How to interpret results from shotgun MS analysis

A quick guide *MQ-PW, 17.07.2013, version 5.4*

Results for shotgun experiments are delivered as an Excel table. These are results from database searches with the software **MASCOT**. For more information and a useful help section on protein identification by MS, please see the MASCOT site (http://www.matrixscience.com/).

In the top part of the table you can find information on the data analysis process such as the database used, taxonomy, mass accuracy and the criteria for validation of peptide and protein identification. Then the list of matched proteins follows, with a description, number of similar matches, accession numbers and a column for every sample that was submitted. These columns contain usually the number of spectra (thus not peptides) that were matched to that particular protein in every sample.

Some important background concepts

1) This is not « true » protein sequencing. What we are doing is matching fragmentation spectra for trypsin fragments of proteins to a database sequence. These spectra do contain sequence information which is however of variable quality and completeness. Since identity is only established by a match to the database, the results can only be good if the correspondence between the database and the organism being studied is good. In other words, if the database does not contain the sequence(s) of the protein(s) you are analyzing nor a close homologue, there will be no match and no results. Even with the best of data. Now, if you are working with one of the common model organisms with sequenced genomes, databases are fairly complete and this is not a concern.

2) Strictly speaking, we are not identifying proteins by mass spectrometry, but peptides. These peptides are mostly between 7 and 20 amino acids long, the average is 11 AA. After having matched peptides, the software we use (Mascot) proceeds to carry out protein inference. This means to derive which sequence(s) in the database contain(s) a given set of peptides. This can yield very univocal identifications if there are enough unique sequences matched. However several database sequences are often matched by the same set of peptides. This can happen because: i) highly homologous protein families exist, and these protein differ only by a few AA (ex. the tubulins) and ii) databases can be redundant and contain several nearly-identical sequences. It is also important to know that the software reports the minimal set of protein sequences which explain the maximum number of identified peptides (principle of parsimony).

3) The data acquisition process is - to a certain degree – random. This means that in a very complex mixture of peptides not always the same ones are chosen for « sequencing », in particular with low abundant peptides. If a certain redundancy of sampling is present, the data can be very reliable, but on the other hand identifications with low number of peptides must be taken with a lot of caution (see below).

Important things to know for interpretation of a list of proteins

1) This table contains the essential results about your experiment. At the same time, you should receive from us a link to download a file containing the full data (but not the raw data). This file will be in a format for the software Scaffold viewer. You can download Scaffold for free from www.proteomesoftware.com. It is a fairly intuitive software which will also allow you to export your data very conveniently (see below). But the advice of an experienced mass spectrometrist can still be necessary to decide to validate or not borderline identifications.

2) The database we use in most cases is UNIPROT (http://www.uniprot.org/), although we call it SPTREMBL because it is essentially a fusion of Swissprot and TrEMBL.

* Swiss-Prot is a curated protein sequence database which strives to provide a high level of annotation (such as the description of the function of a protein, its domains structure, post-translational modifications, variants, etc.), a minimal level of redundancy and high level of integration with other databases.

* TrEMBL is a computer-annotated supplement of Swiss-Prot that contains all the translations of EMBL nucleotide sequence entries not yet integrated in Swiss-Prot.

3) Check the taxonomy of the species we have used for database search. Lets us know if it is not the correct one !

4) Identified proteins: in the « accession number » (2nd column) you can see if more than one database entries was matched by the same set of peptides. Column one («identified proteins ») gives you the annotation available for the accession number shown. Unfortunately the software does not always choose the most annotated database entry to report. So it is worth looking directly at the Scaffold file to see if a more informative sequence is reported.

Example:

Identified proteins	**Accession number**	**Molecular Weight**	**Sample 1**	**Sample 2**
cDNA FLJ78508	A8K7C2_HUMAN (+2)	42 kDa	23	31

In this example a TrEMBL sequence is reported while there are SWISSPROT entries also matched (+2). These usually have a much better annotation.

5) The **Molecular Weight** in the third column of the table is a theoretical one calculated from the database sequence. Of course this does not imply that the real mass of the protein actually corresponds to this value. Databases usually list the precursor (unprocessed) sequence. Also, if the protein detected in your sample is a fragment corresponding to only a portion of the sequence, this can only be deduced by looking in detail at the sequence coverage (i.e. where the matched peptide are located in the sequence). You need to check the full data (Scaffold file) for this.

6) The numbers of assigned spectra give an idea of the confidence of the identification. Although we list matches with only one spectrum, confident identifications need usually a minimum of two distinct peptides. Remember that what you have listed in this table is the number of assigned spectra, not peptides. One and the same peptide can be matched several times. To know exactly how many distinct peptides have been matched you need the full data (Scaffold file). However as a rule of thumb for evaluating your excel table, you should take as good all identifications with 4 or more spectra. Consider with a lot of caution all those underneath this value.

7) Some linearity exists between the number of spectra assigned to a certain protein and its concentration. The numbers of matched spectra can thus be used to make semi-quantitative estimates of protein amount in the sample (this approach is called “spectral counting”). Of course this linearity is good when the numbers of spectra matched are numerous (>10). With lower numbers the relationship is a lot less reliable. In other words if you are comparing Tubulin in samples A and B where it is identified with 100 and 300 spectra, respectively, you can assume that there is a certain difference (2-3x) in the amount of tubulin present. However if the insulin receptor was identified with 3 spectra in sample A and 1 spectrum in sample B, you cannot really conclude that the difference in concentration is significative. The same is true if the numbers are 3 spectra and 0 spectra. You cannot really conclude that the protein is really absent in sample B in this case. In all cases, spectral counting is also much less reliable to compare amount of different proteins in the same sample.

8) For protein pull-down (IP etc) studies : the question of whether an “interactor” protein that you find is really specific is not a trivial one. Pull-down experiments are subjected to a lot of possible artifacts. The CRAPome website (http://www.crapome.org) contains a list of unspecific proteins identified in a large number of negative control experiments. We have also our own list of “common contaminants” which are frequently found in this type of experiments. Please do not hesitate to ask for it.

9) Some links to databases useful to know more about a protein of interest:

General link through EXPASY: http://www.expasy.org/proteomics

Protein Databases: http://www.uniprot.org
http://www.nextprot.org/

Reactome (biological pathways): http://www.reactome.org/

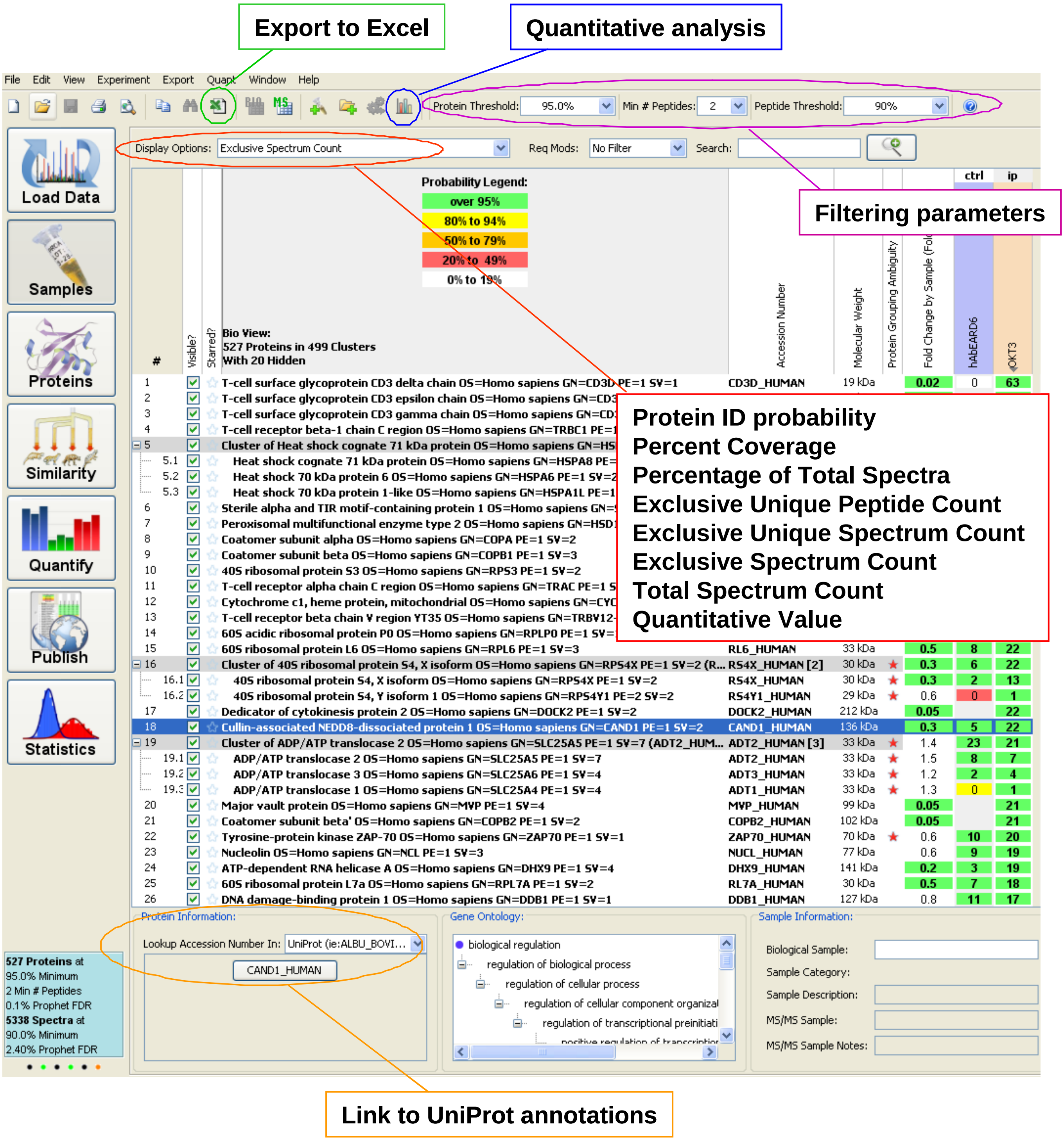
Prediction of protein functional sites: http://elm.eu.org/

Protein-protein interaction databases :

INTACT database :	http://www.ebi.ac.uk/intact/site/index.jsf
HPRD database :	http://www.hprd.org/
BIND database :	http://bond.unleashedinformatics.com/Action?/
DIP:	http://dip.doe-mbi.ucla.edu/
STRING:	http://string-db.org/
Human Proteinpedia:	http://www.humanproteinpedia.org/index_html

Scaffold software

1) Scaffold files capture the most important data produced by MS and the identification process and are an excellent way to distribute the results. Scaffold allows you to further look interactively at your data, and can be downloaded as a free viewer (http://www.proteomesoftware.com). See below a snapshot of the Scaffold output main interface:



Export to Excel

Quantitative analysis

Filtering parameters

Protein ID probability
Percent Coverage
Percentage of Total Spectra
Exclusive Unique Peptide Count
Exclusive Unique Spectrum Count
Exclusive Spectrum Count
Total Spectrum Count
Quantitative Value

Link to UniProt annotations

Protein Threshold: 95.0% Min # Peptides: 2 Peptide Threshold: 90%

Display Options: Exclusive Spectrum Count Req Mods: No Filter Search:

Probability Legend:
- 🟩 over 95%
- 🟨 80% to 94%
- 🟧 50% to 79%
- 🟥 20% to 49%
- ⬜ 0% to 19%

Bio View:
527 Proteins in 499 Clusters
With 20 Hidden

#	Visible?	Starred?	Bio View	Accession Number	Molecular Weight	Protein Grouping Ambiguity	Fold Change by Sample (Fol	ctrl hAbEARD6	ip OKT3
1	☑	☆	T-cell surface glycoprotein CD3 delta chain OS=Homo sapiens GN=CD3D PE=1 SV=1	CD3D_HUMAN	19 kDa		0.02	0	63
2	☑	☆	T-cell surface glycoprotein CD3 epsilon chain OS=Homo sapiens GN=CD3						
3	☑	☆	T-cell surface glycoprotein CD3 gamma chain OS=Homo sapiens GN=CD						
4	☑	☆	T-cell receptor beta-1 chain C region OS=Homo sapiens GN=TRBC1 PE=1						
⊟ 5	☑	☆	Cluster of Heat shock cognate 71 kDa protein OS=Homo sapiens GN=HS						
5.1	☑	☆	Heat shock cognate 71 kDa protein OS=Homo sapiens GN=HSPA8 PE=						
5.2	☑	☆	Heat shock 70 kDa protein 6 OS=Homo sapiens GN=HSPA6 PE=1 SV=2						
5.3	☑	☆	Heat shock 70 kDa protein 1-like OS=Homo sapiens GN=HSPA1L PE=1						
6	☑	☆	Sterile alpha and TIR motif-containing protein 1 OS=Homo sapiens GN=						
7	☑	☆	Peroxisomal multifunctional enzyme type 2 OS=Homo sapiens GN=HSD1						
8	☑	☆	Coatomer subunit alpha OS=Homo sapiens GN=COPA PE=1 SV=2						
9	☑	☆	Coatomer subunit beta OS=Homo sapiens GN=COPB1 PE=1 SV=3						
10	☑	☆	40S ribosomal protein S3 OS=Homo sapiens GN=RPS3 PE=1 SV=2						
11	☑	☆	T-cell receptor alpha chain C region OS=Homo sapiens GN=TRAC PE=1 S						
12	☑	☆	Cytochrome c1, heme protein, mitochondrial OS=Homo sapiens GN=CYC						
13	☑	☆	T-cell receptor beta chain V region YT35 OS=Homo sapiens GN=TRBV12-						
14	☑	☆	60S acidic ribosomal protein P0 OS=Homo sapiens GN=RPLP0 PE=1 SV=						
15	☑	☆	60S ribosomal protein L6 OS=Homo sapiens GN=RPL6 PE=1 SV=3	RL6_HUMAN	33 kDa		0.5	8	22
⊟ 16	☑	☆	Cluster of 40S ribosomal protein S4, X isoform OS=Homo sapiens GN=RPS4X PE=1 SV=2 (R...	RS4X_HUMAN [2]	30 kDa	★	0.3	6	22
16.1	☑	☆	40S ribosomal protein S4, X isoform OS=Homo sapiens GN=RPS4X PE=1 SV=2	RS4X_HUMAN	30 kDa	★	0.3	2	13
16.2	☑	☆	40S ribosomal protein S4, Y isoform 1 OS=Homo sapiens GN=RPS4Y1 PE=2 SV=2	RS4Y1_HUMAN	29 kDa	★	0.6	0	1
17	☑	☆	Dedicator of cytokinesis protein 2 OS=Homo sapiens GN=DOCK2 PE=1 SV=2	DOCK2_HUMAN	212 kDa		0.05		22
18	☑	☆	Cullin-associated NEDD8-dissociated protein 1 OS=Homo sapiens GN=CAND1 PE=1 SV=2	CAND1_HUMAN	136 kDa		0.3	5	22
⊟ 19	☑	☆	Cluster of ADP/ATP translocase 2 OS=Homo sapiens GN=SLC25A5 PE=1 SV=7 (ADT2_HUM...	ADT2_HUMAN [3]	33 kDa	★	1.4	23	21
19.1	☑	☆	ADP/ATP translocase 2 OS=Homo sapiens GN=SLC25A5 PE=1 SV=7	ADT2_HUMAN	33 kDa	★	1.5	8	7
19.2	☑	☆	ADP/ATP translocase 3 OS=Homo sapiens GN=SLC25A6 PE=1 SV=4	ADT3_HUMAN	33 kDa	★	1.2	2	4
19.3	☑	☆	ADP/ATP translocase 1 OS=Homo sapiens GN=SLC25A4 PE=1 SV=4	ADT1_HUMAN	33 kDa	★	1.3	0	1
20	☑	☆	Major vault protein OS=Homo sapiens GN=MVP PE=1 SV=4	MVP_HUMAN	99 kDa		0.05		21
21	☑	☆	Coatomer subunit beta' OS=Homo sapiens GN=COPB2 PE=1 SV=2	COPB2_HUMAN	102 kDa		0.05		21
22	☑	☆	Tyrosine-protein kinase ZAP-70 OS=Homo sapiens GN=ZAP70 PE=1 SV=1	ZAP70_HUMAN	70 kDa	★	0.6	10	20
23	☑	☆	Nucleolin OS=Homo sapiens GN=NCL PE=1 SV=3	NUCL_HUMAN	77 kDa		0.6	9	19
24	☑	☆	ATP-dependent RNA helicase A OS=Homo sapiens GN=DHX9 PE=1 SV=4	DHX9_HUMAN	141 kDa		0.2	3	19
25	☑	☆	60S ribosomal protein L7a OS=Homo sapiens GN=RPL7A PE=1 SV=2	RL7A_HUMAN	30 kDa		0.5	7	18
26	☑	☆	DNA damage-binding protein 1 OS=Homo sapiens GN=DDB1 PE=1 SV=1	DDB1_HUMAN	127 kDa		0.8	11	17

527 Proteins at
95.0% Minimum
2 Min # Peptides
0.1% Prophet FDR
5338 Spectra at
90.0% Minimum
2.40% Prophet FDR

Protein Information:
Lookup Accession Number In: UniProt (ie:ALBU_BOVI...
CAND1_HUMAN

Gene Ontology:
- ● biological regulation
  - regulation of biological process
    - regulation of cellular process
      - regulation of cellular component organizat
        - regulation of transcriptional preinitiati
          - positive regulation of transcriptior

Sample Information:
Biological Sample:
Sample Category:
Sample Description:
MS/MS Sample:
MS/MS Sample Notes:

Please note that it is better to have a reasonably powerful computer with minimum of 4GB RAM to open large Scaffold files. We cannot provide here a full introduction in the capabilities of Scaffold, but the software is quite user-friendly and has a good Help menu. As Scaffold is regularly updated, be sure to have the latest version to be able to open your Scaffold files (.sf3).

2) The table can be sorted based on one particular column by clicking on the header

3) Scaffold allows the validation of peptide & protein IDs, the alignment of samples and the quantitative comparison of samples based on spectral counting. Usually 95% protein probability, 90% peptide probability and min. 2 peptides are used as filtering parameters, but these settings can be relaxed if you are looking for a specific low abundant protein or peptide. Caveat : these recovered identifications should be used with caution: ask for the advice of an experienced mass spectrometrist!

4) Scaffold summarizes the protein list using two levels of hierarchy:

- **Protein Group (PEG)** is a set of protein sequences that are associated with an identical set of peptides. Protein groups are by default represented by the sequence that has the highest probability and the largest associated number of spectra.

- **Protein Cluster** – is a group of **PEGs** created using a hierarchical clustering algorithm. Proteins member of the cluster share some peptides but not all of them. Protein Clusters are by default represented by the protein that shows the highest associated probability. Clusters can be collapsed or expanded directly in the protein list (-/+ buttons on the left).

5) Various display options are available to look at the list of identified proteins (see below). “**Exclusive Spectrum Count**” or “**Total Spectrum Count**“ is often preferred as it is also the classical basis of (semi-)quantitative comparison of proteins in the different samples (see the option quantitative analysis).

• **Protein Identification Probability -** Scaffold’s calculated probability that the protein identification for any of the MS Samples is correct. Results are color-coded to indicate significant differences in protein identification confidence.

• **Percentage Coverage -** The percentage of all the amino acids in the protein sequence that were covered by identified peptides detected in the sample.

• **Percentage of Total Spectra** - The number of spectra matched to a protein, summed over all MS Samples, as a percentage of the total number of spectra in the sample.

• **Exclusive Unique Peptide Count** - The number of different amino acid sequences, regardless of any modification that are associated with a single protein group.

• **Exclusive Unique Spectrum Count** - Number of distinct spectra associated only with a single protein group. Spectra are considered distinct when:

i) they identify different sequences of amino acids or peptides;

ii) they identify different charge states or a modified form of the peptide within the same identified sequences of amino acids.

• **Exclusive Spectrum Count** - The number of spectra, associated only with a single protein group.

• **Total Spectrum Count** - The total number of spectra associated to a single protein group, including those shared with other proteins.

• **Quantitative Value (***Selected quantitative method***) -** Scaffold will display the results of the Quantitative Method selected from the Quantitative Analysis Dialog Box.

6) The option "search", either by name or accession number, is very useful when you are looking at a specific protein in a list of a few thousands hits! It is also easy to sort the list by protein name, molecular weight, number of spectra, etc.

7) For looking in more details at the sequence coverage, assigned peptides, or spectra of a protein, you can select it (click on it) and use the "Proteins" window of the left menu bar. Note that the **GO annotations** of all identified proteins can be downloaded into the Scaffold file (Menu bar –> Experiment -> Add NCBI annotations). You can also look at alternative accession numbers matched as well as at the detailed Uniprot annotations of a selected protein, using the button/link in the Protein Information frame.

8) Various export options are available, the most useful being the Samples Excel export.

9) Samples can be grouped in categories (e.g. in the case of replicate samples) and more sophisticated semi-quantitative analyses and statistical tests can be performed in the quantitative analysis submenu.

10) For more information :

- Searle, B. C. (2010). Scaffold: a bioinformatic tool for validating MS/MS-based proteomic studies. *Proteomics*, *10* (6), 1265–9.

- Scaffold User's Guide (available from the Help menu in Scaffold)

- http://www.proteomesoftware.com

Then go to their Resource library or their FAQ