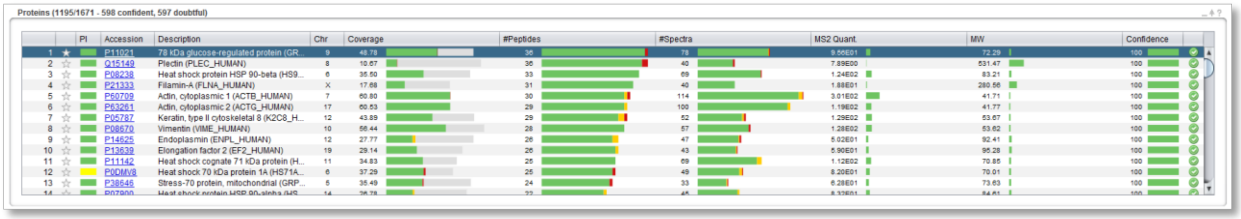
# Peptide and Protein Validation

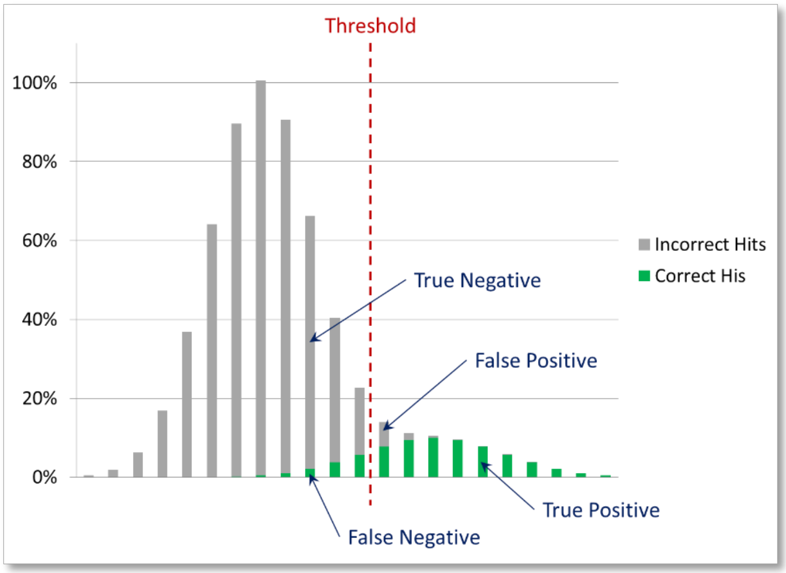
Start by loading the PeptideShaker[1] example project (the same project that has been used in the previous chapters). The project is easily available from the PeptideShaker Welcome Dialog. We are now looking at the protein table:



Proteins (1195/1671 - 598 confident, 597 doubtful)

| | | PI | Accession | Description | Chr | Coverage | #Peptides | #Spectra | MS2 Quant. | MW | Confidence | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | ★ | 🟩 | P11021 | 78 kDa glucose-regulated protein (GR... | 9 | 48.78 | 36 | 78 | 9.56E01 | 72.29 | 100 | ✓ |
| 2 | ☆ | 🟩 | Q15149 | Plectin (PLEC_HUMAN) | 8 | 10.67 | 36 | 40 | 7.89E00 | 531.47 | 100 | ✓ |
| 3 | ☆ | 🟩 | P08238 | Heat shock protein HSP 90-beta (HS9... | 6 | 35.50 | 33 | 69 | 1.24E02 | 83.21 | 100 | ✓ |
| 4 | ☆ | 🟩 | P21333 | Filamin-A (FLNA_HUMAN) | X | 17.68 | 31 | 40 | 1.88E01 | 280.56 | 100 | ✓ |
| 5 | ☆ | 🟩 | P60709 | Actin, cytoplasmic 1 (ACTB_HUMAN) | 7 | 60.80 | 30 | 114 | 3.01E02 | 41.71 | 100 | ✓ |
| 6 | ☆ | 🟩 | P63261 | Actin, cytoplasmic 2 (ACTG_HUMAN) | 17 | 60.53 | 29 | 100 | 1.19E02 | 41.77 | 100 | ✓ |
| 7 | ☆ | 🟩 | P05787 | Keratin, type II cytoskeletal 8 (K2C8_H... | 12 | 43.89 | 29 | 52 | 1.29E02 | 53.67 | 100 | ✓ |
| 8 | ☆ | 🟩 | P08670 | Vimentin (VIME_HUMAN) | 10 | 56.44 | 28 | 57 | 1.28E02 | 53.62 | 100 | ✓ |
| 9 | ☆ | 🟩 | P14625 | Endoplasmin (ENPL_HUMAN) | 12 | 27.77 | 26 | 47 | 5.02E01 | 92.41 | 100 | ✓ |
| 10 | ☆ | 🟩 | P13639 | Elongation factor 2 (EF2_HUMAN) | 19 | 29.14 | 26 | 43 | 5.90E01 | 95.28 | 100 | ✓ |
| 11 | ☆ | 🟩 | P11142 | Heat shock cognate 71 kDa protein (H... | 11 | 34.83 | 25 | 69 | 1.12E02 | 70.85 | 100 | ✓ |
| 12 | ☆ | 🟨 | P0DMV8 | Heat shock 70 kDa protein 1A (HS71A... | 6 | 37.29 | 25 | 49 | 8.20E01 | 70.01 | 100 | ✓ |
| 13 | ☆ | 🟩 | P38646 | Stress-70 protein, mitochondrial (GRP... | 5 | 35.49 | 24 | 33 | 6.28E01 | 73.63 | 100 | ✓ |
| 14 | ☆ | 🟩 | P07900 | Heat shock protein HSP 90-alpha (HS... | 14 | 26.78 | 22 | 45 | 8.32E01 | 84.61 | 100 | ✓ |

If you scroll down the protein list you will note that some of the proteins further down the list are supported by very few identified peptides that in total do not cover much of the corresponding protein sequences. Some of these low-quality hits are likely to be false identifications introduced by errors of the search engines. We are now going to set a threshold to retain only the best scoring hits and control the error rate of the final results.



As displayed in the figure above, depending on your validation threshold, hits can be sorted into four classes: (1) False Positives, (2) True Positives, (3) False Negatives, (4) True Negatives. *Which population do we want to retain? To control?* [1.5a]

*Harald Barsnes (harald.barsnes@uib.no) and Marc Vaudel (marc.vaudel@uib.no)*

Note that PeptideShaker provides a confidence for every protein, peptide and peptide to spectrum match (PSM). These metrics provide an unbiased estimation of the quality of the hits, independent of the sample, the mass spectrometer and the search engine. How is this possible? When using SearchGUI earlier in the tutorial, we actually appended sequences of non-existing proteins (so-called decoy sequences) to the protein database. In fact, these fake sequences are the reversed versions of the actual sequences.

Here is an example from our database (the FASTA file):

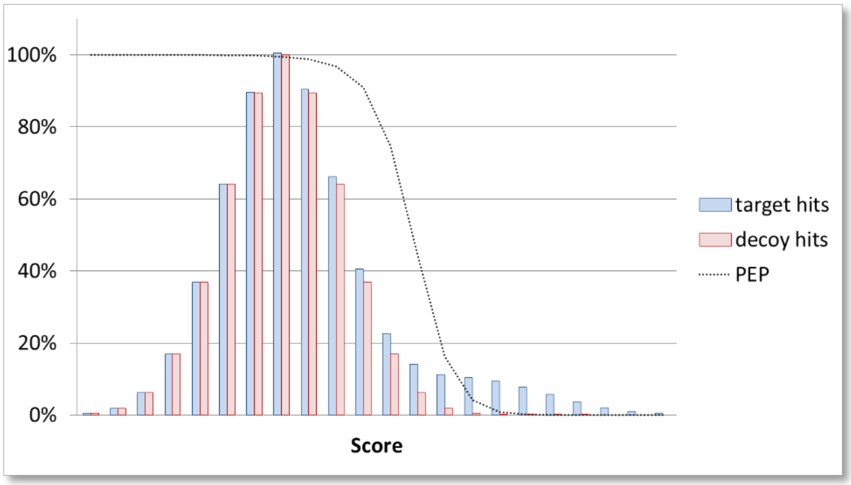
```
>sw|Q8TCZ7|CU074_HUMAN Putative uncharacterized protein encoded by LINC00308 OS=Homo sapiens
GN=LINC00308 PE=5 SV=2
MAYVFNLSCLGSQVERLLEARSSRPTWIIQPSPKKAPEACFSFHSSYERNWA

>sw|Q8TCZ7_REVERSED|CU074_HUMAN Putative uncharacterized protein encoded by LINC00308 OS=Homo
sapiens GN=LINC00308 PE=5 SV=2-REVERSED
AWNREYSSHFSFCAEPAKKPSPQIIWTPRSSRAELLREVQSGLCSLNFVYAM
```

Thus, whenever a mistake is made, it is as likely to happen in the real database (called the target database) as it is in the artificial database (called the decoy database).[3] When a decoy hit is found among five target hits

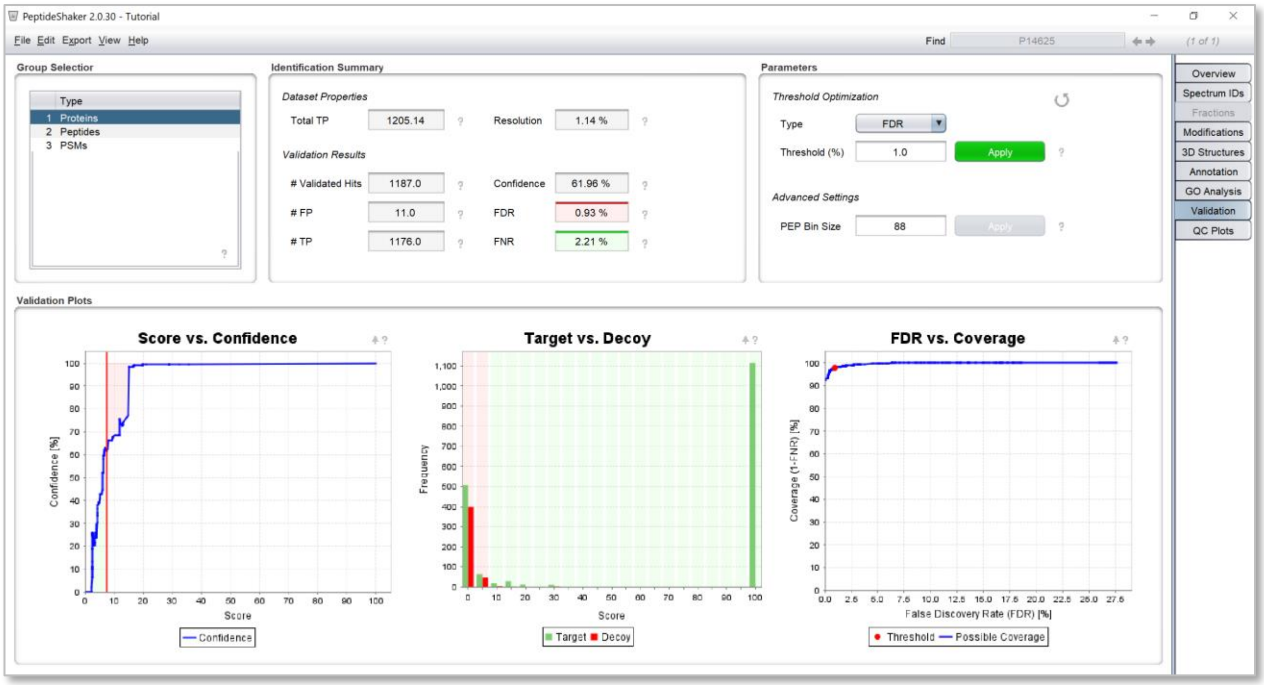
$$(\textit{Target hit}\quad \textit{Target hit}\quad \textit{Target hit}\quad \textit{Decoy hit}\quad \textit{Target hit}\quad \textit{Target hit})$$

We hence assume that there is one false positive among the target hits (20% error). *Do we know which one? Are there other ways to create a decoy database? Which one is the best? [1.5b]*



The decoy hits allow us to estimate the number of false positives in a result set. There are two main ways to control the number of false positives in the validated protein set. First of all, we can set a confidence threshold; typically we would validate protein hits in which we are more than 95% confident. However, scientists usually prefer to control the False Discovery Rate (FDR), the total error share in the result set; typically we allow 1% FDR, meaning that 1% of the retained proteins are expected to be false positives. PeptideShaker already did this standard validation for you. Some may have noticed the symbols at the far right of each row. These indicate that the corresponding protein, peptide or PSM passed a 1% FDR threshold. *On the top of the table you can see that 1187 protein groups were validated out of 1757, how many false positives do we expect? [1.5c]*

The validation threshold can be optimized in the 'Validation' tab of PeptideShaker. Opening the 'Validation' tab you should see this:



Dataset Properties

Total TP 1205.14

Resolution 1.14 %

Validation Results

# Validated Hits 1187.0

Confidence 61.96 %

# FP 11.0

FDR 0.93 %

# TP 1176.0

FNR 2.21 %

Threshold Optimization

Type FDR

Threshold (%) 1.0

Advanced Settings

PEP Bin Size 88

We will now change the validation criteria for our proteins, peptides and PSMs. The group selected in the top-left box should be 'Proteins'. The 'Identification Summary' section provides results from our 1% FDR validation. The 'Parameters' section to the right allows us to customize the estimation, and plots below visualize the results and control their quality. We will now only focus on the main settings.

**Tip:**

*Note the question marks present everywhere to guide you through all the parameters.*

Two metrics can be defined to evaluate the validation procedure: (A) the False Discovery Rate (FDR) indicating the share of retained false positives; (B) False Negative Rate (FNR) indicating the share of false negatives:

| | Number of Validated Hits | Number of Rejected Hits |
|---|---|---|
| Correct Hits | $n_{TP}$ | $n_{FN}$ |
| Incorrect Hits | $n_{FP}$ | $n_{TN}$ |

$$FDR = \frac{n_{FP}}{n_{FP} + n_{TP}} \qquad FNR = \frac{n_{FN}}{n_{FN} + n_{TP}}$$

The identification summary indicates that 1187 proteins were validated including 11 false positives. PeptideShaker estimates that a maximum of 1205.14 true positive proteins could be found in the data set: we are thus including almost all of them.

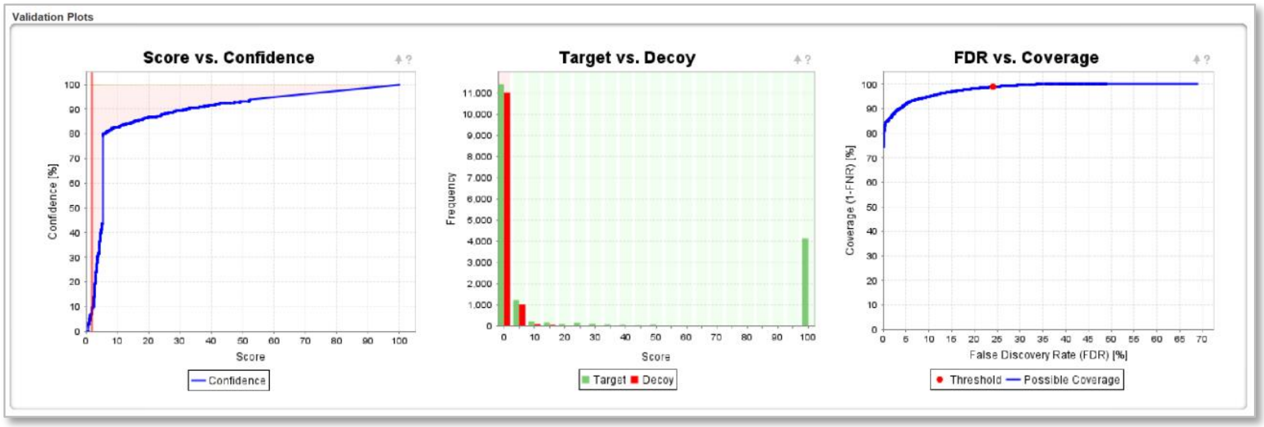
The 'Validation Results' show that the FDR limit used is exactly 0.93%. *Why is it not exactly 1% when using an FDR at 1%? [1.5d]*

The three plots at the bottom display the current threshold settings. The Score vs. Confidence plot shows the variation of the protein confidence against the score and the chosen threshold in red. Note that the green and red areas in this plot represent the number of false negatives and false positives, respectively. These are used to estimate the FDR and FNR values displayed in the Target vs. Decoy plot and in the corresponding green and red boxes. These metrics allow the drawing of a FDR vs. Coverage plot (also called a receiver operating characteristic (ROC) curve) which allows you to optimize your threshold. Note that the current setting is represented by a red point on the curve. Note that the point can deviate from the curve: this is a direct illustration of the confidence estimation imprecision.

As you can see from the Score vs. Confidence plot, our threshold (red line) is set in an area where the confidence is around 62%. *How accurate is the confidence estimation in this case? If we include hundred proteins at 95% confidence, how many false positives do we expect? [1.5e]*

The same operations can also be conducted on Peptides and PSMs when changing the selected population in the top left section. *How do the validation metrics for peptides and PSMs compare to the protein level? [1.5f]*

Select the PSMs category. Note that it is possible to set a threshold at a stringent confidence level. However, we are now going to ask PeptideShaker to focus on quantity and set a False Negative Rate (FNR) of 1%. Select FNR as the threshold type, type 1 and hit Enter:



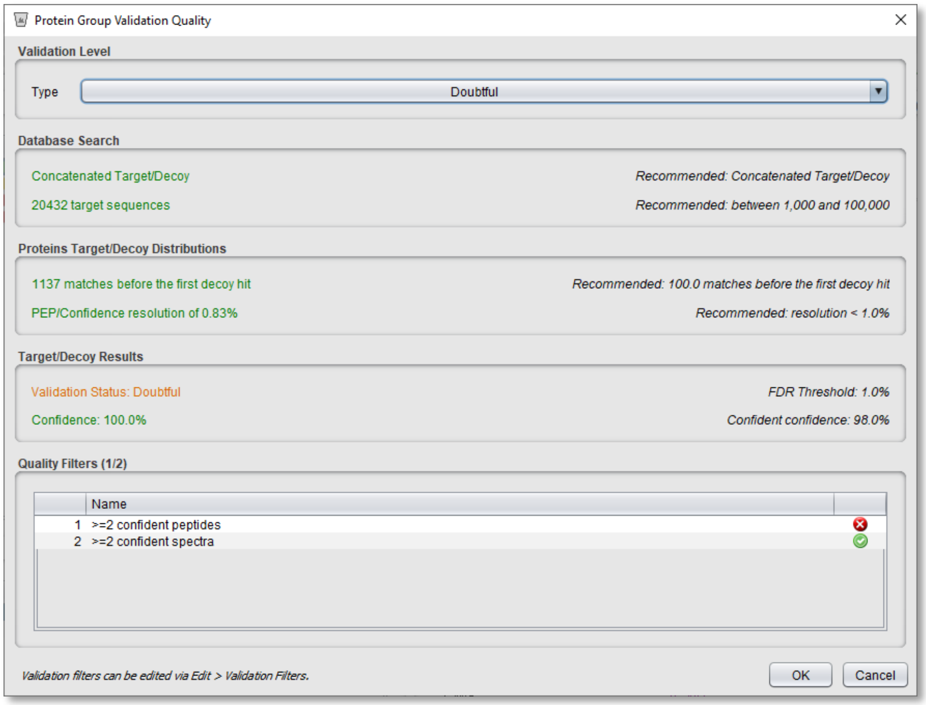
Note that the red line illustrating the threshold and the points indicating the FDR and FNR have moved to the left in the plots. *What are the new FDR and FNR values? What are the values using a 95% confidence threshold? In your opinion, what is the best threshold? [1.5g]*

If you want to apply new validation settings to the actual dataset, click on the green 'Apply' button in the Parameters section. If you go back to the 'Overview' tab, the green and red symbols indicating protein validation will reflect the new validation settings. Note that without clicking the 'Apply' button the new thresholds will be ignored!

> **Tip:**
>
> *Use the 'Apply' button only when you are happy with the selected threshold.*
>
> **Tip:**
>
> *When statistical significance is ensured, PSMs are grouped according to their charges and peptides according to their modification status in order to maximize the identification yield.[2]*

Now go back to the Overview tab, on top of every table, you will see that PeptideShaker marked some of the validated matches as confident and some as doubtful. At 1% FDR 583 protein matches were marked as confident and 604 as doubtful, out of a total of 1187 statistically validated hits. *Why are protein matches doubtful although validated? [1.5h]*

If you scroll down to row 223 you will see the first protein (group) marked as doubtful: O15067. If you click on the yellow warning icon to the right, a diagnostic dialog will open showing why this protein validation was marked as doubtful:



Protein Group Validation Quality

Validation Level

Type: Doubtful

Database Search

Concatenated Target/Decoy — *Recommended: Concatenated Target/Decoy*

20432 target sequences — *Recommended: between 1,000 and 100,000*

Proteins Target/Decoy Distributions

1137 matches before the first decoy hit — *Recommended: 100.0 matches before the first decoy hit*

PEP/Confidence resolution of 0.83% — *Recommended: resolution < 1.0%*

Target/Decoy Results

Validation Status: Doubtful — *FDR Threshold: 1.0%*

Confidence: 100.0% — *Confident confidence: 98.0%*

Quality Filters (1/2)

| | Name | |
|---|---|---|
| 1 | >=2 confident peptides | ✗ |
| 2 | >=2 confident spectra | ✓ |

*Validation filters can be edited via Edit > Validation Filters.*

As you can see here, the protein was marked as doubtful because its identification is supported by only one confident peptide. It is important to note that this does not mean that the protein identification is wrong, only that it should be considered with care.

Similar filters are also used at the peptide and PSM level, and the following colors are always used: green (validated and confident), yellow (validated, but doubtful) and red (not validated).

# References

1. Vaudel, M. et al. PeptideShaker enables reanalysis of MS-derived proteomics data sets. *Nat Biotech* **33**, 22-24 (2015).
2. Vaudel, M., Burkhart, J.M., Sickmann, A., Martens, L. & Zahedi, R.P. Peptide identification quality control. *Proteomics* **11**, 2105-2114 (2011).
3. Elias, J.E. & Gygi, S.P. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat Methods* **4**, 207-214 (2007).