Supplementary Figures

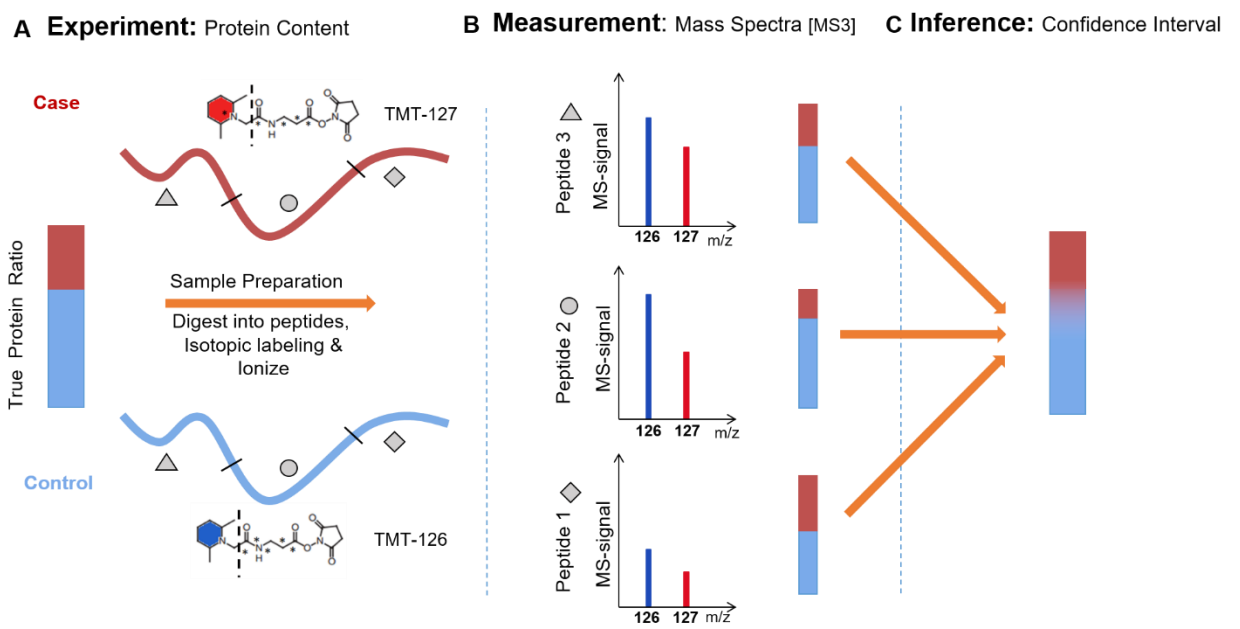


Figure S1. Overview of the challenge to integrate multiplexed proteomics measurements into an estimation of protein ratio between conditions and the associated confidence in this measurement. A) Multiplexed proteomics allows the comparison of protein abundance among multiple conditions. For simplicity, only two conditions and three peptide measurements are shown. A protein with true protein ratio (red-blue bar) is digested into peptides. The peptides are labeled with isobaric tags e.g. TMT to encode the different conditions. The digestion and labeling for different conditions happen in different tubes. The barcoded samples are then combined and ionized. This differential digestion and labeling could introduce disagreement in the peptide ratios. B) The different peptides derived from each protein result in separate spectra, which allow quantification using accurate multiplexed proteomics methods, e.g. MultiNotch MS3 or TMTc+. Each spectrum contains the information of relative abundance and signal strength of a peptide. The limited number of ions measured (proportional to signal strength) introduce yet another source of distortion. C) The challenge is how the information of

different peptide signal and agreement/disagreement between measured peptide ratios can be integrated to accurately estimate the underlying true protein ratio with associated confidence.

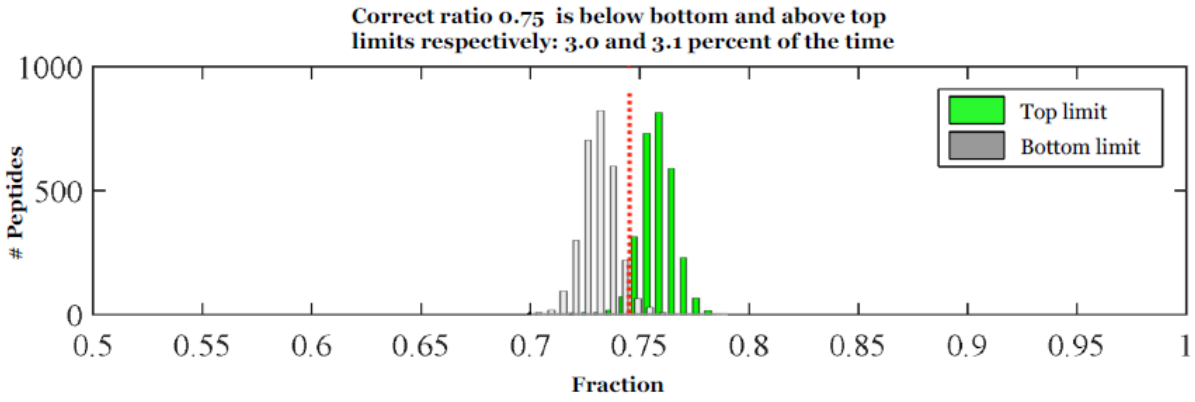


Figure S2: Assignment of confidence intervals on the peptide level is adequate for various ratios. Histogram of the upper and lower bound values for the 95% confidence intervals, where the true ratio of all the peptides is 3:4. The observed percentage of peptides for which the true answer is below bottom limits is 3.0% of the time and above top limits is 3.1% of the time.

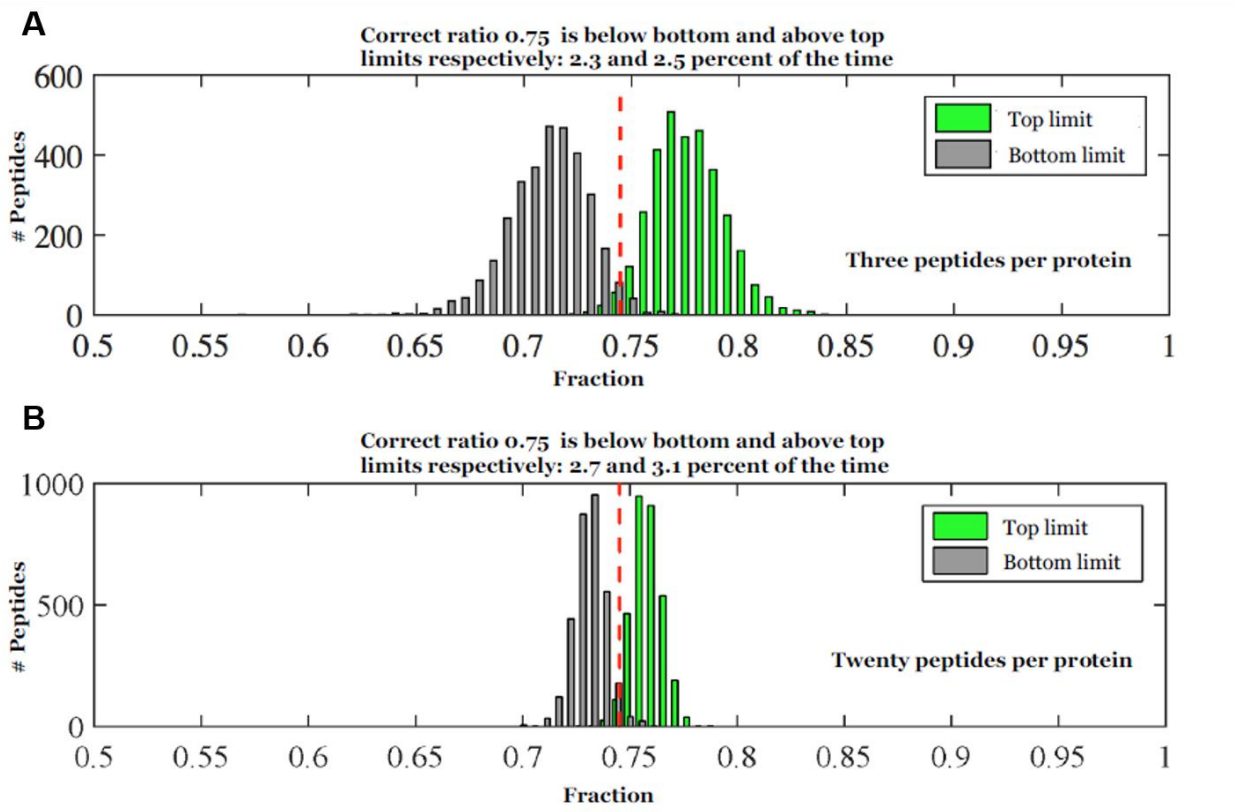


Figure S3: Summing up peptide signal is adequate for the artificial protein case. Artificial proteins are generated by summing from several peptides from a sample in which all peptide ratios are identical. A) Three peptides were selected per artificial protein. The correct mixing ratio for this sample is 0.75. The 95% confidence intervals are below bottom limits 2.7% of the time and above top limits 3.2% of the time. B) 20 peptides were assigned per artificial protein. Bottom limit of 95% confidence intervals is 2.3% of the time above the true ratio, and top limit is 2.5% of the time below true ratio.

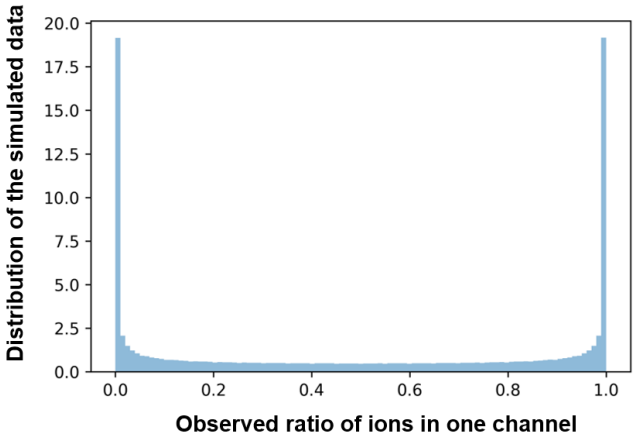


Figure S4: Prior Predictive Checks for the Beta-Binomial model at protein level. The simulated data for the observed ratio in one channel has more mass around extreme values of 0 and 1, which are not observed in the proteomics data, hence the priors are only weakly informative. However, they are informative enough to not produce values outside the domain of (0, 1).

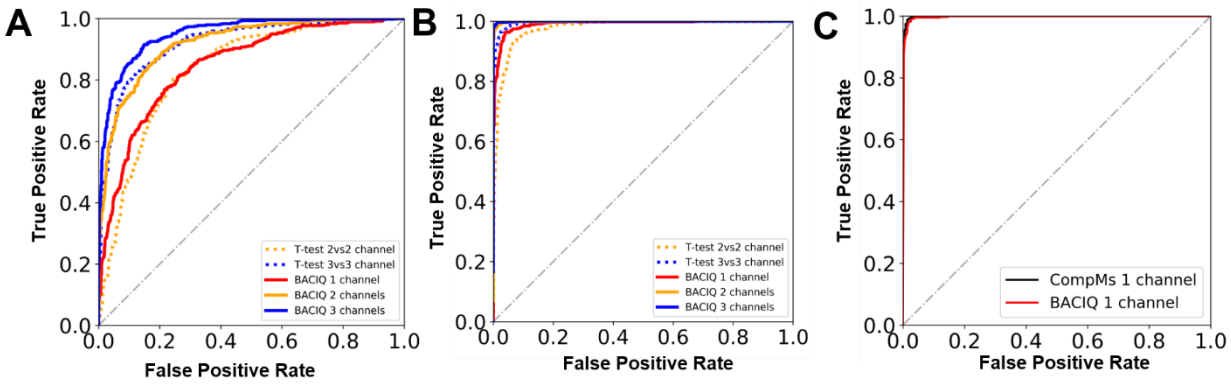


Figure S5. Comparison of BACIQ with other models in its ability to distinguish differential expression. The subpanel A and B were equivalently constructed to figure 6, but the ratio of the spiked-in E. coli proteins was changed to A) 1:1.1 and B) 1:1.4. For all tested ratios BACIQ outperforms the naive t-test in the ability to detect expression changes. C) Analysis of the 32-fold change dataset with constant background from O'Brien *et al.* ROC curves were constructed without replicates. The ROC curves show that both BACIQ and CompMS are nearly perfectly able to distinguish between changing and constant proteins.

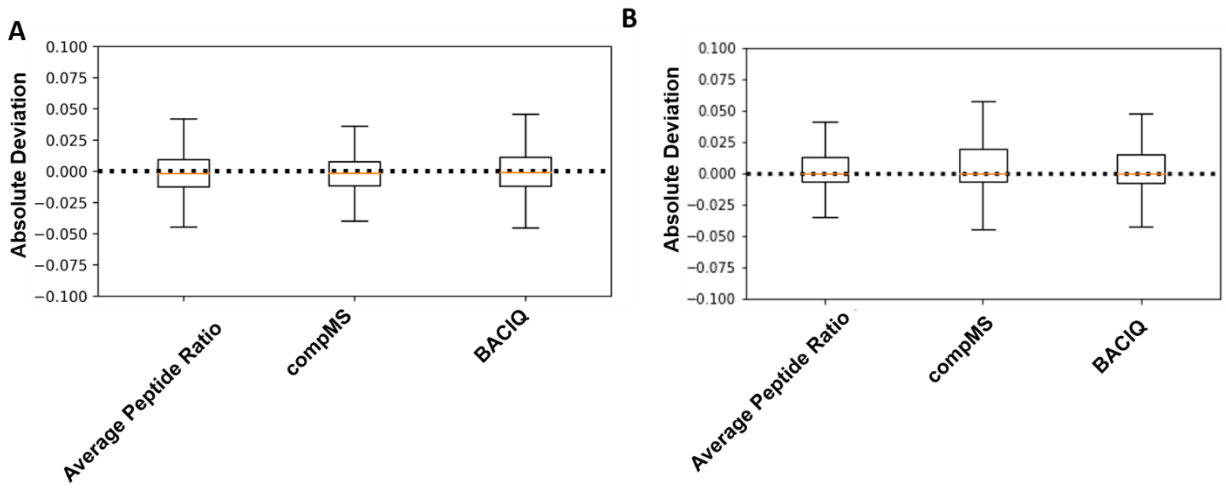


Figure S6: BACIQ shows similar accuracy to alternative methods. The accuracy is reported as the absolute deviation of protein estimate from true known protein ratio. Shown are boxplots with 25th, 50th and 75th percentiles. A) Absolute deviation between predicted and known answer for 1.2-fold changing proteins (Figure 6), analyzed with average peptide ratio, compMS, and BACIQ. B) For this panel, we re-analyzed the 32-fold change dataset with constant background from O'Brien *et al.* Absolute deviation between predicted and known answer for 32-fold changing proteins, analyzed with average peptide ratio, compMS, and BACIQ.

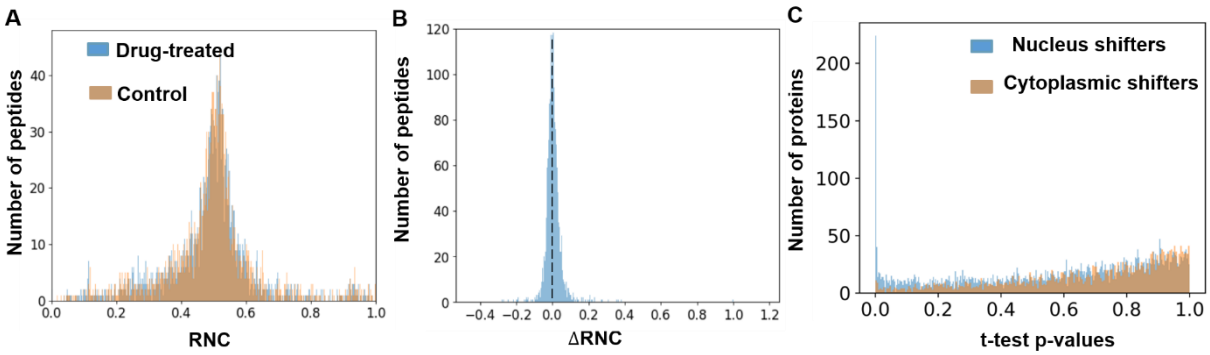
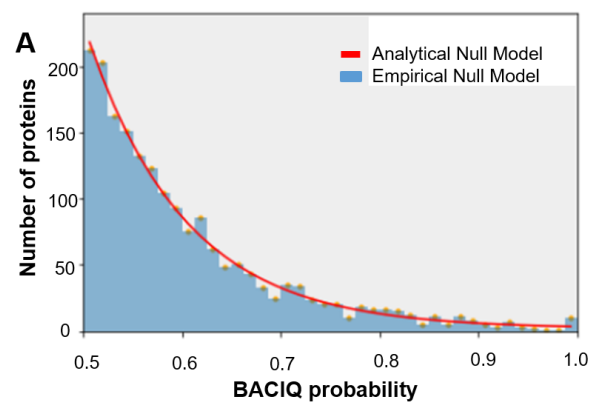
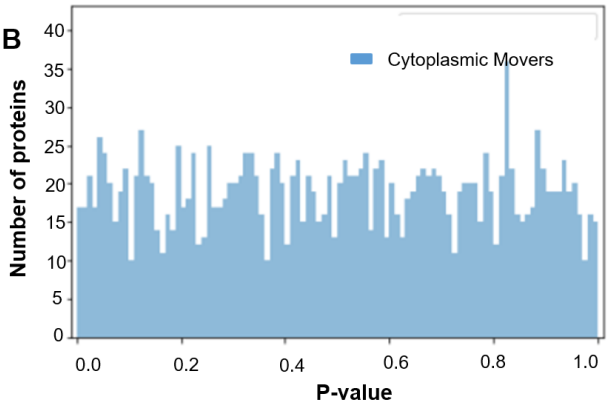
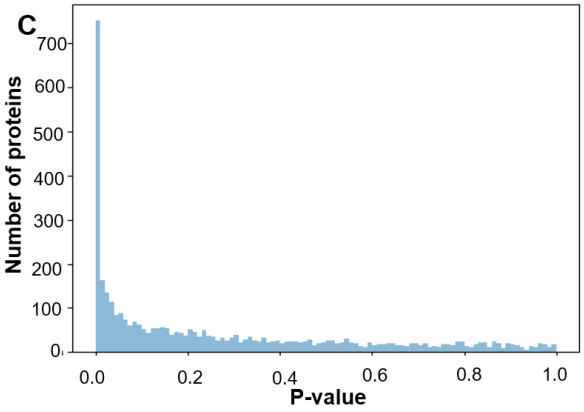


Figure S7: Normalization of the nucleocytoplasmic data. A) Histogram of RNC for the control and drug–treated samples for proteins in complexes of less than 40kDa. These proteins are small enough to diffuse and equilibrate via the nuclear pore and are therefore not expected to respond to LMB treatment. The identical distribution before and after drug treatment indicates that our data is well normalized. B) Histogram of the normalized peptides from small proteins (native MW <40kDA) indicate a median RNC change of zero upon drug treatment. C) We fine-tuned normalization of the drug treated vs control data so that there were equal number of cytoplasmic and nucleus movers for high t-test p-value scores.







D

Significance Level	Bonferroni (FWER)	Benjamin-Hochberg (FDR)
	Total proteins	Total proteins
1%	160	454
5%	200	747

Figure S8: Multiple Hypothesis Corrections to assign FDRs to nucleocytoplasmic movers. A) A histogram of the BACIQ scores of apparent cytoplasmic shifters. The red line is the analytic approximation. B) P-values associated with the BACIQ score of the cytoplasmic movers. The p-values of the cytoplasmic shifters follow a uniform distribution indicating that the cytoplasmic shifters fulfill a key requirement for the true null. C) BACIQ p-values of the nucleus movers. The peak close to 0 is where most of the alternative hypothesis lies along with some potential false positives D) Number of significant movers after applying standard multiple hypothesis correction procedures (Bonferroni, or Benjamini-Hochberg) for 1% and 5% FDR.

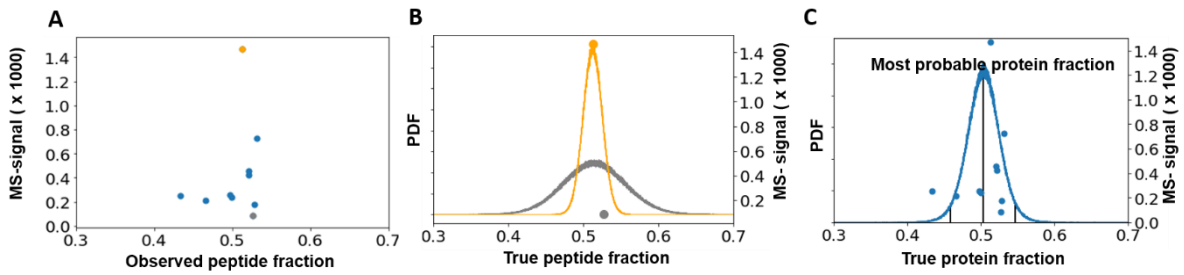


Figure S9 Worked-out example for the model. A) Raw peptide data with two highlighted data-points. B) Posterior distributions of the true peptide fractions for two of the peptides marked as yellow and grey circles in A. Peptides with lower MS-signal have wider distributions. C) Posterior distribution of the true protein fraction that accounts for measurement errors due to poor ion statistics and peptide ratio agreement. The median of this posterior distribution gives the true protein fraction estimate (which is 0.50 in the present case) and the 95% confidence interval is (0.46, 0.54)

Supplementary Material

Beta-Binomial model for an individual peptide

In a complex mixture, the lowest possible concentration of peptides that can be detected with mass spectrometry based proteomics is at least 10 attomole [1-3] or ~6,000,000 molecules. However, not all these ions ionize into the instrument simultaneously. Furthermore, throughout their passage in the instrument, the ions are further lost. Ultimately, a single peptide is quantified with ~ 20-2000 reporter ions for two channels. Sampling of 2000 reporter ions from an initial pool of at least 6,000,000 ions (10 attomole) (drawn "without replacement"), follows a hypergeometric distribution. However, since the sampled number of ions are low (0.03%) compared to the total amount, the hypergeometric distribution can be approximated by a binomial distribution. Please note that this is the worst-case scenario. For the vast majority of peptides, the fraction that is analyzed is much smaller than 0.03% of what is loaded on the mass spectrometer.

We chose a non-informative prior distribution such that the probability of log-odds $\log(\frac{\theta}{1-\theta})$, is constant, which corresponds to the improper Beta (0, 0) density on θ. Using an improper prior makes sense only if the posterior is proper. For Orbitrap data, the FT-noise band provides a natural threshold for signal detection. Therefore, the minimal value used is 1 MS-signal i.e. $\alpha>0$, $\beta>0$.

Fitting the parameter to convert MS-signal to number of ions.

Number of successes α on *n* coin tosses, given its fairness θ follow a binomial distribution $\alpha \sim Bin(n, \theta)$. The functional form of the convergence of fraction of successes to the true underlying fraction (fairness of the coin) on increasing number of coin tosses n is given by the relationship of the CV of binomial distribution with n. The mean of the binomial distribution with parameters n (number of tosses) and θ (probability of success) is $n\theta$ and the standard deviation is $\sqrt{n\theta(1-\theta)}$. Therefore, the coefficient of variation is $\sqrt{(1-\theta)/n\theta}$

We fit a single parameter m as a multiplier to MS-signal value s where $n = ms$ to the binned data as illustrated in Figure 2B. The data of 10534 points is binned by 500 data points into 21 bins, and CV is calculated for each bin. The parameter m is obtained with MATLAB's Nonlinear-Least-Squares fitting method.

Justification of priors for Beta-Binomial model at the protein level

To make no prior assumption about the true protein ratio, we set the prior for the mean (μ) to be a uniform distribution on the domain (0, 1). The range of precision parameter (κ) is $[0, \infty)$. We chose a weakly informative prior of exponential distribution with the rate parameter of 0.05. This is only weakly informative because the exponential distribution has a higher probability mass on lower values, with the expected value of 20 for the current rate parameter of 0.05, which is conservative because we expect κ to much larger than 20 in the posterior. However, it is informative enough to not cause any sampling problems.

To further understand, how prior distributions affects the given model, we simulate fake data from the computational model. This is called "Prior Predictive Checks" [4, 5].

Figure S4 shows the Prior Predictive distribution for BACIQ. The simulated data for the observed ratio in one channel has more mass around extreme values of 0 and 1. This is because of extremely conservative exponential distribution of κ. We believe that these prior choices are still only weakly informative, in that the implied data generating process can generate data that is much more extreme than we would expect from the domain of our data. However, it is informative enough to not produce the data that is outside the domain of (0, 1).

Since, the data generated is broader than the expected distribution of the observed data, we can be sure that the priors comply with the principle of weakly informative priors.

Worked-Out example for Beta-Binomial multi-peptide protein model

1) Let's start with a protein measured with 10 peptides in two different conditions (Fig. S9A). We know that this protein is present in equal amounts in both the conditions. The goal is to get the estimate of the protein fraction from the peptide data and the associated confidence. To be able to apply the model, we first need to convert the peptide MS-signal to ion counts. Since the peptide was shot on the Fusion/Lumos with a resolution of 15k in the Orbitrap, we multiply the signal from each channel with 3.4 (Look up Table 1 for appropriate multiplication factor).
2) We have 20 observations for this case, 10 of those are $(\alpha_1, \alpha_2, \alpha_3 \ldots . \alpha_{10})$ the ion counts of the 10 underlying peptides corresponding to the first channel, and the other 10 are the $(n_1, n_2, n_3 \ldots . n_{10})$ sum of the signal in two channels. Figure S9A shows the raw peptide fractions $({}^{\alpha_1}/_{n_1}, {}^{\alpha_2}/_{n_2} \ldots . {}^{\alpha_{10}}/_{n_{10}})$ with the respective signal strengths $(n_1, n_2, n_3 \ldots . n_{10})$ for all the peptides.
3) For this protein, we have 12 latent parameters $(\boldsymbol{\mu}, \boldsymbol{\kappa}, \theta_1, \theta_2, \theta_3 \ldots . \theta_{10})$. We use MCMC algorithm to draw samples from the posterior distributions. The algorithm draws samples for all the unobserved parameters including true peptide fractions $(\theta_1, \theta_2, \theta_3 \ldots . \theta_{10})$. Figure S9B shows the posterior distributions for two of those peptides.
4) The quantity of interest is the variable $\boldsymbol{\mu}$, which represents the relative protein fraction. The histogram of all the samples drawn for $\boldsymbol{\mu}$ gives us the probability distribution of $\boldsymbol{\mu}$ over the domain of 0, 1. This histogram represents the posterior probability density of $\boldsymbol{\mu}$ marginalized over all the other unobserved parameters. The median of the samples drawn from $\boldsymbol{\mu}$ represent the estimated protein fraction and the 2.5th and 97.5th percentiles represent the 95% confidence interval (Figure S9C)

Comparison of the ability to distinguish differential expression for BACIQ and other approaches

We compare the t-test, which uses 2 or 3 repeats, to BACIQ. The t-test ranks the proteins on their differential expression using p-value. The p-value is the probability that the different sets of peptide measurements for different conditions would be observed in the context of a model constructed under the assumptions of the t-test and the null hypothesis. Therefore, the proteins from most likely to least likely differentially expressed are rank ordered by 1 - (p-value). BACIQ ranks the proteins based on the probability mass falling to the right of 0.5 in the distribution for fraction between two conditions. Sliding threshold on the ranking allows building ROC curve – a way to compare probabilistic classifiers' tradeoff between true positive and false positive rates.

Partially pooled Beta-Binomial Model

For a fair comparison to the compMS model, we partially pooled variances across proteins in our model (BACIQ pooled). Bayesian Models can be partially pooled by adding hierarchy to the variance/precision parameter [6]. Our partially pooled BACIQ model is defined below:

- The unobserved true peptide ratios θ_{ij}, for all peptides of a protein *j* are sampled from a beta distribution parameterized with mean μ_j and precision κ_j. Here, $i = 1,2,3, \ldots, I$ indexes the number of peptides for the protein *j* and $j = 1,2,3, \ldots, J$ is the index running over the proteins.
- Given true underlying peptide ratios θ_{ij}, the number of ions in one channel α_{ij} for all peptides of protein *j* can be modelled using a Binomial distribution, with n_{ij} representing the sum of counts in the two channels.
- The prior for the mean parameter μ_j is Uniform over the domain (0, 1).
- Partial pooling of the variance is accomplished by sampling the precision parameter (κ_j) for every protein from the exponential distribution with a common hierarchical rate parameter(τ).

- The rate parameter (τ) is sampled from a Gamma distribution. Again, the choice of hyperparameters, 3 for the scale and 3 for the rate, was made conservatively. This is because, a-priori $E\ [\tau] = \ 1$ and $E\ [\kappa_j] = \ 1$, which is conservative because we expect κ_j to be much larger than that in the posterior.

Partially pooled Beta-Binomial model:

$$\alpha_{ij}|\theta_{ij}, n_{ij} \ \sim Bin\left(n_{ij}, \theta_{ij}\right)$$

$$\theta_{ij}|\mu_j, \kappa_j \ \sim Beta\left(\mu_j, \kappa_j\right)$$

$$\mu_j \ \sim Uniform(0,1)$$

$$\kappa_j|\tau \sim Exponential\ (\tau)$$

$$\tau \sim Gamma(3,3)$$

Accuracy of reported fold changes.

BACIQ performs very well in distinguishing small fold changes as shown by ROC curves (Figure 6, S5). However, we do not see any significant improvement in accuracy of reported estimates of protein abundances when compared to naive peptide ratio averaging model. Figure S6 A shows the accuracy boxplots for small fold changes (1.2-fold). Figure S6 B shows the accuracy boxplots for large fold changes (32-fold). For the large fold change, we reanalyzed the boundary case experiment raw files from O'Brien *et. al*. We used the two channels that had yeast expressed as 32-fold change on a 1:1 mouse background [7]. As expected, our model should have no effect on accuracy as we do not model "ratio compression" which is the prominent issue affecting the quantification at such large fold changes.

Implementation of BACIQ for the Exportin 1 inhibition data

BACIQ was first applied on the nucleus and the cytoplasm channel to obtain the RNC distribution for the control and again on the nucleus and cytoplasm channel to obtain the RNC

for drug treated sample. We considered the 7841 proteins that have more than one peptide. The samples for the marginal posterior distribution of the true protein ratio were collected by dividing the RNC space (0,1) into 40000 bins, giving the grid size of 2.5e-5 RNC.

Computing the probability values for the Exportin-1 inhibition data

To calculate the probability of shifting from the RNC distributions of control and drug treated samples, we ask the $P(RNC_{drug} > RNC_{control})$ for the proteins with positive shift RNC ($shift\ RNC = \ Median\ RNC_{drug} - Median\ RNC_{control}$) and $P(RNC_{control} > RNC_{drug})$ for the negative shift RNC. These probabilities can be numerically computed using the following formulae: -

$$P(RNC_{control} > RNC_{drug}) = \int_0^1 \ cdf_{drug}(RNC)\ pdf_{control}\ (RNC)\ d\ RNC$$

$$= \ \sum_1^{40000} \ \ cdf_{drug}(i)\ pdf_{control}\ (i)$$

$$= Area\ to\ the\ left\ of\ 0\ in\ shift\ RNC\ curve$$

$$P(RNC_{drug} > RNC_{control}) = \int_0^1 \ cdf_{control}(RNC)\ pdf_{drug}\ (RNC)\ d\ RNC$$

$$= \ \sum_1^{40000} \ \ cdf_{control}(i)\ pdf_{drug}\ (i)$$

$$= Area\ to\ the\ right\ of\ 0\ in\ shift\ RNC\ curve.$$

Normalization and multiple hypothesis correction

To correct for pipetting errors within the control and drug-treated experiment, we normalized corresponding nuclear and cytoplasmic proteins for each experiment, so that the median of proteins in complexes smaller than 40kDA are distributed 1:1 (RNC = 0.5) (Fig. S7A, B) [8]

To counteract detection of apparent LMB movers due to slight off normalization or isotopic impurities of the TMT tag, we fine-tuned normalization for control and drug-treated fractions, so that there were equal number of cytoplasmic and nucleus movers for the proteins that had no significant movement (Fig. S7C) i.e. equal number of high t-test p-value scores.

We used the distribution of the BACIQ probability scores for cytoplasmic movers (before overcorrection) as a model for the null. To accurately calculate the p-value associated with the BACIQ scores for all the proteins, the empirical null model is approximated by an analytical model. The binned data was fit to the exponential distribution using Non-Linear Least Squares method (Fig. S8A). A key property of the null distribution is that its p-values are uniformly distributed. A uniform distribution of the p-values for the BACIQ scores of cytoplasmic movers suggest that they indeed correspond to the true null (Fig. S8B). On the other hand, the p-values for the nucleus movers show a large peak close to 0 (Fig. S8C). Fig. S8D tabulates the number of proteins being pulled out as significant on applying the two standard but conservative multiple hypothesis correcting procedures

Comparison of putative Exportin-1 substrates with other databases.

We compared proteins we identified as significantly moving towards the nucleus to the Exportin-1 binder from Xenopus published by the Görlich group [9] and a database curated from literature (NESDB) [10]. To facilitate comparison of proteins identified with an mRNA derived reference database, we mapped all Xenopus proteins to their human gene symbols as previously described [1].

All the proteins that were called Cargo A, Cargo B, or low abundant Cargo by Görlich were considered as “binders” for the overlap. The 971 binder proteins collapsed down to 960 unique gene symbols. For our study, the 747 proteins at 5% FDR for our dataset collapsed down to 612 unique gene symbols. We found an overlap of 245 unique gene symbols between both the datasets of unique gene symbols. To calculate the statistical significance of this overlap, we

calculated the probability of observing an intersection of at least 245 on overlapping 960 genes and 612 genes from the total of 5531 unique gene symbols identified in our data set, by random chance. The small p-values (5.5 e-29) using the hypergeometric distribution indicate that this overlap is highly significant.

Similarly, for the 230 unique gene symbols of NESDB, only 76 were given any predictions from our dataset. Out of these 76, 26 overlapped with the 612 unique gene symbols from our study. To calculate the statistical significance of this overlap, we asked the probability of observing an intersection of at least 26 on overlapping 76 genes with 612 genes from the total of 5531 unique gene symbols. The small p-values (3.1 e-6) using the hypergeometric distribution indicate that this overlap is highly significant.

Supplemental references

1. Wuhr, M., et al., *Deep proteomics of the Xenopus laevis egg using an mRNA-derived reference database.* Curr Biol, 2014. **24**(13): p. 1467-1475.
2. Presler, M., et al., *Proteomics of phosphorylation and protein dynamics during fertilization and meiotic exit in the Xenopus egg.* Proc Natl Acad Sci U S A, 2017. **114**(50): p. E10838-E10847.
3. Sidoli, S., K. Kulej, and B.A. Garcia, *Why proteomics is not the new genomics and the future of mass spectrometry in cell biology.* J Cell Biol, 2017. **216**(1): p. 21-24.
4. Andrew Gelman, D.S., Michael Betancourt, *The prior can generally only be understood in the context of the likelihood.* 2017.
5. Jonah Gabry, D.S., Aki Vehtari, Michael Betancourt, Andrew Gelman, *Visualization in Bayesian workflow.* J. R. Stat. Soc. A, 2019: p. 389-402.
6. Gelman, A., et al., *Bayesian data analysis*. Vol. 2. 2014: CRC press Boca Raton, FL.
7. O'Brien, J.J., et al., *Compositional Proteomics: Effects of Spatial Constraints on Protein Quantification Utilizing Isobaric Tags.* J Proteome Res, 2018. **17**(1): p. 590-599.
8. Wühr, M., et al., *The Nuclear Proteome of a Vertebrate.* Current biology : CB, 2015. **25**(20): p. 2663-71.
9. Kirli, K., et al., *A deep proteomics perspective on CRM1-mediated nuclear export and nucleocytoplasmic partitioning.* Elife, 2015. **4**.
10. Xu, D., N.V. Grishin, and Y.M. Chook, *NESdb: a database of NES-containing CRM1 cargoes.* Molecular biology of the cell, 2012. **23**(18): p. 3673-6.