**Supplementary Information**

**isobar\textsuperscript{PTM}: A Software Tool for the Quantitative Analysis of Post-translationally Modified Proteins**

Florian P. Breitwieser and Jacques Colinge*

CeMM – Research Center for Molecular Medicine of the Austrian Academy of Sciences, Lazarettgasse 14, AKH-BT. 25.3, 1090 Vienna, Austria

1. **Null distribution**

In the original paper describing the isobar statistical models\cite{1} we showed that unregulated protein ratios follow a heavy tailed distribution. Here, exploiting three datasets already described in this paper to cover multiple MS and labeling techniques (orbitrap iTRAQ and TMT, MALDI-TOF/TOF TMT), we collected peptide ratios from unregulated proteins in these datasets and observed that they also follow a heavy tailed distribution (Suppl. Figures S1-S3).

**Suppl. Figure S1.** Peptide ratios from TS1 dataset (Ref. \cite{1}), LTQ-Orbitrap, iTRAQ 4-plex labeling. Cauchy (blue) and generalized Student’s T (pink) models.
Suppl. Figure S2. Peptide ratios from MALDI-TOF/TOF dataset (Ref. [1]), TMT 6-plex labeling. Cauchy (blue) and generalized Student’s T (pink) models.

Suppl. Figure S3. Peptide ratios from LTQ-Orbitrap, CSF fluid dataset (Ref. [1]), TMT 6-plex labeling. Cauchy (blue) and generalized Student’s T (pink) models.
We then performed the same analysis using the data of Phanstiel et al.[2] collecting phosphorylated peptide ratios between replicates, i.e. modified peptide ratios for unregulated PTMs. The result is identical (Suppl. Figure S4) thus validating isobar null distribution at the modified peptide level.

**Suppl. Figure S4.** Unregulated phosphorylated peptide ratios (Ref.[2]). Cauchy (blue) and generalized Student’s T (pink) models.
2. Validation of the selection model and performance evaluation

Again exploiting further the depleted plasma test sample generated for the original isobar paper, we conducted a novel false/true positive rate evaluation but at the peptide level. Selecting peptide ratios at 5% false positives according to the statistical model actually delivered actual false positive rates close and below this limit when comparing biological replicates of the same sample where no ratio should be selected ideally (Suppl. Table S1).

**Suppl. Table S1.** False positive rates observed when a 5% threshold was imposed on the selection.

| Num spectra | Isobar Cauchy | Isobar general. T | T-test | Fold change |
|-------------|---------------|------------------|--------|-------------|
| 1           | 0.01          | 0.01             | 0.00   | 0.08        |
| 2           | 0.01          | 0.01             | 0.06   | 0.04        |
| 3           | 0.01          | 0.03             | 0.09   | 0.03        |
| 5           | 0.01          | 0.02             | 0.27   | 0.02        |
| 10          | 0.00          | 0.02             | 0.45   | 0.01        |
| 15          | 0.00          | 0.01             | 0.50   | 0.00        |
| 20          | 0.00          | 0.01             | 0.54   | 0.00        |

*a* Isobar significant ratio selection procedure with Cauchy null distribution for modeling biological variability.

*b* Cauchy replaced by the more accurate generalized Student’s T.

*c* Fold change of 1.5 considered as a significant ratio.

True positive rate estimations are dependent on the peptide abundance, the number of spectra available, and the actual ratio magnitude. From the test sample we selected a large number of peptides at different concentrations, known ratios, and we randomly selected different numbers of spectra (when more were available) following the procedure we applied to characterized protein detection performance in Breitwieser et al. (2011). Results are similar to protein level performance (Suppl. Table S2) and also show that isobar selection, which was able to control false positives successfully (Suppl. Table S1), is not always the most sensitive but more sensitive methods are typically the ones yielding far unacceptable false positive rates. We hence conclude that the selection method is appropriately ported to the peptide level.
Suppl. Table S2. True positive rates.

| Num spectra | Isobar Cauchy | Isobar general. T | T-test | Fold change |
|-------------|---------------|-------------------|--------|-------------|
| **Peptide ratio 1.3** | | | | |
| 1 | 0.13 | 0.16 | 0.00 | 0.41 |
| 2 | 0.11 | 0.14 | 0.07 | 0.24 |
| 3 | 0.14 | 0.22 | 0.21 | 0.19 |
| 5 | 0.13 | 0.34 | 0.37 | 0.17 |
| 10 | 0.13 | 0.39 | 0.62 | 0.16 |
| 15 | 0.17 | 0.47 | 0.75 | 0.20 |
| 20 | 0.20 | 0.48 | 0.73 | 0.23 |
| **Peptide ratio 1.5** | | | | |
| 1 | 0.27 | 0.27 | 0.00 | 0.51 |
| 2 | 0.22 | 0.24 | 0.07 | 0.43 |
| 3 | 0.31 | 0.39 | 0.24 | 0.40 |
| 5 | 0.28 | 0.46 | 0.44 | 0.34 |
| 10 | 0.40 | 0.61 | 0.68 | 0.50 |
| 15 | 0.47 | 0.72 | 0.82 | 0.61 |
| 20 | 0.54 | 0.68 | 0.80 | 0.64 |
| **Peptide ratio 2 (abundant)** | | | | |
| 1 | 0.66 | 0.65 | 0.00 | 0.87 |
| 2 | 0.75 | 0.76 | 0.33 | 0.95 |
| 3 | 0.88 | 0.91 | 0.78 | 0.94 |
| 5 | 1.00 | 1.00 | 1.00 | 1.00 |
| 10 | 1.00 | 1.00 | 1.00 | 1.00 |
| 15 | 1.00 | 1.00 | 1.00 | 1.00 |
| 20 | 1.00 | 1.00 | 1.00 | 1.00 |
| **Peptide ratio 2 (low)** | | | | |
| 1 | 0.42 | 0.41 | 0.00 | 0.68 |
| 2 | 0.47 | 0.47 | 0.19 | 0.75 |
| 3 | 0.59 | 0.62 | 0.51 | 0.77 |
| 5 | 0.69 | 0.79 | 0.78 | 0.80 |
| 10 | 0.94 | 0.97 | 0.98 | 0.97 |
| 15 | 0.96 | 0.99 | 1.00 | 0.99 |
| 20 | 0.98 | 1.00 | 1.00 | 1.00 |
3. Modified peptide ratio correction

As discussed in the paper, Wu et al. (Mol Cell Proteomics, 2011) showed that correcting modified peptide ratios by the abundance change ratio of the corresponding proteins much improves their accuracy. When both ratios are available, i.e. modified peptide and protein ratios, we perform this correction and compute an upper bound on the variance of the corrected ratio, which is used the statistical test taking care of the modified peptide selection.

Let $R_n$ be the observed log-ratio of a modified peptide and $R_p$ the log-ratio of the corresponding protein. We estimate the true modified peptide ratio $R_m$ by the following formula:

$$R_m = R_n - R_p.$$  \hspace{1cm} (1)

The variance of $R_m$ is given by

$$\text{Var}(R_m) = \text{Var}(R_n) + \text{Var}(R_p) + 2\text{Cov}(R_n, R_p).$$  \hspace{1cm} (2)

The covariance $\text{Cov}(R_n, R_p)$ of the peptide and the protein ratios, however, is unknown. Omitting the covariance term means assuming independence between $R_n$ and $R_p$ but this is wrong in general since an increase of $R_p$ causes an increase of $R_n$ in the iTRAQ or TMT measurements. For the same reason, a positive correlation can be assumed generally. We can use the Pearson’s correlation coefficient $p$ formula to modify Eq. (2) and obtain an upper bound on $\text{Var}(R_m)$, hence yielding conservative ratio selections. Namely, we have

$$p(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma(X)\sigma(Y)},$$

$$\text{Cov}(X, Y) = p(X, Y)\sigma(X)\sigma(Y),$$

where $\sigma(.)$ denotes the standard deviation. With $p = p(R_n, R_p)$ we further obtain

$$\text{Cov}(R_n, R_p) = p\sigma(R_n)\sigma(R_p),$$

$$\text{Var}(R_m) = \text{Var}(R_n) + \text{Var}(R_p) + 2\sigma(R_n)\sigma(R_p).$$

$p$ is not known for the pair $(R_n, R_p)$ but it is assumed positive and $\text{Var}(R_m)$ is thus bounded by $(p = 1)$:

$$\text{Var}(R_m) \leq \text{Var}(R_n) + \text{Var}(R_p) + 2\sigma(R_n)\sigma(R_p).$$
Suppl. Figure S5. Examples of corrected modified peptide ratios. Observed modified peptide ratios (blue) are corrected according to observed protein ratios (red) to obtain a corrected modified peptide ratio (green).

4. An improved heavy tailed distribution model

In Suppl. Figures S1-S4 we have shown that unregulated peptide ratios followed a heavy tailed distribution that was well modeled by a Cauchy, which is isobar default null distribution for such ratios. In recent work we found the generalized Student’s T distribution to provide a more precise model of the distribution tails. It is visible in Suppl. Figures S1-S4 (pink curves) and this is also valid for protein ratios (Suppl. Figure S6).

To better model the tails of the null provides a more sensitive selection of peptides (or proteins) as can be nicely observed in Suppl. Table 2 without causing false positives beyond the pre-imposed error rate (Suppl. Table 1).

Since the new generalized Student’s T null model has one more parameter than the default Cauchy, we wanted to investigate how much more data is required for a safe parameter fit. We used Phanstiel et al. data and built reference nulls on the entire dataset (8884 data points) for both Cauchy and generalized T curves. We then resampled the data 1000 times using less data points (10, 50, 100, 500, 1000) and compared the proteins called significant at the 5% level with the same selection based on the reference nulls. Overlap of selections was measured by the Jaccard index and is reported in Suppl. Table S3. Clearly, Cauchy-based selections converge quicker (one less parameter to estimate, hence more robust) but generalized Student’s T yields stable selections for >1000 data points (and more sensitivity as just discussed), which is often available in current iTRAQ/TMT datasets.
**Suppl. Figure S6.** Unregulated protein ratios (Ref.[2]). Cauchy (blue) and generalized Student’s T (pink) models.

**Suppl. Table 3.** Jaccard indexes of protein selections.

| Sample size | Generalized Student's T | Cauchy |
|-------------|-------------------------|--------|
|             | median      | mean   | median | mean   |
| 10          | 0.40        | 0.42   | 0.67   | 0.62   |
| 50          | 0.67        | 0.64   | 0.82   | 0.81   |
| 100         | 0.75        | 0.73   | 0.86   | 0.85   |
| 500         | 0.88        | 0.86   | 0.95   | 0.93   |
| 1000        | 0.91        | 0.90   | 0.95   | 0.95   |
| all         | 1.00        | 1.00   | 1.00   | 1.00   |

**References**

1. Breitwieser FP, Muller A, Dayon L, Kocher T, Hainard A, Pichler P, Schmidt-Erfurth U, Superti-Furga G, Sanchez JC, Mechtler K *et al*: General Statistical Modeling of Data from Protein Relative Expression Isobaric Tags. *J Proteome Res* 2011, 10(6):2758-2766.

2. Phanstiel DH, Brumbaugh J, Wenger CD, Tian S, Probasco MD, Bailey DJ, Swaney DL, Tervo MA, Bolin JM, Ruotti V *et al*: Proteomic and phosphoproteomic comparison of human ES and iPS cells. *Nat Methods* 2011, 8(10):821-827.