Supplementary Note 1. MAYU Software.

The **MAYU** software is a perl command line program and can be downloaded together with a detailed manual at:

http://tools.proteomecenter.org/Mayu.php

If a perl interpreter is installed on the system **MAYU** can be run directly after unpacking of the zip file.

Essential input. MAYU utilizes the search results of a target-decoy search and the corresponding database as input. The search results can be provided in three formats: pepXML, Mascot .csv or a format specific for **MAYU**. The **MAYU** format is specified as a comma separated value file with the following columns:

1. scan (run.scannr.scannr.charge)
2. raw peptide sequence of identification
3. protein identifier (decoy ids must have a prefix)
4. modifications (pos1=mass1:pos2=mass2)
   position: position starting with 1, 0 and L+1 for N and C-terminal modifications respectively
   mass: amino acid mass minus water plus modification in dalton
5. discriminant score, where a high score defines a good match (e.g. PeptideProphet probability score)

One line in such a **MAYU** input file could look like:

run1.10.10.2,DTKMLMK,F02H6.4,4=147.192:6=147.192,0.6824

Output. MAYU writes a number of table output files. The **Mayu** file is the major output file and stores information about the global peptide identification FDR (pepFDR), global protein identification FDR (protFDR), single hit FDR (protFDRs) and all but single hit FDR (protFDRns) along with additional information. The **Mayuidout** file contains a list of PSMs that were filtered with a user defined FDR on PSM, peptide or protein level. In the **mFDR** file the discriminant score and its corresponding PSM FDR is saved along with additional statistics on FP and TP PSM. The **bin_protFDR** file contains information on the protein identification FDR/protFDR calculated in the protein size bins along with additional statistics on

FP and TP protein identifications. The **feat_prot** file stores feature information (e.g. number of supporting PSM, PSM alignment type) for each protein identification and can be used to estimate local FDR.

Features. In addition to FDR calculation, **MAYU** allows to monitor and therefore improve experimental design. In order to assess the characteristics of the data, a range of data selection schemas can be applied and the **MAYU** analysis is automatically performed on each of these data subsets:

- cumulative input files (e.g. experiments)
- shuffled cumulative input files (e.g. experiments)
- cumulative runs
- shuffled cumulative runs
- orthogonality (non-redundancy) sorted runs

The orthogonality analysis allows to score the performance of LC-MS/MS runs compared to the rest and the cumulative contribution of each part to the total data set. The **feat_prot** file can be used to derive a range of local FDR.

Supplementary Note 2. R code for MAYU protein identification false discovery rate estimation.

This is a snippet of R code for the estimation of the number of false positive protein identifications (numbers from **Fig. 2b)**. **MAYU** performs such a calculation for each protein size partition (default is to 20 partitions). The block implementing the FDR estimation is highlighted in grey.

```
# numbers from Fig. 2b
target_PID <- 11 # number of target protein identifications
decoy_PID <- 7 # number of decoy protein identifications
total_PID <- 19 # total entries in the protein database
```

```
# vectors of possible TP, FP PID combinations
v_possible_FP_PID <- 0:decoy_PID
v_possible_TP_PID <- target_PID - v_possible_FP_PID
v_possible_not_TP_PID <- total_PID - v_possible_TP_PID

# hypergeometric distribution returns a probability for each number of FP PID
hyper_prob <-
      dhyper(v_possible_FP_PID,v_possible_not_TP_PID,v_possible_TP_PID,decoy_PID)
hyper_prob <- hyper_prob/sum(hyper_prob)

expectation_value_FP_PID <- round( sum( hyper_prob*v_possible_FP_PID ), 0 )
PID_FDR <- round( expectation_value_FP_PID / target_PID, 2 )
```

```
# make a plot
plot(
        hyper_prob ~ v_possible_FP_PID,
        main=expression( paste( italic(MAYU),
                ": protein identification false discovery rate estimation",
                sep="" ) ),
        ylab=expression("P"(h["fp"])),
        xlab=expression(h["fp"]),
        ylim=c(0,1)
)
legend( "topleft", legend=c(
                paste( "protein identification false discovery rate:", PID_FDR ),
                paste( "target protein idenfitications:", target_PID ),
                paste( "decoy protein idenfitications:", decoy_PID ),
                paste( "total proteins in database:", total_PID ) )
)
abline( v = expectation_value_FP_PID, lty=2 )
```

Supplementary Figure 1. MAYU protein identification false discovery rates are little influenced by the choice of decoy database. Protein identification false discovery rate (FDR) estimates are stable with respect to the underlying decoy database. We show this by repeated database searches of the ***C. elegans*** data set, each based on a different decoy database (see **Supplementary Method 1**). Relative standard deviation of the resulting FDR estimates in any case fell below 10% (**a, c**). We observe a slight trend towards larger variability of the corresponding single hit FDR estimates, revealing the limitations of the non-parametric estimates of protein identification property distributions (**c, d**).

Supplementary Figure 2. Protein identification false discovery rate for protein inference excluding ambiguous peptides. From the total data set of 20 experiments all peptide-spectrum matches (PSMs) referring to peptides pointing to more than one (target or decoy) protein, were removed. For the remaining PSMs the protein identification false discovery rate (FDR) was estimated. This protein inference method has no influence on the general behaviour of the protein identification FDR estimates as expected from the underlying model.

Supplementary Method 1. Target-decoy database generation.

We generated ten different decoy databases by sampling from a zeroth order Markov model with amino acid frequencies and protein length distribution gathered from the target database (**Supplementary Fig. 1**). Since randomizing of redundant sequences leads to a decoy database being "effectively" larger, i.e. featuring a larger amount of non-redundant sequences, than the target database [1], we corrected the target database prior to sampling of amino acids. This was done for the splice variants by removing random amino acids from the non main splice variants accordingly (with the main splice variant being the alphabetically first). If there were groups of identical protein sequences all but one of these were deleted.

Supplementary Method 2. Formal derivation of the protein identification FDR estimate.

The set of PSMs produced in the course of a proteomics experiment give rise to protein identifications. A set of PSMs mapping to the same protein sequence defines a protein identification. In the following we refer to the set of all protein identifications as $\boldsymbol{H}$, the subset mapping to the target database $\boldsymbol{P_t}$ as $\boldsymbol{H_t}$ and its complement as $\boldsymbol{H_d}$. We distinguish three types of protein identifications, i.e. (1) TP identifications, which all together we denote $\boldsymbol{H_{tp}}$. A protein identification is considered to be TP, if it contains at least one TP PSM. While the second type (2) covers the set $\boldsymbol{H_{fp}}$ of FP protein identifications mapping to $\boldsymbol{P_t}$, the complementing set with its identifications projecting to the decoy database $\boldsymbol{P_d}$ equals $\boldsymbol{H_d}$. A protein identification is considered to be FP, if all of its PSM are FP. As the third type (3) we introduce the set $\boldsymbol{H_{cf}}$ that is composed of all protein identifications in $\boldsymbol{P_t}$ each containing FP PSM. Note that elements of $\boldsymbol{H_{cf}}$ can be TP as well as FP. The size of the defined sets shall be denoted by lowercase letters, as for instance $|\boldsymbol{H}| = \boldsymbol{h}$.

Making the reasonable assumption that FP PSM equally likely map to either target or decoy database, it is straightforward to estimate the expected value of FP PSM mapping to $\boldsymbol{P_t}$ with the number of PSM pointing to $\boldsymbol{P_d}$. According to the definition of false discovery rates [2], we can estimate the PSM FDR as the ratio of the number

of PSM pointing to P_d and P_t respectively**.** Considering that target and decoy database share the same protein length distribution, the expected value for h_{cf} can be estimated analogously with h_d. Note that h_{cf} does not necessarily equal h_{fp}.

In order to determine the FDR for protein identifications, we firstly calculate the conditional expectation value for $E[\, h_{fp} \mid h_t, h_d, \theta_{exp}]$ for the number of FP protein identifications given the proteomics experiment characterized by parameters θ_{exp} and its outcome h_t, h_{cf}. Amongst others, θ_{exp} particularly includes parameters related to the target protein database, such as the number of protein entries N. By application of Bayes' formula and by assuming $P(h_{tp} \mid h_{cf}, \theta_{exp})$ and $P(h_t \mid h_{cf}, \theta_{exp})$ to be uniform and $h_d = h_{cf}$, $E[\, h_{fp} \mid h_t, h_d, \theta_{exp}]$ evaluates as follows.

$$
\begin{aligned}
E\left[h_{fp} \middle| h_t, h_{cf}, \theta_{\exp}\right] &= \sum_{h_{fp}} h_{fp} \cdot P\left(h_{fp} \middle| h_t, h_{cf}, \theta_{\exp}\right) \\
&\overset{h_{tp}=h_t-h_{fp}}{=} \sum_{h_{fp}} h_{fp} \cdot P\left(h_{tp} \middle| h_t, h_{cf}, \theta_{\exp}\right) \\
&= \sum_{h_{fp}} h_{fp} \cdot \frac{P\left(h_t \middle| h_{tp}, h_{cf}, \theta_{\exp}\right) \cdot P\left(h_{tp} \middle| h_{cf}, \theta_{\exp}\right)}{P\left(h_t \middle| h_{cf}, \theta_{\exp}\right)} \\
&\overset{h_{tp}=h_t-h_{fp}}{=} \sum_{h_{fp}} h_{fp} \cdot \frac{P\left(h_{fp} \middle| h_{tp}, h_{cf}, \theta_{\exp}\right) \cdot P\left(h_{tp} \middle| h_{cf}, \theta_{\exp}\right)}{P\left(h_t \middle| h_{cf}, \theta_{\exp}\right)} \\
&= \sum_{h_{fp}} h_{fp} \cdot P\left(h_{fp} \middle| h_{tp}, h_{cf}, \theta_{\exp}\right) \cdot \frac{N - h_{cf} + 1}{N + 1}
\end{aligned}
$$

Let us assume for a moment that all protein sequences in the target and decoy database have the same size. As the probability of a FP PSM mapping to a certain protein sequence scales linearly with its size, each entry in P_t would be equally likely to be part of H_{cf}. Thus, protein identifications containing FP PSM would be uniformly distributed across P_t. Accordingly, $P(h_{fp} \mid h_{tp}, h_{cf}, \theta_{exp})$ would follow the hypergeometric distribution, where h_{fp} is modelled as a random variable representing the number of successful hits of a "non-TP-identified" protein in a sequence of h_{cf} draws without replacement from the N entries in P_t.

Clearly, the initial assumption about the singular size distribution does not hold for biological protein databases. So as to compile an estimate for $E[\, h_{fp} \mid h_t, h_d, \theta_{exp}]$ from subgroups closely meeting this assumption, we have partitioned $P = P_t \cup P_d$ into

subsets $\boldsymbol{P_i}$ of protein sequences of similar size. In this context, protein sequence size is defined as number of tryptic peptides from in silico digestion (400-6000 Da, ≤ 2 missed cleavages). Variables $\boldsymbol{h_{t,i}}$, $\boldsymbol{h_{cf,i}}$, $\boldsymbol{h_{tp,i}}$, $\boldsymbol{h_{fp,i}}$, $\boldsymbol{N_i}$ are defined for each $\boldsymbol{P_i}$ in analogy to those for $\boldsymbol{P}$. By applying the foregoing argument we approximate $\boldsymbol{E}[\ \boldsymbol{h_{fp}} \mid \boldsymbol{h_t}, \boldsymbol{h_d}, \theta_{exp}]$ as follows

$$\hat{\boldsymbol{E}}\left[\boldsymbol{h_{fp}} \middle| \boldsymbol{h_t}, \boldsymbol{h_{cf}}, \theta_{\exp}\right] = \sum_i \hat{\boldsymbol{E}}\left[\boldsymbol{h_{fp,i}} \middle| \boldsymbol{h_{t,i}}, \boldsymbol{h_{cf,i}}, \theta_{\exp}\right]$$

$$= \sum_i \sum_{h_{fp,i}} \boldsymbol{h_{fp,i}} \cdot \boldsymbol{P}\left(\boldsymbol{h_{fp,i}} \middle| \boldsymbol{h_{cf,i}}, \boldsymbol{h_{tp,i}}, \theta_{\exp}\right) \cdot \frac{\boldsymbol{N_i} - \boldsymbol{h_{cf,i}} + 1}{\boldsymbol{N_i} + 1}$$

where

$$\boldsymbol{P}\left(\boldsymbol{h_{fp,i}} \middle| \boldsymbol{h_{cf,i}}, \boldsymbol{h_{tp,i}}, \theta_{\exp}\right) = \frac{\binom{\boldsymbol{N_i} - \boldsymbol{h_{tp,i}}}{\boldsymbol{h_{fp,i}}} \binom{\boldsymbol{h_{fp,i}}}{\boldsymbol{h_{cf,i}} - \boldsymbol{h_{fp,i}}}}{\binom{\boldsymbol{N_i}}{\boldsymbol{h_{cf,i}}}}$$

We have assessed this approximation for $\boldsymbol{E}[\ \boldsymbol{h_{fp}} \mid \boldsymbol{h_t},\ \boldsymbol{h_d},\ \theta_{exp}]$ by confirming quick convergence in experiments with various partitions featuring increasing size homogeneity within the subsets (**Fig. 3a**).

We obtain the final estimate for FDR by appropriately inserting $\hat{\boldsymbol{E}}\left[\boldsymbol{h_{fp}} \middle| \boldsymbol{h_t}, \boldsymbol{h_{cf}}, \theta_{\exp}\right]$.

$$\hat{\text{pFDR}} = \frac{\hat{\boldsymbol{E}}\left[\boldsymbol{h_{fp}} \middle| \boldsymbol{h_t}, \boldsymbol{h_{cf}}, \theta_{\exp}\right]}{\boldsymbol{h_t}}$$

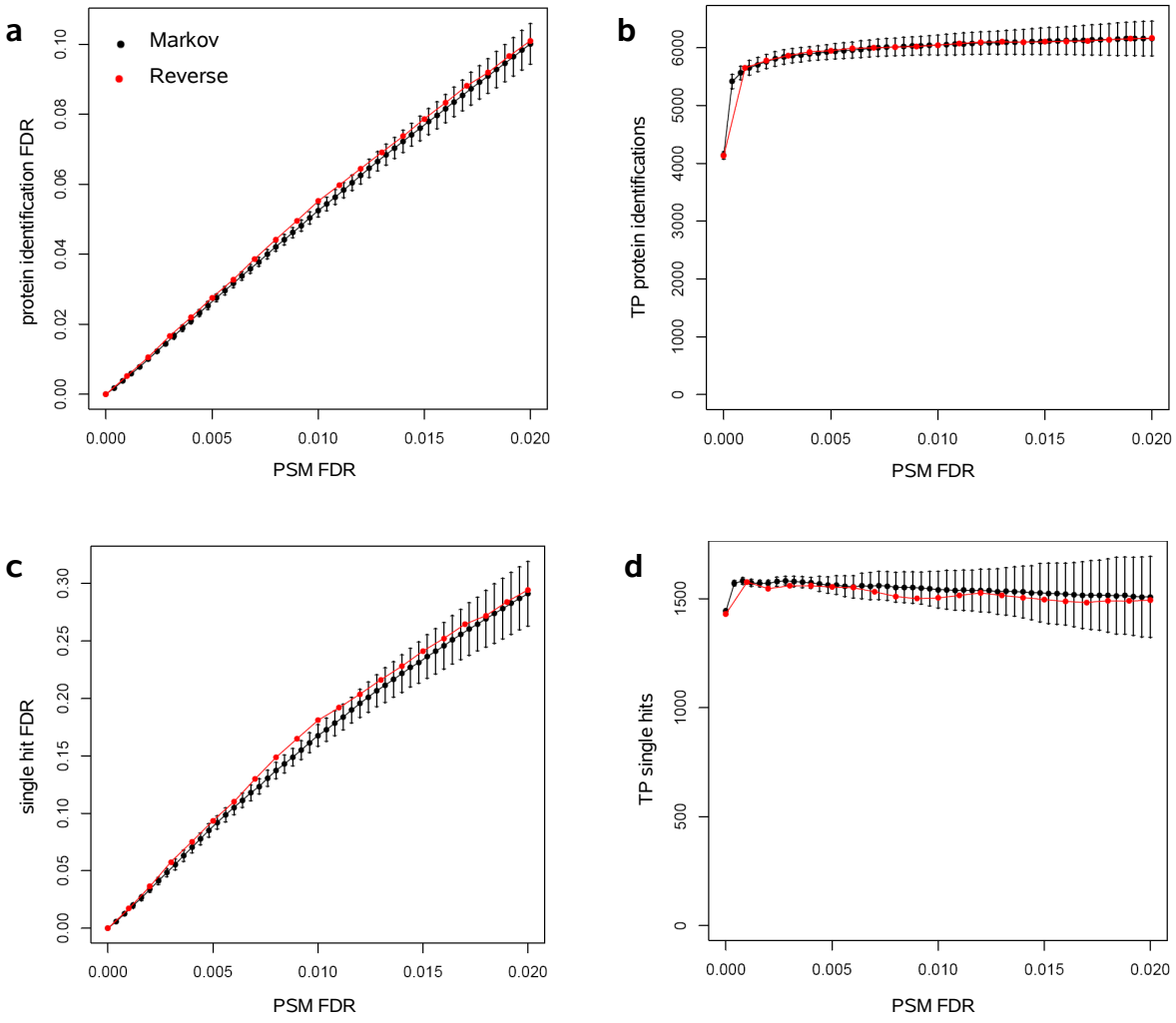
Supplementary Method 3. Figures and tools.

Figures were generated using the R statistical package [3] and OpenOffice.

REFERENCES

1. Elias, J.E. and S.P. Gygi, ***Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry.*** Nat Methods, 2007. **4**(3): p. 207-14.
2. Benjamini, Y. and Y. Hochberg, ***Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing***.
3. ***R Development Core Team. R: A Language and Environment for Statistical Computing.*** 2008.

Supplementary Figure 1. Reiter, Claassen et al.



Supplementary Figure 2. Reiter, Claassen et al.