# Supporting Information

## Proteomics Quality Control – A Quality Control Software for MaxQuant Results

Chris Bielow[1,2,*] (chris.bielow@mdc-berlin.de), Guido Mastrobuoni[1] (guido.mastrobuoni@mdc-berlin.de), Stefan Kempa[1,2,*] (stefan.kempa@mdc-berlin.de)

[1] Max-Delbrück-Centrum for Molecular Medicine Berlin, Robert-Rössle-Straße 10, 13125 Berlin-Buch

[2] Berlin Institute of Health, Kapelle-Ufer 2, 10117 Berlin

*Corresponding authors: Chris Bielow (chris.bielow@mdc-berlin.de) and Stefan Kempa (stefan.kempa@mdc-berlin.de), Tel: +49 30 9406 3114, Fax: +49 30 9406 49164

## Table of Contents

# Summary of datasets

The mass spectrometry proteomics data have been either obtained (dataset 1) or have been deposited to (dataset 2) the ProteomeXchange Consortium via the PRIDE partner repository[1]. The dataset identifiers are given below.

Dataset 1:

**Source:** external

- Pride Archive, ID PXD000427 (http://www.ebi.ac.uk/pride/archive/projects/PXD000427)

**Short description**: 24 fractions of 6-plex TMT samples from three patients with Parkinson's disease vs. three controls.
**MaxQuant version**: 1.5.1.2
**Instrument**: LTQ-Velos Orbitrap
**Label**: TMT
**Fractions**: yes, 24
**Highlighted Problem:**

- false annotation of fraction number 13
- low intensity for a handful of fractions; potential for merging fractions

Dataset 2:

**Source**: in-house

- Pride Archive, ID PXD003133 (http://www.ebi.ac.uk/pride/archive/projects/PXD003133)
- Pride Archive, ID PXD003134 (http://www.ebi.ac.uk/pride/archive/projects/PXD003134)

**Short description**: four samples of a human HEK293 cell line; routinely used for machine performance testing
**MaxQuant version**: 1.5.1.2/1.5.2.8
**Instrument**: LTQ-Velos Orbitrap, Q-Exactive Plus
**Label**: label-free
**Fractions**: no
**Highlighted Problem:**

- match-between-runs performance (alignment and ID-transfer)
- mycoplasma contamination

See PDF report files in the Suppl. Information for the complete reports for each data set.

# Figure S1

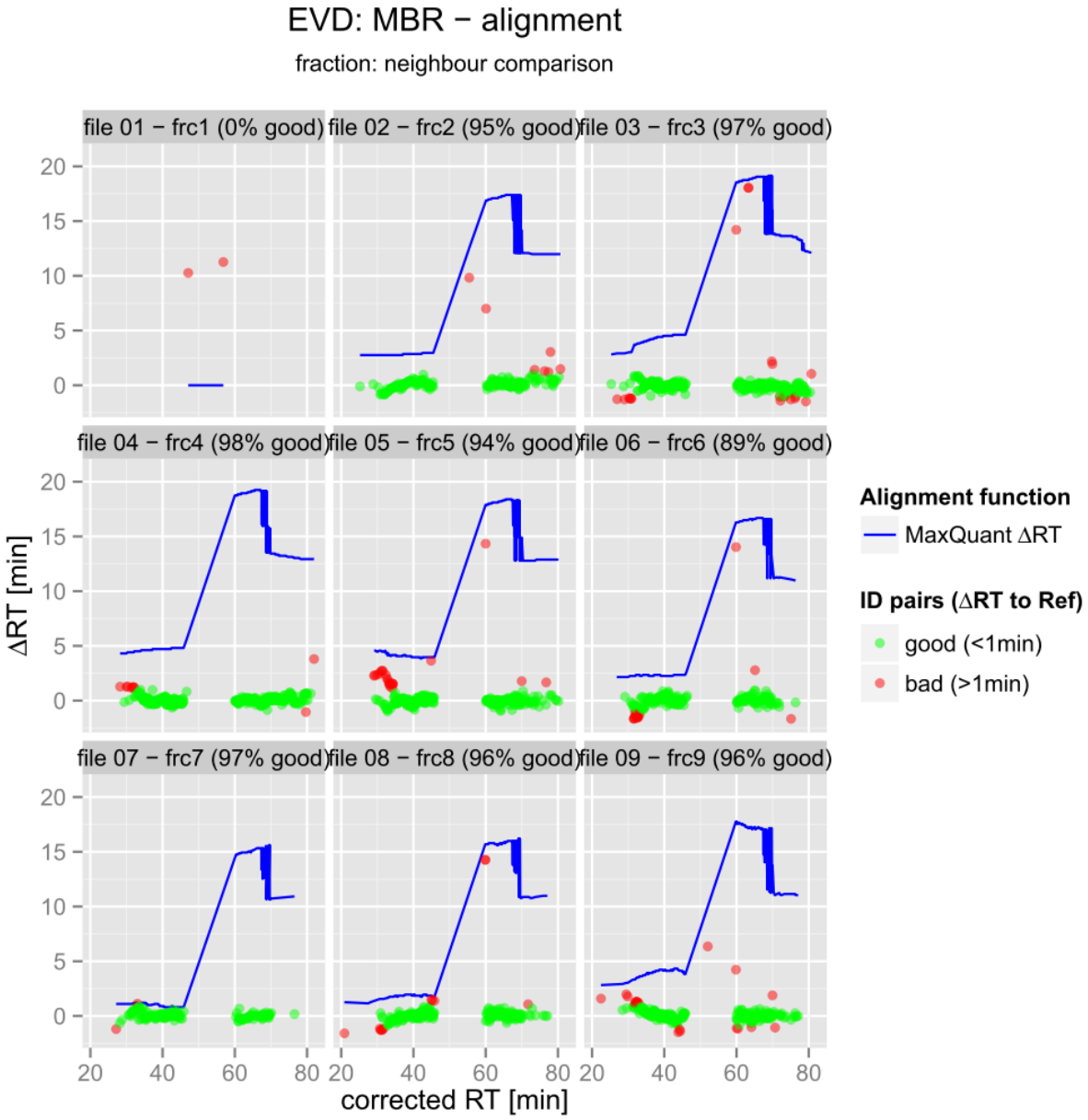


*Figure S1: PTXQC alignment validation plot for data set 1 (TMT). If MaxQuant was supplied with fraction data, PTXQC will indicate this in the sub-title and list the file name and fraction numbers above each subplot panel as "<filename> - frc <fraction>". Each file with fraction 'i' is compared to all other files with fraction between i-1 and i+1. Here, each fraction has only one representative, therefore it is compared to at most two other files, e.g. 'file 02' is compared against 'file 01' and 'file 03'. The first fraction does not align well with fraction 2 ('file 02'), hence there are almost no common ID-pairs. Obviously, the situation is the same vice versa. Fraction two scores higher (95%) than fraction 1 (0%), since it benefits from mostly good pairs with fraction 3. All other fractions can be aligned well to their immediate neighbors and receive high scores.*

# Complete Set of Metrics provided by PTXQC

In the following section we will describe each metric/plot and give an example where appropriate. We have used various data sets as basis for different metrics, to demonstrate all features of the report. Every plot contains the name of the data set it was derived from. A summary of data sets and the complete PTXQC reports in PDF format can be found in the SupportingInformation_2.pdf.

**Input**

The input is the txt folder generated by MaxQuant (see main article for details). By default, up to 24 quality control metrics are calculated from six MaxQuant txt files – see Table 1. Most of them can be scored. Five of them remain unscored since they originate from the proteinGroups.txt, where a 1:1 relationship between groups and Raw files is not guaranteed.

**Parameters/Meta data**

The parameters.txt summarizes the settings used for the MaxQuant analysis. The content is reformatted and presented as the first page of the report. Key parameters are MaxQuant version, Re-quantify, Match between runs and mass search tolerances. A list of protein database files is also provided, allowing to track database completeness and database version information (if given in the filename).

## QC Metrics for Sample Preparation

**Default Contaminants Annotation (unscored)**

External protein contamination should be controlled for, therefore MaxQuant ships with a comprehensive, yet customizable protein contamination database, which is enabled by default. PTXQC generates a contamination plot derived from the proteinGroups table showing the fraction of total protein intensity attributable to contaminants. The plot employs transparency to indicate the total intensity, enabling the user to delineate a high contamination in high complexity samples from a high contamination in low complexity samples (e.g. from in-gel digestion). Note that this plot is based on experimental groups, and therefore may not correspond 1:1 to Raw files.

**Contaminant details**

In addition to the contaminant plot derived from proteinGroups.txt, the evidence table allows inspecting contaminations per Raw file, giving details about which contaminant is contributing how much (see Figure S-2). PTXQC will explicitly show the five most abundant protein contaminants, and summarize the remaining ones as ‘other’.

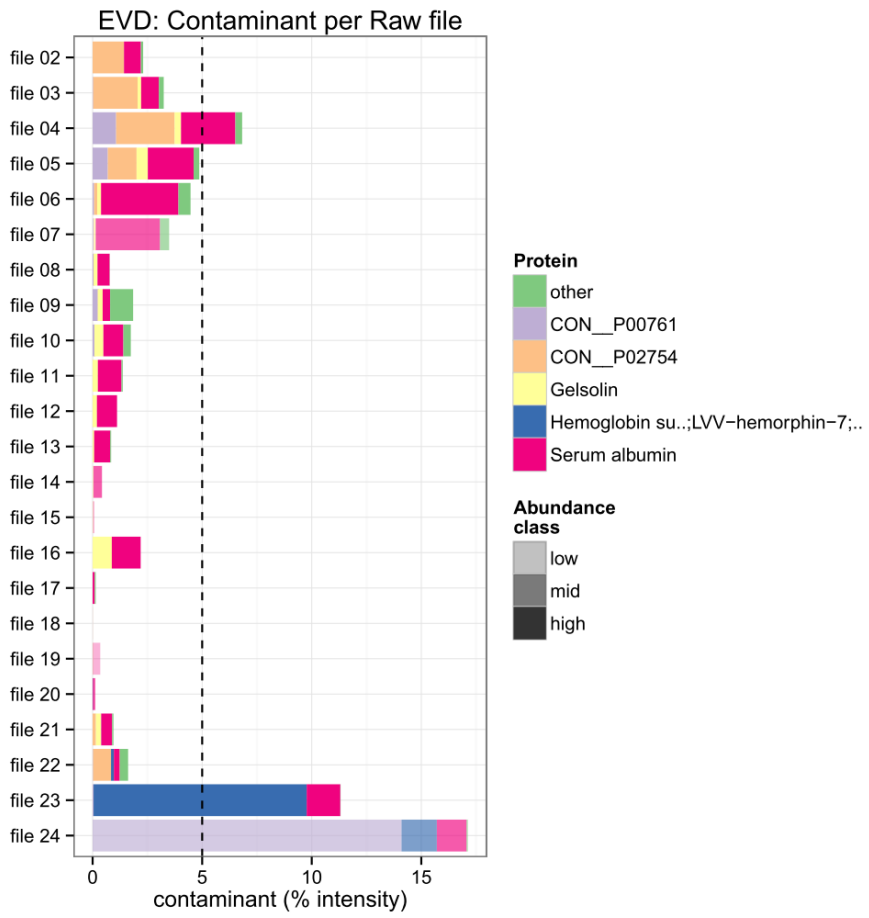


*Figure S-2: Individual contaminant contribution from each Raw file of data set 1 (TMT), based on the evidence table. Similar to contaminants from protein groups table, transparency is used to hint at the total intensity of each Raw file. File 24 shows high contamination of ~18%, while being in the class of lowest abundance ('low', hence having high transparency). File 23 however is in class "high" (about 100 fold more total intensity) with a notable contamination of ~11%. Most files show acceptable contamination below 5% of their total intensity.*

## Digestion: Missed Cleavages

Under optimal digestion conditions, only few missed cleavages (MC) are expected. Multiple studies observed vast differences in digestion performance, depending on enzyme grade and conditions[2]. In general, increased MC counts also increase the number of peptide signals, thus cluttering the available space and potentially provoking overlapping peptide signals, biasing peptide quantification. Thus, low MC counts should be favored, when possible. Interestingly, it has been shown recently that incorporation of peptides with missed cleavages does not negatively influence protein quantification[3]; however this is true only if all samples show the same degree of digestion. High missed cleavage values can indicate for example, either a) failed digestion, b) a high (post-digestion) protein contamination, or c) a sample with high amounts of unspecifically degraded peptides which are not digested by trypsin. PTXQC reports and scores the fraction of fully cleaved peptides per Raw file based on MaxQuant's msms table. Additionally, each Raw file is scored for its deviation from the 'average' digestion state of the current study. Each score represents one column in the heatmap.

**Peptide and Protein Intensity Overview**

The amount of material loaded onto the LC column has a major effect on both quantification and identification. Low column load leads to unfavorable signal-to-noise ratios and usually bad performance in general. Empirically, we found a median peptide intensity target value of about 800 k (23 in $\log_2$) as appropriate target value for LTQ-Velos Orbitrap data. For protein intensity in the proteinGroups table (unscored) a default value of about 33 million (25 in $\log_2$) is used, for both plain intensity and label-free quantification (LFQ) intensity. The algorithmic details of LFQ in MaxQuant have been published recently.[4] In short, LFQ normalizes the global intensity distribution across different groups by using a least squares criterion.
Failing to reach the intensity threshold is usually due to unfavorable column conditions, inadequate column loading or ionization issues. If the study is not a dilution series or pulsed SILAC experiment, we would expect every condition to have about the same median log-intensity (of about 25 [$\log_2$]). We compute and report the relative standard deviation (RSD) as a guidance parameter:

$$\mathrm{RSD}(\mathrm{x}) = \sigma(\mathrm{x})/\bar{x} \times 100 \qquad \text{(Eq. 1)}$$

where x is the vector of medians. Each median in x is computed from all protein intensities from the respective Raw file. Reliable experiments show an RSD below 5% for plain protein intensity. RSD values based on LFQ protein intensity are usually slightly worse, due to a higher count of 0's in the LFQ column (see [4] for details).
Depending on the type of experiment and the available data, a plot is automatically generated for raw protein intensities, LFQ intensities or reporter ion intensities (e.g. for TMT or iTRAQ experiments), allowing to judge channel loading.

**Protein Ratio Distribution (unscored)**

If PTXQC detects a labeling experiment (e.g. SILAC or pulse-chase SILAC) a ratio distribution plot will be generated based on protein ratios from proteinGroups.txt. Similar to protein intensity boxplots, this allows to spot unequal channel mixing during sample preparation. If equal mixing is expected, the distribution should be unimodal and its mode close to 1 (i.e., a 1:1 ratio), as indicated by a visual guidance line. Multimodal distributions are flagged as such automatically. If PTXQC detects ratios deviating strongly from 1:1 (parameterized by default beyond the range between 1:4 and 4:1), PTXQC automatically assumes a pulsed experiment and reports the label incorporation in percent for all groups.

## QC Metrics for Liquid Chromatography

**LC Peak Width**

One parameter of optimal and reproducible chromatographic separation is the distribution of widths of peptide elution peaks, which can be derived from the evidence table. Ideally, all Raw files show a similar distribution, e.g. to allow for equal conditions during dynamic precursor exclusion, RT alignment or peptide quantification.

**Distribution of Peptide Identifications IDs in the Chromatographic Dimension**

From the evidence.txt file a line plot is produced which allows to judge column occupancy over time. Ideally, the LC gradient is chosen such that the number of identifications (here, after FDR filtering) is uniform over time, to ensure consistent instrument duty cycles. Sharp peaks and uneven distribution of identifications over time indicate potential for LC gradient optimization.

## QC Metrics for Mass Spectrometry

**Charge Distribution**

Under common experimental conditions, tryptic peptides are generally expected to carry two charges (one N-terminal and one at the C-terminal R or K residue). However, charge states can also reach 3 (or higher). A single peptide species can occur in multiple charge states, charge two usually being the most abundant. Several factors, such as the presence of additional basic amino acids (e.g. due to missed cleavages), the spray parameters or the pH of the eluents, can shift the distribution towards higher charge states.
Consistent charge distribution is paramount for comparable 3D-peak intensities across samples, thus PTXQC extracts charge information from the evidence table for each Raw file and plots the distribution. To score the charge distribution for each Raw file, PTXQC computes the deviation of the charge 2 proportion from a representative Raw file.

**$MS^1$ Mass Decalibration**

By using the top hits of a tolerant $MS^2$ search for recalibration of $MS^1$ scans, MaxQuant can usually achieve precursor mass accuracy in the sub-ppm range. To verify that the recalibration was successful and the initial ppm tolerance was sufficient, we plot the uncalibrated (before re-calibration) and calibrated mass error distributions based on data from the evidence table. If the search margins for the precursor mass error were wide enough, every Raw file should show a rather narrow distribution of precursor mass errors. To obtain a quality score, PTXQC computes the distance of the average precursor mass error from zero with respect to the search margin. If most precursors were identified close to the margin (20 ppm by default) this indicates that the margin should be increased. If the user changed the default margin in MaxQuant, the margin parameter is matched automatically by PTXQC if the mqpar.xml configuration file provided. PTXQC will issue a warning if this is not the case.
A bug in MaxQuant sometimes leads to excessively high ppm mass errors ($>10^4$) reported in the output data. However, this can be corrected for by re-computing the delta mass error from other data. If this is the case, a warning ("bugfix applied") will be shown.

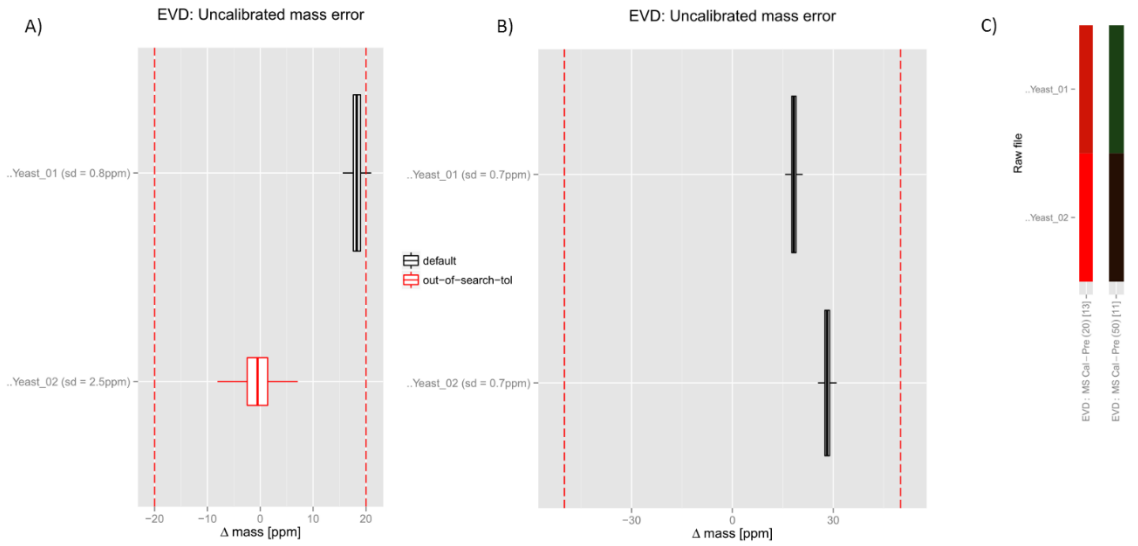


*Figure S-3: Precursor mass error plot before recalibration. Width (in ppm) of the bars on the x-axis shows precision spread; sample names are given on the y-axis. The height of each bar is proportional to the number of peptides identified. The precursor mass search tolerance is shown as vertical dashed bars and extracted from the mqpar.xml automatically. A) Two Raw files are shown. The first is barely inside the search tolerance of 20ppm, and can be calibrated successfully. The second file, is marked in red as "out-of-search-tol", indicating failed calibration. This is also supported by the high standard deviation (2.5ppm) shown on the left. B) The same two files, reanalyzed with a larger search tolerance of 40ppm. File 2 can be recovered and shows a decalibration of ~30ppm. C) Extract from the two heatmaps, corresponding to the analysis in panel A and B respectively. The scores improve for both files: file 1 is now further within the margin (30 ppm instead of 20 ppm), and file 2 is within the margin and fulfills the calibration critieria (standard deviation < 1ppm and id-rate > 1%).*

If the instrument is severely decalibrated beyond the tolerant search thresholds, then MaxQuant's recalibration will fail and subsequently identification rates for $MS^2$ spectra will be extremely low (<1%, after FDR filtering). Additionally, the standard deviation of calibrated precursor masses will be very high (>2 ppm compared to the usual <1 ppm for successful recalibration). PTXQC will detect these signs of extreme decalibration and report the affected Raw files as failed (annotated as "out-of-search-tol"). The corresponding heatmap score will be set to zero (failed, colored in red). An example can be found in Fig. S-3, using two in-house yeast samples. The reason for decalibration was an unexpected temperature rise due to climate control failure.

## $MS^1$ Mass Recalibration

Similarly, a plot for post-recalibration mass errors is shown. Here, the ppm errors should be centered on zero and their spread is expected to be significantly smaller. The variance and centeredness around zero of the calibrated distribution are used to derive a score for the summary heatmap. Figure S-4 shows an example corresponding to Figure S-3 after calibration.

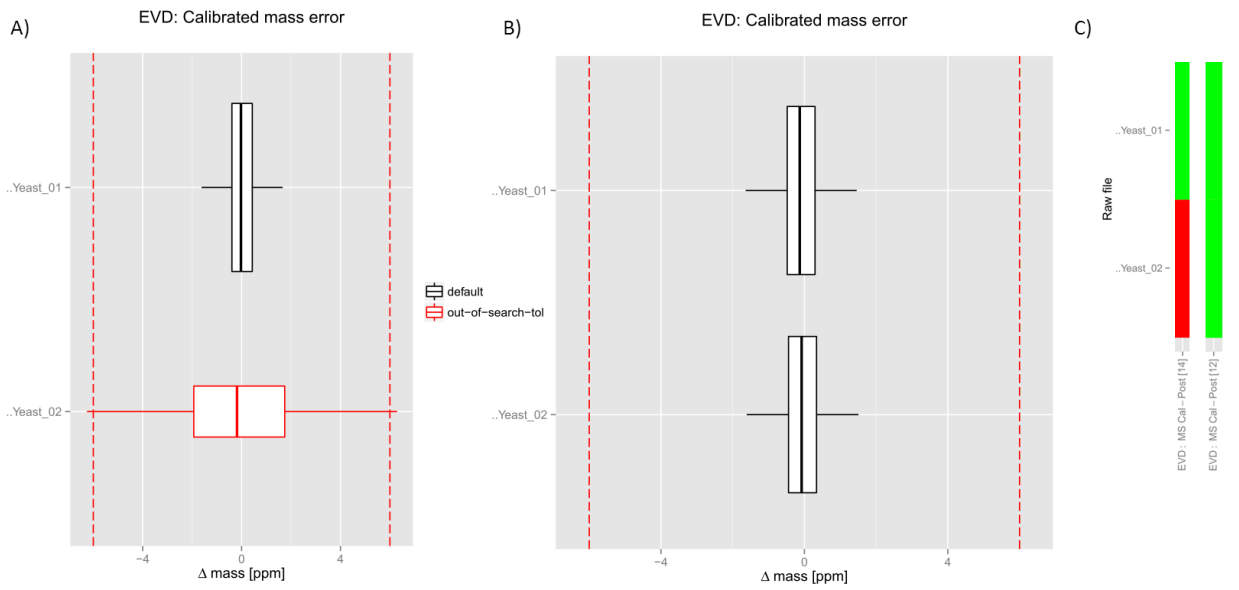


*Figure S-4: Calibrated mass error plot, matching the settings of Fig. S-3. The residual mass error is expected to be centered on zero, with a small spread. Panel A) shows the results for the data shown in Fig. S-3a. PTXQC knows that calibration failed, and marks the second file as such. Also note the larger spread in mass error, and the low bar height indicating few identifications. B) Using a larger search tolerance of 40 ppm, calibration was successful and the calibrated masses show a small spread. Also the number of identifications is on par with file 1 now. C) Extract from the two heatmaps, corresponding to the analysis in panel A and B respectively. The score for file 1 does not change; it was calibrated successful in both cases. File 2 failed calibration in (A), thus received 'red', but succeeded for panel (B), giving an almost perfect score.*

## $MS^2$ Fragment Mass Error

Analogous to the calibration of $MS^1$ precursor data, PTXQC checks on the mass accuracy of $MS^2$ fragments using the msms.txt file. MaxQuant/Andromeda allows for a certain mass tolerance when searching for fragment ions (e.g. 0.5 Da for ion trap-based spectra). If most of the fragments reported are within tighter bounds, the user can optimize the fragment mass tolerance to obtain more identifications under the same FDR. On the other hand, if the fragment mass errors are not centered on zero, a recalibration of the instrument should be performed. The heatmap score is computed by assessing the centeredness on zero with respect to the total matching tolerance.

## $MS^2$ Identification Rate

The summary.txt file provides direct access to the fraction of $MS^2$ fragment scans which were successfully identified and passed FDR thresholds. We provide a scatterplot, which groups performance into three categories (defaulting to, bad: <20%, ok: 20-35%, great: > 35%). Conditions classified as 'bad' are listed separately on the next page of the report for convenient follow-up (plot not shown). The heatmap score reaches 1 (100%) if the threshold for 'great' is reached or exceeded.

## Oversampling Estimation of 3D peaks in $MS^2$

An oversampled 3D-peak is defined as a peak whose peptide ion (same sequence and same charge state) was identified by at least two distinct $MS^2$ spectra in the same Raw file.
For high complexity samples, oversampling of individual 3D-peaks automatically leads to undersampling or even omission of other 3D-peaks, reducing the number of identified peptides. Oversampling occurs in low-complexity samples or long LC gradients, as well as undersized dynamic exclusion windows for data independent acquisitions. PTXQC computes the percentage of oversampled 3D-peaks from evidence.txt file for each Raw file. The percentage of non-oversampled 3D-peaks is used as quality score for the heatmap.

**Scan Event Performance (TopN)**

In data-dependent acquisition mode the instrument schedules a number of $MS^2$ scans dynamically after detecting candidate precursors from a preceding $MS^1$ scan. Under optimal conditions, the number of $MS^2$ scans should be constantly high over the whole chromatogram, while still being informative (i.e. of high-quality to enable identification). Optimal settings for each type of sample depend on sample complexity and other factors like LC gradient length. PTXQC plots the frequency of scan events per Raw file. The maximum number N of scan events between consecutive $MS^1$ scans is denoted TopN. The scoring function returns a high score if the instrument reached N scan events on a regular basis. However, a scheduled $MS^2$ event alone does not guarantee successful identification. Looking at the identification rates per scan event can give hints on how well scheduled precursor peaks could be fragmented and identified. Similar to *ID over RT*, one can spot dense regions in the LC gradient by plotting the maximum number of *TopN over RT*, i.e. the number of scheduled $MS^2$ spectra per time point. Low values over extended periods indicate that the LC gradient could be shortened, high values indicate saturation of the instrument's $MS^2$ capabilities. The scoring of this metric rewards a constant number of identifications over time and punishes sharp peaks.

## QC Metrics for Overall Performance

**Peptide and Protein Counts**

One of the most indicative QC metrics are peptide and protein counts. The number given by MaxQuant within the evidence and proteinGroups table already accounts for the false discovery rate threshold set by the user (1% by default in MaxQuant). The peptide count plot distinguishes between genuine and transferred peptides (found by MQ via Match-between-runs). Both peptide and protein counts are scored against a user-defined target threshold. Reaching this threshold (e.g. 15.000 peptides per Raw file for a LTQ-Velos Orbitrap ) will result in a maximum score.

**PCA Plot (unscored)**

If multiple conditions are compared, a principal component analysis (PCA) can be used to gauge similarities or differences between the analyzed samples. In general, technical replicates (if present) and similar biological states are expected to cluster close together. Potential batch-effects are revealed easily, e.g. if samples cluster by acquisition time, this might indicate a column effect (e.g., column

renewal). The PCA plot (see Figure S-5) is computed based on raw intensities of proteins and (if available) their LFQ equivalent from the proteinGroups table.

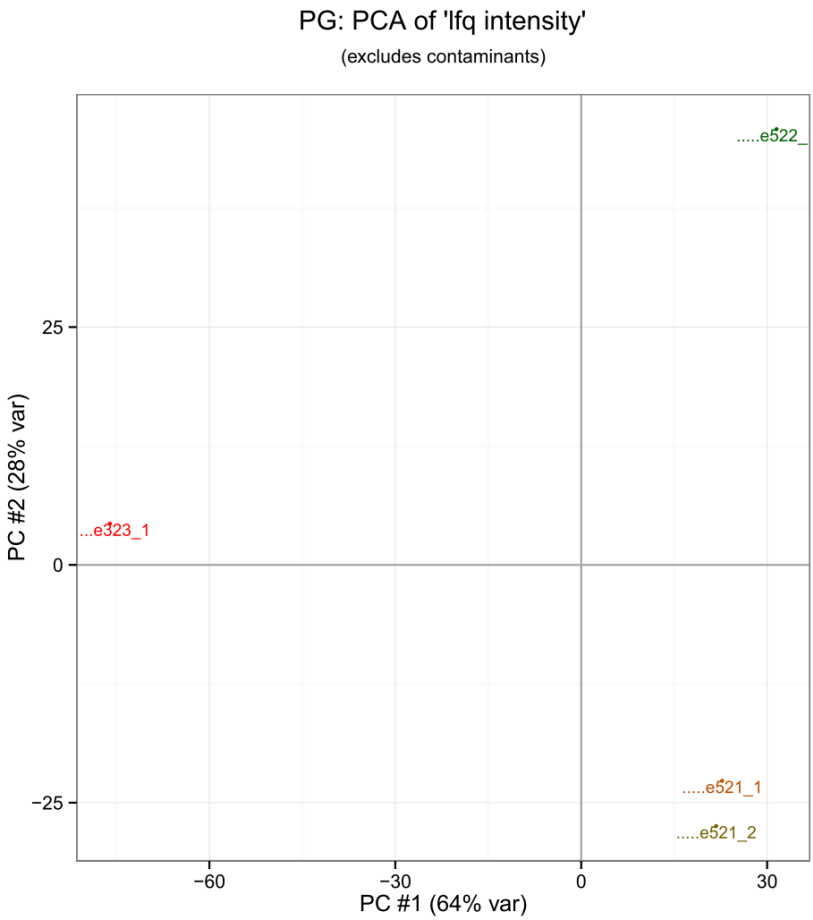


*Figure S-5: Principal component analysis plot, showing the four Raw files from data set 2, which are aliquots from a single biological sample of HEK293 cells. Files clearly cluster by date of measurement, being the main source of variation. Axis annotation is augmented with the percentage of variance explained (64% and 28% respectively). The remaining variance (here: 8%) is hidden in the principal components three and onwards (not shown).*

## References

(1) Vizcaíno, J. A. et al. The PRoteomics IDEntifications (PRIDE) database and associated tools: status in 2013. *Nucleic acids research* **2013**, 41, D1063—9.

(2) Burkhart, J. M., Schumbrutzki, C., Wortelkamp, S., Sickmann, A., and Zahedi, R. P. Systematic and quantitative comparison of digest efficiency and specificity reveals the impact of trypsin quality on MS-based proteomics. *J. Proteomics* **2012**, *75*, 1454 – 1462.

(3) Chiva, C., Ortega, M., and Sabidó, E. Influence of the Digestion Technique, Protease, and Missed Cleavage Peptides in Protein Quantitation. *J. Proteome Res.* **2014**, 13, 3979-86

(4) Cox, J., Hein, M. Y., Luber, C. A., Paron, I., Nagaraj, N., and Mann, M. Accurate Proteome-wide Label-free Quantification by Delayed Normalization and Maximal Peptide Ratio Extraction, Termed MaxLFQ. *Mol. Cell. Proteomics* **2014**, 13, 2513–2526.

# Scoring functions of summary heatmap

One scoring function was assigned to each quality metric. Here we specify the mathematical equations which allow to compute a quality score. All equations below return a value ranging between zero (=fail) and one (=perfect).

**Centered:**
Quality metric for 'centeredness' of a distribution around zero. A median of zero gives the best score of one. The closer the median is to the most extreme value of the distribution, the smaller the score (until reaching zero). The metric can be used for calibrated mass errors, as a measure of how well they are centered on zero. E.g. if the median is 0.1, while the range is [-0.5, 0.5], the score will be 0.8 (punishing 20% deviation). If the range of data is asymmetric, e.g. [-1.5,-0.5] and does not include zero, the score cannot reach 1, since the median can never be zero.

$$q_{centered}(x) = 1 - \frac{|\text{median}(x)|}{\max(|x|)}$$

**CenteredRef:**
Quality metric for 'centeredness' of a distribution around zero with a user-supplied range threshold. See 'centered' metric.

$$q_{centeredRef}(x, ref) = \max\left(0, 1 - \frac{|\text{median(x)}|}{ref}\right)$$

**Uniform:**
This equation computes the deviation from a uniform distribution. Input 'x' is a vector of relative frequencies for equally spaced bins in a histogram. A uniform distribution (e.g. {1/3, 1/3, 1/3} will get a score of 1. The worst possible case (e.g. {1, 0, 0}), will get a score of 0. A linear increasing function (e.g. {1/6, 2/6, 3/6}) will receive a value in between (0.585). In addition, bin values can be weighted (e.g. by their confidence). The total sum of both x and weights is assumed to be equal to 1.

$$q_{worst}(x, w) = w_{max}\sqrt{|1 - \overline{x}|} + (1 - w_{max})\sqrt{|0 - \overline{x}|}$$

$$q_{sc}(x, w) = \sum_{i=1}^{n} w_i \sqrt{|x_i - \overline{x}|}$$

Where $w_{max} = \max(w)$ and $\bar{x} = \sum_{i=1}^{n} x_i w_i$ and $\sum_{i=1}^{n} x = 1$ and $\sum_{i=1}^{i} w = 1$.

$$q_{uniform}(x, w) = \begin{cases} 1, & q_{worst}(x, w) = 0 \\ \dfrac{q_{worst}(x, w) - q_{sc}(x, w)}{q_{worst}(x, w)}, & q_{worst}(x, w) > 0 \end{cases}$$

**MaxN:**
Score an empirical density distribution of values, where the best possible distribution is right-skewed.

$$q_{maxN}(x) = 1 - \frac{(n-1) - \sum_{i=1}^{n} x_i (n - i)}{n - 1}$$

Where n is the length of x.

**MedianDist:**
Quality metric which measures the absolute distance from median. The median is assumed to be a representative target value. I.e. the fraction of peptides in a sample which are perfectly cleaved (e.g. a target value of 80% (=0.8)). Deviations from this are punished (linearly).

$$q_{medianDist}(x, m) = 1 - |x - m| \text{ where } x, m \in [0,1].$$

**BestKS:**
This test allows to spot Raw files whose RT peak width distribution differs from the majority of other peak width distributions. From a list of distributions (here: RT peak widths), compute all vs. all Kolmogorov-Smirnoff distance statistics (D) and report the row 'r' of the matrix which maximizes the sum of KS statistics (i.e contains the best reference distribution). Using the reference distribution, we compute the distance of all other distributions to this reference by using the statistic D of the Kolmogorov-Smirnoff test.

$$\mathrm{r} = \arg\max_{b} \sum_{i=1}^{n} KS_D(d_i, d_b)$$
$$q_{bestKS}(d_i, d_r) = KS_D(d_i, d_r)$$

Where $d_i$ is the distribution of peak widths of Raw file i, and $d_r$ is the reference distribution of peak widths.

**GaussDev:**
Evaluate the probability of a Gaussian at a position 0, with reference to the max obtainable probability of that Gaussian at its center.

$$q_{GaussDev}(\mu, \sigma) = \frac{f(0, \mu, \sigma)}{f(\mu, \mu, \sigma)}$$

Where $f(x, \mu, \sigma)$ is the density function of the normal distribution.
This is useful to estimate how well a Gaussian is centered on 0 (e.g. the distribution of post-calibration precursor $MS^1$ ppm errors).

**AlignDist:**
This is an inter-Raw file measure. After obtaining a distribution of retention time differences of ID-pairs (landmarks) after alignment, we compute the fraction of these distances which are below a given time threshold (e.g. 0.7 min for the latest MaxQuant version 1.5), i.e. are in close proximity across Raw files to be potential candidates for subsequent matching. The rationale is that unless landmarks align perfectly, we cannot expect to successfully transfer IDs in the next step of MBR. A fraction of 100% gives a score of 1 and decreases linearly as the fraction decreases to zero.

**MatchDist:**
An intra-Raw file measure, which examines the retention time range of groups. A group is a set of 3D-peaks, with at least one matched identification. Groups can be assigned to either of two classes: their RT range is either within the average RT peak width or exceeds it. The former case could be explained by annotating a segmented 3D-peak, whereas the latter is most likely a false positive annotation. The

fraction of identifications belonging to ‘out-width’ groups is penalized and thus deducted from the best obtainable score (100%).

**LinRef:**
Quality metric with linear response to input, reaching the maximum score at the given threshold (reference). This measure is beneficial for estimating if a target value, e.g. $MS^2$ identification rate of 40% was achieved: A given ID rate of 20%, would yield a score of $q_{linRef}(0.2, 0.4) = 0.5$. Exceeding the threshold will return a constant score of 1.

$$q_{linRef}(x, ref\ ) = min\left(1, \frac{x}{ref}\right)$$