Chapter 9

Surface Proteome Biotinylation Combined with Bioinformatic Tools as a Strategy for Predicting Pathogen Interacting Proteins

Anita Horvatić, Josipa Kuleš, Nicolas Guillemin, Franjo Martinković, Iva Štimac, Vladimir Mrljak, and Mangesh Bhide

Abstract

Constant advancements in methodology and mass spectrometry instrumentation, genome sequencing and bioinformatic tools have enabled the identification of numerous pathogen proteomes. Identifying the pathogen interacting proteins by means of high-throughput techniques is key for understanding pathogen invasion and survival mechanisms and in such a way proposing specific proteins as pharmaceutical targets. Herein we describe the methodology for the enrichment and identification of pathogen surface proteome using cell surface protein biotinylation followed by LC-MS/MS and bioinformatic analyses of such data. This strategy is to be employed for the determination of protein subcellular localization and prediction of potential pathogen interacting proteins.

Key words Biotinylation, LC-MS, Surface proteome, Bioinformatics, Subcellular localization, Interacting proteins, DAVID, CELLO

1 Introduction

Defining the cell surface proteome has profound importance for understanding host-pathogen interactions. Pathogen plasma membrane proteins (PM) that reside on the cell surface regulate and directly interact with host cells proteins during recognition and invasion process influencing on immune response of host organism [1]. Furthermore, as PMs are involved in ion transport, cell signaling and communication, this makes them ideal targets for various therapeutics and promising vaccine candidates [2]. Owing to their hydrophobic nature, plasma membrane proteins pose analytical challenges and, despite efforts to overcome difficulties, remain under-represented in proteomic studies. The most critical component of the experimental approach is the enrichment and purification of plasma membrane proteins [3]. The most commonly used

Carlos Medina and Francisco Javier López-Baena (eds.), *Host-Pathogen Interactions: Methods and Protocols*, Methods in Molecular Biology, vol. 1734, https://doi.org/10.1007/978-1-4939-7604-1_9,

techniques for enrichment and extraction of membrane proteins are protein shaving, biotinylation followed by (strept)avidin affinity chromatography, and ultracentrifugation. The availability of novel -omics technologies coupled to high-throughput protein expression and purification, and bioinformatic tools together with -omics databases availability enables more rational and faster identification of antigens among large number of pathogen proteins [4]. Antigen identification represents the most important bottleneck in vaccine development against any pathogen, as this was usually achieved through rather empirical, time-consuming, and labor-intensive in vivo and in vitro experiments [5].

Chemical labeling of cell surface proteins is an emerging technology for the isolation of target proteins containing specific residues which can subsequently be resolved from untagged proteins using affinity purification. Biotinylation of cell surface proteins is a method of choice for the selective capture of plasma membrane proteins, but it is limited to pathogens that can be cultivated in protein-free media. The procedure involves selective, covalent labeling of proteins with a biotinylation reagent followed by capture of biotin-conjugated proteins/peptides via an avidin/streptavidin-coated solid support (i.e., resins, magnetic beads, microtiter plates and chips). Unbound components (nontagged proteins) are washed away and captured proteins are eluted or detached under various conditions.

Chemical derivatization of reactive groups in proteins with a biotin moiety is one of the most widely used techniques in protein biochemistry. Biotinylation reagents typically consist of three components: the biotin moiety, a spacer—possibly containing a cleavable linker unit—and a reactive moiety that interacts with the proteins of interest [6]. Selection of the most suitable reagent should consider the following factors: water solubility and membrane impermeability, presence of a cleavable linker, size of the spacer, target functional group on the protein and binding characteristics of the biotin moiety. The highly stable interaction between biotin and avidin ($Kd = 10^5$ M) presents a drawback for this method, as elution of biotin-labeled proteins from the avidin support is difficult. In an attempt to resolve this problem, a disulfide bridge in the linker region of the biotinylation reagent has been introduced (sulfo-NHS-SS-biotin). Under reducing conditions, the disulfide bridge is cleaved, thus removing the biotin label and releasing the captured proteins/peptides.

Low membrane protein concentration, low yield of biotinylation, as well as molecular weight and hydrophobicity of membrane proteins requires very sensitive and high resolution instrumentation. For that reason, nanoLC-MS/MS, as a high-throughput analysis technique, using a bottom-up proteomic approach is the method of choice for the analysis of biotinylated surface proteins. Both strategies, shotgun and gel-based proteomic approach, can be

employed, having in mind protein amount and detergent (originating from lysis buffer) removal prior to LC-MS analysis. Commonly used detergents for the extraction of membrane proteins are Triton X-100, CHAPS, SDS, sodium deoxycholate, NP-40, etc., which cause interferences during LC-MS analysis resulting in low number of identified proteins. Depending on detergent type, methods such as dialysis, ultrafiltration, strong cation exchange and/or reverse phase chromatography, or detergent removal resins can be applied for detergent removal [7]. The other gel-based approach, mostly for detergent removal, includes tube gels or SDS-PAGE followed by in gel digestion and LC-MS analysis [8]. The quality of proteomic data due to the low abundance of biotinylated proteins, inadequate sample preparation or processing can result in false positive or negative results.

High-throughput methodologies, such as LC-MS, produce big datasets and identified proteins might differ in confidence. Among that, due to nonspecific binding not all enriched proteins are actually surface membrane proteins. For that reason, bioinformatics is inevitable for in silico data validation, filtering and database mining. There are different computational programs available for subcellular localization prediction, such as CELLO, BaCeILo, TargetP, and PSORTb, using various algorithms based on a decision tree of several support vector machines (SVMs), protein functional domains and/or the amino acid compositional differences in proteins from different subcellular locations [9–11]. Gene ontology (GO) analysis, interaction prediction and enrichment, as well as pathway analysis can be performed using open access platforms such as Cytoscape and its plugins or DAVID, depending on organism of interest and availability of its databases. Currently available computational approaches for predicting interacting proteins are based on genomic and structural information, use of network topology, literature mining/database search and machine learning algorithms utilizing heterogeneous -omics features [4]. Except for biotinylated proteins, bioinformatic tools can be also applied for the data analysis of any kind of proteomic results (identified proteins from cell lysates, enriched membrane proteins, etc.) in order to predict subcellular localization and interacting proteins (domains).

The isolation of surface membrane proteins of *Leishmania infantum* will be used as an example of how cell surface protein biotinylation with streptavidin affinity separation can be used for assessing pathogen interacting proteins. After subsequent tryptic digestion and LC-MS acquisitions, data can be processed using Proteome Discoverer. Further bioinformatic data filtering by CELLO and DAVID can be employed to determine subcellular localization, gene ontology, and potential interaction proteins using domain prediction.

2 Materials

2.1 Equipment

1. Cooling centrifuge.
2. Dry incubator shaker for small tubes.
3. NanoDrop spectrophotometer.
4. nanoLC-MS system (Dionex Ultimate 3000 RSLS nano flow system; Thermo Scientific Orbitrap Q Exactive Plus mass spectrometer).
5. Rotator.
6. Sonicator.
7. Vacuum concentrator.
8. Vortex.
9. −80 °C freezer.
10. Microscope.

2.2 Chemicals and Consumables

1. Acetonitrile (LC-MS grade).
2. Ammonium bicarbonate.
3. Ammonium hydroxide.
4. Dithiothreitol (DTT).
5. EZ-Link^{TM}Sulfo-NHS-SS-Biotin.
6. Formic acid (LC-MS grade).
7. Iodoacetamide.
8. Water (LC-MS grade).
9. Methanol.
10. NeutrAvidin agarose resin.
11. Spin columns (empty 800 μL spin columns).
12. Trypsin gold, porcine.
13. ZipTips (SCX, RP C18).

2.3 Solutions

1. Phosphate-saline buffer (PBS 1×; for 1 L): 8 g NaCl, 0.201 g KCl, 1.42 g Na_2HPO_4, 0.272 g KH_2PO_4. Adjust pH = 7.4.
2. Lysis buffer: Commercial RIPA buffer.
3. Quenching solution: 100 mM glycine in PBS.
4. Elution buffer: 50 mM DTT in ammonium bicarbonate.
5. Mobile phases for LC-MS (A—0.1% formic acid in water; B—80% acetonitrile/0.1% formic acid in water).
6. Solutions for strong cation exchange (SCX) chromatography: W1—0.1% formic acid in water; W2—50% methanol in water; E1—5% ammonium hydroxide–30% methanol in water.

2.4 Bioinformatic Tools

1. Proteome Discoverer (Thermo Scientific).
2. CELLO: sub**CEL**lular **LO**calization predictor (http://cello.life.nctu.edu.tw/).
3. **D**atabase for **A**nnotation, **V**isualization and **I**ntegrated **D**iscovery (**DAVID**) (https://david.ncifcrf.gov/home.jsp).

3 Methods

3.1 Cell Surface Protein Biotinylation Protocol

Cell surface biotinylation was performed on purified promastigotes from stationary phase culture of *Leishmania infantum*, but can be performed on any other cell type by optimizing cell concentration and lysis buffer/conditions.

1. Wash cells three times with PBS (pH = 7.4) and centrifuge at 1000 × *g* for 1 min to remove any contaminating proteins.
2. Suspend cells at a concentration of 10^6–10^7 cells/mL in PBS.
3. Immediately before use, prepare a 10 mM solution of Sulfo-NHS-SS-Biotin. Add the appropriate volume of the Sulfo-NHS-SS-Biotin solution to the cells suspension (*see* **Note 1**).
4. Incubate reaction mixture at room temperature for 40 min with gently rotation on the rotator or rocking on the orbital shaker (*see* **Note 2**).
5. Quench the reaction by adding 100 μL of 100 mM glycine solution in PBS. Wash cells two more times with ice-cold PBS to remove nonreacted biotinylation reagent (*see* **Note 3**).
6. Centrifuge cells in a benchtop centrifuge 1 min at 500 × *g*, discard the supernatant, and add the lysis buffer of choice to the cell pellet (*see* **Note 4**).
7. Lyse cells by two cycles of freezing at −80 °C and thawing at room temperature, followed by 10 cycles of sonication at maximum amplitude. Check the degree of cell lysis microscopically.
8. Centrifuge cells at 16,000 × *g* for 10 min at 4 °C.
9. Transfer clarified supernatant to a new tube. The cell surface proteins are now biotinylated on exposed lysine residues.

3.2 Affinity Purification of Biotinylated Proteins

1. Measure protein concentration in sample solution (*see* **Note 5**).
2. Pack the NeutrAvidin Agarose Resin into a column (*see* **Note 6**). Place column into a collection tube. Centrifuge at 500 × *g* for 1 min to remove storage solution.
3. Wash the resin with 100 μL of PBS by centrifugation at 500 × *g* for 1 min and discard buffer from collection tube. Repeat this-step three times.

4. Place column in a new collection tube and add biotinylated sample to the column allowing sample to enter the resin bed. Incubate the mixture 1 h at room temperature with gently rotation.
5. Centrifuge for 1 min at 500 × *g* and collect flow-through.
6. Add 100 μL of lysis buffer to the column, centrifuge for 1 min at 500 × *g* and discard. Repeat twice.
7. Add 100 μL of PBS to the column, centrifuge for 1 min at 500 × *g* and discard. Repeat twice.
8. Place column in a new collection tube and add elution buffer. Incubate 30 min at 55 °C with shaking.
9. Centrifuge for 1 min at 500 × *g* and collect the eluate. Sample can be used for downstream proteomic investigations or stored at −20 °C if not used immediately.

3.3 LC-MS/MS Analysis of Biotinylated Proteins

1. Perform alkylation and tryptic digestion of eluted proteins (*see* **Note 7**).
2. Depending on digestion type and detergents used for cell lysis, apply suitable peptide purification (*see* **Note 8**).
3. Analyze peptides on suitable nanoLC-MS system (*see* **Note 9**).

3.4 Data Analysis

The LC-MS raw data can be analyzed using different programs, such as Proteome Discoverer, MaxQuant, Progenesis LC-MS, and Protein Pilot. In our proteomic workflow we use Proteome Discoverer and database search using SEQUEST, followed by Percolator validation (FDR based confidence scoring) in order to obtain confident protein identities (*see* **Note 10**). Each identified protein has its Protein card (Fig. 1) in Proteome Discoverer containing information about gene ontology, pathways and diseases involved, as well as links to available external data resources for that specific protein, such as STRING, NCBI map, KEGG, UniGene, and SNPs. Except SEQUEST search, Proteome Discoverer enables MASCOT and MS Amanda database searches.

3.5 Prediction of Protein Subcellular Locations

Identified protein usually contain remain of some cellular or other nonspecifically bound proteins. Prediction of subcellular localization can be also performed using CELLO [9] which uses the relationship between sequence similarity (sequence alignment) and identity in subcellular localization to predict subcellular localization, and it is based on multiclass SVM classification system.

1. Go to CELLO: subCELlularLOcalization predictor.
2. Load FASTA file(s) and chose suitable organism (eukaryotes) and sequence (proteins). For each subcellular localization software calculates the reliability (Fig. 2). List outer membrane proteins.

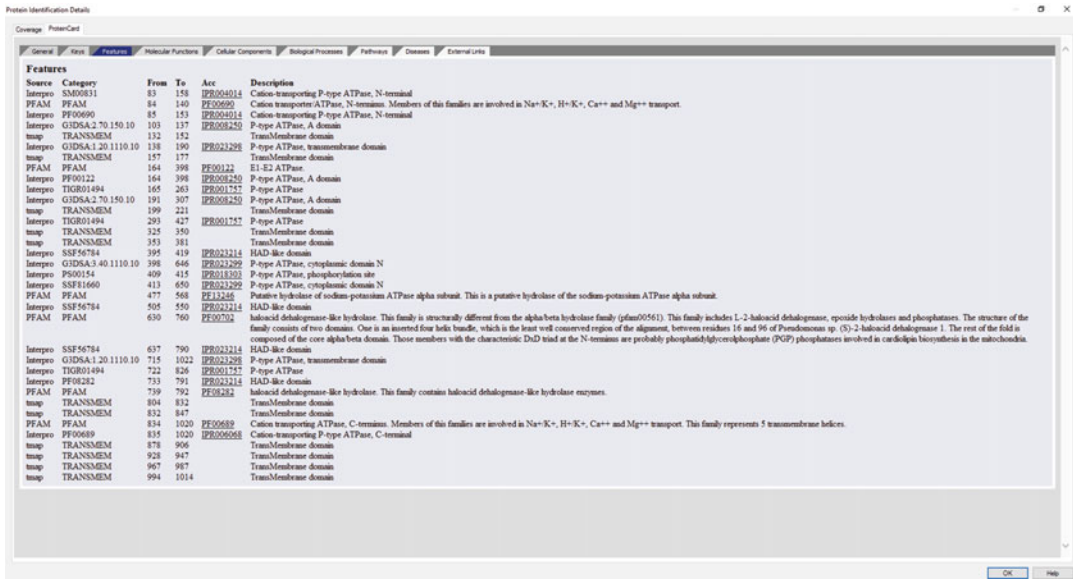


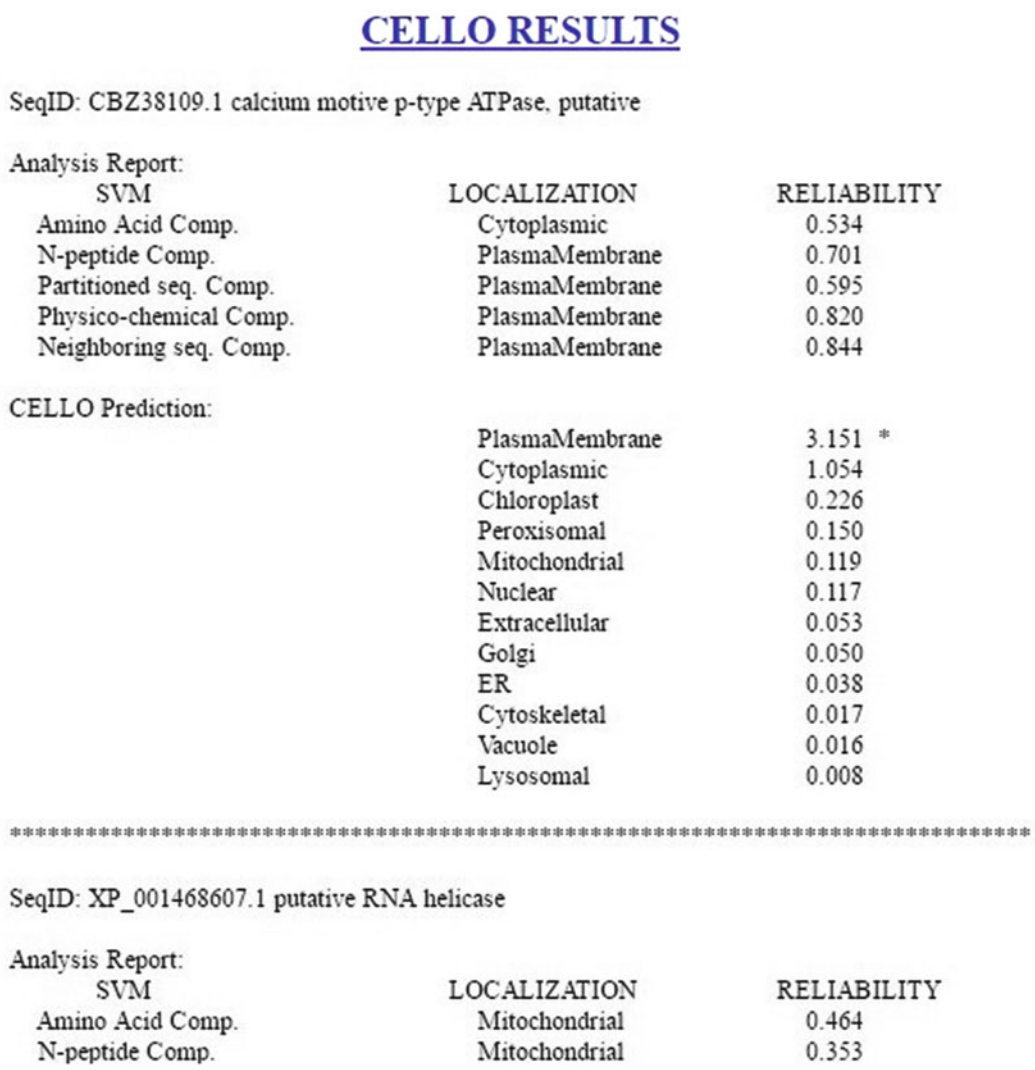
Fig. 1 Example of protein card of identified membrane protein in Proteome Discoverer

Other available databases and computational programs for subcellular localization (together with belonging links) can be found at http://www.geneinfinity.org/sp/sp_proteinloc.html.

3.6 Filtering Data Trough Bioinformatics to Identify Potential Interacting Proteins

List of identified proteins can be applied to The **D**atabase for **A**nnotation, **V**isualization and **I**ntegrated **D**iscovery (**DAVID**) [12, 13] database in order to obtain GO data and filter database to obtain gene ontology data and the list of potential interacting proteins. DAVID represents a set of data-mining and visualization tools that enable functional classification, biochemical pathway maps, and conserved protein domain architectures [14].

1. Copy the list of EntrezGeneID to a new Spread sheet.
2. Go to DAVID Bioinformatics Resources.
3. Go to "Start analysis" tab.
4. Paste the list of EntrezGeneID under the A section (Fig. 3) (**step 1**).
5. In **step 2**, choose the ENTREZ_GENE_ID as identifier.
6. Check "Gene list" in **step 3** and click on "Submit" to start the analysis.
7. Specify the targeted species, or all the proposed species for low information species (Fig. 4).
8. Click on "Functional analysis tool" on the right panel.
9. Click on "Clear all" to deactivate all analysis.



CELLO RESULTS

SeqID: CBZ38109.1 calcium motive p-type ATPase, putative

Analysis Report:

SVM	LOCALIZATION	RELIABILITY
Amino Acid Comp.	Cytoplasmic	0.534
N-peptide Comp.	PlasmaMembrane	0.701
Partitioned seq. Comp.	PlasmaMembrane	0.595
Physico-chemical Comp.	PlasmaMembrane	0.820
Neighboring seq. Comp.	PlasmaMembrane	0.844

CELLO Prediction:

	PlasmaMembrane	3.151 *
	Cytoplasmic	1.054
	Chloroplast	0.226
	Peroxisomal	0.150
	Mitochondrial	0.119
	Nuclear	0.117
	Extracellular	0.053
	Golgi	0.050
	ER	0.038
	Cytoskeletal	0.017
	Vacuole	0.016
	Lysosomal	0.008

****

SeqID: XP_001468607.1 putative RNA helicase

Analysis Report:

SVM	LOCALIZATION	RELIABILITY
Amino Acid Comp.	Mitochondrial	0.464
N-peptide Comp.	Mitochondrial	0.353

Fig. 2 Prediction of protein subcellular localization obtained as CELLO result

10. Select "GOTERM_BP_DIRECT", "GOTERM_CC_DIRECT", "GOTERM_MP_DIRECT" from Gene_ontology tab, and "Interpro", "Pfam", and "Prosite" from the "Protein_Domain" tab (Fig. 5).
11. Click on "Functional annotation table".
12. On the pop-up windows, select "Download the file" and save it as text.
13. Open the file with a spread sheet editor, with "Tab delimited" option.
14. Remove all protein/gene entries (rows) which are not concerned by GO Cell location terms related with "membrane" (column GOTERM_CC_DIRECT).

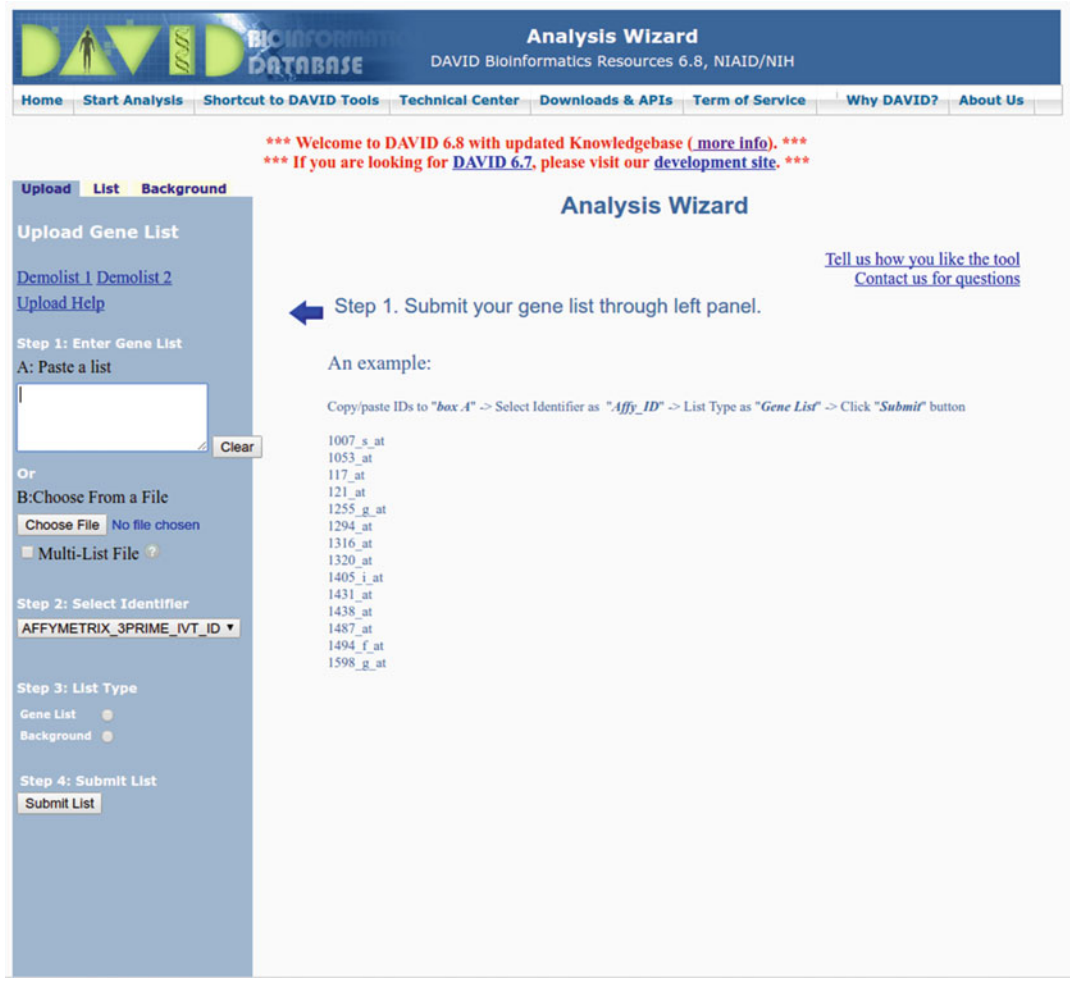


Fig. 3 The Start Analysis tab in DAVID

15. Remove all protein/gene entries (rows) which are implied in known not-related membrane process by GO Biological Process (column GOTERM_BP_DIRECT), like for example "translation" or "protein folding".
16. Remove all protein/gene entries (rows) which are implied in known not-related membrane functions by GO Molecular Functions (column GOTERM_MF_DIRECT), like for example "structural constituent of ribosome" or "DNA binding".
17. For all steps of removal, GO terms can be checked on http://www.geneontology.org/.
18. Using the PFAM, PROSITE and INTERPRO columns, proteins which have domains not related with protein–protein interaction can be removed, like "PF00166:Chaperonin 10 Kd subunit". Every protein domains can be checked on

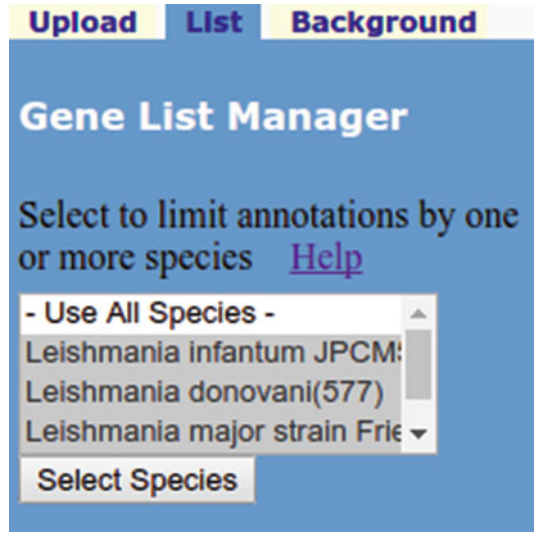


Fig. 4 Species selection screen

PFAM (http://pfam.xfam.org/), PROSITE (http://prosite.expasy.org/) and INTERPRO (https://www.ebi.ac.uk/interpro/) websites.

19. After this step, remaining proteins can be blasted to have more information if available. Manual screening of remaining proteins can be done using both protein domain analysis and BLAST results.
20. Proteins which have passed all those steps of selection are potentially membrane proteins which can interact with other proteins. Protein domain analysis can indicate if such proteins could have interspecies interaction, especially if the domain identified is find in target organism, like SAM domain.

4 Notes

1. Scale the concentration of biotinylation reagent up or down based on cell concentration, size or type. By using the appropriate molar ratio of biotin to the protein, the extent of labeling can be controlled. When labeling diluted protein solutions, a greater molar fold excess of biotin is used compared to a concentrated protein solution. A 100-fold molar excess of biotinylating reagents over the protein amount yields a better degree of cell surface proteins biotinylation as compared to other ratios.
2. Operating at 4 °C throughout the entire procedure helps reduce uptake of biotinylating reagents into the cell.

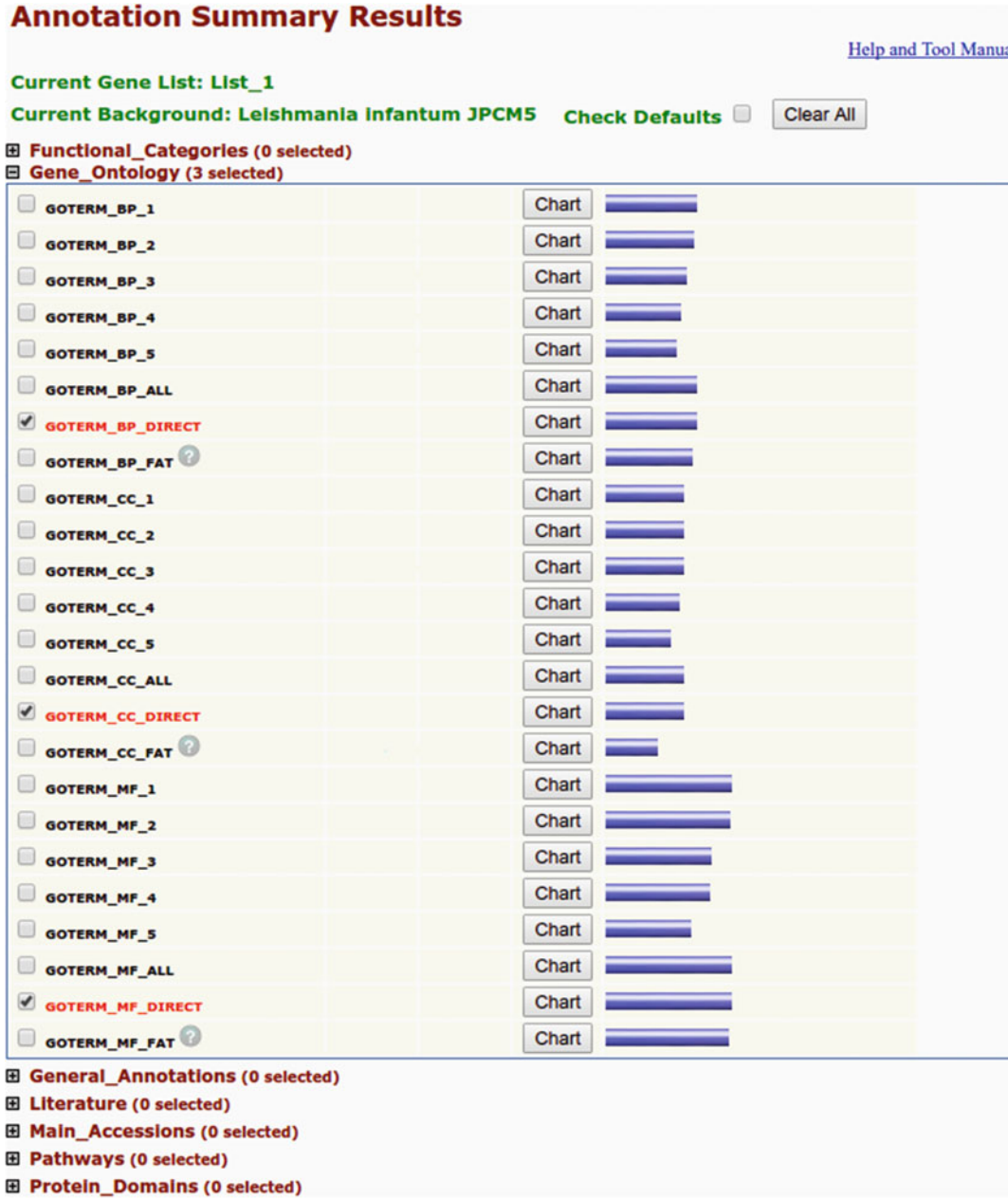


Fig. 5 Gene ontology options

3. A primary amine containing buffer solution as Tris-Cl, ammonium salts, or sodium azide is also commonly used to quench unreacted biotinylating reagent.
4. The choice of lysis buffer depends on the aim of the experiment and specific protocol applied, but also upon considerations bound to the downstream application. Adapt cell lysis buffer and protocol to specific cell type.

5. For protein concentration determination, use Bradford, BCA assay, NanoDrop, or other method compatible with your protein mixture.
6. Based on the protein concentration in the biotinylated sample, calculate the amount of sample and resin needed for affinity purification.
7. Digestion can be performed using different strategies. We recommend FASP protocol [15] using flat bottom filters with 10 kDa cutoff membranes that can be used for up to 200 μg of total protein containing detergents for alkylation (with iodoacetamide) and digestion using trypsin gold (in ratio 1:30). No reduction is needed since DTT is used for elution of biotinylated proteins. Although Triton X-100 cannot be removed by FASP, it does not interfere with FASP digestion. Samples can be also alkylated and digested in solution. Because of low protein amount, overnight ice cold acetone precipitation (four volumes of acetone) can be used. After that pellets should be dissolved in 8 M urea and diluted to 2 M urea with 50 mM ammonium bicarbonate buffer pH 7.6 to final concentration prior alkylation and digestion. Pellets can be also dissolved in sample loading buffer and loaded onto SDS-PAGE gel. Electrophoresis should be performed for approximately 10 min, just to ensure that proteins enter the gel and accumulate into one protein band for salt and detergent removal. Furthermore, standard in gel digestion [16] should be performed, having in mind the yield of tryptic peptide extraction from the gel.
8. If you use Triton X-100 based buffers (such as RIPA), we devised a strategy using strong cation exchange (SCX) chromatography to successfully remove detergents from the peptide sample prior to LC-MS analysis. For the purification up to 10 μg of proteins/peptides, strong cation exchange ZipTips can be used according to following procedure: wash with solution W1 and then load sample diluted in 0.1% formic acid onto SCX ZipTips by aspirating the sample ten times. Wash three times with solution W1, wash five times with solution W2, and elute in 10 μL of elution solvent E1. Finally dry out ammonia and methanol in a Speed-Vac centrifuge and resuspend the sample in 10 μL of 0.1% formic acid. For SDS or CHAPS detergents removal after FASP digestion or in gel digestion, purification with RP C18 ZipTips can be used according to manufacturer procedure. Although high concentrations of CHAPS can interfere with LC-MS analysis, low concentrations can be detected in MS spectrum that do not significantly influence the analysis result and can be easily removed from the nanoLC-MS system after a few sample loop washes and water injections.

9. For LC-MS analysis we usually inject 1 μg of proteins/peptides onto 15 cm nano RP C18 column. Peptides are separated through 3 h gradient from 5–40% mobile phase B followed by gradient increase to 90% B for 5 min. Gradient can be adjusted according to obtained chromatogram.
10. For protein identification in Proteome Discoverer we use SEQUEST to search FASTA files downloaded from NCBI database. As criteria for the search, among standard modifications (oxidation of methionine and carbamidomethylation of cysteine) we use thioacyl (K) as variable modification. The false discovery rate values in Percolator node were set to 1% (strict) and 5% (medium), respectively.

Acknowledgments

The authors acknowledge the European Commission for funding the VetMedZg ERA chair team (ERA Chair Initiative). We also acknowledge Croatian Science Foundation (project 3421) for supporting FM, HRZZ (project 4135) for supporting VM; and APVV-14-218, VEGA1/0258/15, and VEGA 1/0261/15 for supporting MB.

References

1. Finlay BB, McFadden G (2006) Anti-immunology: evasion of the host immune system by bacterial and viral pathogens. Cell 124:767–782
2. Walters MS, Mobley HLT (2009) Identification of uropathogenic *Escherichia coli* surface proteins by shotgun proteomics. J Microbiol Methods 78:131–135
3. Elschenbroich S, Kim Y, Medin JA et al (2010) Isolation of cell surface proteins for mass spectrometry-based proteomics. Expert Rev Proteomics 7:141–154
4. Horvatic A, Kules J, Guillemin N et al (2016) High-throughput proteomics and the fight against pathogens. Mol Biosyst 12:2373–2384
5. Kules J, Horvatic A, Guillemin N et al (2016) New approaches and omics tools for mining of vaccine candidates against vector-borne diseases. Mol Biosyst 12:2680–2694
6. Elia G (2008) Biotinylation reagents for the study of cell surface proteins. Proteomics 8:4012–4024
7. Smith SM (2011) Strategies for the purification of membrane proteins. Methods Mol Biol 681:485–496
8. Lu X, Zhu H (2005) Tube-gel digestion: a novel proteomic approach for high throughput analysis of membrane proteins. Mol Cell Proteomics 4:1948–1958
9. Yu CS, Chen YC, CH L et al (2006) Prediction of protein subcellular localization. Proteins 64:643–651
10. Gardy JL, Laird MR, Chen F et al (2005) PSORTb v.2.0: expanded prediction of bacterial protein subcellular localization and insights gained from comparative proteome analysis. Bioinformatics 21:617–623
11. Wang J, Sung W-K, Krishnan A et al (2005) Protein subcellular localization prediction for Gram-negative bacteria using amino acid subalphabets and a combination of multiple support vector machines. BMC Bioinformatics 6:174–174
12. Huang DW, Sherman BT, Lempicki RA (2008) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc 4:44–57
13. Huang DW, Sherman BT, Lempicki RA (2009) Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res 37:1–13

14. Dennis G, Sherman BT, Hosack DA et al (2003) DAVID: database for annotation, visualization, and integrated discovery. Genome Biol 4:R60–R60
15. Wisniewski JR, Zougman A, Nagaraj N et al (2009) Universal sample preparation method for proteome analysis. Nat Methods 6:359–362
16. Shevchenko A, Tomas H, Havlis J et al (2007) In-gel digestion for mass spectrometric characterization of proteins and proteomes. Nat Protocols 1:2856–2860