Supporting Information

Compositional Proteomics: The Effects of Spatial Constraints on Protein Quantification Utilizing Isobaric Tags

Jonathon J. O'Brien[1*], Jeremy D. O'Connell[1], Joao A. Paulo[1], Sanjukta Thakurta[1], Christopher M. Rose[1], Michael P. Weekes[2], Edward L. Huttlin[1] and Steven P. Gygi[1*]

[1]Department of Cell Biology, Harvard Medical School, Boston, MA, 02115, USA

[2]Cambridge Institute for Medical Research, University of Cambridge, Hills Road, Cambridge CB2 0XY, UK

[*]Correspondence: obrienj@hms.harvard.edu (J.J.O), steven_gygi@hms.harvard.edu (S.P.G)

Table of Contents

Supporting Information includes:

Supplementary Tables:

Supplementary Methods and Discussion

Compositional Data Analysis

Ignoring the compositional effects of ion sampling can lead to nonsensical comparisons as discussed in the main text. However, more subtle errors also exist. In the Figure 1C example, when estimating the ratio between channels 126 and 128, averaging the intensities prior to estimating the ratio, while not advised, will still provide reasonable results. The ratio after averaging would be $\frac{x+y}{5x+5y} = \frac{1}{5}$, which is on target regardless of how greatly x and y diverge. Accordingly, the most important aspect of an analysis is correctly defining the relative quantities of interest. However, this sort of analysis is still suboptimal as lower intensities will contribute less to the average, even if the ratios are of identical quality.

Similarly, it should be noted that log linear models when used to estimate contrasts between conditions, will also agree with a basic log-ratio compositional analysis. This is because the models often have the same expected values, e.g. $E(\log_2 y_1) - E(\log_2 y_2) = E(\log_2 \frac{y_1}{y_2})$. However, this equivalence disappears when considering error estimation or models that include covariates, such as the MS2 model in this paper or a longitudinal model that seeks to estimate a time effect. Importantly, the improvement to accuracy seen with our MS2 model could not have been achieved by simply adding SSN as a covariate in regular log linear model. This is because the parameters of interest are contrasts and peptide parameters create a blocking structure. Consequently, the contrast is taken within each peptide block, and since each peptide has the same SSN across conditions, the effect on the contrast estimate will be non-existent.

The best way to avoid the mistakes discussed above is to use tools from compositional data analysis. The relevant theory is based on the concept of transforming from a constrained geometric space to unconstrained real space, performing a usual analysis, and transforming back as needed. Appropriate transformations include the additive log-ratio (ALR)[1] and the isometric log-ratio transformation (ILR)[2]. The ILR has the advantage of being isometric but at the cost of interpretability (transforming back to the proportions becomes essential). The ALR, while not isometric, provides valid inferences when using likelihood-based methods[1] and enables direct estimation of log-ratios between conditions, which are often of interest in a proteomics experiment.

From the simplex, we transform our data into real space with the ALR. Then we fit a linear model to the log-ratios of all peptides from a single protein and included a regression on isolation specificity (IS), where IS is defined as the proportion of signal in the MS1 isolation window belonging to the targeted mass (or its isotopes). At this point each outcome is defined as a log-ratio to an arbitrary reference channel. Notice that the ALR transformation reduces the dimension of the problem to two. This is because a composition with three parts only had two degrees of freedom (because of the constraint, knowing two parts tells us everything about the composition). With results obtained from two regressions in real space, the inverse of the ALR can then be used (if necessary) to create a single regression line in the simplex. Doing so here is informative.

Points in the ternary diagram seem to be pulled towards the center of the triangle and fitting a regression line on IS appears to explain much of the trend towards unity (grey dashed line). This suggests that predicting where the protein would have been if IS equaled one might mitigate some of the compression problems in isobaric tag proteomics data[3]. In Figure 3, the projected estimate is about half-way between the true value and the average intensity. It should be noted that adding isolation specificity to standard linear models for proteomics[4] would not achieve the same effect as estimating a slope that directly affects the log-ratios.

In a full dataset, making use of peptide-level covariates becomes substantially more difficult because of the unbalanced structure of the data. While certain proteins will have many peptides, others will have very few observations, making the estimation of reliable regression lines extremely difficult. Furthermore, as shown in Figure 2a, the relationship between our compression surrogate and the peptide ratios is not only a function of IS (or summed signal-to-noise), it also depends on the true protein fold-change. Proteins that do not change have zero compression, while proteins with large changes can be dramatically pulled towards unity.

Sharing information with Bayesian modelling

In proteomics it is common to analyze data for each protein independently. However, this can be problematic for proteins that only had a few peptide observations. In these cases regression lines will be fairly unreliable and using them to project relative protein abundance, for example when IS equals one, proves to often be detrimental. However, Figure 2D and E suggest that there may be a consistent relationship between our peptide level covariates and ratio compression that can be modelled across the whole dataset. We can see that the slope of the relationship should always be positive, it should be zero when the true change is zero, and should increase proportionally to the magnitude of the true change. These assumptions motivate the use of a single data-wide parameter that defines the relationship between a peptide level covariate and ratio compression. We incorporate this information into a non-linear Bayesian model for MS2 data defined below.

The unbalanced structure of the data also exacerbates the importance of error estimation. For proteins with only a single observed peptide, standard errors cannot be computed. With a small number of peptides, standard errors can be computed but are not robust to outliers. Both of these problems can be addressed by sharing information across proteins.

One solution to the error estimation problem would be to use a pooled variance estimate. In this case we would estimate one variance component that represented experimental error for every protein. This would immediately solve the problem of unreliable error estimates that will occur when no pooling is used. However, such an approach simply trades one problem for another. While error estimates for proteins with little information may benefit from a pooled variance component, this approach would overestimate the error for proteins with lots of data and little variation.

For this reason we make use of partially pooled variance components[5] so that proteins with few peptides rely almost entirely on the average experimental error, but as more peptides are observed, the variation converges to the within protein variance. This can be accomplished by creating a hierarchical model for the variance components.

The hierarchical model for variance components provides substantially improved variance estimation which was responsible for the improved signal detection shown throughout the paper. Beyond signal detection, an improved assessment of variability can protect against some well-known sources of danger in proteomics data. Post translational modifications and protein isoforms can result in peptides that are assigned to the same protein in the database, but which have very different ratios. There is nothing in our modeling that accounts for this situation. However, since we are improving the estimation of variance it may be possible to detect these scenarios by looking for proteins with many peptides but unusually high variance.

The Bayesian compositional models

The compositional MS2 model that utilizes partially pooled variance estimation and a single compression parameter is defined as follows.

$$y_{ijk} \mid \beta_{ij}, \delta, \sigma_{ij}^2 \sim N\big(\beta_{ij}\big(1 + SSN_{ijk}\delta\big), \sigma_{ij}^2\big)$$

$$\beta_{ij} \sim N(0, 10)$$

$$\delta \sim Uniform(0, \infty)$$

$$\sigma_{ij}^2 \mid \tau \sim InverseGamma(1, \tau)$$

$$\tau \sim HalfNormal(0, 5)$$

Where $i = 1, \dots, n_p$ indexes the n_p unique proteins. $j = 1, \dots, n_c - 1$ indexes the number of relative comparisons made, so that n_c is the number of conditions prior transforming into log ratios. $k = 1, \dots, m_i$ indexes the m_i peptides observed within protein i.

y_{ijk} is the additive log2 ratio of the kth peptide observed within protein i in condition j. β_{ij} represents the average log ratio when summed-signal-to-noise equals zero. The prior was selected with a mean of zero to reflect the experimental assumption that most proteins will not be changed across conditions. The standard deviation of 10 was selected to be a sufficiently large distribution (these are on the log2 scale!) to refrain from effecting our estimation while still being informative enough to avoid sampling problems caused by sampling from unrealistic parameter values. This is often referred to as a weakly informative prior.

SSN_{ijk} is the peptide level covariate describing the observed summed signal-to-noise across all channels. The analysis in our paper contains predictions of the outcome when SSN is at the 99th percentile of observed summed signal-to-noise values. In theory, we would be interested in the value at the maximum, but these datasets often contain extreme outliers in summed signal-to-noise values which motivated a projection to the 99th percentile. δ is a parameter used across the whole dataset to define the relationship between summed signal-to-noise and true fold change. Estimating a single δ parameter allows us to model compression even for proteins that have a small number of observed peptides. The uniform prior was selected since we did not know a priori what a reasonable range for this parameter might be, and after fitting the model the non-informative approach did not cause any problems with convergence.

Regarding the SSN adjustment in the MS2 model, it is worth noting that this approach is substantially different than previous efforts to reverse MS2 compression effects. Others have sought to reverse compression by adjusting the data with learned factors[6–8]. In general, these approaches eliminate compression while sacrificing precision by multiplying the data with a learned quantity. Our approach, while indirectly related to compression, does neither. Using a peptide-level covariate is a fundamentally different approach. It is more general, as peptide-level covariates do not necessarily relate to compression (any observed peptide level quantity that we expect to affect ratios could be added to our model). Even when the covariate used clearly relates to compression, as was the case with SSN, adjusting for the observed relationship only slightly mitigated the compression phenomenon.

But perhaps the most important difference is that using a covariate does not require altering the underlying data. Instead we redefine our target parameter based on the observed surrogate measure and estimate it accordingly. Consequently, a covariate adjustment does not cause the dramatic losses to precision that occur when multiplying data by a compression factor.

The compositional MS3 model is similar only we do not make an adjustment based on summed-signal-to-noise since MS3 experiments were designed to remove interference experimentally and we do not expect the same sort of relationships to hold. The use of SSN provided a noticeable improvement to MS2 estimation accuracy and finding similar quantities that might improve MS3 estimation is certainly a valuable goal of future research. However, appropriate statistics have not yet been studied and the model definition used for MS3 throughout this paper is as follows:

$$y_{ijk} \mid \beta_{ij}, \sigma_{ij}^2 \sim N\left(\beta_{ij}, \sigma_{ij}^2\right)$$

$$\beta_{ij} \sim N(0, 10)$$

$$\sigma_{ij}^2 \mid \tau \sim \ InverseGamma(1, \tau)$$

$$\tau \sim HalfNormal(0, 5)$$

Where β_{ij} now represents the expected log ratio for the i'th protein between the j'th condition and the reference condition.

In both models partially pooled variance estimation is accomplished by sampling variance components, σ_{ij}^2, from a distribution of variances where the shape of the distribution is determined by the hierarchical parameter τ. Once again the hyperparameters were selected to be weakly informative. However, the choice of the half-normal and inverse gamma distributions was made for convergence considerations which are detailed below. It should be noted that this is likely the modeling decision most responsible for the dramatic improvements seen in signal detection. Notice that the accuracy between compositional MS3 and old MS3 models remains unchanged, while the ROC plots show a dramatic jump in performance. With the point estimates unchanged, it stands to reason that the driver of the improvement is the error estimation.

This model provides inference regarding what was seen in the samples being studied. Conclusions drawn from this model should be confined to statements about what was in the samples studied and would not be appropriate for population level inference. While the model can easily be adapted for population level studies, we have not done so here, as the motivating examples were not designed for population level inference.

Bayesian models provide great flexibility in how we predict signal detection as we can directly compute the posterior probability that a parameter lies in a specified interval. For example, a researcher could compute the probability that a protein fold-change was greater than 2. However, in our dilution experiments we do not wish to give our methodology an unfair advantage by using the true ratios as part of our decision rule. Consequently, we instead ranked the proteins by the posterior mean divided by the posterior coefficient of variation. This statistic ranks proteins by the magnitude of the posterior mean but diminishes or strengthens the evidence based on the CV. This was used as the predictive statistic in ROC plots throughout the paper.

Credible intervals are a feature of Bayesian modelling. They can be thought of as analogous to confidence intervals, however they differ in a few important aspects. In a Bayesian model we have the ability to directly asses the probability, conditional on the observed data, that a model parameter (protein fold-change) lies in a particular interval. Confidence intervals lack this clear probabilistic interpretation. This is advantageous as we can ask questions with a Bayesian model that are simply not possible in a frequentist framework. For example, if we wanted to know the probability that a protein fold-change was very small (between -.1 and .1) we could do so directly. This is profoundly different than looking for large p-values which in no way suggests evidence in favor of a null hypothesis.

The compositional transformation, peptide level covariate adjustments, and partially pooled variance estimation are all implemented in our publicly available R package which contains pre-compiled Stan models to make use of efficient Bayesian simulation algorithms[9].

Assessing Model Convergence

Bayesian modeling typically depends on simulations to characterize posterior distributions. The Stan programming language makes use of Hamiltonian Monte Carlo techniques to efficiently explore posterior distributions. However, whenever estimation is performed with non-deterministic methods concerns about model convergence need to be considered. This concern is especially relevant for models as large and complex as the ones proposed in this paper.

Distributional assumptions and parametrizations all play a role in obtaining convergence. In particular the hierarchical variance components require the non-centered reparameterization described in the Stan case studies (http://mc-stan.org/users/documentation/case-studies/divergences_and_bias.html). Furthermore, we found that many distributional assumptions resulted in computational difficulties. Consequently, the distributions described in this paper, and implemented in our package were often selected out of computational necessity.

Stan offers a number of tools to assess model convergence. By default, in Stan and our compMS package, every sampling chain is independently run four times from varying sets of initial conditions. This means that we can observe trace plots of each parameter to visually inspect whether the parameters converged, and whether not each different starting points still converged to the same place. As an example, a trace plot for δ from the MS2 model is shown in Figure S1, A.

Our proteomics models contain thousands of parameters and examining every trace plot is not realistic. Fortunately, there are some options for systematically assessing convergence properties. One way is to compare the variance between and within each of the independent chains. This is done with the $\hat{R}$ statistic[10] which should be close to 1 if the model has converged. A histogram of all the $\hat{R}$ statistics can be generated by calling the stan_rhat() function on any stan model fit object (Figure S1, B). For reference, Stan best practices suggest that Rhat should be less than 1.1 (https://github.com/stan-dev/stan/wiki/Stan-Best-Practices).

Another way to assess overall performance is to inspect the distributions of the parameters of interest. The compMS function caterpillar() creates a plot of the credible intervals of protein fold changes (Figure S1, C). If the model did not converge at all we might expect to see unusually large distributions for every parameter. Notice that the variability of protein estimates truly spans an order of magnitude when the parameters have converged. Consequently, if all of the credible intervals are similar and large, it would be highly suggestive that the model did not converge.

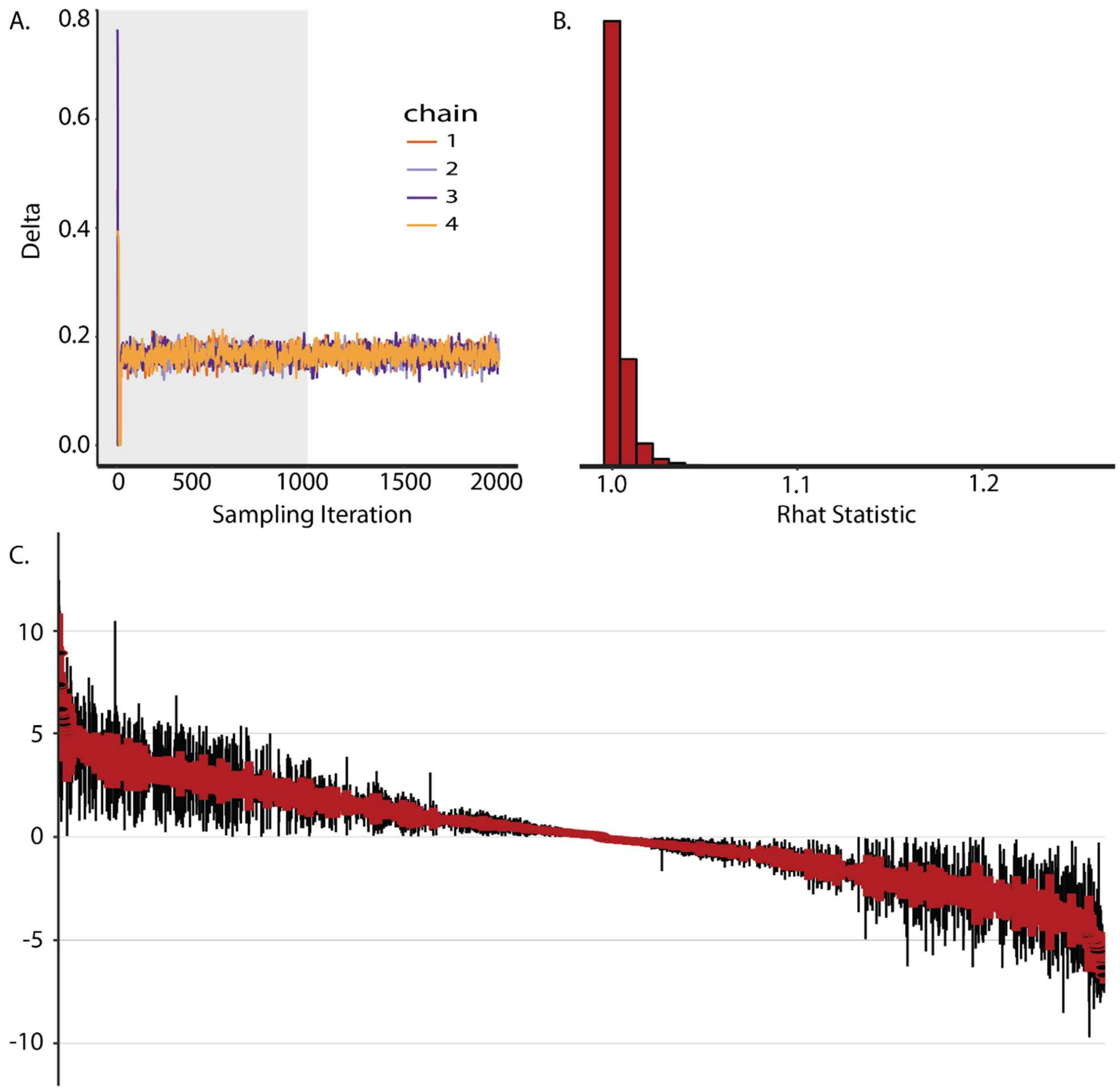

Figure S1. Assessing the convergence of a Stan model. A) A trace plot for the parameter δ from the MS2 model in the Boundary Case experiment. This plot shows connects the points from each sampling iteration for the parameter. Each of the independent chains is plotted in a different color. The first thousand iterations are shown with a grey background because they were defined as part of the warmup period and are not used for inference. **B)** Histogram of the Rhat statistic. This shows the relative frequency of Rhat for all parameters in the MS2 Boundary Case experiment. **C)** Caterpillar plot for log2 protein fold-changes in the MS2 Boundary Case experiment. This plot is rank ordered by the posterior mean. Red intervals represent 80% credible intervals and the black tails represent 95% credible intervals. Only 95% credible intervals that do not contain zero are shown here.

Boundary Case Experiment

Mouse whole brain tissue lysate and yeast whole cell lysate was prepared as described previously [11,12]. Protein digestion and TMT labeling was performed as described previously[13]. After labeling, we diluted the yeast samples at ratios of 1,1, 1.25, 1.5, 1.75, 2, 4, 8, 16 and 32. The dilutions were done in both directions (diluting from the 32, and diluting from the 1). We then repeated this setup with ratios of 1, 1, 2, 4, 0, 0, 0, 0, 32, and 100. The mouse peptide background was constant across channels, so that it is 1:1 at the highest level of yeast peptide. Approximately 1 μg of unfractionated sample was analyzed by mass spectrometry.

Samples were analyzed on an Orbitrap Fusion Lumos mass spectrometer (Thermo Fisher Scientific, San Jose, CA) coupled to a Proxeon EASY-nLC 1200 liquid chromatography (LC) pump (Thermo Fisher Scientific). Peptides were separated on a 100 μm inner diameter microcapillary column packed with 35 cm of Accucore C18 resin (2.6 μm, 150 Å, ThermoFisher). Peptides were separated using 60 min gradient of 3 to 25% acetonitrile in 0.125% formic acid with a flow rate of 450 nL/min.

Each analysis used an MS3-based TMT method[3,14], which has been shown to reduce ion interference compared to MS2 quantification[15]. The scan sequence began with an MS1 spectrum (Orbitrap analysis, resolution 120,000, 400−1400 Th, automatic gain control (AGC) target 2E5, maximum injection time 100 ms). The top ten precursors were then selected for MS2/MS3 analysis. MS2 analysis consisted of: collision-induced dissociation (CID), quadrupole ion trap analysis, automatic gain control (AGC) 4E4, NCE (normalized collision energy) 35, q-value 0.25, maximum injection time 150 ms), and isolation window at 0.7. Following acquisition of each MS2 spectrum, we collected an MS3 spectrum in which multiple MS2 fragment ions are captured in the MS3 precursor population using isolation waveforms with multiple frequency notches [22]. MS3 precursors were fragmented by HCD and analyzed using the Orbitrap (NCE 55, AGC 2.5E5, maximum injection time 150 ms, resolution was 50,000 at 400 Th). For MS3 analysis, we set the isolation window to 1.2 Th.

For MS2-only analysis, scan sequence began with an MS1 spectrum (Orbitrap analysis, resolution 120,000, 400−1400 Th, automatic gain control (AGC) target 1E6, maximum injection time 100 ms). The top ten precursors were then selected for MS2 analysis. MS2 analysis consisted of: high energy collision-induced dissociation (HCD), quadrupole ion trap analysis, automatic gain control (AGC) 1E5, NCE (normalized collision energy) 37, maximum injection time 150 ms), and isolation window at 0.7.

Mass spectra were processed using a Sequest-based in-house software pipeline[16]. Database searching included all entries from the species-appropriate database (mouse database downloaded from UniProt on July 2, 2014 and yeast database downloaded from The Saccharomyces Genome Database (SGD) on March 24, 2014. The database was concatenated with one composed of all protein sequences in the reversed order. Searches were performed using a 50 ppm precursor ion tolerance for total protein level analysis. The product ion tolerance was set to 0.9 Da. These wide mass tolerance windows were chosen to maximize sensitivity in conjunction with Sequest searches and linear discriminant analysis[16,17]. TMT tags on lysine residues and peptide N termini (+229.163 Da) and carbamidomethylation of cysteine residues (+57.021 Da) were set as static modifications, while oxidation of methionine residues (+15.995 Da) was set as a variable modification.

Peptide-spectrum matches (PSMs) were adjusted to a 1% false discovery rate (FDR)[18,19]. PSM filtering was performed using a linear discriminant analysis, as described previously[16], while considering the following parameters: XCorr, ΔCn, missed cleavages, peptide length, charge state, and precursor mass accuracy. For TMT-based reporter ion quantitation, we extracted the signal-to-noise (S:N) ratio for each TMT channel and found the closest matching centroid to the expected mass of the TMT reporter ion. For protein-level comparisons, PSMs were identified, quantified, and collapsed to a 1% peptide false discovery rate (FDR) and then collapsed further to a final protein-level FDR of 1%. Moreover, protein assembly was guided by principles of parsimony to produce the smallest set of proteins necessary to account for all observed peptides.

Channels were randomly permuted within each protein to expand the full range of possible outcomes then peptide level two-way ANOVA's similar to those used by Oberg et. al.[20] were used to estimate log2 protein level fold-changes for both the MS2 and MS3 data. Note that the more common approach of discarding peptide level variation and performing a t-test on protein level estimates is not an option here as the experiment had no replicates. For each protein, the following model was fit in R with the lm function.

$$y_{jk} = \mu + \beta_j + \alpha_k + \epsilon_{jk}$$

Where y_{jk} is the log2 signal-to-noise ratio where $j = 1, \ldots, n_c$ indexes the number of conditions. $k = 1, \ldots, K$ indexes the K peptides observed within the protein. Reference cell coding is used so that $\beta_1 = 0, \alpha_k = 0.$ μ represents the expected value of peptide 1 in condition 1. β_j, for $j \neq 1$, represents the expected log2 contrast for the protein between condition j and condition 1. α_k, for $k \neq 1$, represents the average effect difference between peptide k and peptide 1.

For proteins with only a single observed peptide, the model is reduced to a one-way anova without a peptide effect.

P-values for the hypothesis that each fold change is zero, were taken directly from the model objects. The purpose of our hypothesis testing was to use the p-values to generate ROC plots. Since these plots depend only on the rank order of the p-values, and multiple hypothesis test corrections do not change the rank ordering, we did not perform any corrections.

ROC plots were generated with the ROCR package in R[21]. True positives are defined as yeast proteins that are known to change by given amount, and false positives are the yeast proteins known to not change.

Both of the compositional models were coded in the Stan programming language and are provided in our R package. To simplify the computational complexity each dataset analyzed was split up into randomly selected sets of 1000 proteins. This provides a substantial reduction in processing time.

Note that the point estimates from this model are very similar to the difference in the average log2 intensities between the conditions. This is true for any model that shares the mean structure of the described ANOVA, including the mixed model, and the GLM discussed in a recent review paper[22]. This also holds for the compositional model when no covariate adjustments are made, which explains why the MS3 accuracy was the same between the ANOVA model and the Compositional model in the Boundary Case Experiment. The Compositional MS3 Model gave virtually identical point estimates to the ANOVA model but provided improved error estimation which led to a dramatic improvement in signal detection.

Raw files have been uploaded to ProteomeXchange (Accession Number – PXD008259).

Common Case Experiment

Yeast cells were diluted to ratios of 1:2:3, with human cell lines added to compensate for lost material.

Cell lysis and protein digestion. Yeast cultures were harvested by centrifugation, and resuspended in lysis buffer - 50 mM HEPES pH 8.5, 8 M urea, 75 mM NaCl, protease (complete mini, EDTA-free) inhibitors (Roche, Basel, Switzerland) at 4°C. Yeast cells were lysed using the MiniBeadbeater (Biospec, Bartlesville, OK) in microcentrifuge tubes with 1ml zirconium oxide beads at maximum speed for five cycles of 30 sec each, with 1 min pauses between cycles to minimize heating the lysates. Human cells (SHSY-5Y cell line) were harvested, resuspended in lysis buffer, and lysed by 20 pumps through a a 21-gauge needle. After centrifugation, lysates were transferred to new tubes, spun to pellet cell debris, and the supernatants saved. We determined the protein concentration in the lysate using the bicinchoninic acid (BCA) protein assay (Thermo Fisher Scientific, Waltham, MA).

Proteins disulfide bonds were reduced with 5 mM tris (2-carboxyethyl) phosphine (TCEP), (room temperature, 25 min) and alkylated with 10 mM iodoacetamide (room temperature, 30 min in the dark). Excess iodoacetamide was quenched with 15 mM dithiotreitol (room temperature, 15 min in the dark). Methanol-chloroform precipitation was performed prior to protease digestion. In brief, four parts neat methanol was added to each sample and vortexed, one part chloroform was added to the sample and vortexed, and three parts water was added to the sample and vortexed. The sample was centrifuged at 4000 RPM for 15 min at room temperature and subsequently washed twice with 100% methanol, prior to air-drying. Samples were resuspended in 50 mM HEPES pH 8.5 and digested at room temperature for 12 hrs with LysC protease at a 100:1 protein-to-protease ratio. Trypsin was then added at a 100:1 protein-to-protease ratio and the reaction was incubated 6 hrs at 37°C. Peptide concentrations in the digests were measured using the Quantitative Colorometric Peptide assay kit (Pierce). Peptides from a Lyc/trypsin digest from human or yeast lysates were mixed to create the ratios shown in Figure 1 in at least triplicate (n=3,4,4)

TMT pipeline. The eleven TMT-labeled samples were mixed into a single sample and separated by basic pH RP HPLC We used an Agilent 1100 pump equipped with a degasser and a photodiode array (PDA) detector (set at 220 and 280 nm wavelength) from Thermo Fisher Scientific (Waltham, MA). Peptides were subjected to a 50 min linear gradient from 5% to 35% acetonitrile in 10mM ammonium bicarbonate pH 8 at a flow rate of 0.6 mL/min over an Agilent 300Extend C18 column (3.5 μm particles, 4.6 mm ID and 220 mm in length). The peptide mixture was fractionated into a total of 96 fractions which were consolidated into 24 fractions. Samples were subsequently acidified with 1% formic acid and vacuum centrifuged to near dryness. Each eluted fraction was desalted via StageTip , dried via vacuum centrifugation, and reconstituted in 5% acetonitrile, 5% formic acid for LC-MS/MS processing.

For MS3 analysis, eleven pooled fractions were analyzed using 3-hr gradient separations using the instrument parameters described above for the boundary case experiment, injecting roughly 1 ug per fraction for 11 pooled fractions on a Samples were analyzed on an Orbitrap Fusion Lumos mass spectrometer (Thermo Fisher Scientific, San Jose, CA) coupled to a Proxeon EASY-nLC 1200 liquid chromatography (LC) pump (Thermo Fisher Scientific).

For the MS2 analysis, the same 11 samples were analyzed on the same Orbitrap Fusion Lumos mass spectrometer (Thermo Fisher Scientific, San Jose, CA) coupled to a Proxeon EASY-nLC 1200 liquid chromatography (LC) pump (Thermo Fisher Scientific). Instrument parameters were preserved for MS1 analysis, but for MS2 precursors were fragmented by HCD and analyzed using the Orbitrap (NCE 55, AGC 2.5E5, maximum injection time 150 ms, resolution was 50,000 at 400 Th). For MS2 analysis, we set the isolation window to 1.4 Th.

The data generated in this experiment is offered available with the manuscript titled "Proteome-Wide Evaluation of Two Common Protein Quantification Methods" which is currently under review.

Data analysis. Samples were searched with the Sequest algorithm against a combined yeast (downloaded from SGD on March 24, 2014) and human database (downloaded from UniProt on February 4, 2014) which was concatenated with their reversed sequences as decoys for FDR determination. Results were filtered to a 1% FDR at the peptide and protein levels using linear discriminant analysis and the target-decoy strategy. MS3 spectra were processed as signal-to-noise ratios for each reporter ion based on noise levels within a 25 Th window. Proteins were quantified by summing reporter ion intensities across all matching PSMs using in-house software, as described previously. PSMs with low isolation specificity (<0.7), MS3 spectra with more than eight TMT reporter ion channels missing, MS3 spectra with TMT reporter summed signal to noise ratio that is less than 200, or no MS3 spectra were excluded from quantitation. Equal human protein starting amounts was enforced by normalizing to the sum of all human peptides for each of the 11 channels. The normalization is slightly different for the compositional modelling. Instead using a multiplicative factor to equalize protein level intensities, additive factors are used to ensure that the average PSM log ratio to the reference channel are equivalent for all channels.

In this experiment with multiple replicates, t-tests were possible so we chose to analyze the data in accordance with the most commonly used method. Protein estimates were obtained from the standard in house software and t-tests were performed for hypothesis testing.

The compositional modelling was identical to the procedures described for the Boundary Case Experiment.

Once again, ROC plots were generated with ROCR[21]. True positives are defined as the yeast proteins that are known to change by a certain amount (2- or 3-fold). False positives are defined by the unchanging human proteins.

Dual multiplexed viral infection experiment with infinite changes

Samples analyzed for the Human Cytomegalovirus Time Course Experiment were prepared as described previously[23] . Two TMT experiments were designed as follows: 1) a TMT10-plex in the following order from 126-131: mock infection 1, mock infection 2, 6h post-infection, 12 h post-infection, 12 h post-infection (irradiated), 18h post-infection, 24h post-infection, 48h post-infection, 72h post-infection, and 96h post-infection and 2) a TMT2-plex with mock 1 (126) and 48h post-infection (130N). For both TMT experiments, the TMT-labeled peptides were pooled at a 1:1 ratio across all samples prior to off-line basic pH reversed-phase (BPRP) fractionation. The combined sample was vacuum centrifuged to near dryness and subjected to C18 solid-phase extraction (SPE) via Sep-Pak

(Waters, Milford, MA). We fractionated the pooled TMT-labeled peptide sample using the Pierce High pH Reversed-Phase Peptide Fractionation Kit (cat. # 84868). Eight fractions were collected using: 12.5%, 15%, 17.5%, 20%, 22.5%, 25% and 50% acetonitrile. Samples were subsequently acidified with 1% formic acid and vacuum centrifuged to near dryness. Each fraction was desalted via StageTip[24], dried again via vacuum centrifugation, and reconstituted in 5% acetonitrile, 5% formic acid for LC-MS/MS processing. Samples were analyzed on an Orbitrap Fusion Lumos mass spectrometer (Thermo Fisher Scientific, San Jose, CA) coupled to a Proxeon EASY-nLC 1200 liquid chromatography (LC) pump (Thermo Fisher Scientific) in a manner like that described above. However, here peptides were separated over a 150 min gradient. Data was analyzed with the same two-way ANOVA and compositional MS3 models described above.

Raw files have been uploaded to ProteomeXchange (Accession Number – PXD008259).

Notes on Software

An R package, compMS, that fits pre-compiled Stan models for compositional proteomics can be installed from www.github.com/ColtoCaro/compMS

The RStan package should be installed prior to installation of compMS. Special instructions for windows users are available at https://github.com/stan-dev/rstan/wiki/Installing-RStan-on-Windows

The compMS package is designed to make fairly complicated Bayesian statistical models accessible to non-experts. Accordingly, most decisions about the modeling are made by altering a header file of a spreadsheet. For details on the features of the package and how to use them, please refer to the readme file on the github page as this will be updated along with new versions of the package.

ROC Plots from the Boundary Case Experiment

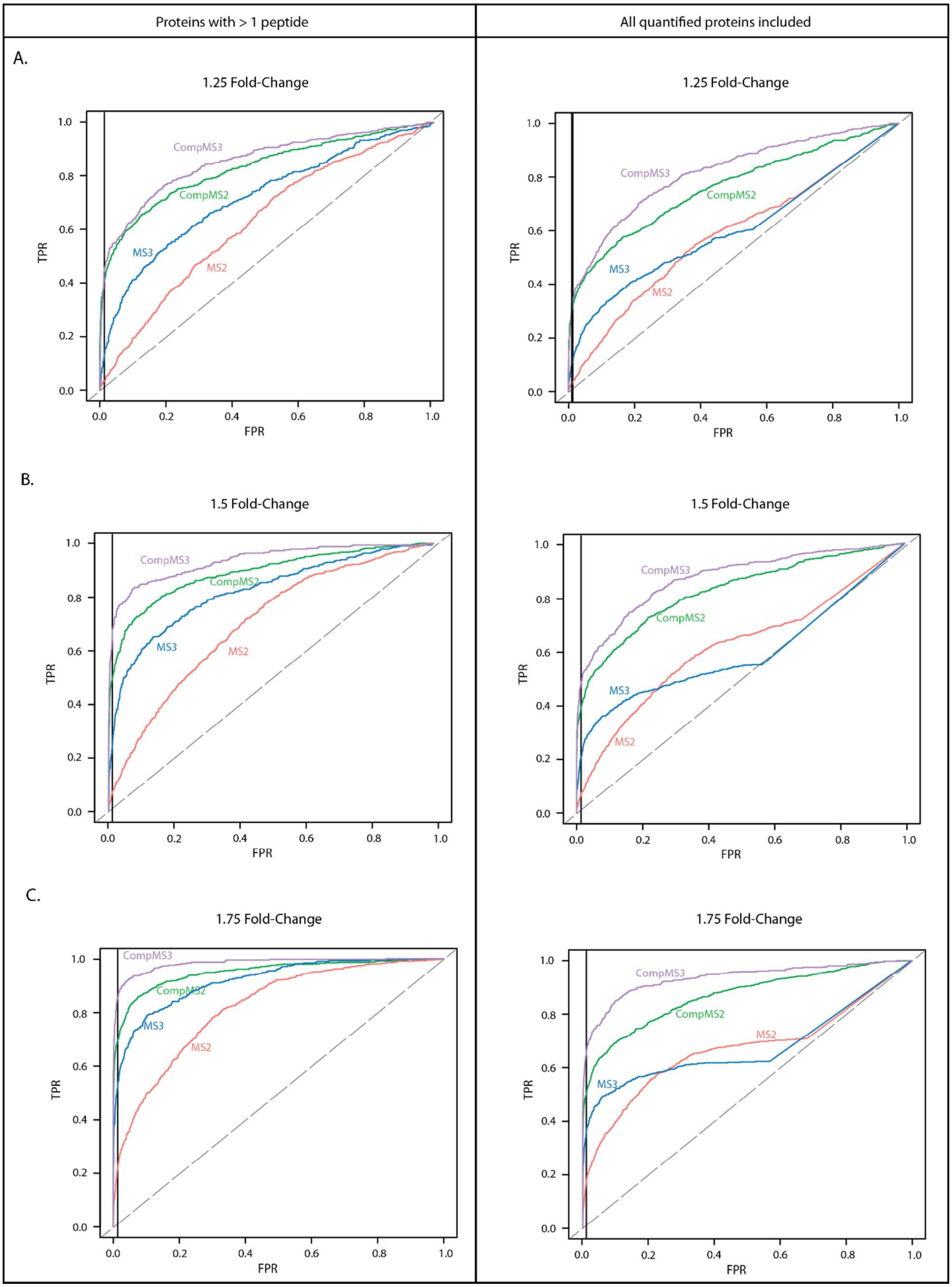

Figure S2. MS3 methods provide enhanced signal detection for small changes

ROC plots from the Boundary Case Experiment for small fold-changes of 1.25 (A), 1.5 (B) and 1.75 (C). The plots on the left do not include proteins for which only a single peptide was quantified, since the ANOVA methodology used cannot provide p-values in these cases. On the right we include all proteins quantified by assigning a value of 1 to all of the missing p-values. These plots show that in this range of changes, MS3 signals are easier to detect than MS2. We also see that methodology plays a very large role in our ability to pick out these small signals, with the compositional modelling providing substantial gains in all cases. Finally, it is worth noting that including proteins with only a single measurement hurts the performance of MS3 methods more than MS2. This is because the MS3 data contains, as a percentage, far more proteins with only one observation.

ROC Plots from the Boundary Case Experiment

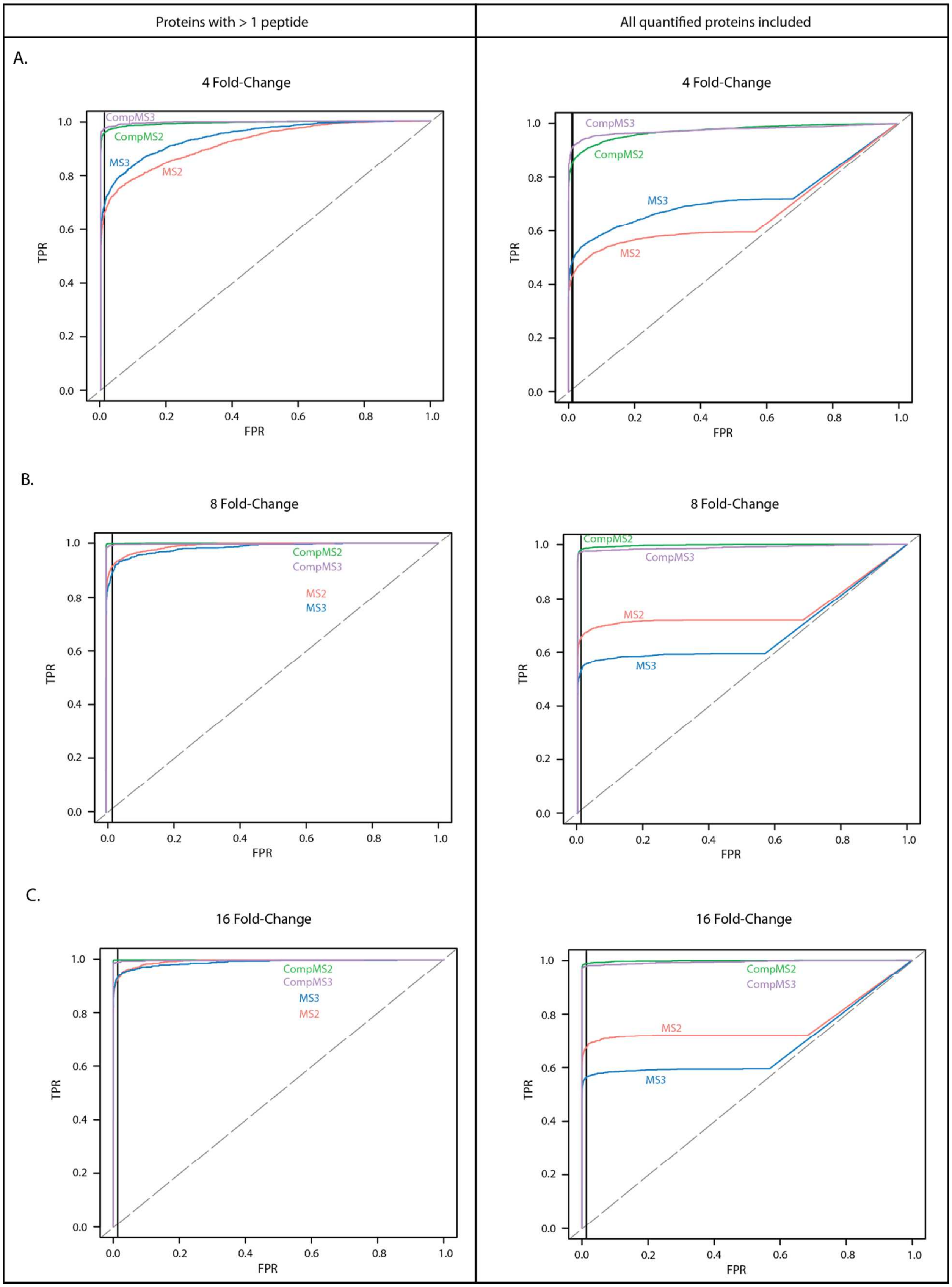

Figure S3. Methodological advantages diminish as the true fold-changes increase

ROC plots from the Boundary Case Experiment for fold-changes of 4 (A), 8 (B) and 16 (C). The plots on the left do not include proteins for which only a single peptide was quantified, since the ANOVA methodology used cannot provide p-values in these cases. On the right we include all proteins quantified by assigning a value of 1 to all of the missing p-values. These plots show that in this range of changes, the gains of compositional modelling are still substantial. However, the improved signal detection of MS3 over MS2 begins to disappear. In fact, when including proteins with only a single observation, the MS2 methods begin to outperform MS3 for fold-changes of 8 and 16. MS2 methods provide more data for each protein but with less accuracy. Presumably, the shift in performance occurs because large enough changes do not require great accuracy to reject a hypothesis test that the change was zero.

ROC Plots from the Boundary Case Experiment

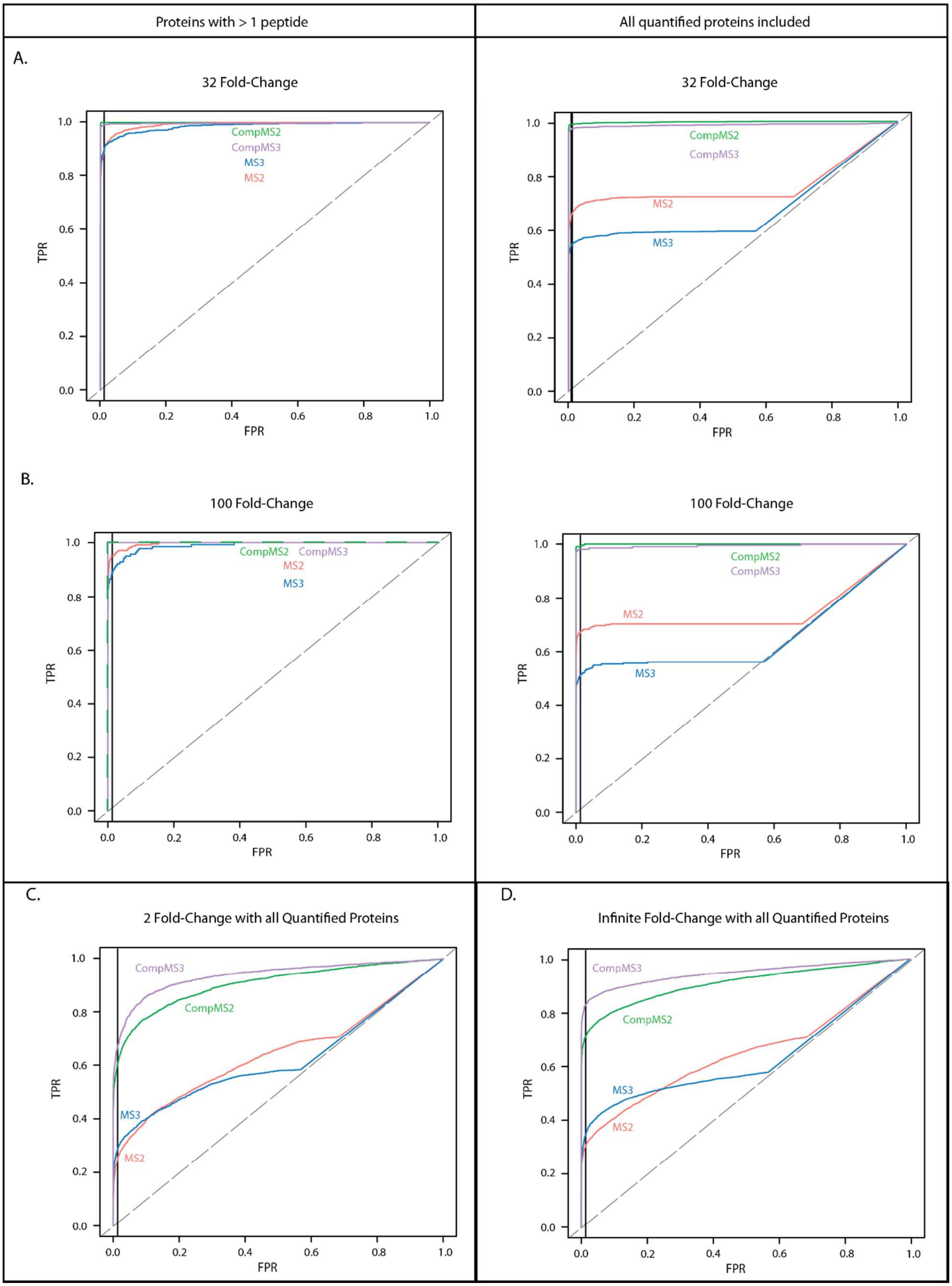

Figure S4. Methodological advantages diminish for large, but not infinite, fold changes

A-B) ROC plots from the Boundary Case Experiment for fold-changes of 32 (A) and 100 (B). The plots on the left do not include proteins for which only a single peptide was quantified, since the ANOVA methodology used cannot provide p-values in these cases. On the right we include all proteins quantified by assigning a value of 1 to all of the missing p-values. These plots show that in this range of changes, there are still some gains due to compositional modelling however and MS2 continues to outperform MS3 however, all of the advantages continue to shrink as all methods do a good job of detecting large changes. That said, many of the single observation proteins in this category are real changes and the ANOVA methodology cannot detect them, providing a substantial hit, especially for MS3.

C) ROC plot for 2-fold changes from the Boundary Case Experiment. This is the plot corresponding to one presented in Figure 2. The difference is that here we include all quantified proteins, even those with only one observation. The consequence is that the methodological gains from compositional modelling grow more substantial.

D) ROC plot for Infinite fold-changes from the Boundary Case Experiment. This is the plot corresponding to one presented in Figure 3A. The difference is that here we include all quantified proteins, even those with only one observation. The gains from compositional modelling are substantial in this category. The reason for this is that many of the infinite changes appear relatively small. Consequently, the detection performance parallels the patterns seen with smaller changes. Modelling provides enormous gains, and the improved accuracy for MS3 once again takes on more importance than the added number of observations from MS2 (of course we are not here considering the total number of observations, only the sensitivity and specificity of detecting a single signal).

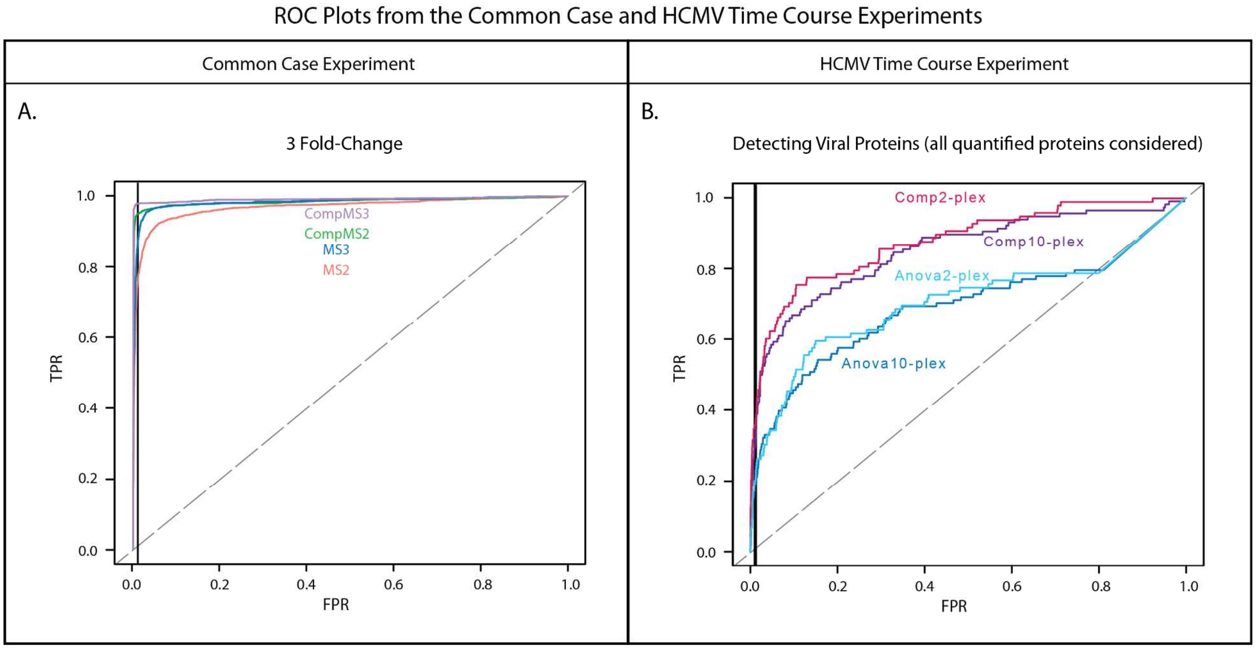

Figure S5. Further analyses confirm performance patterns

A) ROC plot from the Common Case Experiment for 3-fold changes. This plot reinforces the rank order of performance seen for 2-fold changes. The lesson regarding absolute numbers of true positives also remains similar. At a one-percent false positive rate MS2 and MS3 technologies with t-tests gave us 1412 and 1155 true 3-fold changes. Using compositional modelling these numbers increased to 1779 and 1306 respectively.

B) ROC plot showing the ability to detect infinite viral protein fold-changes from a background of mostly unchanged human proteins. This is the plot corresponding to one presented in Figure 3E. The difference is that here we include all quantified proteins, even those with only one observation. The consequence is that the methodological gains from compositional modelling grow more substantial. As expected, the ROC curves for compositional modeling in both the 2- and 10-plex have dropped (categorizing single observations is a more difficult problem). However, many of these prove to be true positives resulting in a greater divergence between methodologies when considering all observed proteins.

References

(1) Aitchison, J. The Statistical Analysis of Compositional Data. *J. R. Stat. Soc. Ser. B. Methodol.* **1982**, *44* (2), 139–177.

(2) Egozcue, J. J.; Pawlowsky-Glahn, V.; Mateu-Figueras, G.; Barceló-Vidal, C. Isometric Logratio Transformations for Compositional Data Analysis. *Math. Geol.* **2003**, *35* (3), 279–300.

(3) Ting, L.; Rad, R.; Gygi, S. P.; Haas, W. MS3 Eliminates Ratio Distortion in Isobaric Multiplexed Quantitative Proteomics. TL - 8. *Nat. Methods* **2011**, *8 VN-re* (11), 937–940.

(4) Hill, E. G.; Schwacke, J. H.; Comte-Walters, S.; Slate, E. H.; Oberg, A. L.; Eckel-Passow, J. E.; Therneau, T. M.; Schey, K. L. A Statistical Model for iTRAQ Data Analysis. *J. Proteome Res.* **2008**, *7* (8), 3091–3101.

(5) Gelman, A.; Hill, J. *Data Analysis Using Regression and Multilevel/Hierarchical Models*; Cambridge University Press, 2006; pp 251-276.

(6) Ahrné, E.; Glatter, T.; Viganò, C.; Von Schubert, C.; Nigg, E. A.; Schmidt, A. Evaluation and Improvement of Quantification Accuracy in Isobaric Mass Tag-Based Protein Quantification Experiments. *J. Proteome Res.* **2016**, *15* (8), 2537–2547.

(7) Savitski, M. M.; Mathieson, T.; Zinn, N.; Sweetman, G.; Doce, C.; Becher, I.; Pachl, F.; Kuster, B.; Bantscheff, M. Measuring and Managing Ratio Compression for Accurate iTRAQ/TMT Quantification. *J. Proteome Res.* **2013**, *12* (8), 3586–3598.

(8) Mertins, P.; Udeshi, N. D.; Clauser, K. R.; Mani, D.; Patel, J.; Ong, S. -e.; Jaffe, J. D.; Carr, S. A. iTRAQ Labeling Is Superior to mTRAQ for Quantitative Global Proteomics and Phosphoproteomics. *Mol. Cell. Proteomics* **2012**, *11* (6), M111.014423-M111.014423.

(9) Carpenter, B.; Gelman, A.; Hoffman, M. D.; Lee, D.; Goodrich, B.; Betancourt, M.; Brubaker, M.; Guo, J.; Li, P.; Riddell, A. Stan : A Probabilistic Programming Language. *J. Stat. Softw.* **2017**, *76* (1), 1–32.

(10) Gelman, A.; Carlin, J.; Stern, H.; Dunson, D.; Vehtari, A.; Rubin, D. *Bayesian Data Analysis*, 3rd ed.; Chapman & Hall/CRC, 2013; pp 281-286.

(11) Paulo, J. A.; O'Connell, J. D.; Everley, R. A.; O'Brien, J.; Gygi, M. A.; Gygi, S. P. Quantitative Mass Spectrometry-Based Multiplexing Compares the Abundance of 5000 S. Cerevisiae Proteins across 10 Carbon Sources. *J. Proteomics* **2016**, *148*, 85–93.

(12) Paulo, J. A.; McAllister, F. E.; Everley, R. A.; Beausoleil, S. A.; Banks, A. S.; Gygi, S. P. Effects of MEK Inhibitors GSK1120212 and PD0325901 in Vivo Using 10-Plex Quantitative Proteomics and Phosphoproteomics. *Proteomics* **2015**, *15* (2–3), 462–473.

(13) Paulo, J. A.; Gygi, S. P. Nicotine-Induced Protein Expression Profiling Reveals Mutually Altered Proteins across Four Human Cell Lines. *Proteomics* **2017**, *17* (1–2), 1600319.

(14) McAlister, G. C.; Nusinow, D. P.; Jedrychowski, M. P.; Wühr, M.; Huttlin, E. L.; Erickson, B. K.; Rad, R.; Haas, W.; Gygi, S. P. MultiNotch MS3 Enables Accurate, Sensitive, and Multiplexed Detection of Differential Expression across Cancer Cell Line Proteomes. *Anal. Chem.* **2014**, *86* (14), 7150–7158.

(15) Paulo, J. A.; O'Connell, J. D.; Gygi, S. P. A Triple Knockout (TKO) Proteomics Standard for

Diagnosing Ion Interference in Isobaric Labeling Experiments. *J. Am. Soc. Mass Spectrom.* **2016**, *27* (10), 1620–1625.

(16) Huttlin, E. L.; Jedrychowski, M. P.; Elias, J. E.; Goswami, T.; Rad, R.; Beausoleil, S. A.; Villén, J.; Haas, W.; Sowa, M. E.; Gygi, S. P. A Tissue-Specific Atlas of Mouse Protein Phosphorylation and Expression. *Cell* **2010**, *143* (7), 1174–1189.

(17) Beausoleil, S. A.; Villén, J.; Gerber, S. A.; Rush, J.; Gygi, S. P. A Probability-Based Approach for High-Throughput Protein Phosphorylation Analysis and Site Localization. *Nat. Biotechnol.* **2006**, *24* (10), 1285–1292.

(18) Elias, J. E.; Gygi, S. P. Target-Decoy Search Strategy for Increased Confidence in Large-Scale Protein Identifications by Mass Spectrometry. *Nat Methods* **2007**, *4* (3), 207–214.

(19) Elias, J. E.; Gygi, S. P. Target-Decoy Search Strategy for Mass Spectrometry-Based Proteomics. *Methods Mol. Biol.* **2010**, *604* (2), 55–71.

(20) Mahoney, D. W.; Therneau, T. M.; Heppelmann, C. J.; Higgins, L.; Benson, L. M.; Zenka, R. M.; Jagtap, P.; Nelsestuen, G. L.; Bergen, H. R.; Oberg, A. L. Relative Quantification: Characterisation of Bias, Variability and Fold Changes in Mass Spectrometry Data from iTRAQ-Labelled Peptides. *J. Proteome Res.* **2011**, *10*, 4325–4333.

(21) Sing, T.; Sander, O.; Beerenwinkel, N.; Lengauer, T. ROCR: Visualizing Classifier Performance in R. *Bioinformatics* **2005**, *21* (20), 3940–3941.

(22) D’Angelo, G.; Chaerkady, R.; Yu, W.; Hizal, D. B.; Hess, S.; Zhao, W.; Lekstrom, K.; Guo, X.; White, W. I.; Roskos, L.; et al. Statistical Models for the Analysis of Isobaric Tags Multiplexed Quantitative Proteomics. *J. Proteome Res.* **2017**, acs.jproteome.6b01050.

(23) Fielding, C. A.; Weekes, M. P.; Nobre, L. V; Ruckova, E.; Wilkie, G. S.; Paulo, J. A.; Chang, C.; Suarez, N. M.; Davies, J. A.; Antrobus, R.; et al. Control of Immune Ligands by Members of a Cytomegalovirus Gene Expansion Suppresses Natural Killer Cell Activation. *Elife* **2017**, *6* (e22206).

(24) Rappsilber, J.; Ishihama, Y.; Mann, M. Stop And Go Extraction Tips for Matrix-Assisted Laser Desorption/ionization, Nanoelectrospray, and LC/MS Sample Pretreatment in Proteomics. *Anal. Chem.* **2003**, *75* (3), 663–670.