HHS Public Access
Author manuscript
J Proteome Res. Author manuscript; available in PMC 2018 December 20.

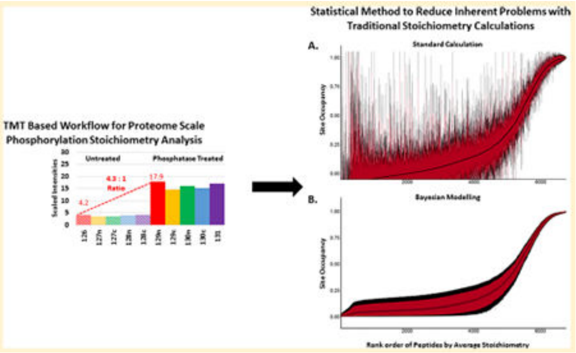
Published in final edited form as:
J Proteome Res. 2017 November 03; 16(11): 4217–4226. doi:10.1021/acs.jproteome.7b00571.

Improved Method for Determining Absolute Phosphorylation Stoichiometry Using Bayesian Statistics and Isobaric Labeling

Matthew Y. Lim[†], **Jonathon O'Brien**[†], **Joao A. Paulo**[†], and **Steven P. Gygi**[*,†]

[†]Department of Cell Biology, Harvard Medical School, Boston, Massachusetts 02115, United States

Abstract



Phosphorylation stoichiometry, or occupancy, is one element of phosphoproteomics that can add useful biological context (Gerber et al. *Proc. Natl. Acad. Sci. U. S. A.* 2003, *100*, 6940–5). We previously developed a method to assess phosphorylation stoichiometry on a proteome-wide scale (Wu et al. *Nat. Methods* 2011, 8, 677–83). The stoichiometry calculation relies on identifying and measuring the levels of each nonphosphorylated counterpart peptide with and without phosphatase treatment. The method, however, is problematic in that low stoichiometry phosphopeptides can return negative stoichiometry values if measurement error is larger than the percent stoichiometry. Here, we have improved the stoichiometry method through the use of isobaric labeling with 10-plex TMT reagents. In this way, five phosphatase treated and five untreated samples are compared simultaneously so that each stoichiometry is represented by five ratio measurements with no missing values. We applied the method to determine basal stoichiometries of HCT116 cells growing in culture. With this method, we analyzed five biological replicates simultaneously with no need for phosphopeptide enrichment. Additionally, we developed a Bayesian model to estimate phosphorylation stoichiometry as a parameter confined to an interval between 0 and 1

[*]**Corresponding Author**: steven_gygi@hms.harvard.edu.

Supporting Information
The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jproteome.7b00571.
Assessment of phosphorylation stoichiometry by motif peptide type; traceplot of precision parameter from Bayesian modeling; change in estimated stoichiometry when using Bayesian modeling method; error intervals constrained between 0 and 100% (PDF)
Table of all phosphopeptides localized to phosphorylation sites identified during phosphopeptide library experiment (XLSX)
Relative abundances and stoichiometries of peptides identified and quantified during phosphatase experiment (XLSX)
R Script used for executing Stan code (TXT)
Stan code used for Bayesian modelling (TXT)

The authors declare no competing financial interest.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

implemented as an R/Stan script. Consequently, both point and interval estimates are consistent with the plausible range of values for stoichiometry. Finally, we report absolute stoichiometry measurements with credible intervals for 6772 phosphopeptides containing at least a single phosphorylation site.

Keywords

TMT; SPS-MS3; phosphorylation; stoichiometry; phosphatase; Bayesian modeling; human cell lines; mass spectrometry; global proteome; error intervals

INTRODUCTION

Phosphorylation is one of the most common post-translation modifications found in cells. By chemically attaching a phosphate group to amino acid residues such as serine, threonine, and tyrosine, cells can change a protein's function, localization, or degradation in addition to other important cellular activities including signal transduction.[3,4] Because of its many cellular functions, a variety of experiments have been designed to probe different elements of phosphorylation. Quantitative experiments such as Western blotting and phosphopeptide mass spectrometry analysis are often implemented to measure phosphorylation dynamics.[3] While generally useful, these methods are often limited to identifying fold changes which may not provide sufficient information to fully understand the underlying biological mechanism.

One facet of quantitative phosphorylation proteomics that can have potential biological insight is phosphorylation stoichiometry, or occupancy.[1] A measured fold change of 2 for a phosphopeptide's levels can be caused a by a multitude of different cellular processes: a doubling in a protein production, a doubling in a phosphorylation occupancy, a decrease in protein degradation of the nonphosphorylated version, or any number of other cellular events. Additionally, a two-fold increase in relative phosphorylation levels can mean anything from an increase of 2% to 4% overall occupancy to an increase of 50% to 100% occupancy. Such stark differences in the absolute amount of phosphorylation occupancy could suggest that different cellular processes are activated in response to stimuli at different phosphorylation stoichiometries.[2,5–8]

Traditionally, phosphorylation stoichiometry has been measured using low throughput methods such as quantitative Western blotting. In 2001, we used AQUA peptides as absolute internal standards to measure the absolute amounts of both the phosphorylated and nonphosphorylated forms of site Ser-1126 on the protein separase.[1] We showed that this site was held at very high stoichiometry until the anaphase-metaphase transition, where-upon it became dephosphorylated, releasing its protease activity to finish mitosis. In recent years, we and others have developed high-throughput whole proteome techniques to assess phosphorylation stoichiometry *en masse*.[2–5,9] All current global proteome methods suffer from the same unavoidable drawback–their inability to distinguish phosphorylation stoichiometry of individual sites in multiply phosphorylated peptides. As such, it can only be claimed that for a multiply phosphorylated peptide, this is the maximum possible stoichiometry considering all sites. One such method utilizes stable isotope labeling with

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

amino acids in cell culture (SILAC) to measure three distinct ratios to generate a phosphorylation stoichiometry measurement.[5] However, SILAC can only be utilized where heavy amino acids can be doped into cell culture. Additionally, the SILAC method for assessing phosphorylation stoichiometry can only detect stoichiometry for sites that undergo a change in stoichiometry based on two conditions.[5]

We have previously published a method utilizing phosphatase treatment to assess phosphorylation stoichiometry.[2] A sample is divided into two aliquots, chemically labeled with a unique label, and one is treated with phosphatase. Phosphorylation stoichiometry can be assessed by analyzing the increase in signal of the nonphosphorylated form of phosphopeptides after phosphatase treatment.[2,8] Calculated stoichiometries are then assigned to phosphopeptides by matching the stoichiometries of these nonphosphorylated peptides to their known phosphorylated form from a phosphopeptide database or a previously generated phosphopeptide library. This indirect measurement circumvents issues of phosphorylation enrichment efficiencies as well as ionization efficiency for phosphorylated peptides, and potential digestion problems related to analyzing phosphorylated peptides.[2,6,8] Importantly, no comparison to a second condition is required allowing for the basal phosphorylation stoichiometry of a cell to be assessed. Others have adapted our method further for iTRAQ labeling or kinase treatment to improve this phosphatase-based method.[6,7]

This method can, however, report negative stoichiometries. For example, if the true occupancy level is 2% but the measurement error is 5%, it is possible to calculate negative values. To our knowledge, no group has successfully addressed the negative stoichiometries resulting from measurement error. Furthermore, previous attempts at analyzing phosphorylation stoichiometry relied on sample standard deviations to calculate confidence intervals for each stoichiometry measurement.[2,5–7] These intervals frequently include stoichiometry values below 0% or above 100%, which are not possible. Fortunately, these issues can be resolved by carefully defining a statistical model with appropriate distributions and ranges.

TMT reagents are a conduit for sample multiplexing in quantitative proteomics.[10–13] TMT chemically modifies the N-terminus and all free lysine residues of a peptide and is commercially available as a 2-, 6-, 8-, and 10-plex.[10,12,14] Each label is divided into two regions, a reporter ion region and a mass balance. All labels have the same nominal mass but differ in the placement of heavy ^{13}C and ^{15}N atoms, distributed between the reporter ion and mass balance regions.[10] TMT labeled peptides are, thus, indistinguishable during chromatographic separations and even via MS1 analysis.[10,12,14,15] However, during peptide fragmentation in a mass spectrometer, the balance remains attached to the peptide while the reporter region falls off as a reporter ion. Each label, or channel, has a unique reporter ion mass. Quantitation is performed by assessing the relative ratios of reporter ions.[10,12] A multinotch MS3 method can be used to collect accurate reporter ion ratios, greatly reducing or removing completely interference caused by coeluting and cofragmenting peptides.[15,16]

Here we have extended the TMT workflow to include stoichiometry analysis. We determined absolute stoichiometry from five biological replicates of asynchronously

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

growing HCT116 cells under basal conditions. We used statistical modeling to address negative stoichiometries in our data set. We treated stoichiometry as an estimable parameter rather than a directly calculated statistic. Finally, we provide occupancy measurements for 6772 unique phosphopeptides containing at least one phosphorylation site in HCT116 cells.

MATERIALS AND METHODS

Cell Culture

HCT116 cells were cultured in DMEM (Gibco) supplemented with 10% (v/v) fetal bovine serum (Hyclone) and 50 μL/mL penicillin and 50 μL/mL streptomycin (Gibco) in a 15 cm dish as described previously.[13,17] Cells were incubated at 37 °C at 5% CO_2 until approximately 80% confluent. Cells were then washed with ice cold phosphate buffered saline (Gibco) and lysed on plate with 1 mL of an 8 M urea lysis buffer containing a protease and phosphatase inhibitor cocktail (Roche). Lysate was collected and stored at −80 °C until sample preparation for mass spectrometry.

Sample Preparation

HCT116 lysate was homogenized by passing the lysate through a 21-gauge needle followed by sedimentation by centrifugation at 21 000*g* for 15 min.[13] The supernatant was transferred to a new tube, and protein concentration was determined by a bicinchoninic acid (BCA) assay (ThermoFisher Scientific). The proteins were then reduced and alkylated to block reactive cysteine groups and chloroform–methanol precipitated. Proteins were resuspended in 200 mM EPPS pH 8.5 and digested with Lys-C (Wako) overnight at room temperature and subsequently digested with sequencing grade trypsin (Promega) for 6 h at 37 °C. Digests were then desalted using C18 solid-phase extraction (SPE) (Sep-Pak, Waters) and dried down in a vacuum centrifuge.

Phosphatase Experiment To Generate Stoichiometry

We adapted our previous phosphatase method[2] to make use of TMT. Briefly, five dried down desalted digests were resuspended in 100 mM EPPS pH 8.5 and separated into two equivalent 50 μg aliquots. Each digest corresponded to a biological replicate. Each aliquot was labeled with a TMT10 reagent for 90 min at room temperature and then quenched with hydroxylamine. The quenched reaction was flash frozen and dried down in a vacuum centrifuge and then resuspended in CutSmart Buffer (New England Biolabs) and one labeled aliquot from each replicate was treated with 200 units of calf intestinal phosphatase (New England Biolabs) while the other aliquot from the replicate was treated with water. All aliquots were incubated at 37 °C for 3 h and then acidified with formic acid to a final concentration of 1%. All aliquots were then combined at a 1:1:1:1:1:1:1:1:1:1 ratio.[11] The pooled sample was then subjected to C18 SPE (Sep-Pak, Waters) and then dried down in a vacuum centrifuge before resuspension in 10 mM ammonium bicarbonate and 5% acetonitrile for off-line basic pH reversed-phase (BPRP) fractionation.

Phosphopeptide Enrichment Experiment

A separate phosphopeptide enrichment experiment was performed on HCT116 cell lysates to generate a phosphopeptide library as previously described.[18] Briefly, 10 mg of protein

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

from HCT116 lysates was digested and subjected to enrichment with immobilized metal affinity chromatography with Fe^{3+} (Fe-IMAC). The phosphopeptide enriched digest was then labeled with a TMT10 reagent as described above. The sample was then dried down in a vacuum centrifuge, resuspended in 1% formic acid, and subjected to C18 solid phase extraction (SPE) (Sep-Pak, Waters). The desalted phosphopeptide enrichment was dried down in a vacuum centrifuge before resuspension in 10 mM ammonium bicarbonate and 5% acetonitrile for off-line basic pH reversed-phase (BPRP) fractionation.

BPRP Fractionation

Off-line BPRP HPLC was performed on an Agilent 1100 pump with a degasser and photodiode array detector.[11] A gradient of 13%–37% acetonitrile in 10 mM ammonium bicarbonate was used over 50 min. The pooled TMT-labeled sample and the phosphopeptide enriched sample were each separated into 96 fractions by the instrument. For each fractionation experiment, fractions were collected in a 96-well plate and combined into 24 fractions as previously described. The 24 fractions were acidified to 1% formic acid and dried down in a vacuum centrifuge. Dried down fractions were resuspended in 5% acetonitrile and 5% formic acid for LC–MS/MS analysis.

Liquid Chromatography and Tandem Mass Spectrometry (LC–MS/MS)

Data for all LC–MS/MS experiments were collected on an Orbitrap Fusion Lumos mass spectrometer (Thermo Fisher Scientific, San Jose, CA) with LC separation performed on an attached Proxeon EASY-nLC 1200 liquid chromatography (LC) pump (Thermo Fisher Scientific). LC–MS/MS method was modified from a previous study.[11] A 100 μm inner diameter microcapillary column packed with 35 cm of Accucore C18 resin (2.6 μm, 150 Å, ThermoFisher) was used to separate peptides. Approximately 2 μg of peptide were loaded onto the column for analysis.

A 150 min gradient of 6% to 25% acetonitrile in 0.125% formic acid was used at a flow rate of ~450 nL/min to separate peptides from the pooled TMT-labeled samples: MS1 spectra (Orbitrap resolution, 120 000; mass range, 350–1400 *m/z*; automatic gain control (AGC) target, 5×10^5; maximum injection time, 100 ms). We then used a Top10 method to select precursors for further downstream analysis. MS2 spectra were collected after collision-induced dissociation (CID) (AGC target, 2×10^4; normalized collision energy (NCE), 35%; maximum injection time, 120 ms; and isolation window, 0.7 Th). MS2 analysis was performed in the ion trap. We performed an MS3 analysis for each MS2 scan acquired by isolating multiple MS2 fragment ions that were used as precursors for the MS3 analysis with a multinotch isolation waveform. We detected the MS3 analysis in the Orbitrap (resolution 50 000) after high energy collision induced dissociation (HCD) (NCE, 65% with instrument parameters: AGC target, 2.5×10^5; maximum injection time, 150 ms; and isolation window, 1.3 Th).

For the phosphopeptide enriched sample, a high-resolution MS2 method was utilized for analysis as there was no quantitation to perform. Peptides were again separated by a 150 min gradient. MS1 spectra were obtained in the Orbitrap (resolution, 120 000; mass range, 350–1400 *m/z*; AGC target, 5×10^5; maximum injection time, 100 ms). We selected precursors

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

for MS2 analysis using a TopSpeed method of 3 s. MS2 analysis occurred in the Orbitrap as well (HCD fragmentation; NCE, 38%; AGC target, 1×10^5; maximum injection time, 150 ms; isolation window, 1.6 Th).

Data Analysis

Spectra acquired from LC–MS/MS experiments for the TMT-pooled phosphatase experiments were processed using a Sequest-based software pipeline.[11,19] First a modified version of ReAdW.exe converted spectra to the mzXML format. These files were then searched against a database which contained the human proteome (Uniprot Database ID: 9606, downloaded February 4, 2014) concatenated to a database of all protein sequences reversed.[20] A precursor ion tolerance of 50 ppm and a product ion tolerance of 0.9 Da were used as search parameters. Static modifications for TMT tags (+229.163 Da) on lysine residues and the peptide's N termini and carbamidomethylation (+57.021 Da) on cysteine residues were used in conjunction with a variable modification for oxidation (+15.995 Da) on methionine.

Peptide-spectrum matches (PSMs) were then filtered using linear discriminant analysis to a false discovery rate (FDR) of 1% as described previously.[21] XCorr, ΔCn, missed cleavages, peptide length, charge state, and precursor mass accuracy were used as parameters for the LDA. The false discovery rate was estimated by using the target-decoy method. Peptides were identified and collapsed using principles of parsimony to a final protein-level FDR of 1%.

For quantitation, we extracted the signal-to-noise (S:N) ratio of the closest matching centroid to the expected mass of the TMT reporter ion for each TMT channel from MS3 scans triggered by MS2 scans. MS3 spectra were filtered for a minimum TMT reporter ion sum S:N of 200 and an isolation specificity of at least 0.5.

Data from the phosphopeptide enrichment were processed similarly except an additional variable modification of phosphate (+79.966) on serine, threonine, and tyrosine residues was included as a Sequest search parameter. Additionally, because the analysis was a high-resolution MS2 scan, product ion tolerance was tightened to 0.03 Da. Site localization was performed using Ascore.[22] No quantitation was performed. The generated localized phosphopeptide list was filtered to remove any duplicate phosphopeptides to create a unique-matchable list.

Filtered PSMs from the phosphatase experiment were then matched to the unphosphorylated form of peptides from the unique-matchable phosphopeptide list. A TMT-based reporter ion quantitation method was then performed on these matched PSMs utilizing the S:N ratios for each reporter ion channel from the phosphatase experiment. To calculate stoichiometry we compared S:N ratios for reporter ion channels corresponding to the same biological replicate. This was done with three different computational approaches, which we will refer to as the standard stoichiometry, 0% lower limit, and Bayesian method. We defined the standard stoichiometry calculation as

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

$$\text{Stoichiometry\%} = \frac{\text{TMT S:N}_{\text{phosphatasetreated}} - \text{TMT S:N}_{\text{untreated}}}{\text{TMT S:N}_{\text{phosphatasetreated}}} \times 100$$

For our 0% lower limit method, the calculation of stoichiometry was identical except that any negative stoichiometry calculated was replaced with 0%. Arithmetic means and sample standard deviations were calculated for both methods across the five biological replicates for each peptide.

An in-house Bayesian modeling program in R/Stan treated stoichiometry as an estimable parameter rather than a statistic. Briefly, to prevent negative estimation of stoichiometry and to generate credible intervals that contain only physically possible numbers (i.e., stoichiometry estimations constrained between 0 and 100%) we chose to model stoichiometry as a beta distribution–a distribution naturally constrained to the unit interval. Additionally, instead of calculating stoichiometry as a statistic directly from the raw data, we calculated the fraction of S:N contributed by the untreated channel and used this statistic to make inferences about the phosphorylation:

$$\text{S: N Contribution}_{\text{untreated}} = \frac{\text{TMT S: N}_{\text{untreated}}}{\text{TMT S: N}_{\text{phosphatasetreated}} + \text{TMT S: N}_{\text{untreated}}}$$

This statistic is calculated for each pair of TMT channels corresponding to a biological replicate.

All calculated values are then fed into our in-house software, which then fits the following Bayesian model,

$$y_{ij} \sim Beta\left(\phi_i, \lambda\right)$$

$$\phi_i \sim Pareto(0.1, 1.5)$$

$$\lambda = \text{log}it^{-1}\left(\mu_i + \beta_j\right)$$

$$\mu_i \sim halfNormal(0, 5)$$

$$\beta_j \sim Normal(0, 5)$$

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

$$S_i = 1 - \frac{\mathrm{log}it^{-1}\left(\mu_i\right)}{1 - \mathrm{log}it^{-1}\left(\mu_i\right)}$$

where $i = 1,\ldots, n_p$ indexes the n_p phosphorylation sites. $j = 1,\ldots, n_t$ indexes the n_t tubes/replicates. y_{ij} represents the observed untreated signal-to-noise contribution, which was defined above, and the Beta distribution here is defined in terms of mean parameters, ϕ_i, and a precision parameter, λ. μ_i represents the true contribution of untreated signal to the i'th site and the β_j's represents tube effects (pipetting error). Finally, S_i represents the true phosphorylation of the ith site. This is the main parameter of interest. Notice that it is the use of a half-Normal distribution for μ_i that forces stoichiometry between 0 and 1. All prior distributions were selected to be weakly informative.

Bayesian methods gave us the flexibility to pick distributions and domains that place stoichiometry within the correct interval. It is not clear how this would be achieved with frequentist methods. Our Bayesian method is not deterministic and requires simulations to describe the posterior distributions of our parameters. In the domain specific programming language Stan, Markov chain Monte Carlo simulations using Hamiltonian Dynamics, also known as a Hamilton Monte Carlo, achieve this goal. After executing a predefined number of simulated draws, 2000 is the default in Stan, we discard the first half (since convergence may not have been achieved) and use the latter to describe the distributions of interest. Here we aim to determine the probability distribution of each stoichiometry, given the observed data. We summarize this distribution with the posterior mean and percentiles that correspond with 80% and 95% credible intervals for each peptide. Additionally, the posterior mean of λ provides a measurement of how much overall variation is seen in the data.

Convergence can be assessed by looking at traceplots which show the values of a parameter after each iteration. In our experiment, we always observed convergence within the first few hundred iterations.

RESULTS

Experiment Workflow

The TMT10-plex workflow for determining phosphorylation occupancy is shown in Figure 1. We chose to implement the workflow using five biological replicates of HCT116 cells (Figure 1A). To minimize variability, samples were only subjected to individual desalting columns once, after digestion (before splitting). TMT10 reagent usage was optimized by dividing each replicate into 2 aliquots such that each aliquot received a unique TMT tag. After TMT labeling, all aliquots were dried down in a vacuum centrifuge before reconstitution in phosphatase buffer. All aliquots were recombined for a single desalting step before off-line BPRP fractionation prior to mass spectrometer analysis. For each biological replicate, phosphorylation stoichiometry was calculated for each peptide whose phosphorylated version could be found in a known library (Figure 1A,B). TMT10 enabled us to analyze all five biological replicates simultaneously, which was not possible previously.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Generation of Phosphopeptide Library Found in HCT-116 Cells

The first iteration of our phosphatase-based method used a database of known phosphorylation sites found in the literature.[2] Instead of using a literature-based database, we created our own by performing a Fe-IMAC enrichment on confluent HCT-116 cells (Figure 2A). We enriched phosphopeptides from 10 mg of protein and then separated the enriched sample into 24 fractions by off-line BPRP HPLC. Each fraction was subjected to high-resolution MS2 analysis using HCD fragmentation. We identified over 42 000 unique phosphopeptides that were localized to 24028 sites categorized by type (acidic, basic, proline-directed, other) based on our lab's previous algorithm (Figure 2B, Supplementary Table 1).[19] This data set was then utilized as the known peptide library (Figure 1A). Forty percent of observed phosphorylation sites were of the proline-directed type, 26% acidic, 19% basic, and 16% did not fall any of the listed categories (Figure 2B). After assigning stoichiometry to the matched sites, we observed that sites with an acidic motif were found at higher average stoichiometries (Figure S1).[2]

Phosphatase Experiment Observes 25% from Generated Phosphopeptide Database

We analyzed all 24 fractions of our TMT10 labeled phosphatase-based stoichiometry experiment on an Orbitrap Fusion Lumos instrument. Over 124 000 total peptides were identified, corresponding to 8351 proteins (Figure 3). For consistent quality, we then filtered our data set for peptides with precursor isolation specificity of at least 0.5 and a sum S:N ratio of 200 across the 10 TMT reporter ion channels. This resulted in 72 074 unique peptides being passed for quantification (Figure 3). After matching our identified peptides to their phosphorylated forms in our phosphopeptide library, we assigned 6772 unique peptides a phosphorylation stoichiometry value (Figure 3, Supplementary Table 2). The stoichiometries for these peptides were then calculated in the standard method, 0% lower limit method, and our Bayesian modeling method.

Calculating Stoichiometries Directly from Raw Data Can Result in Negative Values

We first proceeded to calculate stoichiometries for our phosphopeptides using the standard method (Figure 1B). We looked at six examples of the phosphorylation calculation that were found in targeted studies according to previous literature (Figure 4A). While the averages of the five replicates were all physically possible (between 0% and 100% stoichiometry), we noticed that the individual stoichiometry measurements for each replicate could be calculated as negative values (Figure 4A,B). An example is nucleophosmin, NPM1, which had a positive average stoichiometry near 0%, but had individual replicates that were assigned negative stoichiometries using our standard method of stoichiometry calculation (Figure 4A,B).[23] We then attempted to address these negative stoichiometry issues by either setting the lower limit of stoichiometry to 0% and by developing our Bayesian model.

Boundary Conditions of the Stoichiometry Measurement Affect Its Distribution

Previously, peptide phosphorylation stoichiometry was treated as statistic and calculated directly from the raw data.[2,6,7] As such, we initially calculated this stoichiometry statistic and plotted a histogram of the results. The resulting distribution was centered near 0% resulting in substantial negative stoichiometries being calculated (Figure 5A). Additionally,

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

a second population of stoichiometries near 100% was observed. Both observations are in line with previous data from our lab.[2]

To address the issue of negative stoichiometries, we then calculated stoichiometry but only allowed the lowest value to be 0%, as reported previously.[2] This resulted in all negative stoichiometries being set to 0%. The resulting histogram showed little to no change in bins containing average stoichiometries of 30% or more but showed an increase in the bin height of the bins containing 6% or less average stoichiometries (Figure 5B). While this solved the issue of negative stoichiometries, it created a new problem of artificially reducing our error estimates, as discussed later.

As an alternative to limiting the lowest stoichiometry to 0%, we created a Bayesian model that would treat phosphorylation stoichiometry as an unobserved parameter defined on the interval 0 to 1. In doing so, we can utilize all of the observed measurements to estimate stoichiometry and inform our error and precision. We ran our statistical model on the data set and observed that it converged rapidly with a precision value of 94 (Figure S2). The precision value is inversely related to the variance of a beta distribution given a specific expected mean. As such, increasing precision results in decreasing variance. Plotting the distribution of the stoichiometries as a histogram highlighted large increases in the bins containing average expected stoichiometries between 6 and 10% (Figure 5C).

To assess how the Bayesian modeling was affecting the stoichiometries obtained by traditional methods, we compared the differences between the standard method for calculating stoichiometry and the 0% lower limit method with the Bayesian model. We found that a majority of stoichiometry values did not change dramatically (Figure S3). Additionally we observed that most changes to the stoichiometry when going from average measurements to expected means from the Bayesian model resulted in a 5–10% increase. These data agree with the change in the distribution of the histograms, implying that our statistical method preferentially affects the calculations yielding negative or low stoichiometries (Figure 5 and Figure S1).

The increase in the observed stoichiometry value when utilizing the Bayesian modeling method suggests that measuring a 0% stoichiometry is extremely difficult with the current instrumentation and that perhaps the lower end of our reliable estimation of stoichiometries is approximately 5–10%. This was further confirmed when we assessed how phosphorylation site motifs affect stoichiometry by utilizing the stoichiometries obtained from our 0% lower limit and Bayesian modeling method. When using the 0% lower limit method, peptides assigned a phosphorylation stoichiometry containing an acidic motif peptide were more likely to be observed at a higher stoichiometry, with ~20% of all peptides kept at 0% (Figure S1A). This is in line with our previous findings in yeast whole cell lysate.[2] When utilizing the Bayesian modeling method, this trend is preserved; however, we noticed that peptides estimated to be at 0% by the 0% lower limit method were pushed off the x-axis in the Bayesian modeling method (Figure S1A,B). This seemingly implied that the cell maintains a low level of phosphorylation for peptides thought to be kept at 0% stoichiometry. However, both the cumulative distribution plots from Supplemental Figure 1 and the histograms from Figure 5 only visualize the point estimate of each stoichiometry

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

distribution for each peptide. Large variance could render these point estimation worthless and necessitate the investigation of the error intervals surrounding each stoichiometry point estimator. As such, it cannot be inferred that a majority of the proteome is kept at 5–10% stoichiometry without first looking at the error intervals.

Proper Modeling Prevents Error Intervals from Containing Senseless Results

To assess the variation of the stoichiometry distributions by the different estimation methods, we plotted the rank ordered peptide phosphorylation averages with their 80% and 95% confidence intervals. When calculating the stoichiometry value using the standard method, we observed a noticeable number of the average phosphorylation stoichiometries that fell below 0%; furthermore, a majority of peptides had confidence intervals that included negative values or values exceeding 100% (Figure 6A). Additionally, we noted the abundance of relatively large confidence intervals throughout the data set.

We then assessed how setting the lower limit of stoichiometry to 0% would affect this plot. About 12% of the data were incorrectly reported as having no variance while about half of the peptides displayed a trend of increasing interval size as the peptide's stoichiometry average increased (Figure 6B). This linked relationship between increasing average and standard deviation coupled with the region of no variance caused us to question the validity of this method. By artificially clipping the negative stoichiometries we calculated to 0%, measurements of variability were artificially reduced, with greater reductions occurring the closer the stoichiometry average was to 0%. Furthermore, we still had problems with nonsensical error intervals containing values outside of the 0 to 1 range.

When utilizing our Bayesian model to estimate stoichiometry, the program additionally generates credible intervals around the expected stoichiometry value. We performed the same plotting method as above, which shows that all peptides have credible intervals corresponding to physical reality (Figure 6C). We observed additionally a vertical shift at the low end of the graph indicating that most peptides previously thought to be at 0% phosphorylation stoichiometry now had stoichiometry point estimators slightly higher (Figures 6C, 5C, and Figure S3). Overall, while the general shape and trend of the plots remain unchanged, the error intervals improved dramatically when utilizing the Bayesian model. This is further highlighted by the observation that approximately 2000 peptides with confidence intervals containing only physically possible results when utilizing the standard method and approximately 3000 with the 0% lower limit method (Figure S4). Additionally, the credible intervals using the Bayesian method suggest that, for a majority of peptides, even though the point estimator suggests 5% stoichiometry the true stoichiometry lies anywhere between 0% and 20% stoichiometry.

DISCUSSION

In this study, five biological replicates of HCT116 cells were analyzed to gain insight into the basal level of phosphorylation stoichiometry of this colorectal cancer cell line. Prior to determining stoichiometry we collected a reference database of 24028 phosphorylation events under basal conditions which served as a library of sites to attempt stoichiometry assessment. Our occupancy analysis was performed using TMT labeling which increased the

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

sample multiplexing capacity to allow simultaneous analysis of all five biological replicates. In addition, by utilizing TMT, there were no missing values in that all five measurements were determined for all peptides in the data set. In total, we assigned 6772 unique peptides, from our generated reference library, stoichiometry values.

As stoichiometry is defined as the fractional occupancy, its values should, ideally, reside within the unit interval [0, 1]. Despite quantifying peptides across five replicates, by following the standard method of calculating phosphorylation stoichiometry values, we initially obtained some negative stoichiometry values.[2] This occurred stochastically when sites were present at low stoichiometries such that the error in the five measurements was greater than the % occupancy. Additionally, our initial attempts at calculating stoichiometry resulted in confidence intervals containing values greater than 1 suggesting over 100% occupancy. Both phenomena are physically impossible.

Previous iterations of this phosphatase method estimated stoichiometry from a stoichiometry statistic calculated from the raw data rather than treating stoichiometry solely as an estimable parameter.[2,6–8] As the raw data are the S:N values collected from the instrument ranging from 1 to positive infinity, nothing constrains a stoichiometry statistic calculated with the formula in Figure 1b to the unit interval. If we treat stoichiometry as a constrained parameter we wish to estimate, rather than a statistic calculated from the raw data, we can utilize novel approaches to estimate the true stoichiometry of a peptide utilizing alternative statistics that leverage the raw data's properties.

Furthermore, negative stoichiometries traditionally have been dealt with by replacing the negative stoichiometries with 0% or discarding those measurements.[2,6,7] However, as mentioned above, the raw data can be transformed into a meaningful statistic from which a stoichiometry parameter can be estimated. This alternative statistic is the proportion of the sum S:N of the paired TMT channels corresponding to a replicate contributed by the untreated channel. Furthermore, the statistic, which is based on the proportionality of the data, can easily be converted into the traditional stoichiometry measurement thus allowing us to use this statistic to estimate phosphorylation stoichiometry as a parameter.

We implemented our Bayesian modeling by developing an R/Stan script included in the Supporting Information Files 4 and 5. The software samples mean peptide effects on stoichiometry, sample handling effects, and overall experimental precision. We chose to utilize an overall experimental precision due to the low sample size when analyzing precision per peptide. This measurement of precision provides a quantitative measure of the global experimental variance while still providing individual peptide variance as the variance of a beta distribution is governed by the mean and the precision term. This can be preferable to using a perpeptide error as we only acquired five measurements per peptide, one for each biological replicate, resulting in unstable estimations of error. The trade-off is that an overall experimental error gives a coarse overview of the error may not accurately represent each peptide. A further benefit is that a single experimental precision provides a quick and quantitative factor to compare multiple experiments.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

On the basis of the amount of uncertainty surrounding many of the stoichiometry point estimators, we found it was more effective to bin point estimators based into low (0–25%), medium (25–70%), and high (70–100%) categories. Similar to the Wu et al. paper, we found that acidic residues are phosphorylated at a higher stoichiometry than sites with other phosphorylation motifs (Figure S1).[2] This specific phosphorylation of acidic motifs is likely due to the high activity of Casein kinase II, which targets the motif SxxE/D.[24]

CONCLUSION

We simultaneously compared the basal phosphorylation stoichiometry of five biological replicates of HCT116 using a TMT based workflow eliminating previous problems involving missing data. We then presented a novel statistical method to address negative stoichiometries from using the phosphatase based phosphorylation stoichiometry experiment. While the credible intervals were larger than we had hoped, the global phosphorylation can be binned into low, medium, and high phosphorylation stoichiometry categories, which allow for a quick first-pass assessment of the phosphorylation state of the cell. Further study into improving measurement precision by utilizing targeted approaches or real time search may further narrow these bins. Overall, our study provides a methodical way to make sense of complex phosphorylation occupancy experiments and a quantitative read out for experimental error.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGMENTS

We thank Dr. Ed Huttlin, Dr. Jeremy O’Connell, Dr. Devin Schweppe, and Dr. David Nusinow for encouraging discussions. This work was funded in part by an NIH grant (HG67945 to S.P.G. and DK098285to J.A.P.).

REFERENCES

(1). Gerber SA; Rush J; Stemman O; Kirschner MW; Gygi SP Absolute quantification of proteins and phosphoproteins from cell lysates by tandem MS. Proc. Natl. Acad. Sci. U. S. A 2003, 100, 6940–5. [PubMed: 12771378]

(2). Wu R; et al. A large-scale method to measure absolute protein phosphorylation stoichiometries. Nat. Methods 2011, 8, 677–683. [PubMed: 21725298]

(3). Humphrey SJ; James DE; Mann M Protein Phosphorylation: A Major Switch Mechanism for Metabolic Regulation. Trends Endocrinol. Metab 2015, 26, 676–687. [PubMed: 26498855]

(4). Newman RH; Zhang J; Zhu H Toward a systems-level view of dynamic phosphorylation networks. Front. Genet 2014, 5, 1–22. [PubMed: 24567736]

(5). Olsen JV; et al. Quantitative Phosphoproteomics Reveals Widespread Full Phosphorylation Site Occupancy During Mitosis. Sci. Signaling 2010, 3, ra3–ra3.

(6). Glibert P; et al. Phospho-iTRAQ: Assessing Isobaric Labels for the Large-Scale Study Of Phosphopeptide Stoichiometry. J. Proteome Res 2015, 14, 839–849. [PubMed: 25510630]

(7). Tsai C-F; et al. Large-scale determination of absolute phosphorylation stoichiometries in human cells by motif-targeting quantitative proteomics. Nat. Commun 2015, 6, 6622. [PubMed: 25814448]

(8). Domanski D; Murphy LC; Borchers CH Assay development for the determination of phosphorylation stoichiometry using multiple reaction monitoring methods with and without

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

phosphatase treatment: application to breast cancer signaling pathways. Anal. Chem 2010, 82, 5610–5620. [PubMed: 20524616]

(9). Horinouchi T; Terada K; Higashi T; Miwa S Using Phos-Tag in Western Blotting Analysis To Evaluate Protein Phosphorylation. Kidney Research: Experimental Protocols, Methods in Molecular Biology; Springer Science+Business Media, 2009; 1397, 267–277.

(10). McAlister GC; et al. Increasing the multiplexing capacity of TMTs using reporter ion isotopologues with isobaric masses. Anal. Chem 2012, 84, 7469–7478. [PubMed: 22880955]

(11). Paulo JA; O'Connell JD; Gygi SP A Triple Knockout (TKO) Proteomics Standard for Diagnosing Ion Interference in Isobaric Labeling Experiments. J. Am. Soc. Mass Spectrom 2016, 27, 1620–1625. [PubMed: 27400695]

(12). Thompson A; et al. Tandem mass tags: a novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS. Anal. Chem 2003, 75, 1895–1904. [PubMed: 12713048]

(13). Paulo JA; Mancias JD; Gygi SP Proteome-Wide Protein Expression Profiling Across Five Pancreatic Cell Lines. Pancreas 2017, 46, 690. [PubMed: 28375945]

(14). Ross PL; et al. Multiplexed Protein Quantitation in Saccharomyces cerevisiae Using Amine-reactive Isobaric Tagging Reagents. Mol. Cell. Proteomics 2004, 3, 1154–1169. [PubMed: 15385600]

(15). Ting L; Rad R; Gygi SP; Haas W MS3 eliminates ratio distortion in isobaric multiplexed quantitative proteomics. Nat. Methods 2011, 8, 937–940. [PubMed: 21963607]

(16). Mcalister GC; et al. MultiNotch MS3 Enables Accurate, Sensitive, and Multiplexed Detection of Differential Expression across Cancer Cell Line Proteomes Graeme C. McAlister, 1 David P. Nusinow, 1. Anal. Chem 2014, 86, 7150–7158. [PubMed: 24927332]

(17). Huttlin EL; et al. The BioPlex Network: A Systematic Exploration of the Human Interactome. Cell 2015, 162, 425–440. [PubMed: 26186194]

(18). Villén J; Gygi SP The SCX/IMAC enrichment approach for global phosphorylation analysis by mass spectrometry. Nat. Protoc 2008, 3, 1630–1638. [PubMed: 18833199]

(19). Huttlin EL; et al. A tissue-specific atlas of mouse protein phosphorylation and expression. Cell 2010, 143, 1174–1189. [PubMed: 21183079]

(20). Wasmuth EV; Lima CD UniProt: the universal protein knowledgebase. Nucleic Acids Res 2017, 45, 1–12. [PubMed: 27899559]

(21). Elias JE; Gygi SP Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. Nat. Methods 2007, 4, 207–214. [PubMed: 17327847]

(22). Beausoleil SA; Villén J; Gerber SA; Rush J; Gygi SP A probability-based approach for high-throughput protein phosphorylation analysis and site localization. Nat. Biotechnol 2006, 24, 1285–1292. [PubMed: 16964243]

(23). Hornbeck PV; et al. PhosphoSitePlus, 2014: Mutations, PTMs and recalibrations. Nucleic Acids Res 2015, 43, D512–D520. [PubMed: 25514926]

(24). Ahmed K; Issinger O Protein Kinase CK2 Cellular Function in Normal and Disease States 2015, 1.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

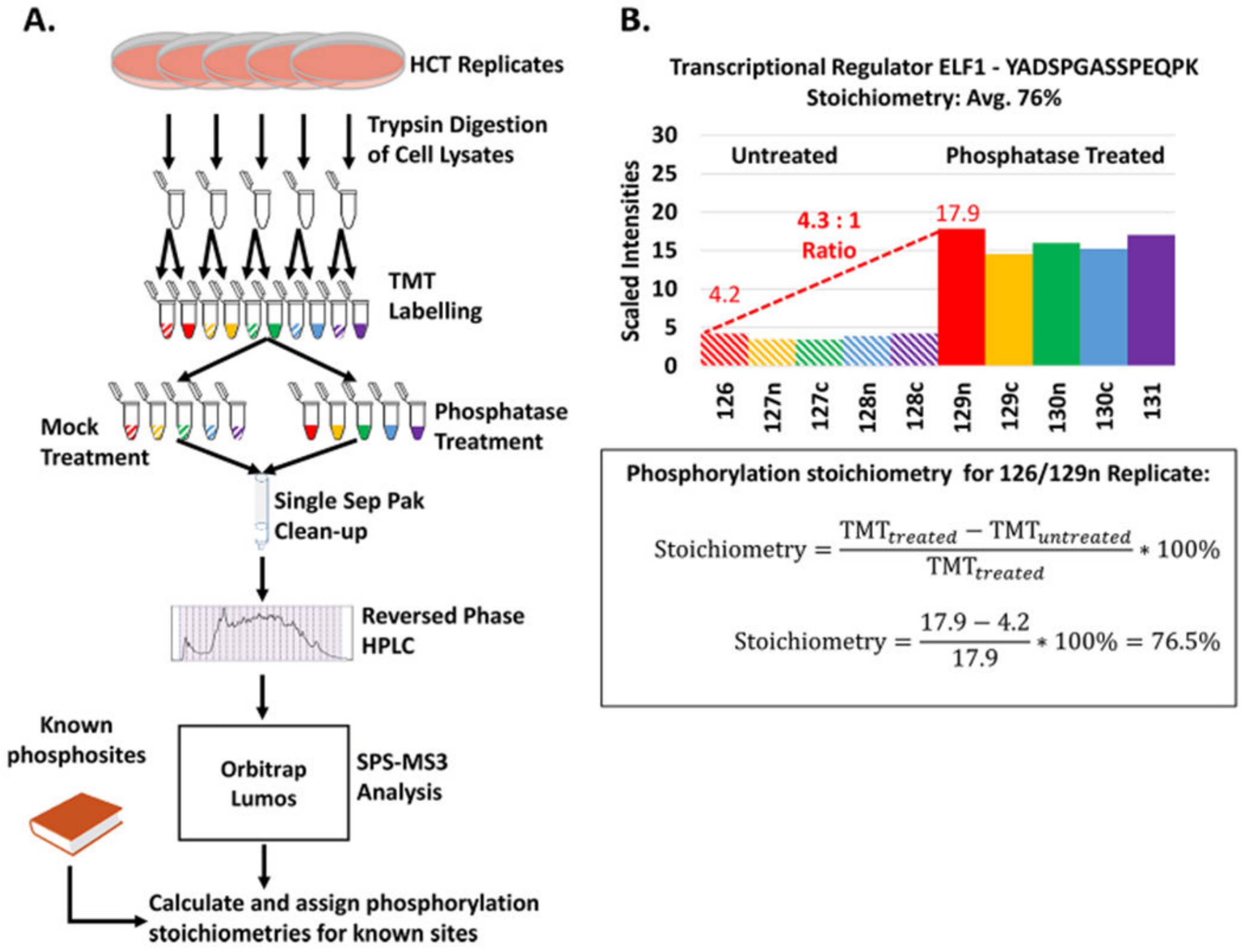

A.
HCT Replicates
Trypsin Digestion of Cell Lysates
TMT Labelling
Mock Treatment
Phosphatase Treatment
Single Sep Pak Clean-up
Reversed Phase HPLC
Known phosphosites
Orbitrap Lumos
SPS-MS3 Analysis
Calculate and assign phosphorylation stoichiometries for known sites
B.
Transcriptional Regulator ELF1 - YADSPGASSPEQPK
Stoichiometry: Avg. 76%
Scaled Intensities
30
25
20
15
10
5
0
Untreated
Phosphatase Treated
4.3 : 1 Ratio
17.9
4.2
126
127n
127c
128n
128c
129n
129c
130n
130c
131
Phosphorylation stoichiometry for 126/129n Replicate:
$\text{Stoichiometry} = \frac{TMT_{treated} - TMT_{untreated}}{TMT_{treated}} * 100\%$
$\text{Stoichiometry} = \frac{17.9 - 4.2}{17.9} * 100\% = 76.5\%$

Figure 1.

(A) Workflow for phosphorylation stoichiometry experiment. Briefly, reduced and alkylated cell lysate from five biological replicates of HCT-116 cells were separately digested with trypsin, and each sample was split into two aliquots for TMT-10 labeling. One labeled aliquot from each sample was subjected to phosphatase treatment while its sister aliquot underwent a mock treatment. All 10 aliquots were combined for Sep-Pak cleanup and subjected to reversed phase HPLC and then analyzed by SPS-MS3 on a Thermo Orbitrap Fusion Lumos. Stoichiometries were then calculated for each peptide and assigned to phosphopeptides from a previous independent phosphopeptide identification experiment. (B) Sample calculation of how stoichiometry is calculated for an observed peptide from our experiment. The stoichiometry for each sample is calculated. In the example shown, the stoichiometry is calculated for the red sample. An equivalent formula is to use the ratio of treated to untreated to calculate the stoichiometry: $1 - \frac{1}{T:U}$.

Author Manuscript
Author Manuscript
Author Manuscript
Author Manuscript

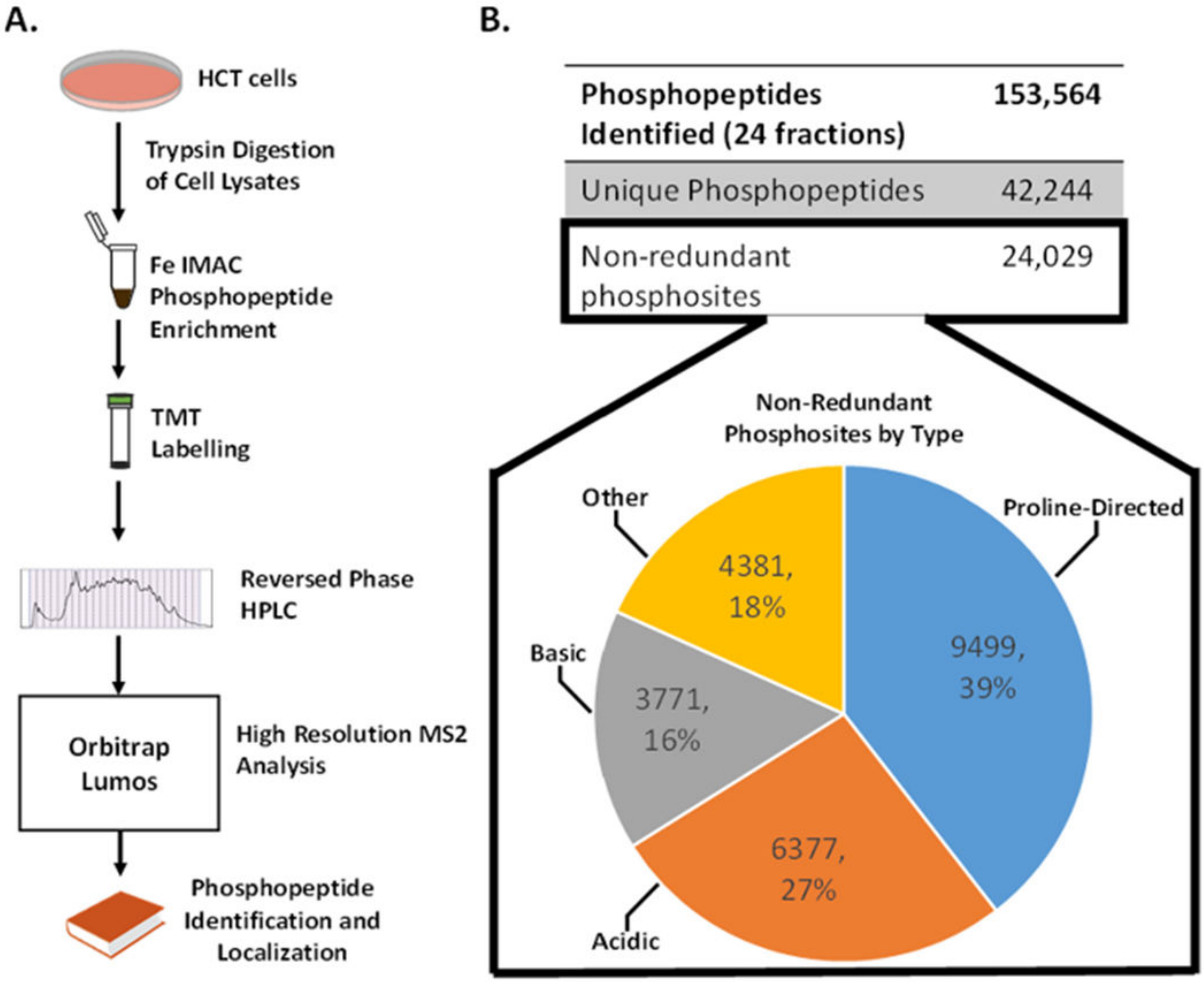

A.
HCT cells
Trypsin Digestion of Cell Lysates
Fe IMAC Phosphopeptide Enrichment
TMT Labelling
Reversed Phase HPLC
Orbitrap Lumos
High Resolution MS2 Analysis
Phosphopeptide Identification and Localization
B.
Phosphopeptides Identified (24 fractions) 153,564
Unique Phosphopeptides 42,244
Non-redundant phosphosites 24,029
Non-Redundant Phosphosites by Type
Other
4381, 18%
Proline-Directed
9499, 39%
Basic
3771, 16%
6377, 27%
Acidic

Figure 2.

(A) Workflow for independent phosphopeptide identification experiment. Fe-IMAC enrichment was performed on the digested cell lysate from HCT-116 cells. The phosphopeptide enriched digest was then TMT-labeled to account for chemical changes caused by TMT-labeling and subjected to fractionation by reverse-phase HPLC. Fractions were analyzed by high resolution MS2 analysis. Resulting phosphopeptide identifications were localized to sites using a modified A-score to generate the known phosphorylation sites library used in Figure 1A. (B) Summary of phosphopeptides identified during this experiment. Pie chart breaks down the phosphopeptides by type: acidic, basic, proline-directed, and other. Sites were assigned a type based on a previously described algorithm.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

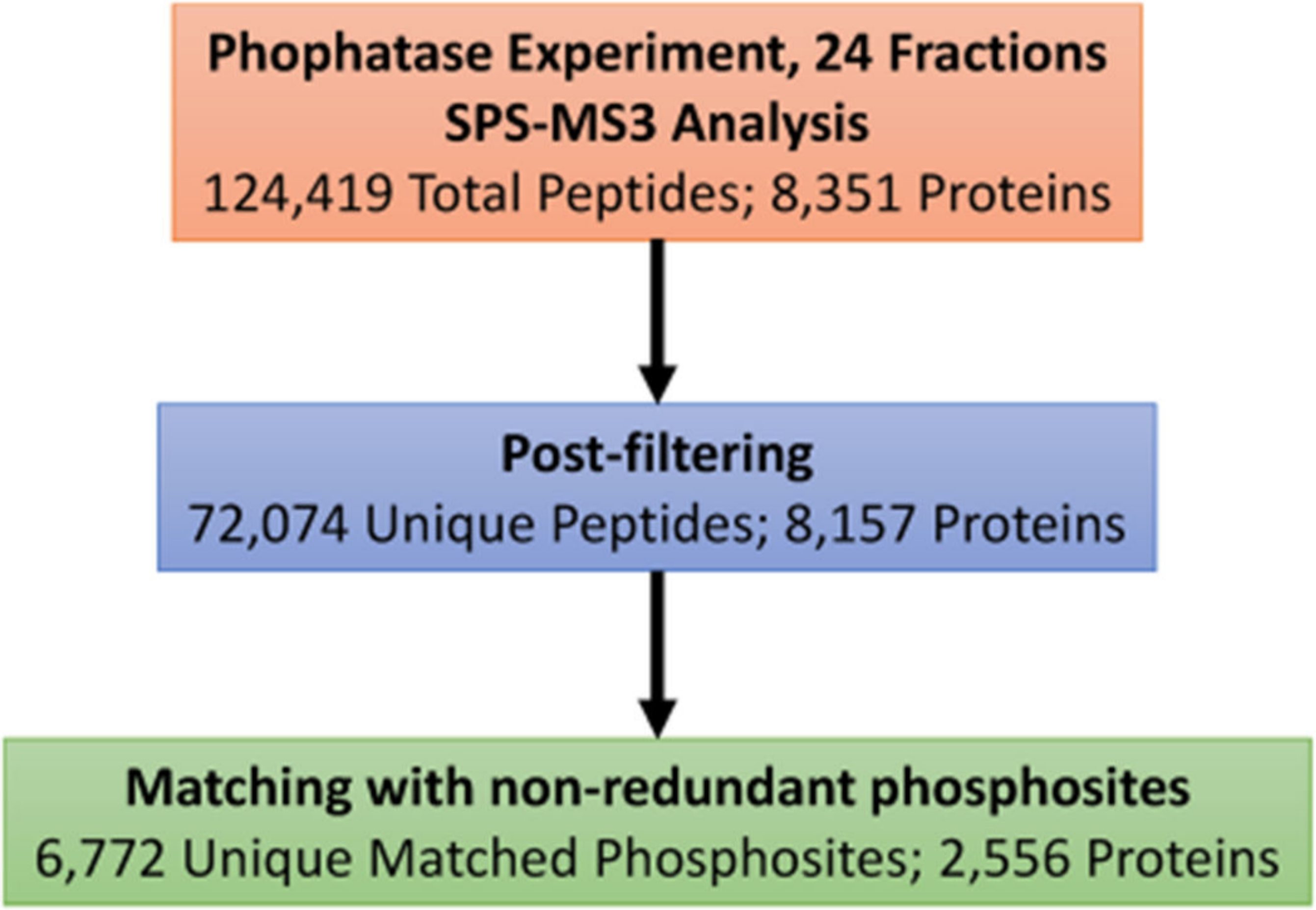
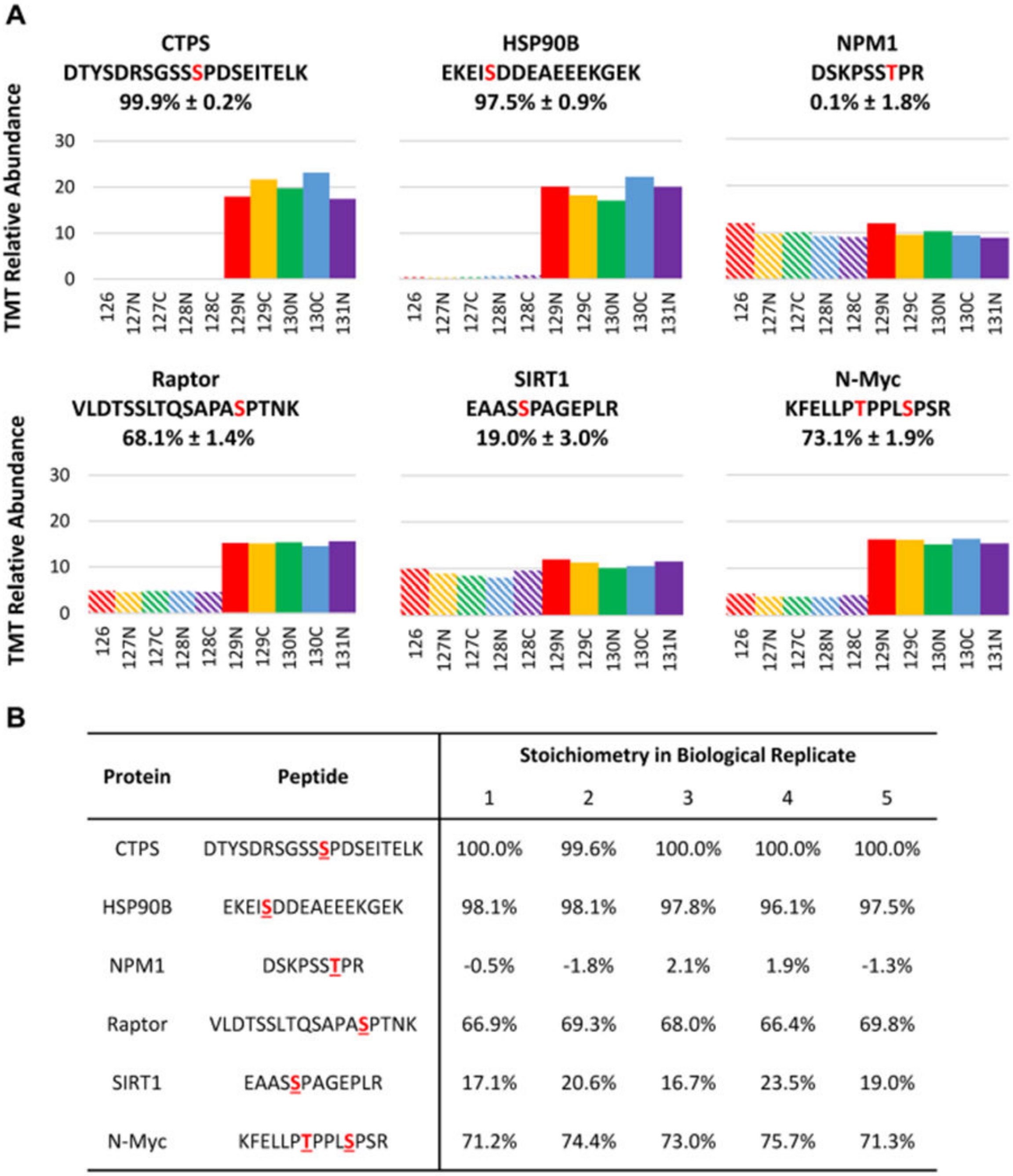

Phophatase Experiment, 24 Fractions
SPS-MS3 Analysis
124,419 Total Peptides; 8,351 Proteins
Post-filtering
72,074 Unique Peptides; 8,157 Proteins
Matching with non-redundant phosphosites
6,772 Unique Matched Phosphosites; 2,556 Proteins

Figure 3.

Summary of phosphorylation stoichiometry experiment results. A total of 124 419 peptides corresponding to 8351 proteins were identified. A total of 6772 unique peptides (2556 proteins) were matched to phosphorylation sites identified in our phosphopeptide enrichment experiment.



Protein	Peptide	Stoichiometry in Biological Replicate				
		1	2	3	4	5
CTPS	DTYSDRSGSSSPDSEITELK	100.0%	99.6%	100.0%	100.0%	100.0%
HSP90B	EKEISDDEAEEEKGEK	98.1%	98.1%	97.8%	96.1%	97.5%
NPM1	DSKPSSTPR	-0.5%	-1.8%	2.1%	1.9%	-1.3%
Raptor	VLDTSSLTQSAPASPTNK	66.9%	69.3%	68.0%	66.4%	69.8%
SIRT1	EAASSPAGEPLR	17.1%	20.6%	16.7%	23.5%	19.0%
N-Myc	KFELLPTPPLSPSR	71.2%	74.4%	73.0%	75.7%	71.3%

Figure 4.

(A) Example TMT-data for peptides known to harbor phosphorylation sites. Stoichiometries were calculated for each sample (red, yellow, green, blue, and purple), the average and standard deviation are reported. Solid colors represent channels where the aliquot was treated with phosphatase while the striped colors represent channels where the aliquot was mock treated. (B) Table displaying the individual sample phosphorylation stoichiometries calculated for each peptide in panel A). Red characters represent the expected

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

phosphorylation site. All sites chosen were identified as regulatory phosphorylation events through targeted studies based on the phosphositeplus.org database.[23]

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

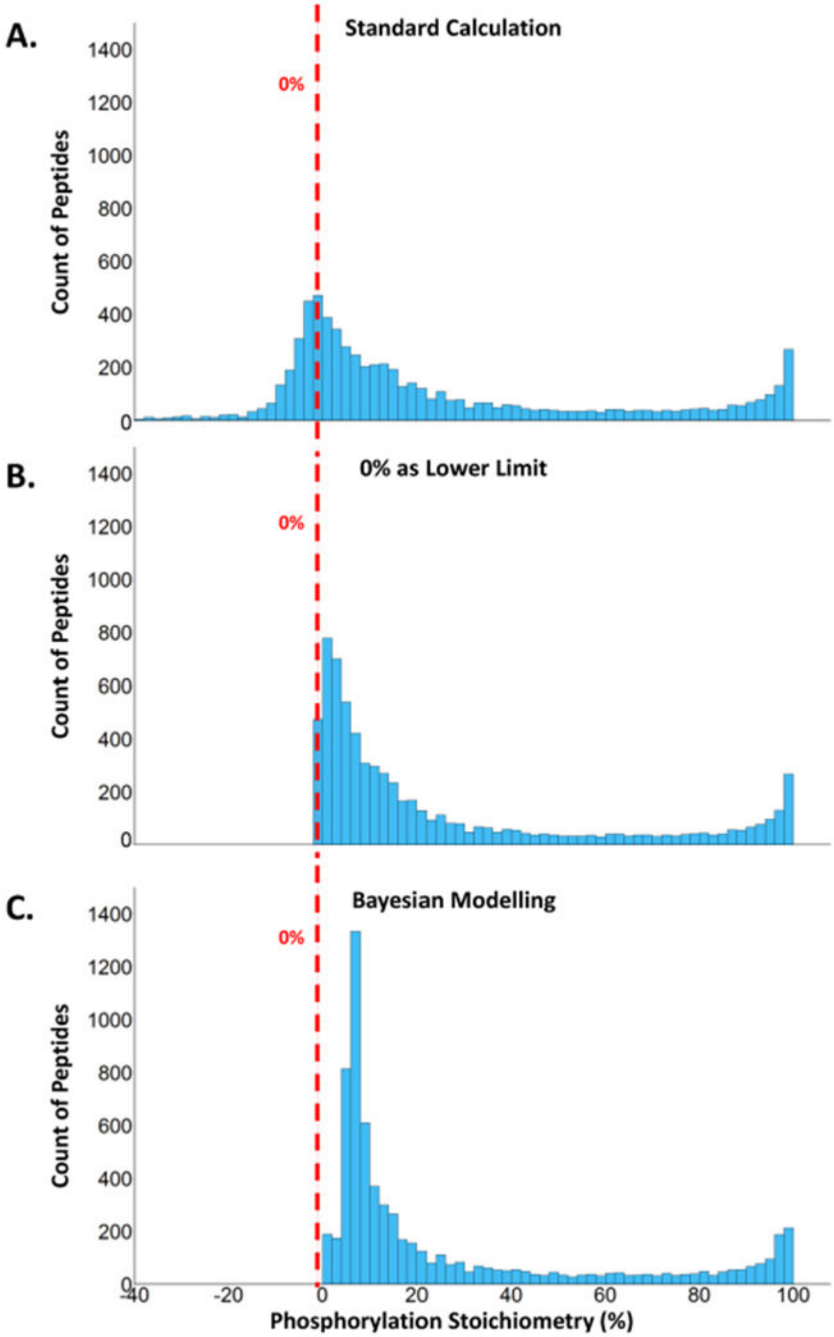


Figure 5.

Histograms of the phosphorylation stoichiometry for each estimation method. (A) Histogram when no correction is performed. (B) Histogram where each negative stoichiometry is replaced with 0. (C) Histogram when stoichiometry is estimated using the Bayesian modeling approach. The red dashed line represents 0%.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

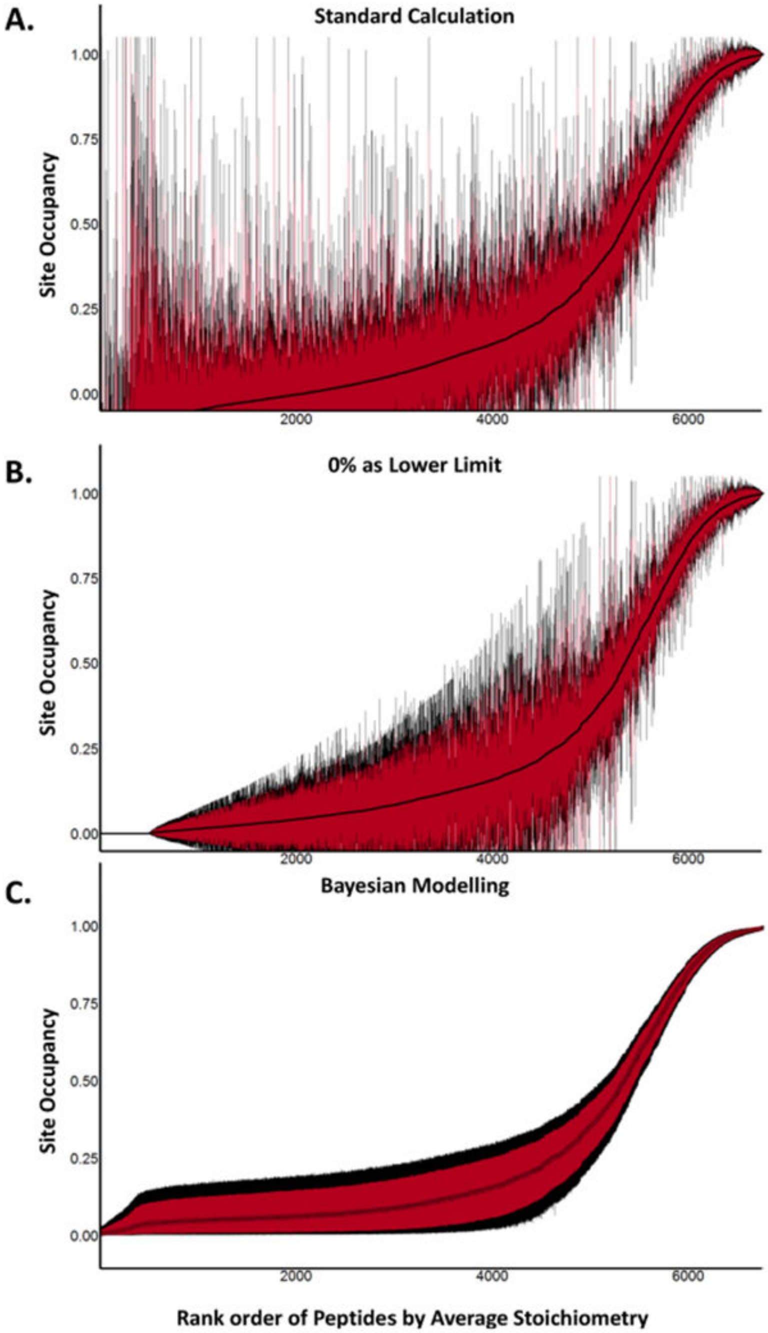


Figure 6.

Peptides were rank ordered (lower values first) by their estimated stoichiometry. 80% confidence intervals (red bars) and 95% confidence intervals (black bars) were drawn around each point. The *y*-axis represents the phosphorylation stoichiometry as a fraction instead of a percent. Resulting caterpillar plots are shown for each method. (A) Standard method with no corrections performed. (B) All negative stoichiometry calculations were replaced with 0. (C) Stoichiometry values were estimated using our Bayesian model.