* **

Example of properly formatted file:

Master Protein Accessions	Annotated Sequence
Q9Y6X5	[K].YGPEDKEnMSR.[V]
Q9Y6X5	[K].YGPEDKEnMSR.[V]
Q9Y6X5	[K].YGPEDKEnMSR.[V]
Q9Y6X5	[R].IQPIILVADEGWTIVLnESSQK.[L]
Q9Y6X5	[R].IQPIILVADEGWTIVLnESSQK.[L]

Q9Y6N7	[S].HnASLEVAILR.[D]
Q9Y6N7	[K].nDGnGTAILVSWQPPPEDTQnGMVQEYK.[V]
Q9Y6N7	[R].NYLGEAVSHnASLEVAILR.[D]
Q9Y6N7	[R].NYLGEAVSHnASLEVAILR.[D]
Q9Y6N7	[R].NYLGEAVSHnASLEVAILR.[D]
P32942	[S].GmGWAAFnLSnVTGNSR.[I]
P32942	[R].VELAPLPPWQPVGQnFTLR.[C]
P32942	[R].VEPQNPVLSAGGSLFVncSTDcPSSEK.[I]
P32942	[R].VEPQNPVLSAGGSLFVncSTDcPSSEK.[I]
P32942	[K].ELVASGmGWAAFnLSnVTGNSR.[I]
P32942	[R].VELAPLPPWQPVGQnFTLR.[C]
Q96NT5	[R].FSADLGYnGTR.[Q]
P13760	[R].FLEQVKHEcHFFnGTER.[V]
P21163	[K].SSIDGVPYGKAH.[T]

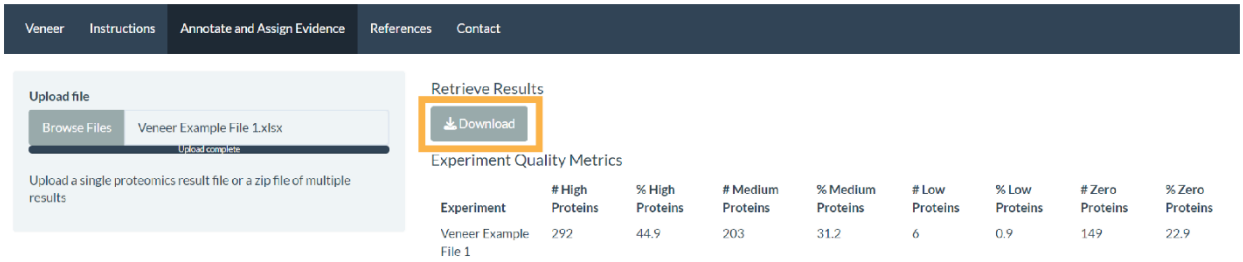
1. From the Home Page of ***Veneer***, click on the ***Annotate and Assign Evidence*** tab in the header bar.



2. Using the 'Browse Files' button, in the 'Upload file' section, navigate to and select the data file to be imported.

3. Once the data file has been imported and processed, the original file names with the number of high, medium, low, and zero proteins and other metrics identified will be visible for manual inspection in the 'Experiment Quality Metrics' pane on the website.

4. Click on 'Download' button to download results.



5. For each experimental file uploaded, there will be two output files annotated as *Filename*_protter and *Filename*_Veneer. The next section of this guide 3.1 "Veneer Output" explains what is contained in each file.

3. Veneer Output – *Filename*_Veneer

3.1. Abbreviations

Abbreviations	Explanation
MPA	Master protein accession
GO	Gene ontology
TM	Transmembrane
HPA	Human Protein Atlas
OMIM	Online Mendelian Inheritance in Man
CD	Cluster of differentiation
SPC	Surface prediction consensus
CIRFESS	Compiled Interactive Resource for Extracellular and Surface Studies
CSC	Cell surface capture

3.2. Terminology

3.2.1. Surface prediction consensus (SPC) score

Surface prediction consensus (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. For more details on the SCP score and predictive datasets used, see www.cellsurfer.net/surfacegenie.

3.2.2. *N*-gly-CIRFESS score

N-gly-CIRFESS score is a predictive measure that sums the total of all predicted non-cytoplasmic *N*-glycopeptides that could be detected in an *N*-glycocapture experiment. For more details on CIRFESS, see https://www.cellsurfer.net/cirfess. *N*-gly-CIRFESS score indicates whether the protein sequence includes a non-cytoplasmic (*i.e.,* portion of a protein that is predicted to be oriented toward a non-cytoplasmic space within the cell, including, but not limited to the extracellular, lumenal ER/Golgi, lysosomal, vesicular, and nuclear spaces) tryptic peptide of suitable size for MS that contains the N-X-S/T/C/V motif.

3.3. Annotation sources

Column Name	Source	Date Accessed (if applicable)	Reference (if applicable) and/or Weblink
Primary Gene Name	UniProt	10/2018	www.uniprot.org
SPC score	SurfaceGenie	N/A	PMID: 32053146 www.cellsurfer.net/surfacegenie
N-gly-CIRFESS score	CIRFESS	N/A	PMID: 32212654 www.cellsurfer.net/cirfess
CD Annotation	UniProt	10/2018	PMID: 3294157
Number of TM Domains	UniProt	10/2018	www.uniprot.org
SOSUI TM Domain Prediction	SOSUI v1.1	03/2019	PMID: 9632836
TMHMM TM Domain Prediction	TMHMM Server v2.0	03/2019	PMID: 11152613
Phobius TM Domain Prediction			PMID: 15111065
Number of Glycosites	UniProt	03/2019	www.uniprot.org/
Number of CRAPome Experiments	CRAPome v1.1	03/2019	PMID: 23921808
Secondary Gene Names	UniProt	10/2018	www.uniprot.org
Protein Names	UniProt	10/2018	www.uniprot.org
Keywords	UniProt	10/2018	www.uniprot.org
Gene Ontology	UniProt	10/2018	www.uniprot.org
Transmembrane	UniProt	10/2018	www.uniprot.org
Subcellular Location	UniProt	10/2018	www.uniprot.org
Ensembl Transcript	UniProt	10/2018	www.uniprot.org
Glycosylation	UniProt	10/2018	www.uniprot.org
GlyConnect	GlyConnect	08/2019	PMID: 30574787 https://glyconnect.expasy.org/
Signal Peptide - PrediSi	PrediSi	N/A	http://www.predisi.de/
Signal Peptide - SignalP	SignalP-5.0	N/A	PMID: 30778233 https://services.healthtech.dtu.dk/service.php?SignalP-5.0
Signal Peptide - Phobius			PMID: 15111065
DrugBank	DrugBank;	10/2018	PMID: 18048412

	UniProt		https://go.drugbank.com/
OMIM	OMIM; UniProt	10/2018	https://www.ncbi.nlm.nih.gov/omim
HPA Weblink	Human Protein Atlas	10/2018	PubMed: 16127175 https://www.proteinatlas.org/
UniProt Website	UniProt	10/2018	www.uniprot.org
CellMarker Website	CellMarker	10/2018	PMID: 30289549 http://biocc.hrbmu.edu.cn/CellMarker/

3.4. Veneer Worksheet Tabs

3.4.1. High, Medium, Low, and Zero proteins

Column header	Explanation
MPAnoIso	MPA with isoform information removed.
MPA	MPA with isoform information.
numPep	Total number of peptides identified per MPA.
numPSM	Total number of PSM identified per MPA.
psmExclusive	Total number of PSM identified exclusively per MPA.
pctExclusive	Ratio of total number of PSM identified exclusively per MPA (psmExclusive) divided by total number of PSM identified per MPA (numPSM).
PSMwSequon	Total number of PSM identified with deamidation in sequon.
pctPSMwSequon	Ratio of total number of PSM identified with deamidation in the sequon (PSMwSequon) divided by total number of PSM identified per MPA (numPSM).
SequononePSM	If MPA was identified by only one PSM with deamidation in the sequon, it will be marked “1”. If MPA was identified by more than one PSM, it will be marked “0”.
countNG	Total number of PSM identified where G = X in sequon with deamidation (N-X-S/T/C/V).
pctNG	Ratio of total number of PSM identified where G = X in sequon with deamidation (countNG) divided by total number of PSM identified with deamidation in sequon (PSMwSequon).
countMultiN	Total number of PSM that contained two asparagine (includes both deamidated and non-deamidated).
countNXT	Total number of PSM identified with deamidation in N-X-T sequon.
pctNXT	Ratio of total number of identified with deamidation in N-X-T sequon (countNXT) divided by total number of PSM identified with deamidation in sequon (PSMwSequon).
countNST	Total number of PSM identified with deamidation in N-X-S sequon.

pctNST	Ratio of total number of identified with deamidation in N-X-S sequon (countNXS) divided by total number of PSM identified with deamidation in sequon (PSMwSequon).
countNCT	Total number of PSM identified with deamidation in N-X-C sequon.
pctNCT	Ratio of total number of identified with deamidation in N-X-C sequon (countNXC) divided by total number of PSM identified with deamidation in sequon (PSMwSequon).
countNVT	Total number of PSM identified with deamidation in N-X-V sequon.
pctNVT	Ratio of total number of identified with deamidation in N-X-V sequon (countNXV) divided by total number of PSM identified with deamidation in sequon (PSMwSequon).
Level of Evidence	Each protein is categorized based on levels of evidence (1, 2, 3). See section 2.1.2.
Assignment	Each protein is assigned a surface protein assignment tier of high, medium, low, or zero based on the levels of evidence present. See section 2.1.3.
Primary.Gene.Name	Primary gene name related to MPA as annotated by UniProt.
Secondary.Gene.Names	Secondary gene name related to MPA as annotated by Uniprot.
Protein.Names	Protein name related to MPA as annotated by UniProt.
Keywords..Uniprot.	Keywords related to MPA as annotated by UniProt.
Gene.Ontology..GO...Uniprot.	GO terms related to MPA as annotated by UniProt.
Transmembrane..Uniprot.	Transmembrane domain as annotated by UniProt.
Number.of.TM.Domains..Uniprot.	Number of transmembrane domains as annotated by UniProt.
Topological.domain	Topology domain related to MPA as annotated by UniProt.
Subcellular.Location..Uniprot.	Subcellular location related to MPA as annotated by UniProt.
Ensembl.Transcript..Uniprot.	Ensemble gene annotation.
DrugBank	DrugBank codes exported as annotated by UniProt.
Glycosylation..Uniprot.	Position of documented or predicted glycosylation as annotated by UniProt.
Number.of.Glycosites..Uniprot.	Number of glycosylation sites annotated in UniProt.
Drug.Target...HPA.	If MPA was identified as a drug target in HPA, it will be marked “Yes”. If MPA was not identified as a drug target in HPA, it will be marked “No”.
HPA.Weblink	Weblink to HPA for a given MPA.
Uniprot.Website	Weblink to UniProt for a given MPA.
CellMarker.Website	Weblink to CellMarker for a given MPA.

OMIM	Identification numbers for a given MPA as annotated by OMIM.
SOSUI.TM.Domain.Prediction	Number of TM domains predicted using SOSUI algorithm.
TMHMM.TM.Domain.Prediction	Number of TM domains predicted using the TMHMM algorithm.
Phobius.TM.Domain.Prediction	Number of TM domains predicted using the Phobius algorithm.
Number.of.Crapome.Experiments	The number of experiments in which a protein was identified in the Contaminant Repository for Affinity Purification database (CRAPome). If this number is high, it can be suggestive of a protein being particularly 'sticky' and a common non-specific binder to streptavidin resin that is used for glycopeptide enrichment.
Number.of.Crapome.Experiments..Streptavidin.	The number of experiments in which a protein was identified in the Contaminant Repository for Affinity Purification database (CRAPome). If this number is high, it can be suggestive of a protein being particularly 'sticky' and a common non-specific binder to streptavidin resin that is used for glycopeptide enrichment.
GlyConnect.ID	Identification numbers for a given MPA as annotated by Glyconnect.
Signal.Peptide..PredSi.	Indicates whether the MPA contains a predicted signal peptide that is common for proteins destined towards the secretory pathway. Prediction from PrediSi algorithm.
Signal.Peptide..SignalP.	Indicates whether the MPA contains a predicted signal peptide that is common for proteins destined towards the secretory pathway. Prediction from SignalP algorithm.
Signal.Peptide..Phobius.	Indicates whether the MPA contains a predicted signal peptide that is common for proteins destined towards the secretory pathway. Prediction from Phobius algorithm.
CD.Number	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 are shown to bind to the molecule.
SPC	See above section *3.2. Terminology* (SPC score).
cirfessScore	See above section *3.2. Terminology* (*N*-gly-CIRFESS score).

3.4.2. High, Medium, Low, and Zero peptides

Column header	Explannation
pepSeq	Amino acid sequence of the peptide identified including flanking residues of the trypsin cleavage site.
MPA	MPA with isoform information.
MPAnoIso	MPA with isoform information removed.
MPAnonSplit	Original MPA assignment(s) of the PSM the peptide was derived from.
numMPA	Indicates whether the peptide was mapped to a single MPA or multiple MPA.
protPSMs	Total number of PSM identified per MPA.
protPctPSMs	Ratio of unique PSM identified per MPA divided by the total number of PSM identified per MPA (protPSMs).
protExclusive	Total number of PSM identified exclusively per MPA.
protPctExclusive	Ratio of PSM identified exclusively per MPA (protExclusive) divided by total number of PSM identified per MPA (protPSMs).
protPSMsSequon	Total number of PSM identified with deamidation in the sequon per MPA.
protPctPSMsSeuqon	Ratio of PSM identified with deamidation in the sequon per MPA (protPSMsSequon) divided by total number of PSM identified per MPA (protPSMs).
protSequononePSM	If MPA was identified by only one PSM with deamidation in the sequon, it will be marked “1”. If MPA was identified by more than one PSM, it will be marked “0”.
pepPSM	Total number of PSM for unique peptides.
pepPSMwSequon	Total number of PSM for unique peptides with deamidation in sequon.
pctPepPSMwSequon	Ratio of total number of PSM for unique peptide with deamidation in sequon (pepPSMwSequon) divided by total number of PSM for unique peptides with Sequon (pepPSM).
hasSequon	Indicates whether the peptide sequence contains a deamidation in sequon.
counNG	Total number of PSM identified where G = X in sequon with deamidation (N-X-S/T/C/V).
countMultiN	Total number of PSM that contained two asparagine (includes both deamidated and non-deamidated).

3.4.3. High, Medium, Low, and Zero PSMs

Column header	Explanation
Master Protein Accessions	MPA as entered in the input file uploaded to Veneer.
Annotated Sequence	Annotated sequence as entered in the input file uploaded to Veneer.
hasSequon	Indicates if PSM has deamidation in sequon. If PSM has deamidation sequon, it will be marked “1”. If PSM does not have deamidation in sequon, it will be marked “0”.
motifs	List which motif was identified in PSM.
hasNG	Indicates if PSM contains G = X in sequon with deamidation (N-X-S/T/C/V). If yes, it will be marked “1”. If no, it will be marked “0”.

hasMultiN	Indicates if PSM contains contained two asparagine (includes both deamidated and non-deamidated). If yes, it will be marked “1”. If no, it will be marked “0”.
pepSeq	Peptide sequence not containing flanking amino acids in brackets.
annSeq	Annotated sequence as entered in the input file uploaded to Veneer.
MPAnonSplit	Original MPA assignment(s) of the PSM the peptide was derived from.
numMPA	Indicates whether the peptide was mapped to a single MPA or multiple MPA.
MPA	MPA with isoform information.
MPAnoIso	MPA with isoform information removed.

3.4.4. Reagent Analysis

This worksheet reports information regarding three proteins used in *N*-glycocapture experiments: trypsin, streptavidin, and PNGase F. These are not cell surface proteins but are included in the database search to promote accurate false-discovery rate calculation because their peptides are commonly present in the final sample. The information in this worksheet can be helpful when troubleshooting *N*-glycocapture experiments as it provides a metric to assess how successfully trypsin is inactivated prior to the peptide mixture being added to the streptavidin beads. For this feature to work, FASTA files containing sequences for these reagents need to be included in the database search. Veneer will consider peptides matched to the following accession numbers when calculating reagents:

P21163: PNGase F (Elizabethkingia miricola)
Q9XBM8: PNGase F (Elizabethkingia meningoseptica)
P22629: Streptavidin (Streptomyces avidinii)
P00761: Trypsin (Sus scrofa)

3.4.5. Sequon Analysis

This worksheet provides an overview of the sequence motifs observed within a dataset. The worksheet reports the total number of PSM containing a deamidated sequon within the dataset. Also listed are the number of observations of each sequon type (*i.e.,* N-X-S, N-X-T, N-X-C, N-X-V, where X ≠ P) across the dataset. The calculation is made at the PSM level. A PSM is included in the count if at least one PSM with a deamidation in the sequon is found within the corresponding peptide sequence. If a PSM contains two deamidated sequons, each motif will be counted.

3.4.6. Specificity

Specificity of an *N*-glycocapture experiment is calculated at the protein and PSM level (number PSMs with deamidation in the sequon / total number of PSM) and protein level (number of proteins identified by at least one PSM with a deamidation in the sequon / total number of proteins). In an *N*-glycocapture experiment, the de-*N*-glycosylated peptides are indicative of those peptides that were localized to the cell surface at the time of labeling. This is calculated for each surface protein assignment tier. It also highlights the total number of proteins identified by just one PSM containing a deamidation in the sequon and the total number of PSM where G = X in the sequon (N-X-S/T/C/V).

3.4.7. GO Terms (UniProt)

This worksheet contains a parsed list of all GO Terms described for high, medium, and low proteins in a convenient list format that can be used for further downstream analyses and generation of graphs.

3.4.8. Keywords (UniProt)

This worksheet contains a parsed list of all keywords described for high, medium, and low proteins in a convenient list format that can be used for further downstream analyses and generation of graphs.

4. Veneer Output – *Filename*_Protter

4.1. Protter Worksheet

This is a tsv file that can be directly uploaded into Protter (https://wlab.ethz.ch/protter/start/) to visualize the experimentally identified peptides within the protein sequence as it relates to topology, thereby, providing a rapid strategy to view the extracellular domains of proteins and inform epitope selection for antibody or other targeting.