

Contents

1. Veneer Overview	1
2. Curate Data – Basics and Tutorial	1
3. Veneer Output – <i>Filename_Veneer</i>	5
4. Veneer Output – <i>Filename_Protter</i>	10

1. Veneer Overview

Veneer is an automated web-based bioinformatic tool to rapidly assess and curate data from cell surface capture (CSC; *e.g.*, μ CSC, 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-PEG₄-biotin) that release captured glycopeptides by peptides-N(4)-(N-acetyl-beta-D-glycosaminyl)asparagine amidase F (PNGase F) treatment. **Veneer** provides enhanced functional annotations contributing added biological insights with high relevance for cell surface proteins. All inputs and outputs are Excel 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. Sequence consensus motif (SCM)

The term "sequence consensus motif" (SCM) used in this workflow is analogous to "sequence motif" and "consensus motif" found in literature. SCM refers to the *N*-glycosylation motif (nXS/T/C/V where X is any amino acid except P). De-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 within the SCM.

2.1.2. Non-specific binder (NSB)

The term "non-specific binders" (NSB) used in this workflow refers to proteins not identified by formerly *N*-glycosylated peptides.

2.2. Overview of Filter and Annotate Usage

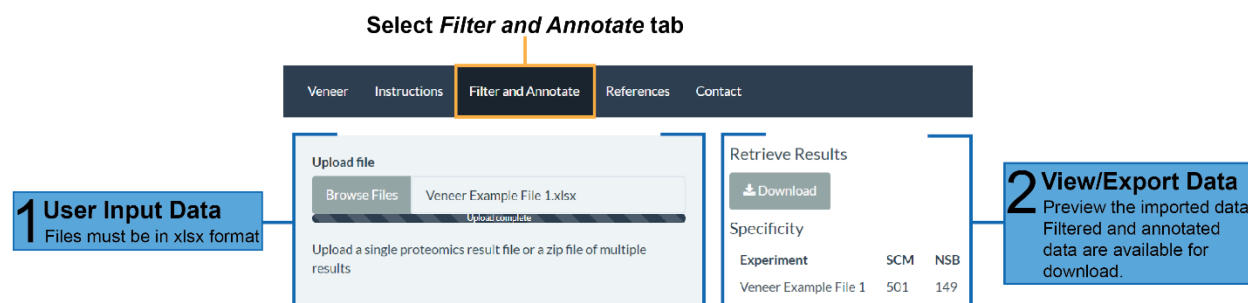
2.2.1. Input

Filter and Annotate accepts xlsx 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. Multiple xlsx files can be combined into a zip file. The file cannot exceed over 50 MB. An example single and zip file can be downloaded from the Instructions page of Veneer.

2.2.2. Output

Users can see an overview of the number of identified SCM and NSB proteins. User can export an xlsx file containing curated and annotated data.

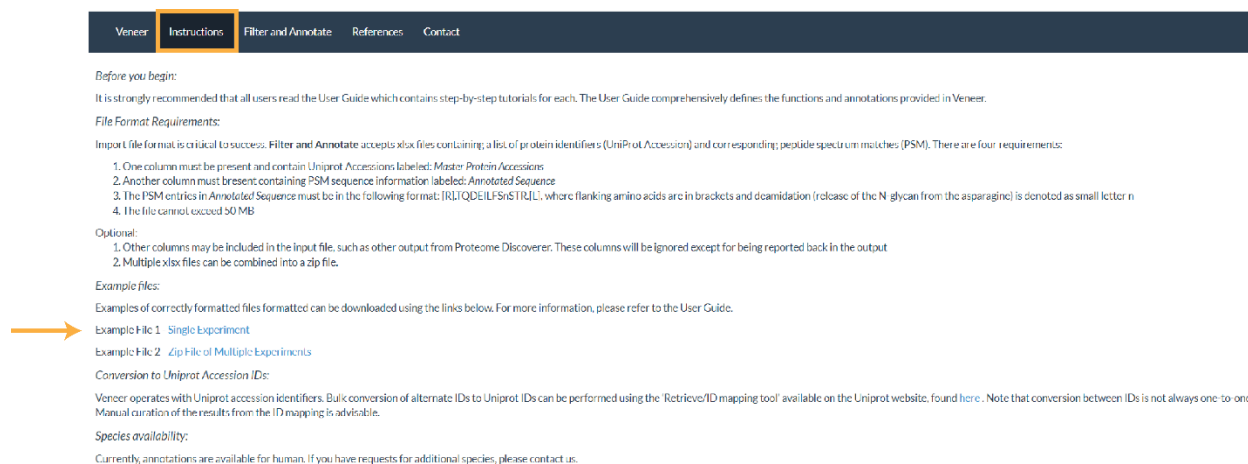
2.3. Filter and Annotate Quick-start Guide



2.4. Filter and Annotate Tutorial

Before you begin:

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



Before you begin:
It is strongly recommended that all users read the User Guide which contains step-by-step tutorials for each. The User Guide comprehensively defines the functions and annotations provided in Veneer.

File Format Requirements:
Import file format is critical to success. Filter and Annotate accepts xlsx files containing a list of protein identifiers (UniProt Accession) and corresponding peptide spectrum matches (PSM). There are four requirements:

1. One column must be present and contain UniProt Accessions labeled: *Master Protein Accessions*
2. Another column must be present containing PSM sequence information labeled: *Annotated Sequence*
3. The PSM entries in *Annotated Sequence* must be in the following format: [R].TQDEILFSnSTR.[L], where flanking amino acids are in brackets and deamidation (release of the N glycan from the asparagine) is denoted as small letter n
4. The file cannot exceed 50 MB

Optional:

1. Other columns may be included in the input file, such as other output from Proteome Discoverer. These columns will be ignored except for being reported back in the output
2. Multiple xlsx files can be combined into a zip file.

Example files:
Examples of correctly formatted files formatted can be downloaded using the links below. For more information, please refer to the User Guide.

Example File 1 [Single Experiment](#)

Example File 2 [Zip File of Multiple Experiments](#)

Conversion to UniProt Accession IDs:
Veneer operates with UniProt accession identifiers. 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, annotations are available for human. If you have requests for additional species, please contact us.

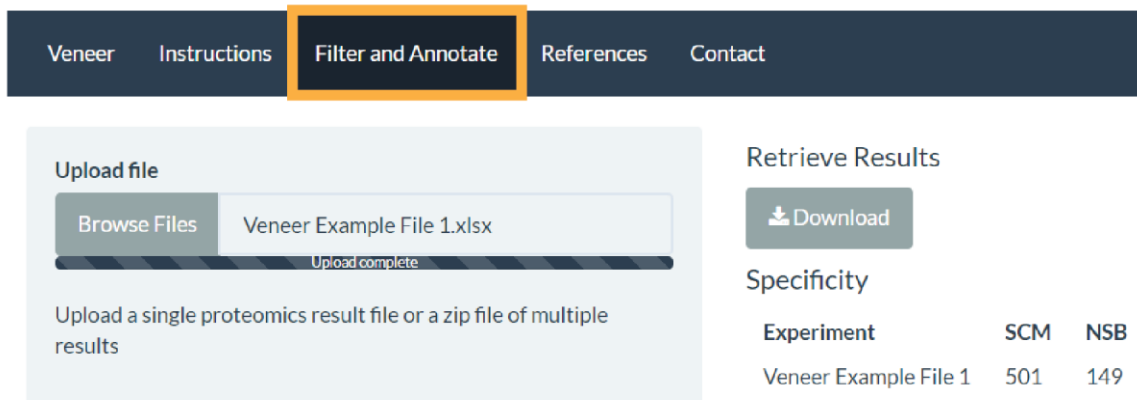
Alternatively, users can follow the steps with their own data provided it conforms to the following specifications:

- File type: xlsx
- 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
P01871	[K].THTnISESHpNATFSaVGEASlcEDDWNSGER.[F]
P32942	[S].GmGWAAFnLSnVTGNSR.[I]
Q96NT5	[R].FSADLGYNgTR.[Q]
P13760	[R].FLEQVKHEcHFFnGTER.[V]
P21163	[K].SSIDGVpYgKAH.[T]

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



Experiment	SCM	NSB
Veneer Example File 1	501	149

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

Veneer Instructions **Filter and Annotate** References Contact

Upload file

Browse Files

Veneer Example File 1.xlsx

Upload complete

Upload a single proteomics result file or a zip file of multiple results

Retrieve Results

Download

Specificity

Experiment	SCM	NSB
Veneer Example File 1	501	149

- Once the data has been imported and processed, the original file names with the number of SCM and NSB proteins identified will be visible for manual inspection in the 'Data Preview' pane.

Veneer Instructions **Filter and Annotate** References Contact

Upload file

Browse Files

Veneer Example File 1.xlsx

Upload complete

Upload a single proteomics result file or a zip file of multiple results

Retrieve Results

Download

Specificity

Experiment	SCM	NSB
Veneer Example File 1	501	149

- Click on 'Download' button to download results.

Veneer Instructions **Filter and Annotate** References Contact

Upload file

Browse Files

Veneer Example File 1.xlsx

Upload complete

Upload a single proteomics result file or a zip file of multiple results

Retrieve Results

Download

Specificity

Experiment	SCM	NSB
Veneer Example File 1	501	149

- For each experimental file uploaded, there will be two output files annotated as *Filename_protter* and *Filename_Veneer*. In the section "Veneer Output", we will explain 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

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 the sum total of all predicted extracellular N-glycopeptides that could be detected in a CSC experiment. For more details on CIRFESS, see <https://www.cellsurfer.net/cirfess>.

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			
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			
DrugBank	DrugBank; UniProt	10/2018	PMID: 18048412 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. SCM proteins, SCM- filtered, and NSB proteins

Column header	Explanation
MPAnolso	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	Percentage of PSM identified exclusively per MPA
PSMwSCM	Total number of PSM identified with SCM
pctPSMwSCM	Percentage of PSM identified with SCM
SCMonePSM	If MPA has only one PSM with SCM, it will be marked "1". If MPA has more than one PSM, it will be marked "0".
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
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. <i>Terminology</i>
N-gly-CIRFESS	See above section 3.2. <i>Terminology</i>

3.4.2. SCM and NSB peptides

Column header	Explanation
pepSeq	Amino acid sequence of the peptide identified including flanking residues of the trypsin cleavage site.
MPA	MPA with isoform information
MPAnolso	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	Percentage of PSM identified per MPA
protExclusive	Total number of PSM identified exclusively per MPA
protPctExclusive	Percentage of PSM identified exclusively per MPA

protPSMsSCM	Total number of PSM identified with SCM per MPA
protPctPSMsSCM	Percentage of PSM identified with SCM per MPA
protSCMonePSM	If MPA has only one PSM with SCM, it will be marked "1". If MPA has more than one PSM, it will be marked "0".
pepPSM	Number of PSM for unique peptides.
pepPSMwSCM	Number of PSM for unique peptides with SCM.
pctPepPSMwSCM	Percentage of unique peptide PSMs with SCM.
hasSCM	Indicates whether the peptide sequence contains a SCM.

3.4.3. SCM and NSB 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
hasSCM	Indicates if PSM has SCM. If PSM has SCM, it will be marked "1". If PSM does not have SCM, 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
MPAAnonSplit	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
MPAAnIso	MPA with isoform information removed

3.4.4. Reagent Analysis

This worksheet reports information regarding three proteins used in CSC 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 in troubleshooting CSC experiments as it provides a metric to assess how successfully trypsin is inactivated prior to the peptide mixture being added to the streptavidin beads.

3.4.5. Motif Analysis

This worksheet provides an overview of the sequence motifs observed within a dataset. The worksheet reports the number of observations of each SCM type (*i.e.*, nXS, nXT, nXC, nXV) across the dataset. The calculation is made at the PSM level. A PSM is included in the count if at least one SCM is found within the corresponding peptide sequence. A PSM is not counted twice if it has two of the same SCM.

3.4.6. Specificity

Specificity of a CSC experiment is calculated at the protein and PSM level (# PSMs with deamidation within SCM / total number of PSM) and protein level (# proteins identified by at least one PSM with a deamidation within SCM / total number of proteins). In a CSC experiment, the de-glycosylated peptides are indicative of those peptides that were localized to the cell surface at the time of labeling (biotinylation).

3.4.7. GO Terms (UniProt)

This worksheet contains a parsed list of all GO Terms described for Proteins (SCM) 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 Proteins (SCM) 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.