SUPPORTING INFORMATION TO
Statistical approach to protein quantification

Authors and affilitations

Sarah Gerster[a,g], Taejoon Kwon[b], Christina Ludwig[c], Mariette Matondo[c], Christine Vogel[d,b], Edward Marcotte[b,e], Ruedi Aebersold[c,f] and Peter Bühlmann[a,f]

[a] Seminar for Statistics, Eidgenössische Technische Hochschule (ETH) Zurich, Rämistrasse 101, 8092 Zurich, Switzerland.

[b] Center for Systems and Synthetic Biology, Institute for Cellular and Molecular Biology, University of Texas at Austin, Austin, Texas 78712, United States.

[c] Institute of Molecular Systems Biology, Eidgenössische Technische Hochschule (ETH) Zurich, Wolfgang-Pauli Strasse 16, 8093 Zurich, Switzerland.

[d] Center for Genomics and Systems Biology, New York University, New York 10003, United States.

[e] Department of Chemistry and Biochemistry, University of Texas at Austin, Austin, Texas 78712, United States.

[f] Competence Center for Systems Physiology and Metabolic Diseases, 8092 Zurich, Switzerland.

[g] current address: Bioinformatics Core Facility, Swiss Institute of Bioinformatics, Quartier Sorge, Genopode, 1015 Lausanne, Switzerland.

Keywords: absolute and relative protein quantification, statistical modeling, shared peptides, bipartite graph, tandem mass spectrometry

Corresponding Author:

Sarah Gerster, Bioinformatics Core Facility, Swiss Institute of Bioinformatics, Quartier Sorge, Genopode, 1015 Lausanne, Switzerland

tel: +41 21 692 4096, fax: +41 21 692 4065, email: `sarah.gerster@isb-sib.ch`

Supporting Information

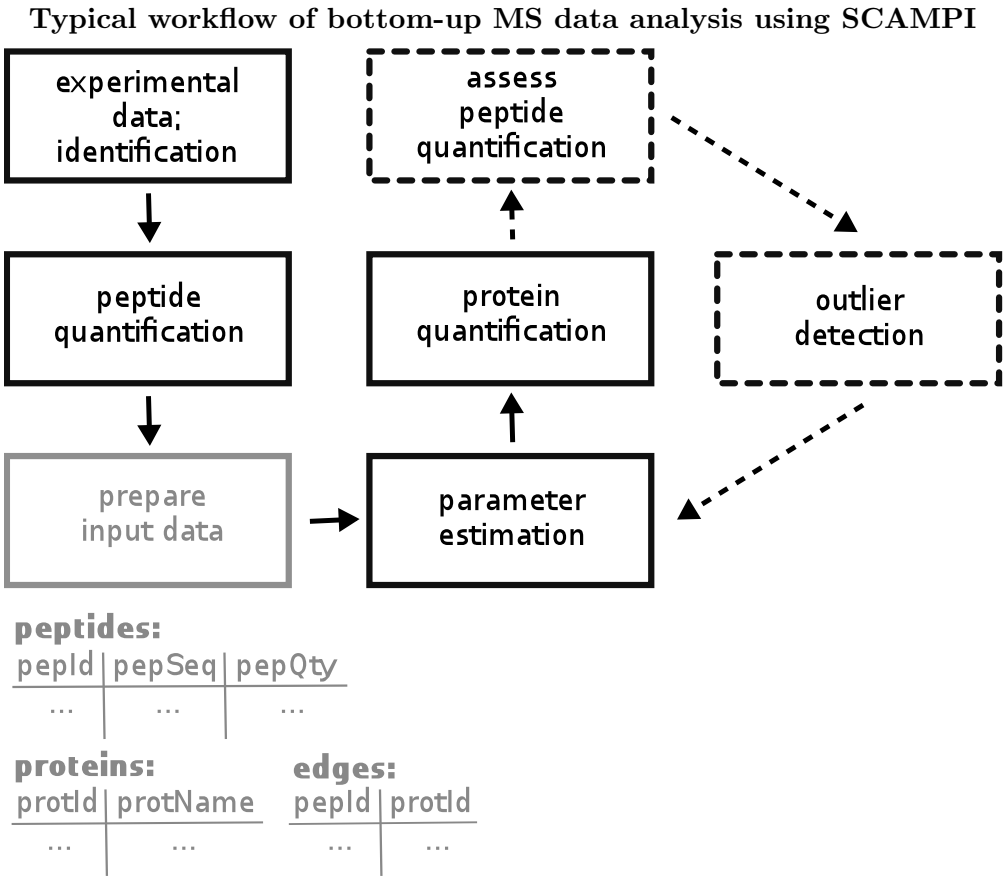
Additional information, mostly technical details, to complete the description of SCAMPI:

- *Supporting Information 1 – Workflow and input data*
- *Supporting Information 2 – Covariances*
- *Supporting Information 3 – Reassessing peptide abundances*
- *Supporting Information 4 – Selecting outliers in the measured peptides*
- *Supporting Information 5 – ILSE parameter estimates*
- *Supporting Information 6 – Input data preparation for SCAMPI*
- *Supporting Information 7 – Computation times*
- *Supporting Information 8 – Parameter estimates*
- *Supporting Information 9 – Additional Figures*
  - *SI 9.1 – Interpretation of diagnostic plots*
  - *SI 9.2 – SRM experiment on Leptospira interrogans*
  - *SI 9.3 – Directed MS human data [4]*
  - *SI 9.4 – SILAC labeled human shotgun proteomics data*
- *Supporting Information 10 – Materials and Methods for the SILAC data set*
- *Supporting Information 11 – SCAMPI on TOP3 input*
- *Supporting Information 12 – SCAMPI results compared to MaxQuand output*

In addition, the compressed file `SupportingInformation13_inputData.zip` contains the input files for SCAMPI for all discussed datasets.

Supporting Information 1 – Workflow and input data

SCAMPI is implemented in the R package `'protiq'` [1]. This section gives an overview of the typical workflow and the required input data.



The peptides are identified (and typically filtered at 1% FDR) and quantified by using a method of choice. The three data frames required as input for the protein quantification procedure are prepared. The model parameters can then be estimated on the data and the protein concentrations are predicted. Optionally, peptide abundances can be reassessed, and the model can be re-run after having eliminated some outliers in the input data and adapted the graph structure (dashed boxes/arrows in the above figure).

Structure of the input data

- `peptides` – Each row describes one (unique) experimentally identified peptide sequence. The mandatory columns are:
  - peptide identification number (`pepId`)
  - peptide sequence (`pepSeq`)
  - peptide abundance score (`pepQty`), typically log-transformed intensity measurement
- `proteins` – Each row defines one (unique) protein sequence matching to at least one experimentally identified peptide sequence. The mandatory columns are:
  - protein identification number (`protId`)
  - protein name (`protName`)

  Ensuring that this table holds each protein sequence only once requires particular attention when some sequences are described by several accession numbers.
- `edgespp` – Each row defines one edge of the bipartite graph. The mandatory columns are:
  - peptide identification number (`pepId`)
  - protein identification number (`protId`)

A convenient way to provide the input is to pack the required input into three separate CSV files.

Consistency checks

Before starting with the computations a couple of constraints are checked to make sure that the provided input is suitable for SCAMPI:

- `peptides`:
  - the columns `pepSeq` and `pepId` are present and contain unique entries
  - each `pepId` occurs in at least one edge
  - the column `pepQty` is present and holds well-defined numeric values

  - the values in `pepId` range from 1 to n (where n is the number of peptides)

- `proteins`:
  - the columns `protName` and `protId` are present and contain unique entries
  - each `protId` occurs in at least one edge
  - the values in `protId` range from $n + 1$ to $n + m$ (where m is the number of proteins)

- `edgespp`:
  - holds the columns `pepId` and `protId`
  - holds only `id`s present in `peptides` or `proteins`
  - each row is unique, i.e. each edge is recorded exactly once

If the input data does not pass the consistency checks SCAMPI stops with an error message.

The input files for all datasets presented in this publication are available in `SupportingInformation12_inputData.zip`.

User's choices

Data preprocessing – It is up to the user to decide which peptides and proteins have been reliably identified and should be used for the quantification process. Furthermore, peptides can be quantified by the user's method of choice.

It is also up to the user to decide how to handle issues such as modifications, different charge states or semi-tryptic peptides. As an example, one could decide to have each charge state of a peptide contributing separately by providing separate input lines `PEPSEQ_2`, `PEPSEQ_3` and `PEPSEQ_4` in the data frame `peptides`. Alternatively, the contributions from the different charge states could be combined during the preprocessing steps, and a single value for `PEPSEQ` would then be included in the data frame `peptides`.

SCAMPI options – Details about the options of the different functions can be found in the documentation of the R package `'protiq'` [1]. The list below provides merely a brief overview.

- parameter estimation method: can be either ILSE, MLE or both (see *Parameter estimation*)
- rescaling the input data: if it has not already been done in the pre-processing, it probably makes sense to ask SCAMPI to logarithmize the input peptide abundances
- SCAMPI can be run iteratively, removing outlier peptides after each iteration step, to recursively improve the abundance predictions

Supporting Information 2 – Covariances

Covariance between peptide scores

The diagonal elements of $\mathbb{\Sigma}$ correspond to the variance of U:

$$\begin{aligned}
\left(\mathbb{\Sigma}_{\underline{U}^{(r)}}\right)_{ii} &= \mathrm{Var}\left(U_i^{(r)}\right) \\
&= \mathrm{Var}\left(\alpha + \beta \sum_{j \in Ne(i)} C_j + \epsilon_i^{(r)}\right) \\
&= \mathrm{Var}\left(\beta \sum_{j \in Ne(i)} C_j\right) + \tau^2 \\
&= \beta^2 \, \mathrm{Var}\left(\sum_{j \in Ne(i)} C_j\right) + \tau^2 \\
&= \beta^2 \, D_{ii} + \tau^2.
\end{aligned}$$

The off diagonal elements can be computed as follows:

$$\begin{aligned}
\left(\mathbb{\Sigma}_{\underline{U}^{(r)}}\right)_{ik} &= \mathrm{Cov}\left(U_i^{(r)}, U_k^{(r)}\right) \\
&= \mathrm{Cov}\left(\alpha + \beta \sum_{j \in Ne(i)} C_j + \epsilon_i^{(r)}, \, \alpha + \beta \sum_{j' \in Ne(k)} C_{j'} + \epsilon_k^{(r)}\right) \\
&= \mathrm{Cov}\left(\beta \sum_{j \in Ne(i)} C_j, \, \beta \sum_{j' \in Ne(k)} C_{j'}\right) \\
&= \beta^2 \, \mathrm{Cov}\left(\sum_{j \in Ne(i)} C_j, \, \sum_{j' \in Ne(k)} C_{j'}\right) \\
&= \beta^2 \, D_{ik}^{(r)}
\end{aligned}$$

The fact that the ϵ_i are i.i.d. and independent of the protein scores C_j was used for these derivations.

Combining the obtained results leads to the matrix in Equation 5 in the manuscript.

Covariance between peptide and protein scores

$\mathrm{Cov}\left(C_j, U_i\right)$ is zero when there is no edge between peptide i and protein j (Markovian assumption).

When there is an edge between peptide i and protein j, the covariance between U_i and C_j can be computed as

$$\begin{aligned}
\left(\Gamma_{C_j \underline{U}^{(r)}}\right)_i &= \mathrm{Cov}\left(C_j, U_i^{(r)}\right) \\
&= \mathrm{Cov}\left(C_j,\, \alpha + \beta \sum_{j' \in Ne(i)} C_{j'} + \epsilon_i\right) \\
&= \mathrm{Cov}\left(C_j,\, \beta \sum_{j' \in Ne(i)} C_{j'} + \epsilon_i\right) \\
&= \mathrm{Cov}\left(C_j,\, \beta \sum_{j' \in Ne(i)} C_{j'}\right) \\
&= \beta\ \mathrm{Var}\left(C_j\right) \\
&= \beta,
\end{aligned}$$

which leads to the result in Equation 6 in the mauscript. The fact that the ϵ are independent of the protein scores C_j was used.

Supporting Information 3 – Reassessing peptide abundances

Equation 1 (in the manuscript) models the peptide abundances as a function of the abundance scores of their neighboring proteins. In the process of estimating the latter values, $\widehat{C}_j$s, all peptide data are used. Therefore, simply computing $\mathbf{E}\left[U_i\right]$ would lead to over-fitting and to over-optimistic results. Hence, in order to estimate the abundance of peptide i, $\widehat{U}_i$, we want to use all peptide

measurements, except for the i^{th} one. The peptide abundance estimate $\widehat{U}_i$ is then defined as

$$
\begin{aligned}
\widehat{U}_i &= \mathbf{E}\left[U_i | \{U_{k \setminus i}\}\right] \\
&= \mathbf{E}\left[\alpha + \beta \sum_{j \in Ne(i)} C_j + \epsilon_i \Big| \{U_{k \setminus i}\}\right] \\
&= \alpha + \beta \sum_{j \in Ne(i)} \mathbf{E}\left[C_j | \{U_{k \setminus i}\}\right] + 0 \\
&= \alpha + \beta \sum_{j \in Ne(i)} \left(\mu + \left(\underline{U}^{(r)}_{\setminus i} - \alpha \underline{1}^{(r)}_{\setminus i} - \beta \mu \, \mathrm{diag}(D^{(r)})_{\setminus i} \right)^{\intercal} \boldsymbol{\Sigma}^{-1}_{\underline{U}^{(r)}_{\setminus i}} \boldsymbol{\Gamma}_{C_j \underline{U}^{(r)}_{\setminus i}} \right) \\
&= \alpha + \beta \, \mu \, D_{ii} + \beta \sum_{j \in Ne(i)} \left(\underline{U}^{(r)}_{\setminus i} - \alpha \underline{1}^{(r)}_{\setminus i} - \beta \mu \, \mathrm{diag}(D^{(r)})_{\setminus i} \right)^{\intercal} \boldsymbol{\Sigma}^{-1}_{\underline{U}^{(r)}_{\setminus i}} \boldsymbol{\Gamma}_{C_j \underline{U}^{(r)}_{\setminus i}}
\end{aligned}
$$

Note that the summation $\left(\sum_{j \in Ne(i)} \dots \right)$ in the equation above is zero for connected components holding a single peptide. In these cases, $\widehat{U}_i = \alpha + \beta \, \mu \, D_{ii}$. The abundances of the matching proteins do not contribute to $\widehat{U}_i$ (since there is no further evidence to estimate $\widehat{C}_j$ without using peptide i). However, the parameters learned on the whole dataset as well as the number of matching proteins allow to provide an estimate for U_i even in these cases.

Supporting Information 4 – Selecting outliers in the measured peptides

A threshold is needed to decide which residuals ($R_i = U_i - \widehat{U}_i$) are "too large", i.e. at which values the two dashed horizontal lines should be drawn in Figure 7 B (in the manuscript). A criteria based on the inter-quartile range ($iqr = Q_3 - Q_1$) is used, where Q_1 is the first quartile in the distribution of the residuals and Q_3 the third one. Concretely, a peptide i is labeled as outlier (marked by gray stars in Figure 7 B in the manuscript) if $R_i \notin [Q_1 - k \cdot iqr, \, Q_3 + k \cdot iqr]$, with $k = 2$.

Supporting Information 5 – ILSE parameter estimates

The solution to Equation 13 in the manuscript can be written in closed form:

$$\hat{\beta}^2 = \frac{\sum_{r,i\neq k} \left(\widehat{\mathbf{\Sigma}}_{\underline{U}^{(r)}}\right)_{ik} D_{ik}^{(r)}}{\sum_{r,i\neq k} \left(D_{ik}^{(r)}\right)^2}$$

with $r = 1, \ldots, R$ and $i, k = 1, \ldots, n_r$.

The solution to Equation 14 in the manuscript has a closed form:

$$\hat{\tau}^2 = \frac{1}{R}\sum_{r=1}^{R}\frac{1}{n_r}\sum_{i=1}^{n_r}\left(\left(\widehat{\mathbf{\Sigma}}_{\underline{U}^{(r)}}\right)_{ii} - \hat{\beta}^2\right)$$

Supporting Information 6 – Input data preparation for SCAMPI

This section holds information about:

- *SRM experiment on* Leptospira interrogans
- *Directed MS human data*
- *SILAC labeled human shotgun proteomics data*

The SCAMPI input files for each of the discussed datasets are available in `SupportingInformation12_inputData.zip`

SRM experiment on *Leptospira interrogans*

The data published in [2] was used. The peptide quantities were averaged over the three technical replicates as follows: (i) select the three most intense transitions (on average over the three

replicates) for each peptide sequence, (ii) compute the peptide intensity by averaging these three numbers and then (iii) average the values for each peptide over the three replicates. The $\log_{10}$ of these quantities are used as input scores U_i. Charge states were not treated as specific peptide identifications. Since the data come from an SRM experiment, all peptides are proteotypic. Modifications and semi-tryptic peptides were not considered.

Directed MS human data

The *progenesis* [3] output with peptides filtered at a 1% FDR is used to prepare the input data for SCAMPI. Charge states and modifications were not treated as specific peptide identifications. Hence, all *progenesis* values matching to a same peptide sequence were summed up and then logarithmized (log_{10}) to obtain the input peptide abundance scores (U_i).

The data published in [4] hold 53 anchor proteins. However, in the provided data, the experimentally measured concentrations for eight of these proteins are based on shared peptides. There are two possible explanations for this: (i) changes in the used database and (ii) protein grouping. Beck et al. [4] worked with ProteinProphet [5], which groups indistinguishable proteins. If an AQUA peptide uniquely maps to such a group, this is counted as a unique match; SCAMPI, however, strives at providing a quantification score for each identified protein sequence, and hence requires AQUA peptides uniquely matching to a single protein sequence. These 8 sequences were excluded from the set of ground truth for this study. Among the 45 remaining anchor proteins, three cannot be quantified based on the provided *progenesis* data (no matching peptide has been quantified for any of these three protein sequences). Hence, finally the performance of the quantification results is assessed on a set of 42 anchor proteins.

SILAC labeled human shotgun proteomics data

The MaxQuant [6] output files `evidence.txt` and `proteinGroups.txt` (filtered at 1% FDR) are used to extract the input data for SCAMPI. Different charge states and modifications were not considered as specific peptide identifications. Hence, all MaxQuant peak intensities corresponding to a same peptide sequence were summed up. These values were logarithmized (log_{10}) and then used as input peptide abundance scores in SCAMPI.

Supporting Information 7 – Computation times

The table below provides a summary of the dataset sizes and the approximate computation times required to run SCAMPI (MLE and ILSE) on each of them. The provided timings were obtained with the R function `system.time()`. We report the time used to run the function `runScampi()` from the R package `'protiq'` on the datasets provided in `SupportingInformation12_inputData.zip` on a single CPU.

	# peptides	% shared	# edges	# proteins	# cc	ILSE time	MLE time
L. interrogans	151	0	151	39	39	$<$ 1s	$<$ 1min
human (directed MS)	49190	6	54720	6257	4984	$<$ 2h	$<$ 5h
human (SILAC, control)	30323	17	38019	3892	2659	$<$ 1h	$<$ 2h
human (SILAC, treated)	30326	17	38025	3890	2658	$<$ 1h	$<$ 2h

In the table above, "# cc" stands for the number of connected components in the bipartite graph. "% shared" stands for the percentage of the peptides matching to at least two proteins.

Supporting Information 8 – Parameter estimates

SCAMPI's parameter value estimates for all datasets presented in the manuscript and supporting information are provided below:

	MLE				ILSE			
	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\mu}$	$\hat{\tau}$	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\mu}$	$\hat{\tau}$
L. interrogans (SRM)	2.81	0.67	4.39	0.39	5.88	0.6	0	0.48
human (directed MS)	5.81	0.43	0	0.56	5.96	0.4	0.16	0.56
human (SILAC, shotgun, control)	6.56	0.53	0	0.55	6.77	0.51	0.09	0.61
human (SILAC, shotgun, treated)	6.41	0.51	0	0.55	6.62	0.51	0.07	0.59

Since the first dataset, *L. interrogans*, does not include any shared peptides, the four parameters are not identifiable. The estimates of $\widehat{C}_j$ are still well-defined, though. The reported values for the MLE parameter estimates correspond to a (potentially local) minimum. In the case of ILSE, $\hat{\mu}$ was constrained to zero.

Supporting Information 9 – Additional Figures

This section holds information about the interpretation of diagnostic plots (*SI 9.1 – Interpretation of diagnostic plots*) as well as additional figures for:

- *SI 9.2 – SRM experiment on Leptospira interrogans*
- *SI 9.3 – Directed MS human data [4]*
- *SI 9.4 – SILAC labeled human shotgun proteomics data*

SI 9.1 – Interpretation of diagnostic plots

Some of the additional figures in this section are diagnostic plots applied on the peptide level. They allow to graphically check whether some assumptions are fulfilled or not. We consider two plots: the Tukey-Anscombe Plot and the Normal Plot. Brief descriptions of these plots are given below. Further information can be found in statistics books about regression, for example in [7, 8]. Furthermore, we provide Bland-Altman plots [9] (also known as Tukey mean-difference plot or MA plot) as a further comparison between computed protein abundance scores.

Tukey-Anscombe Plot : plot of the residuals versus the fitted values (on the x-axis). Ideally, the points in the Tukey-Anscombe plot are "randomly distributed" around the horizontal line through zero. Typical deviations which indicate a problem with the model and/or data are:

- non-constant variability around the reference line,
- trend in the plot indicating that the linear model assumption is not correct.

Normal Plot : quantile-quantile (Q-Q) plot where the empirical quantiles of the residuals are plotted versus the theoretical quantiles of a $\mathcal{N}(0,1)$ distribution (on the x axis). If the residuals are normally distributed with expectation μ and variance σ^2, the Normal Plot shows approximately a straight line. The plot allows for example to visually recognize long-tailed and skewed distributions.

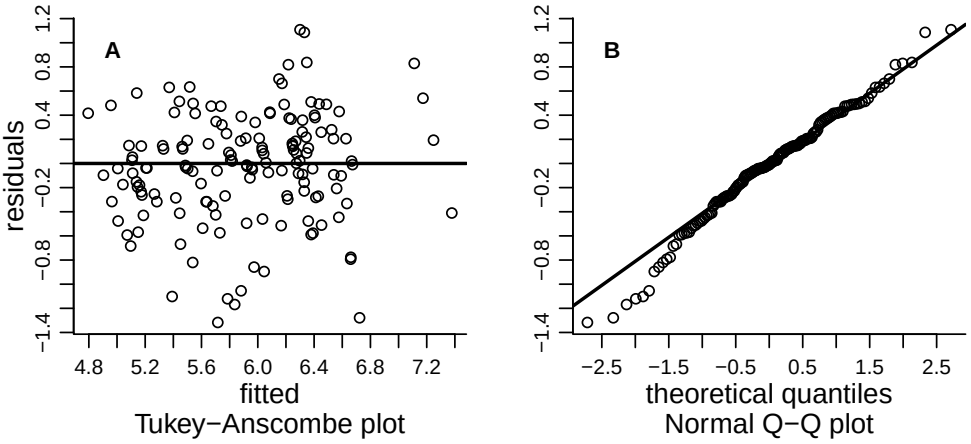
Bland-Altman Plot : plot of the difference between two measurements versus the mean of two measurements (on the x axis). Bland-Altman plots go a step further than correlation anaylsis by checking how well two measurements of a same characteristic agree. In our case it can be used to compare the predicted (log-transformed) concentrations $\widetilde{C}$. The plots are generated as follows:

- take the computed abundance scores: $\widehat{C}_A$ and $\widehat{C}_B$
- use the reference proteins to fit a linear model between the log-scaled (base-10) known concentrations and the computed scores (for $\widehat{C}_A$ and $\widehat{C}_B$)
- apply the trained parameters to all samples for the respective score ($\rightarrow \widetilde{C}_A$ and $\widetilde{C}_B$)

- only consider proteins for which the predicted (log-transformed) concentration lies within the abundance range of the reference proteins
- draw the Bland-Altman plot using the predicted (log-transformed) concentrations
- highlight reference proteins (subject to overfitting since they were used to fit the linear model)

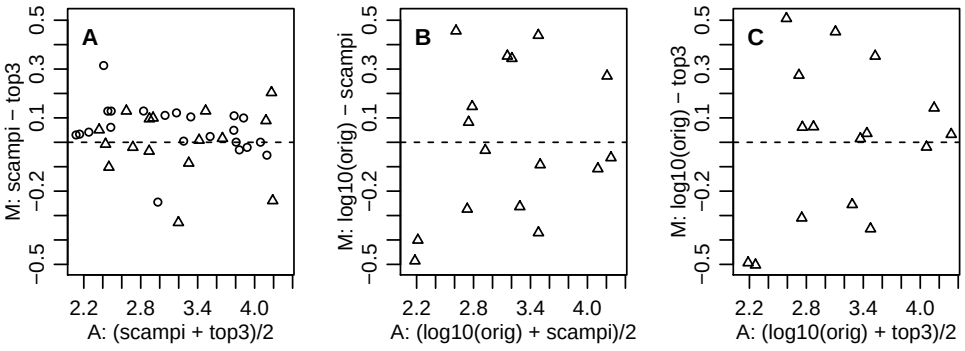
SI 9.2 – SRM experiment on Leptospira interrogans

SRM – Diagnostic plots for peptide residuals with ILSE parameter estimation



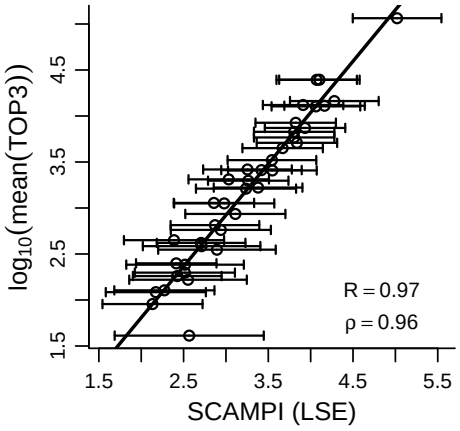
The diagnostic plots (residual plot in panel A and normal Q-Q plot in panel B) for the peptide abundance residuals show no major violation of the assumptions on the noise term ϵ (see Equation 1 in the manuscript).

SRM – Bland-Altman plots for protein scores (ILSE parameter estimation)



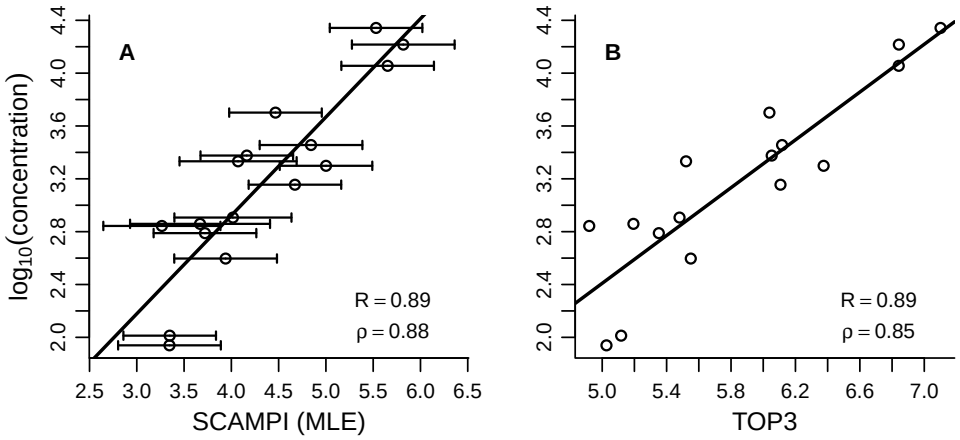
Panel A compares the (log-scaled) concentrations predicted by SCAMPI and by TOP3. Panels B and C compare the (log-transformed) concentrations predicted by SCAMPI and TOP3, respectively, to the known (log-scaled) ground truth concentrations. There is no major difference between TOP3 and SCAMPI. The triangles correspond to the reference proteins (used to learn the parameters of the linear model).

SRM – Comparing the SCAMPI scores to the scores published in [2]



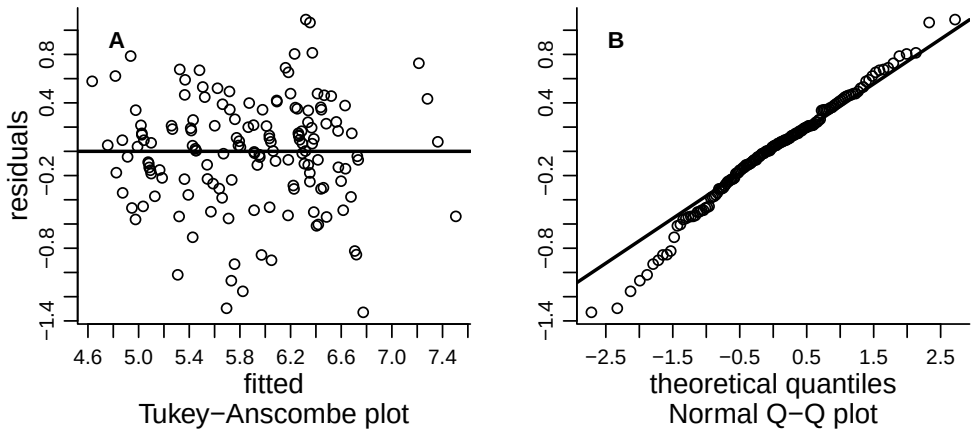
The x-axis shows the predicted concentrations (log_{10}) computed by SCAMPI. The y-axis shows the mean of the results published for the three technical replicates in [2] (also log_{10} of the predicted protein concentrations). The latter scores were computed with a "flexible" TOP3 approach, where proteins are also quantified when they only have one or two experimentally quantified unique peptides. The correlation between the two approaches is very high, even though the input is not exactly the same for the two approaches (different number of transitions considered for the peptide abundances).

SRM – Protein quantification results with MLE parameter estimation



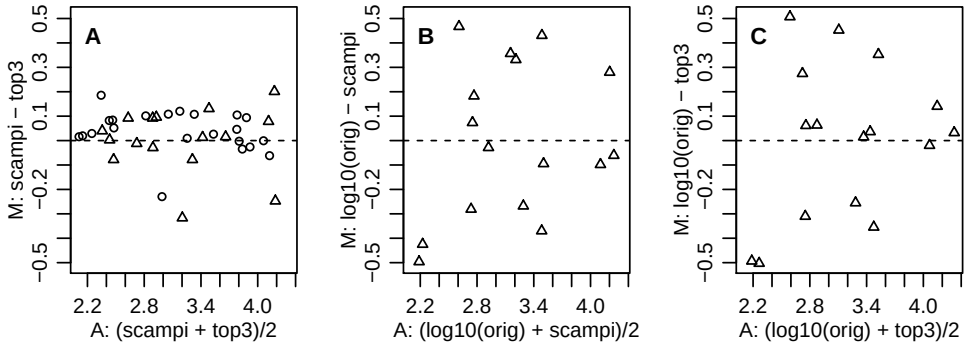
The correlation coefficients for SCAMPI and the TOP3 approach are very similar. The error bars in panel A show the 95% prediction intervals for the computed protein abundance scores. Note that the scale on the x-axis is different in the two panels. The range of the computed scores depends on the underlying model. We cannot compare the scores from SCAMPI and from TOP3 directly, but can look at correlations with a reference score, as presented in this figure.

SRM – Diagnostic plots for peptide residuals with MLE parameter estimation



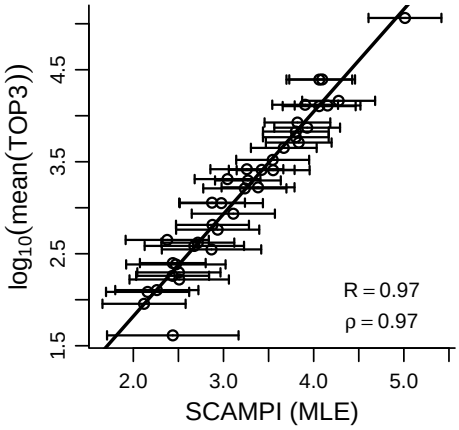
The diagnostic plots (residual plot in panel A and normal Q-Q plot in panel B) for the peptide abundance residuals show no major violation of the assumptions on the noise term ϵ (see Equation 1 in the manuscript).

SRM – Bland-Altman plots for protein scores (MLE parameter estimation)



Panel A compares the (log-scaled) concentrations predicted by SCAMPI and by TOP3. Panels B and C compare the (log-transformed) concentrations predicted by SCAMPI and TOP3, respectively, to the known (log-scaled) ground truth concentrations. There is no major difference between TOP3 and SCAMPI. The triangles correspond to the reference proteins (used to learn the parameters of the linear model).

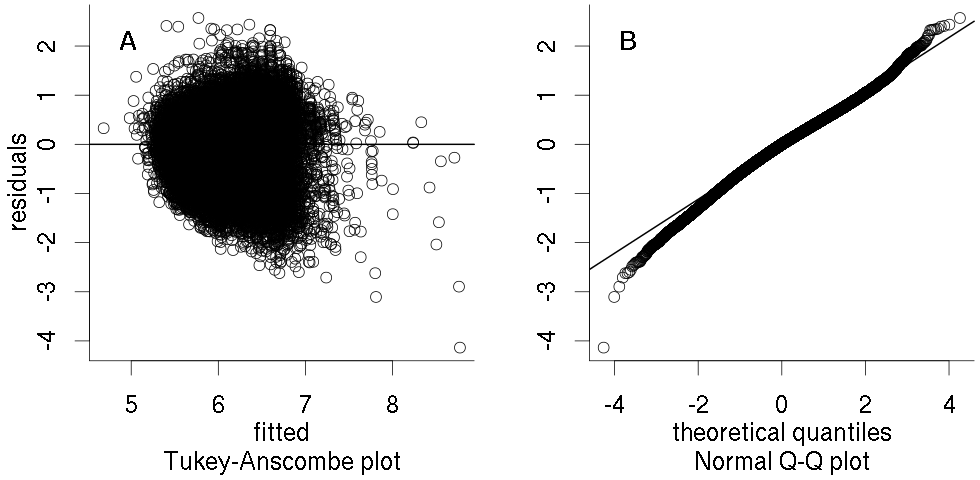
SRM – Comparing MLE results to the scores published in [2]



Again, the correlation between the two approaches (SCAMPI with MLE parameter estimates and "flexible" TOP3) is very high, even though the input is not exactly the same for the two approaches (different number of transitions considered for the peptide abundances).

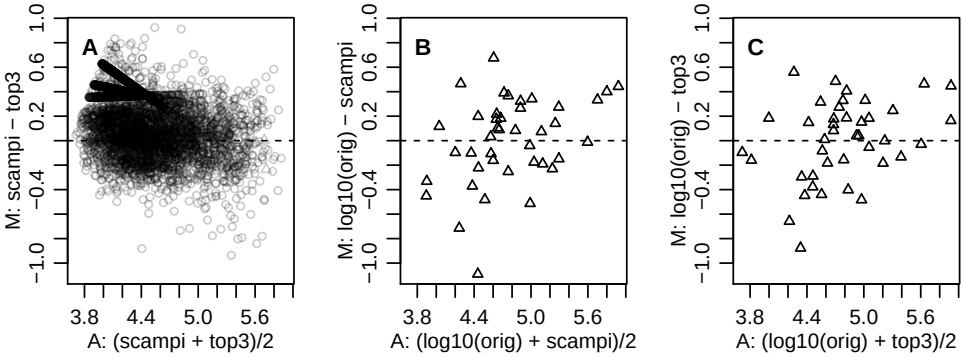
SI 9.3 – Directed MS human data [4]

Directed MS – Diagnostic plots for peptide residuals with ILSE parameter estimation



The diagnostic plots (residual plot in panel A and normal Q-Q plot in panel B) for the peptide abundance residuals show no major violation of the assumptions on the noise term ϵ (see Equation 1 in the manuscript).

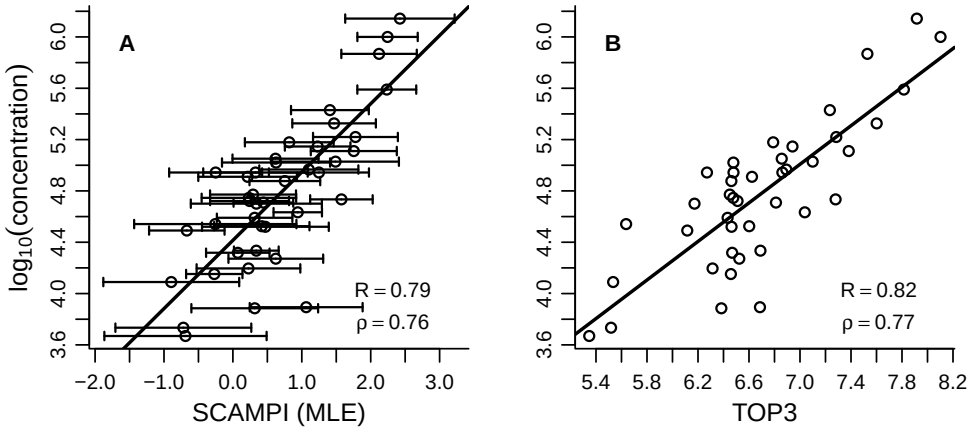
Directed MS – Bland-Altman plots for protein scores (ILSE parameter estimation)



Panel A compares the (log-scaled) concentrations predicted by SCAMPI and by TOP3. Panels B and C compare the (log-transformed) concentrations predicted by SCAMPI and TOP3, respectively,

to the known (log-scaled) ground truth concentrations. There is no major difference between TOP3 and SCAMPI. The triangles correspond to the reference proteins (used to learn the parameters of the linear model). The strong line patterns appearing in panel A correspond to proteins with a single, two or three matching unique peptide(s) as only input. We discuss the linear relationship between SCAMPI and top3 in these cases in *Supporting Information 11 – SCAMPI on TOP3 input.*

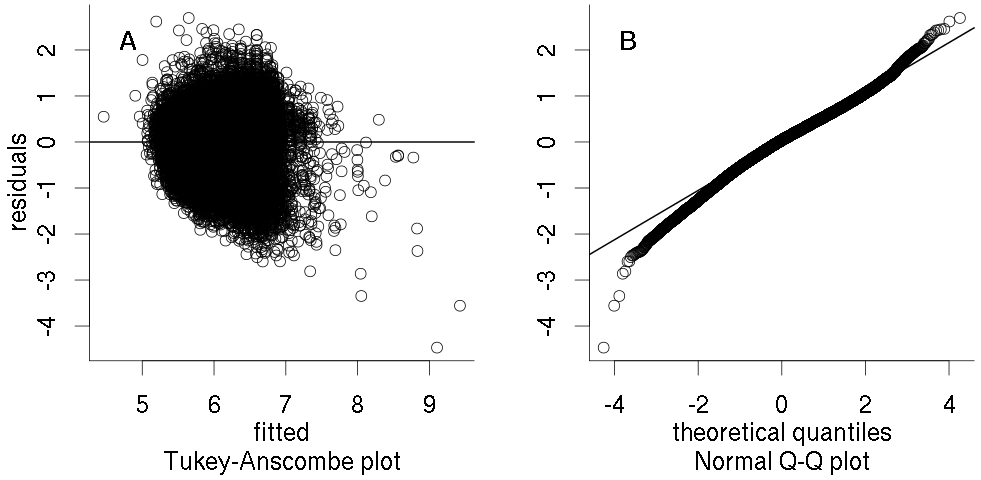
Directed MS – Protein quantification results with MLE parameter estimation



The correlation coefficients for SCAMPI and the TOP3 approach are similar. The error bars in panel A show the 95% prediction intervals for the computed protein abundance scores. Note that the scale on the x-axis is different in the two panels. The range of the computed scores depends on the underlying model. We cannot compare the scores from SCAMPI and from TOP3 directly, but can look at correlations with a reference score, as presented in this figure.

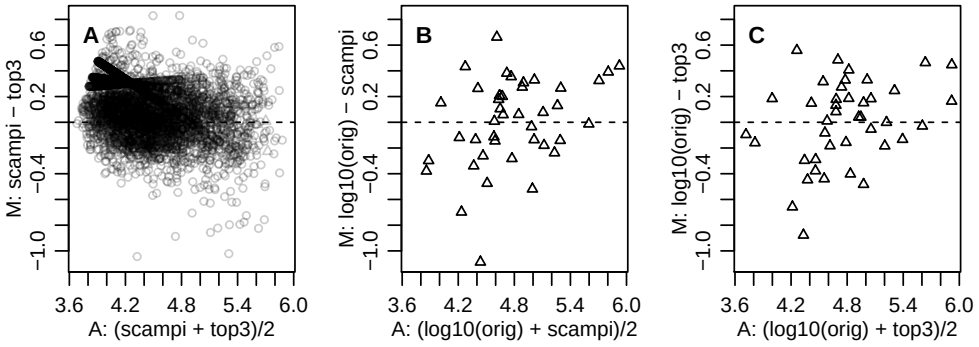
Directed MS – Diagnostic plots for peptide residuals

with MLE parameter estimation



The diagnostic plots (residual plot in panel A and normal Q-Q plot in panel B) for the peptide abundance residuals show no major violation of the assumptions on the noise term ϵ (see Equation 1 in the manuscript).

Directed MS – Bland-Altman plots for protein scores (MLE parameter estimation)

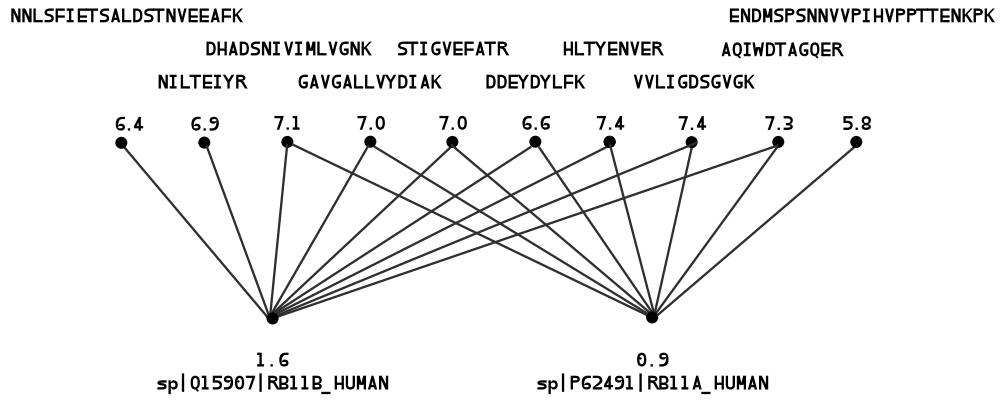


Panel A compares the (log-scaled) concentrations predicted by SCAMPI and by TOP3. Panels B and C compare the (log-transformed) concentrations predicted by SCAMPI and TOP3, respectively, to the known (log-scaled) ground truth concentrations. There is no major difference between TOP3 and SCAMPI. The triangles correspond to the reference proteins (used to learn the parameters of the linear model). The strong line patterns appearing in panel A correspond to proteins with a

single, two or three matching unique peptide(s) as only input. We discuss the linear relationship between SCAMPI and top3 in these cases in *Supporting Information 11 – SCAMPI on TOP3 input.*

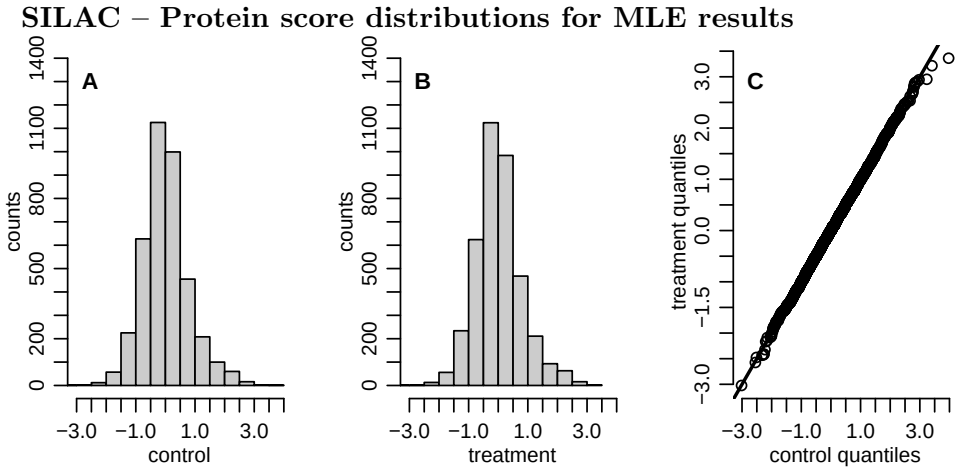
Directed MS – Concrete example of SCAMPI's handling of shared peptides

The picture below shows a connected component with 2 proteins and 10 peptides from the "Directed MS" data set. The names of the proteins and sequences of the peptides are provided. The scores correspond to the input peptide abundances and computed protein abundance scores, respectively.

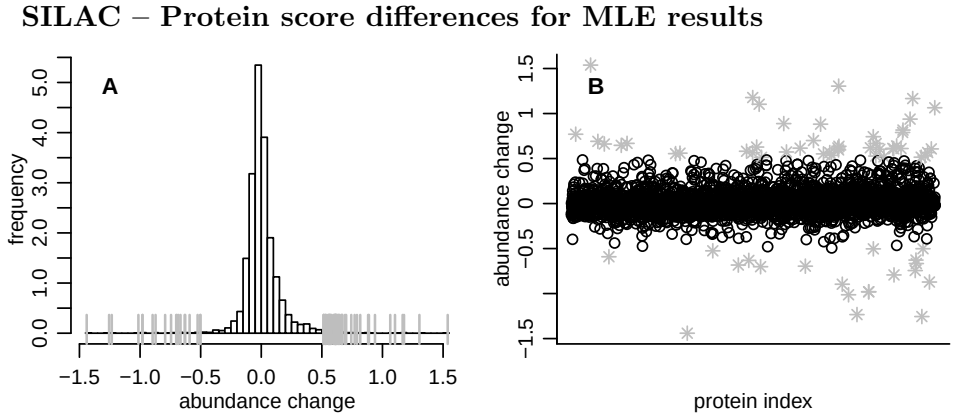


The mean abundance of the peptides matching uniquely to protein `Q15907` is 6.62, while protein `P62491` has a single unique peptide with an abundance score of 5.77. The mean abundance of the shared peptides is 7.10. Both proteins could also be quantified with the (non-strict) TOP3 approach. However, SCAMPI can use the additional information in the shared peptides to gain a better understanding of the data. In this case, this leads to additional confidence (i) in the computed protein abundance scores and (ii) in the fact that both proteins are present in the sample.

SI 9.4 – SILAC labeled human shotgun proteomics data



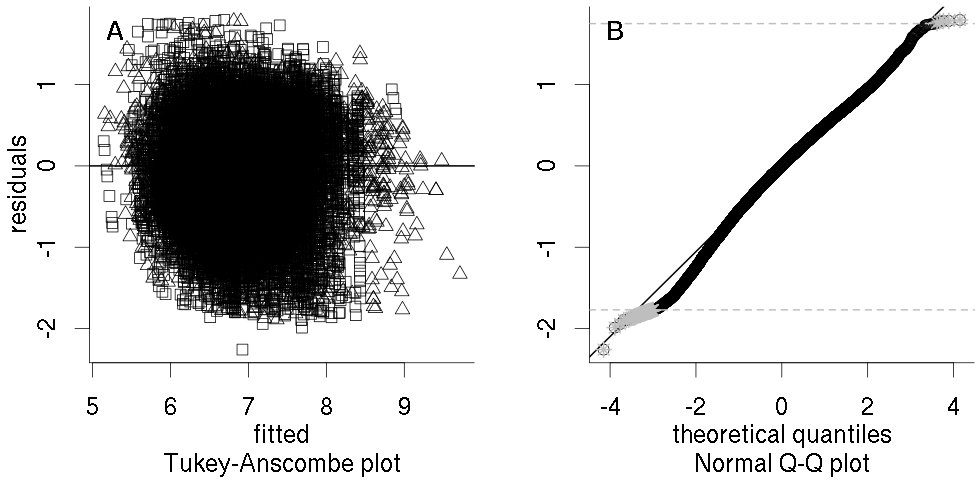
The distributions are comparable (similar median and quartiles).



SCAMPI with the MLE parameter estimates identifies a set of 65 differentially abundant proteins. All 59 differentially abundant proteins found by SCAMPI with the ILSE parameter estimates belong to this set.

SILAC – Running SCAMPI (ILSE estimates) iteratively:

diagnostic plots after second iteration



The R package `'protiq'` provides an option to run SCAMPI iteratively. This means:

- After running SCAMPI on the original data, the peptide outliers are selected with an inter-quartile range criterion (constant can be set by the user). These outliers are the 234 peptides labeled by stars in panel B of Figure 7 in the manuscript.
- All peptides marked as outliers are removed from the dataset for the second run. The data frames holding the proteins and edges are adapted accordingly.
- SCAMPI runs again on this modified dataset. Again, it performs peptide outlier selection.

Basically this iterative procedure can be repeated until none of the peptides are selected as outliers anymore. There are however two important points to keep in mind:

- The points which are removed are outliers with respect to the implemented model. There is no guarantee that they correspond to real biological outliers. Hence, such an iterative approach can be useful in an explorative analysis to get a list of potential outliers, but it is important

to validate these sequences manually before removing them definitely from the final analysis. For this purpose, SCAMPI returns the list of outliers it discards (after each iteration).

- There is a potential risk of over-fitting. If the number of discarded peptides becomes too large compared to the total number of peptides in the original dataset, one should be cautious.

In the example presented in this paper, the peptide diagnostic plots after the second iteration of SCAMPI (with ILSE parameter estimates) look much better (see plot above). The major outliers are gone. The normal Q-Q plot (panel B in the figure above) indicates that only a few peptides would still be selected as outliers to be removed in a third iteration step.

Supporting Information 10 – Materials and Methods for the SILAC data set

Cell lines and culture conditions

Human leukemic cell lines KG1a were cultured in cultured in Iscove's modified Dulbecco's medium (Invitrogen) supplemented with 20% fetal bovine serum (PAA Laboratories), 2mM L-glutamine, 100units/ml penicillin, 100mg/ml streptomycin and containing either 0.5 mM each of L-Lysine-2HCl and L-Arginine-HCl or 13C6 L-Lysine-2HCl and 13C615N4 L-Arginine-HCl. All reagents for isotope metabolic protein labeling of the cells were from Pierce (Rockford, IL) except that L-glutamine and penicillin/streptomycin were from Invitrogen (Carlsbad, CA). Cells were grown in a humidified atmosphere at 37°C and 5% CO2. Cell viability was assessed by standard trypan blue dye exclusion assay. KG1a cells were treated with 5uM MG132 for up to 6 hours. A sample of 3 $\times$ 106 cells from the "heavy" and "light" isotope protein-labeled populations were treated separately with 5uM MG132 or dimethyl sulfoxide (DMSO) (0.05% v/v final concentration) for 6 hours, respectively. Cells were mixed equally and washed three times with ice cold PBS. The harvested cells were resuspended in lysis buffer (10mM HEPES, pH 7.5, 10mM KCl, 1mM MgCl2) containing protease inhibitors (Roche, Indianapolis, Indiana, USA). Crude cell extracts were centrifuged for

10min at 800g and the resulting supernatants were centrifuged at 100 000g for 1h. The latter supernatants correspond to the cytosolic extracts and protein concentration was determined using the BioRad Protein Assay (BioRad, Hercules, CA, USA).

1D SDS-PAGE Fractionation and Nano-LC-MS/MS Analysis

One hundred micrograms of proteins were diluted in Laemmli buffer and boiled for 5 min before being separated on a 12% acrylamide SDS-PAGE gel. Proteins were visualized by Coomassie Blue staining. Each lane was cut into 20 homogenous slices that were washed in 100mM ammonium bicarbonate for 15min at 37°C followed by a second wash in 100mM ammonium bicarbonate, acetonitrile (1:1) for 15min at 37°C. Reduction and alkylation of cysteine residues were performed by mixing the gel pieces in 10mM DTT for 35min at 56°C followed by 55mM iodoacetamide for 30min at room temperature in the dark. An additional cycle of washes in ammonium bicarbonate and ammonium bicarbonate/acetonitrile was then performed. Proteins were digested by incubating each gel slice with 0.6μg of modified sequencing grade trypsin in 50mM ammonium bicarbonate overnight at 37°C. The resulting peptides were extracted from the gel by three steps: a first incubation in 50mM ammonium bicarbonate for 15min at 37°C and two incubations in 10% formic acid, acetonitrile (1:1) for 15min at 37°C. The three collected extractions were pooled with the initial digestion supernatant, dried in a SpeedVac, and resuspended with 14μl of 5% acetonitrile, 0.05% trifluoroacetic acid.

The peptide mixtures were analyzed by nano-LC-MS/MS using an Ultimate3000 system (Dionex) coupled to an LTQ-Orbitrap mass spectrometer (Thermo Fisher Scientific, Bremen, Germany). Five microliters of each sample were loaded on a C18 precolumn (300-μm inner diameter $\times$ 5mm; Dionex) at 20μl/min in 5% acetonitrile, 0.05% trifluoroacetic acid. After 5min of desalting, the precolumn was switched on line with the analytical C18 column (75-μm inner diameter $\times$ 15cm; PepMap C18, Dionex) equilibrated in 95% solvent A (5% acetonitrile, 0.2% formic acid) and 5% solvent B (80%

acetonitrile, 0.2% formic acid). Peptides were eluted using a 5-50% gradient of solvent B during 80min at a 300nl/min flow rate. The LTQ-Orbitrap was operated in data-dependent acquisition mode with the Xcalibur software. Survey scan MS spectra were acquired in the Orbitrap on the 300-2000m/z range with the resolution set to a value of 60,000. The five most intense ions per survey scan were selected for CID fragmentation, and the resulting fragments were analyzed in the linear trap (LTQ). Dynamic exclusion was used within 60s to prevent repetitive selection of the same peptide.

Supporting Information 11 – SCAMPI on TOP3 input

A strict implementation of TOP3 simplifies the problem by discarding all shared peptides, and by focusing on proteins with enough unique peptide evidence. Furthermore, if a protein has "too much" peptide evidence, TOP3 only uses the three most intense matching peptides to quantify the protein. The question arises how well SCAMPI would actually perform if used on exactly the same input data as a strict TOP3 approach, namely a bipartite graph holding only:

- proteins with at least three matching quantified unique peptides
- the three most intense unique peptides for each of these proteins

In particular, the graph does not include any shared peptides and each connected component holds exactly one protein and three peptides. The formula to estimate the protein abundance is given in Equation 3 in the manuscript.

$$\widehat{C}_j \;=\; \mu \;+\; \begin{pmatrix} U_1^{(j)} \;-\; \alpha \;-\beta\mu \\ U_2^{(j)} \;-\; \alpha \;-\beta\mu \\ U_3^{(j)} \;-\; \alpha \;-\beta\mu \end{pmatrix}^{\intercal} \begin{pmatrix} \beta^2+\tau^2 & \beta^2 & \beta^2 \\ \beta^2 & \beta^2+\tau^2 & \beta^2 \\ \beta^2 & \beta^2 & \beta^2+\tau^2 \end{pmatrix}^{-1} \begin{pmatrix} \beta \\ \beta \\ \beta \end{pmatrix}$$

for all proteins.

This leads to $\widehat{C}_j = \dfrac{\left(U_1^{(j)} + U_2^{(j)} + U_3^{(j)} - 3\alpha\right) \cdot \beta + \mu\tau^2}{3\beta^2 + \tau^2}$.

In the case of TOP3, the protein abundance estimate is $\widehat{C}_j^{\text{TOP3}} = \dfrac{U_1^{(j)} + U_2^{(j)} + U_3^{(j)}}{3}$.

Combining both equations leads to $\widehat{C}_j = \dfrac{3\beta}{3\beta^3 + \tau^2}\, \widehat{C}_j^{\text{TOP3}} + \dfrac{\mu\tau^2 - 3\alpha\beta}{3\beta^2 + \tau^2}$.

Hence, in this particular case, the SCAMPI protein abundance scores can be obtained from the strict TOP3 scores by linear transformation. In other words, if (i) one is willing to discard all information, except the input required by a strict TOP3, and (ii) one is only interested in protein abundance estimates, then there is no advantage in running SCAMPI. It is simpler and quicker to take the average of the three selected peptide intensities for each protein. The (potential) drawbacks of such a choice are:

- Proteins matched by less than three unique peptides do not get quantified,
- additional information contained in shared peptides is not taken into account,
- no model is available to reassess/validate the peptide abundance scores, and
- no model is available to compute prediction intervals for the protein abundance scores.

Supporting Information 12 – SCAMPI results compared to MaxQuand output

There are several reasons to explain the differences between the list of high abundant proteins returned by SCAMPI and the one from MaxQuant:

- Part of the high abundant proteins quantified by SCAMPI did not get an abundance estimate in MaxQuant, because they only have a single unique peptide (the corresponding peptide has been quantified, but no summary score for the protein is available in the MaxQuant output).

- MaxQuant uses further data bases for contaminants. For example, MaxQuant does not return the protein `REFSEQ:NP_004684` in its protein level output (recognizing it as a contaminant) while we did not yet filter it out.

From a biological point of view, both output lists seem to have some interesting aspects. In the presented data set (*SILAC labeled human shotgun proteomics data*), at least one of the proteins discarded by MaxQuant (due to too little evidence), but recognized as high abundant by SCAMPI, is of biological interest.

References

[1] Sarah Gerster and Peter Bühlmann. *protiq: Protein (identification and) quantification based on peptide evidence*, 2012. R package version 1.1.

[2] Christina Ludwig, Manfred Claassen, Alexander Schmidt, and Ruedi Aebersold. Estimation of absolute protein quantities of unlabeled samples by selected reaction monitoring mass spectrometry. *Molecular & Cellular Proteomics*, 11(3), 2012.

[3] Nonlinear Dynamics Ltd. Progenesis LC-MS. `http://www.nonlinear.com/products/progenesis/lc-ms/overview/`.

[4] Martin Beck, Alexander Schmidt, Johan Malmstroem, Manfred Claassen, Alessandro Ori, Anna Szymborska, Franz Herzog, Oliver Rinner, Jan Ellenberg, and Ruedi Aebersold. The quantitative proteome of a human cell line. *Mol Syst Biol*, 7, 2011.

[5] Alexey I. Nesvizhskii, Andrew Keller, Eugene Kolker, and Ruedi Aebersold. A statistical model for identifying proteins by tandem mass spectrometry. *Analytical Chemistry*, 75(17):4646–4658, 2003.

[6] Jurgen Cox and Matthias Mann. Maxquant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat Biotech*, 26(12):1367–1372, 2008.

[7] John O. Rawlings, Sastry G. Pantula, and David A. Dickey. *Applied Regression Analysis – A Research Tool*. Springer, 2nd edition, 1998.

[8] Simon J. Sheather. *A Modern Approach to Regression with R*. Springer, 2009.

[9] D. G. Altman and J. M. Bland. Measurement in medicine: The analysis of method comparison studies. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 32(3):pp. 307–317, 1983.