Spectrum Counting Quantification

Spectrum counting based quantification is the simplest quantification method existing in Proteomics. It relies on the rationale that highly abundant peptides will trigger more MS2 spectra measurements, also likely to present a higher signal to noise ratio. As a result, peptides from abundant proteins are more likely to be identified and in more spectra[1](#_ENREF_1). Two approaches were hence followed: count the number of peptides identified for a given protein – like in the so-called emPAI index[2](#_ENREF_2) – method and count the number of spectra ascribed to a protein[3](#_ENREF_3).

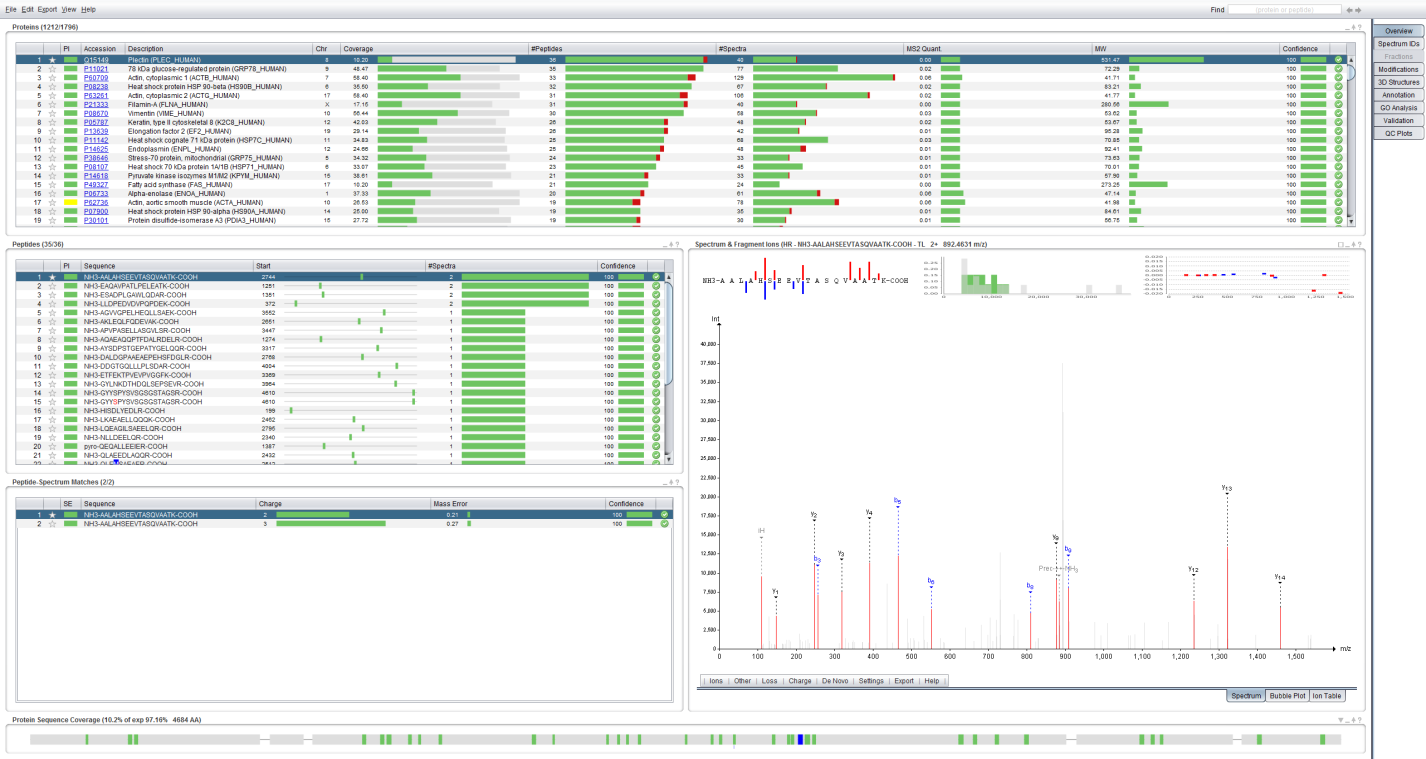
*In your opinion, what are the advantages/shortcomings of this approach?*

The main advantage of this technique is obviously its simplicity: for a given protein, one simply needs to count spectra – no need for advanced experiments, signal processing or data mining. As a result, abundance indexes are straightforwardly available from any measurement in virtually no processing time.

This however comes at costs in quantification accuracy. Indeed, in terms of quantification performance, an abundance metric based on spectrum counts can simply not compare to intensity based metrics. For example, a protein where only one spectrum is validated will hence be seen two times down-regulated when compared to a similar one where two spectra are validated – although we saw in the precedent chapters how subjective spectrum matches validation is.

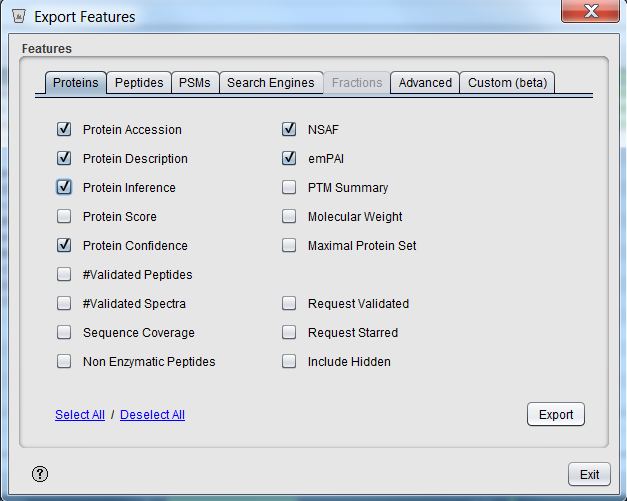
Spectrum counting quantification accuracy is hence dramatically sensitive with regards to the identification process. Other factors can impair accuracy: different proteins do not have the same probability to generate observable tryptic peptides; these do not have the same probability to generate a validated identification. Also, by design, the protein inference problem tackled in the identification chapter dramatically biases the index. Finally, modern mass spectrometers are actually tuned toward the reduction of redundant peptide acquisition, hence biasing the base assumption of spectrum counting indexes.

Spectrum counting indexes can be exported from all identification software, we will now see an example with PeptideShaker: load the example dataset as explained in the identification chapter. You should see the following screen:

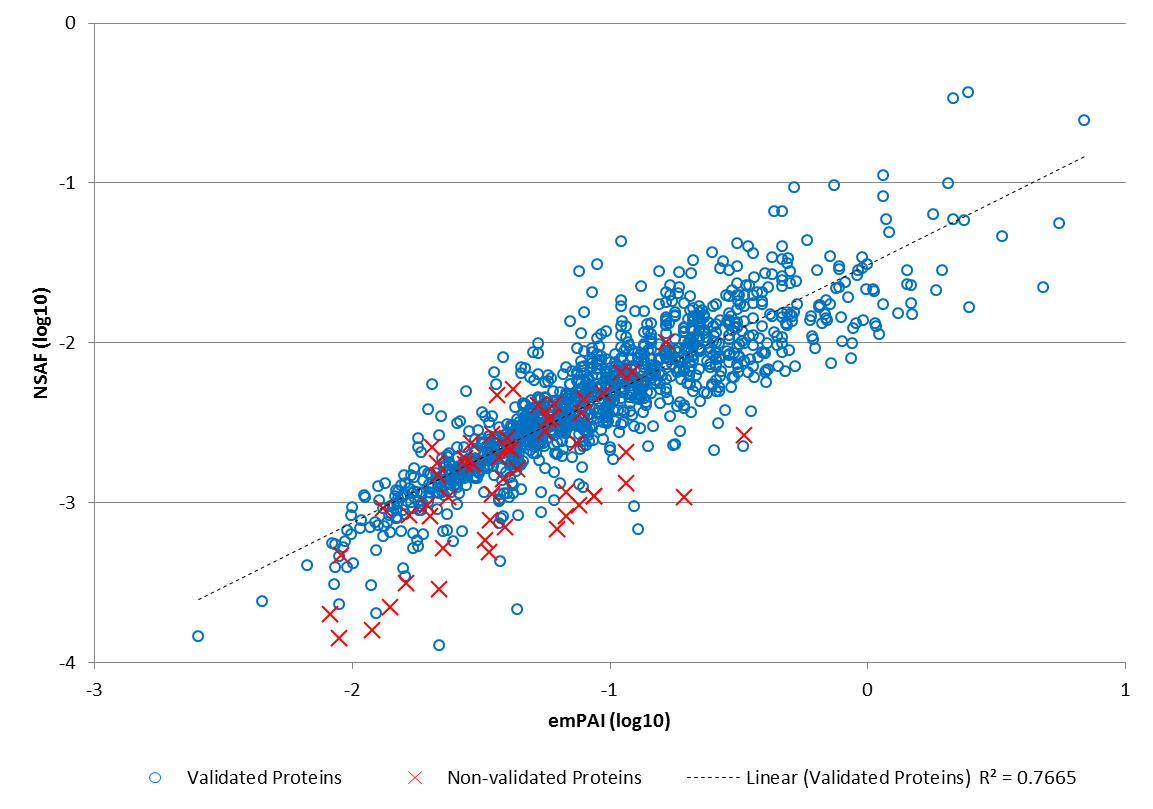


Note that the spectrum counting quantification index has already been calculated for you and is readily displayed in the ‘MS2 Quant.’ column of the protein table. By default the NSAF index is selected as it was found to be more reliable[4](#_ENREF_4). Note also that PeptideShaker corrects the NSAF index automatically for sections of the protein sequence which cannot generate observable peptides and for protein inference problems using the methods detailed in the identification chapter[5](#_ENREF_5).

Export the results using the ‘Identification Features’ menu of the ‘Export Menu’, select ‘Protein Accession’, ‘Protein Description’, ‘Protein Inference’, ‘Protein Confidence’, ‘NSAF’ and ‘emPAI’. Make sure that not only validated proteins are exported by unchecking ‘Request Validated’:



The export can be imported into Microsoft Excel or Perseus for post-processing. If you plot the NSAF index against the emPAI indexes of the target validated and non-validated protein matches, you should get the following:



*How do the indexes compare? Why are we using a log scale?*

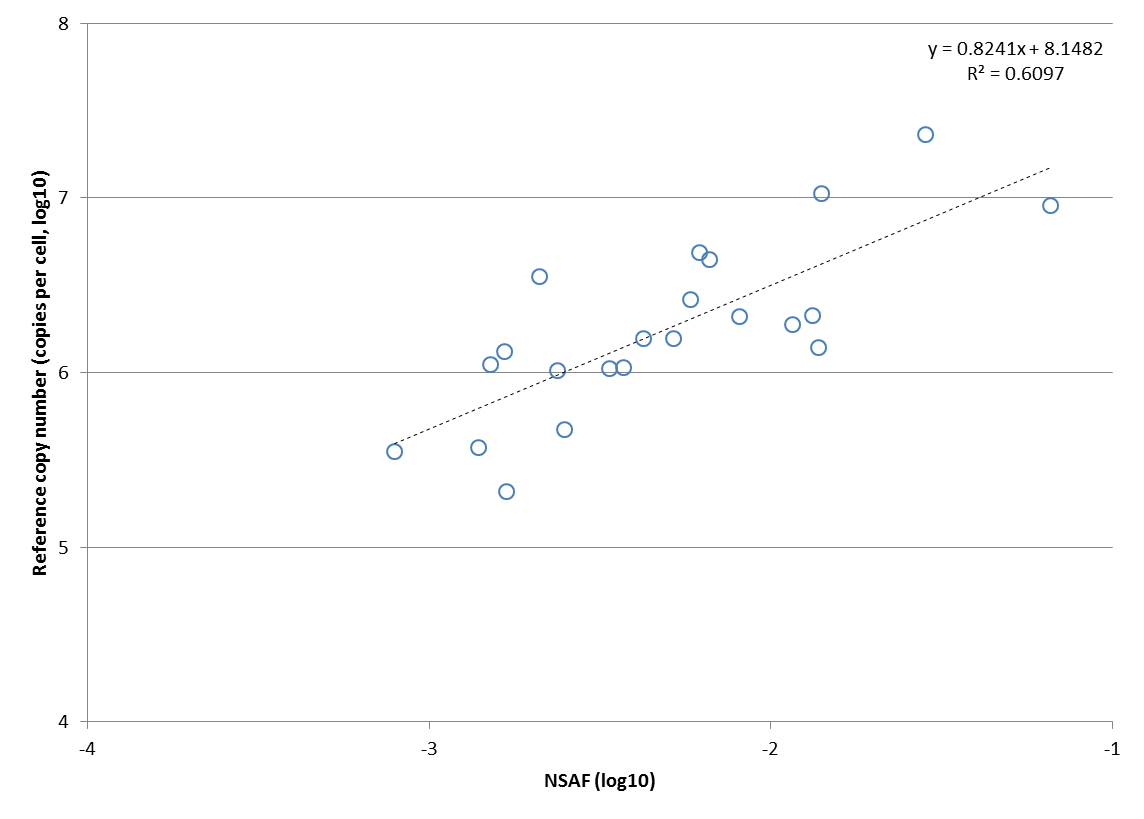
*The results of both indexes correlate well with an R2 value of 0.7665 which is not bad compared to the expected reliability of the method. Note that we cover four orders of magnitude, in order to fairly compare protein abundances on such a large range, we use the logarithm of the value.*

In order to give an illustrative value to these spectrum counting indexes, we will calibrate these values according to abundance levels from the literature[6](#_ENREF_6). In the file ‘spectrum\_counting.xlsx’, you will find these reference values and in the column ‘spectrum counting 1h’, the NSAF values exported by PeptideShaker.

*Why are some values missing? Can we infer a detection limit for our experiment?*

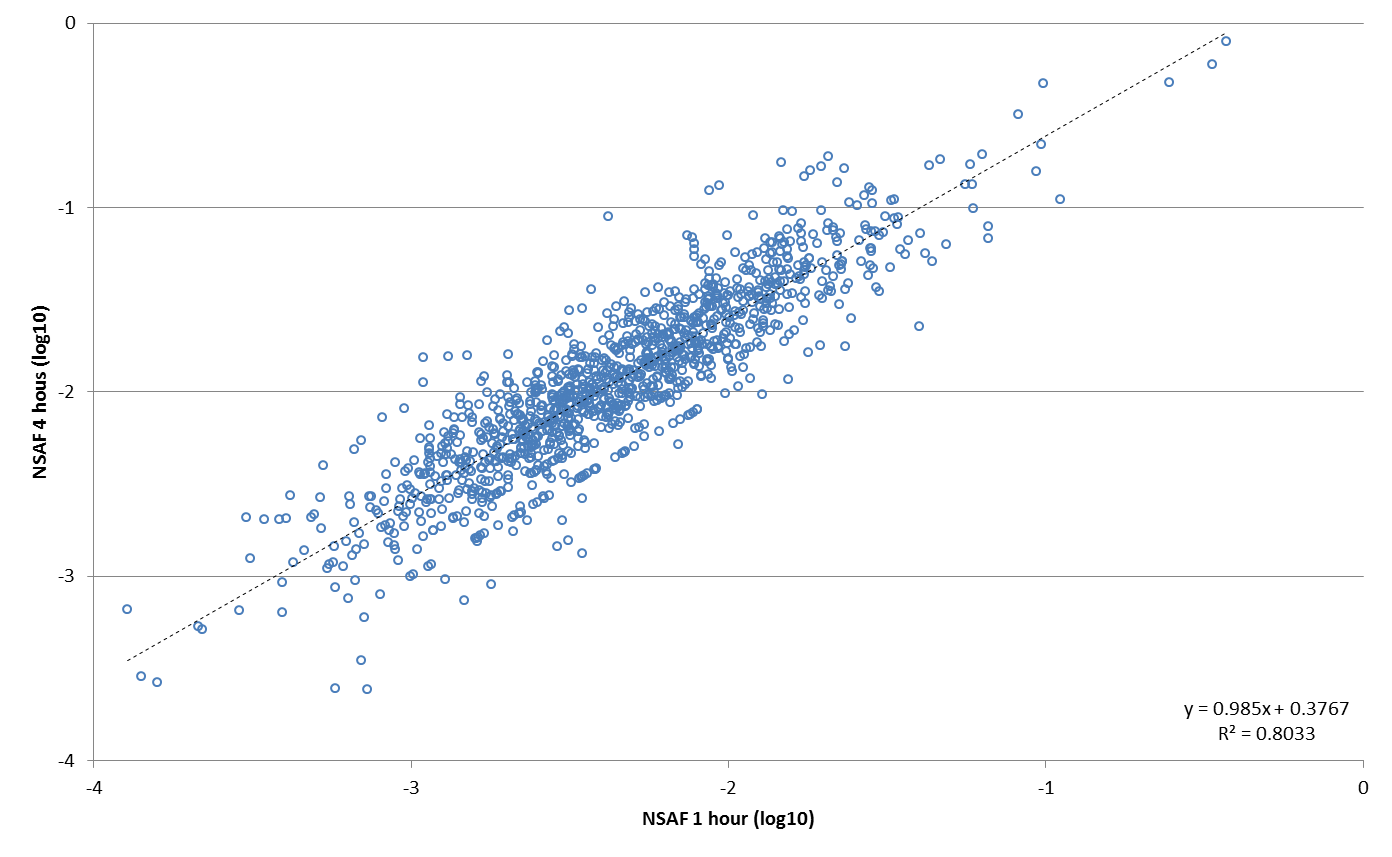
*This example consists of a single Q Exactive run (1 hour gradient) which does not allow full proteome coverage. In fact, only the most abundant proteins of the list were identified. We can clearly estimate our detection limit to 3E5 copies per cell. Note also that the probability of detecting a protein does not only depend on its abundance but also chemical properties. Complementary experimental methods might help increasing proteome coverage.*

If you plot the reference numbers against the NSAF value, you should see the following:

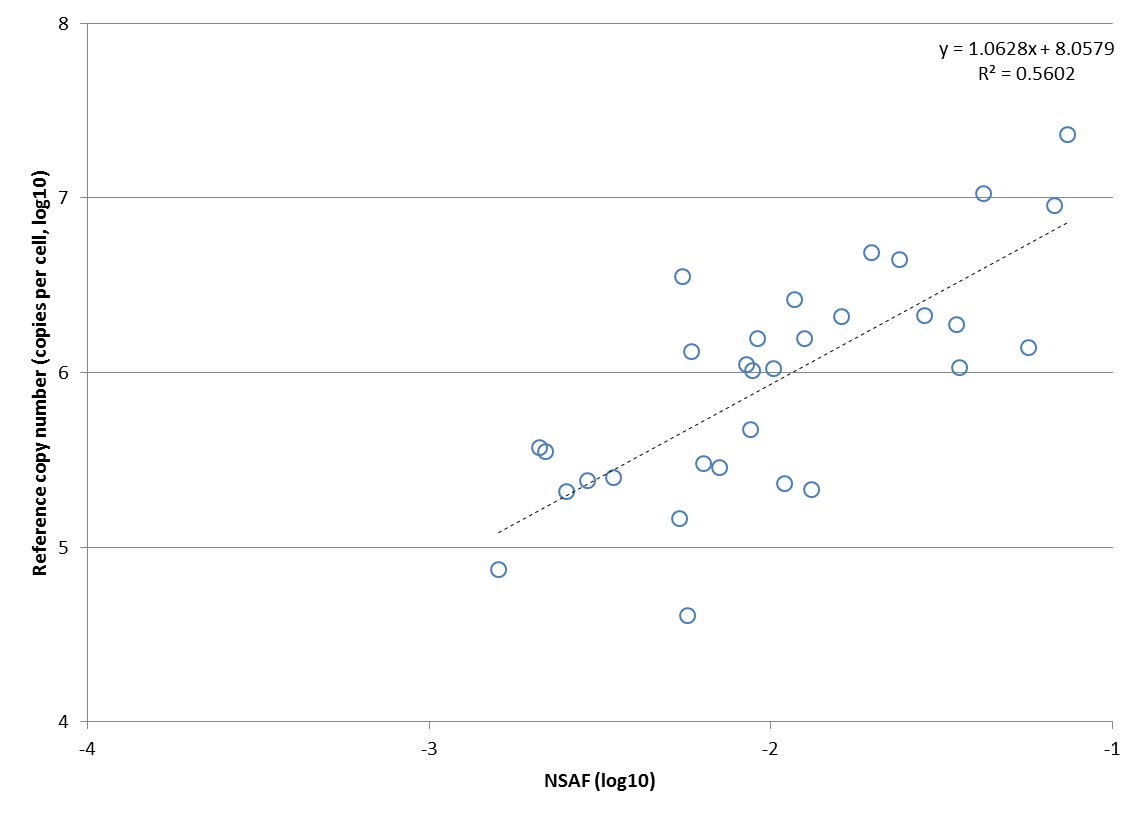


As you can see, we can use a simple formula to translate the NSAF index into a protein count using a linear regression (R2=0.6097).

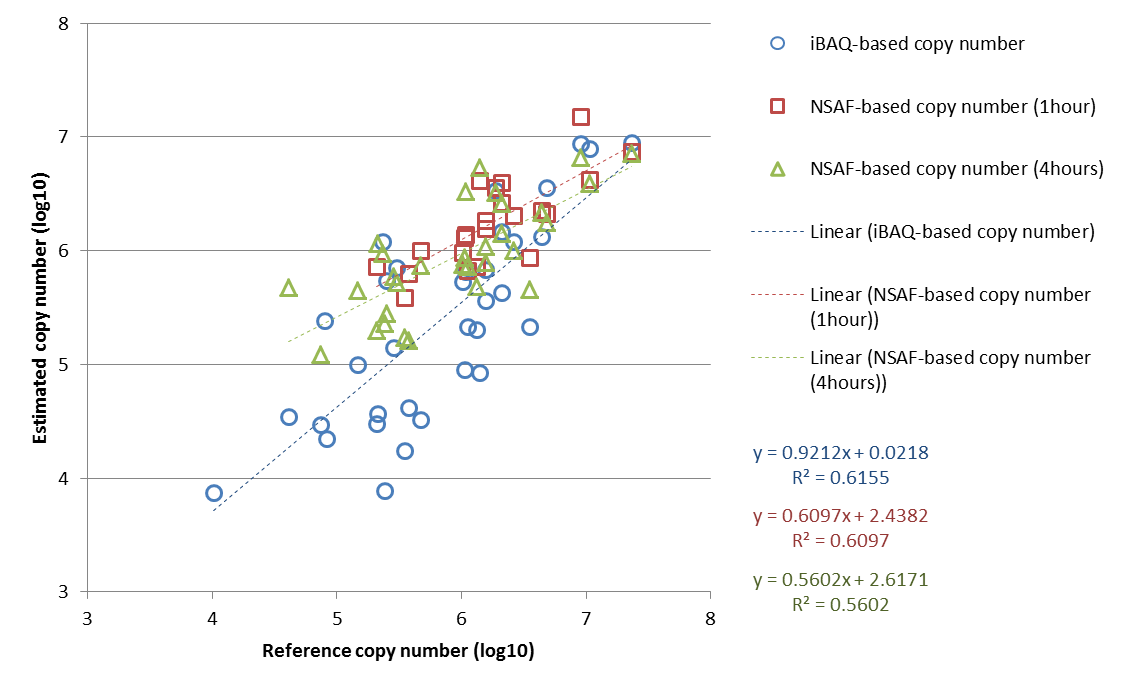
In order to increase our proteome coverage, we measured the same sample with a longer gradient (4 hours) - the PeptideShaker export is provided in the excel sheet. As you can see, 3,480 protein matches were validated for this run, three times more than in the previous dataset (1,212). If you plot the commonly identified proteins, you will observe the good correlation between the datasets:



Conduct the same normalization as for the previous dataset in order to obtain copy numbers. You should obtain the following:



In the ‘iBAQ-based copy number’, we provide the protein abundances obtained from MS1 intensity-based quantification[7](#_ENREF_7). If you plot the values obtained on the three datasets, you should see the following:

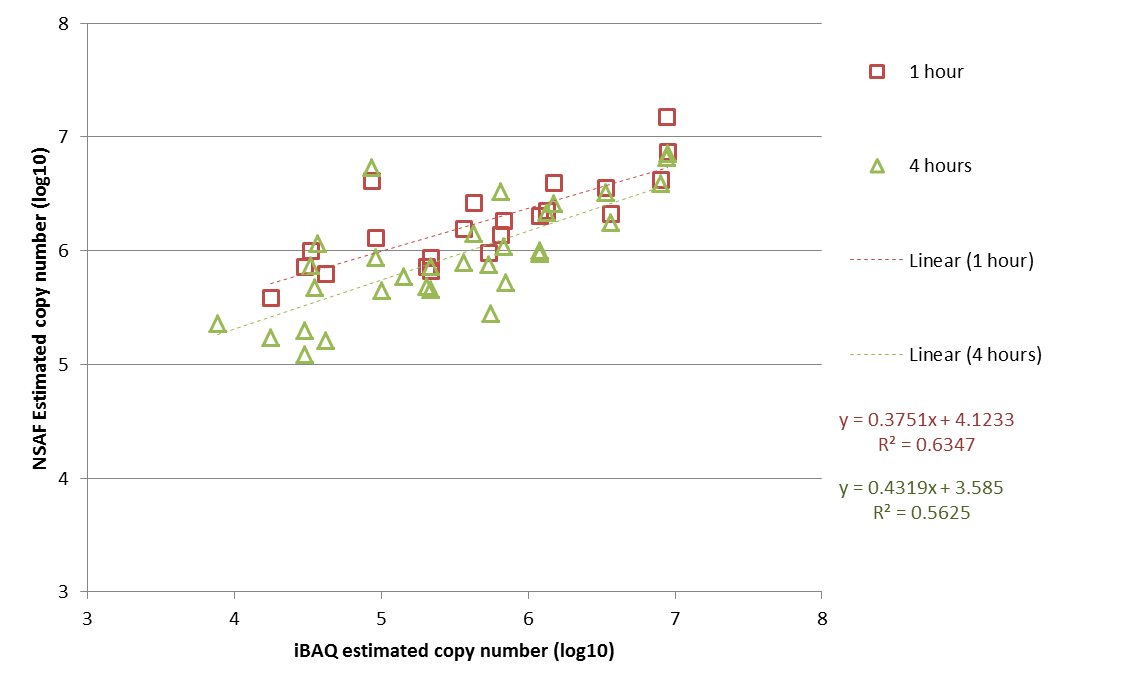


*What are the performances of the quantification on the different datasets?*

First, the different studies present different detection limit. The intensity based study identified very low abundant proteins (few thousands per cell) covering more than 10,000 proteins. This is mainly due to the experimental setup, much more complex than in our tutorial examples.

In terms of quantification however, the variability of the intensity based quantification R2=0.6155 is very similar to the one obtained with spectrum counting, respectively 0.6091 and 0.5602 for a one and four hours gradient.

When comparing the spectrum counting approach to the intensity based quantification, you will obtain the following:



Note that the R2 values for spectrum counting compared to intensity based quantification are very similar as the ones compared to the reference. We can hence assume that this variability is due to the quantification error and not the reference.

*What are the advantages and shortcomings of spectrum counting in comparison to intensity based quantification?*

Spectrum counting quantification presents the advantage of being extremely simple and fast to compute. His main shortcoming is the limited accuracy and lack of robustness as pointed by the dependence on the identification procedure. Hence, it typically serves as fast and rough abundance estimator. For more accurate results, it is preferable to set-up an SRM experiment.

References

1. Liu, H., Sadygov, R.G. & Yates, J.R., 3rd A model for random sampling and estimation of relative protein abundance in shotgun proteomics. *Analytical chemistry* **76**, 4193-4201 (2004).

2. Ishihama, Y. et al. Exponentially modified protein abundance index (emPAI) for estimation of absolute protein amount in proteomics by the number of sequenced peptides per protein. *Molecular & cellular proteomics : MCP* **4**, 1265-1272 (2005).

3. Powell, D.W. et al. Cluster analysis of mass spectrometry data reveals a novel component of SAGA. *Molecular and cellular biology* **24**, 7249-7259 (2004).

4. Colaert, N., Gevaert, K. & Martens, L. RIBAR and xRIBAR: Methods for reproducible relative MS/MS-based label-free protein quantification. *Journal of proteome research* **10**, 3183-3189 (2011).

5. Burkhart, J.M. et al. The first comprehensive and quantitative analysis of human platelet protein composition allows the comparative analysis of structural and functional pathways. *Blood* **120**, e73-82 (2012).

6. Zeiler, M., Straube, W.L., Lundberg, E., Uhlen, M. & Mann, M. A Protein Epitope Signature Tag (PrEST) library allows SILAC-based absolute quantification and multiplexed determination of protein copy numbers in cell lines. *Molecular & cellular proteomics : MCP* **11**, O111 009613 (2012).

7. Nagaraj, N. et al. Deep proteome and transcriptome mapping of a human cancer cell line. *Molecular systems biology* **7**, 548 (2011).