Published in final edited form as:

J Proteome Res. 2013 October 4; 12(10): . doi:10.1021/pr400678m.

Nonparametric Bayesian evaluation of differential protein quantification

Oliver Serang,

Thermo Fisher Scientific Bremen, Bremen, Germany (+49)04215493,

Oliver.Serang@ThermoFisher.com

A. Ertugrul Cansizoglu,

Department of Neurobiology, Harvard Medical School, Boston Children's Hospital, Boston, MA, USA

Lukas Käll,

Royal Institute of Technology (KTH), Science for Life Laboratory, School of Biotechnology, Solna, Sweden

Hanno Steen, and

Department of Pathology, Harvard Medical School, Boston Children's Hospital, Boston, MA, USA

Judith A. Steen

Department of Neurobiology, Harvard Medical School, Boston Children's Hospital, Boston, MA, USA, +1 617-919-2450, Judith.Steen@Childrens.harvard.edu

Abstract

Arbitrary cutoffs are ubiquitous in quantitative computational proteomics: maximum acceptable MS/MS PSM or peptide q -value, minimum ion intensity to calculate a fold change, the minimum number of peptides that must be available to trust the estimated protein fold change (or the minimum number of PSMs that must be available to trust the estimated peptide fold change), and the "significant" fold change cutoff. Here we introduce a novel experimental setup and nonparametric Bayesian algorithm for determining the statistical quality of a proposed differential set of proteins or peptides. By comparing putatively non-changing case-control evidence to an empirical null distribution derived from a control-control experiment, we successfully avoid some of these common parameters. We then apply our method to evaluating different fold change rules and find that, for our data, a 1.2-fold change is the most permissive of the plausible fold change rules.

Introduction

High-throughput biological profiling tools (*e.g.* microarrays, high-throughput DNA and RNA sequencing, and mass spectrometry) are essential to the shift toward quantitative

Correspondence to: Oliver Serang; Judith A. Steen.

Principal contribution: An example experimental setup for producing an empirical null distribution (this setup is applicable to other types of analysis, *e.g.* SILAC and label-free quantification), and a statistical algorithm that uses this empirical null distribution to quantify technical variation and agnostically evaluate differential significance without commonly used fold change cutoffs.

Availability The source code from this study can be downloaded from <http://steenlab.org/software/npCI> and the data has been posted on PeptideAtlas (dataset identifier PASS00264).

Author Contributions O.S. proposed the nonparametric method for evaluating differential proteins using the npCI with an empirical null distribution. E.C. cultured the cells and then performed TMT-LC-MS/MS analysis and Scaffold search. O.S., E.C., L.K., H.S., and J.S. proposed the experiments in the paper.

hypothesis generation experiments. In particular, mass spectrometry has found favor due to its ability to directly identify and quantify the proteome (including the post-translational modifications)¹⁻⁴. However, to our knowledge there is currently no high-throughput quantitative data analysis technique that makes use of an empirical null distribution and that does not assume that the distributions are known (*e.g.* the frequent assumption that distributions are normal in order to apply the *t*-test). Instead, current methods use parametric methods to perform multiple statistical tests, such as *t*-tests^{5,6} or ANOVAs⁷, by rank-based comparison⁸ (which is insensitive to changes in the distribution tails^{9,10}), or by simply employing arbitrary cutoffs that establish what constitutes a significant fold change. Currently, a 1.2-fold change or higher (relative to a control sample) averaged over at least 3 constituent peptides is commonly regarded as indicative of a significantly varying protein¹¹⁻¹⁴. Furthermore, such approaches may result in insidious multiple testing problems, because peptides shared between proteins (sometimes called “degenerate peptides”) result in testing multiple hypotheses that are dependent due to shared data. Correlated hypotheses may result in one differential protein “dragging along” other proteins with which it shares peptides, and thus can cause such statistical testing to incorrectly conclude that all of these proteins significantly differential. However, systematically excluding such shared data also has its problems: despite the fact that excluding shared data facilitates the testing of multiple independent hypotheses (meaning that they are appropriate for standard multiple testing analysis), it results in a substantial loss of information, and may introduce significant biases because shared peptides may not be uniformly distributed among all proteins (for example, excluding shared peptides may bias against finding proteins from genes with substantial splice variation, because these proteins are more likely to have peptides shared with other splice variants).

More rigorously choosing the appropriate fold change cutoff has been impeded by the field’s inability to validate the results; after all, validating with a “gold standard” derived from another parametric method or a method with arbitrarily chosen parameters will simply check the degree with which the evaluated method’s assumptions agree with the parametric evaluation method. More generally, this inability to validate differential protein quantification results has obstructed the creation of rigorous and practically useful parametric statistical methods, *e.g.* Bayesian networks for modeling quantitative proteomics.

To date, empirically validating quantitative case-control (*i.e.* treated-untreated) differentials with ground-truth data sets (*i.e.* using case and control samples where the differential proteins are known) has likewise proven infeasible. Many proteins are known to play multiple roles within an organism: such pleiotropic gene action makes assigning *all* proper gene ontology (GO) terms a difficult task^{15,16}, and as a result GO terms are not reliable as a gold standard for differential quantification. For example, in the data analyzed in this experiment, using the GO terms containing “mitosis” to distinguish proteins likely to change as a response to prometaphase arrest would be incomplete, because some proteins may be labeled with terms such as “DNA repair”, but not “mitosis”, despite the plausibility that such a gene would be differentially regulated during the rapid DNA synthesis and proofreading that takes place during mitosis.

Using a subset of well-established proteins with very well-characterized functions, many proteotypic peptides and dramatic fold changes yields a data set that is not only limited in size, but which is also biased: trusted positive and negative controls are respectively enriched for very significant (*i.e.* fold change $\gg 1$) and strongly insignificant (*i.e.* fold change ≈ 1) results. For this reason investigators are generally limited to using noisy labels or employing “spike-in” data sets, which have neither the number of significantly varying proteins, the complexity, nor the noise found in real data.

Microarray analysis suffered from similar problems, and so researchers proposed the “self-self hybridization” (*i.e.* a control-control comparison)^{17;18}. These techniques quantified technical variation by analyzing the fold change between two samples with no biological variation of interest. The resulting distribution of technical variation was visualized by creating a ratio-intensity plot of the results (generally higher “outlier” ratios are more frequent where the average intensity was low, because the denominator may fluctuate to be very close to zero). Intensity-specific fold change distributions were computed by fitting a normal density within a sliding window enclosing each intensity of interest. These distributions are used to compute a *p*-value for each case-control intensity and ratio pair by looking up the intensity-specific normal distribution and performing a *t*-test. These windowed *t*-test approaches from the microarray literature also inspired direct application to proteomics, also using multiple windowed *t*-tests^{5;6}.

Unfortunately, there are reasons this windowed *t*-test procedure is not ideal for application to mass spectrometry-based proteomics data: First, due to complex network dependencies (*i.e.* proteins depend on their constituent peptides and peptides depend on the spectra that match them to create PSMs), the hypotheses tested do not only suffer from multiple testing, they are also correlated because they share data¹⁹, and as a result, are not truly appropriate for independent statistical tests as performed by the microarray analysis procedure. Second, mass spectrometry data is notoriously difficult to parametrically model, and score distributions may unexpectedly diverge from normality as sample sizes increase²⁰ due to extreme value phenomena when matching peptides to spectra. Third, applying this parametric method to mass spectrometry data would require estimating free parameters (*e.g.* the sliding window size, which loosely corresponds to degree of smoothing), meaning that it still needs heuristics in order to be used in practice.

In this paper we propose a method that uses a nonparametric approach^{9;10;21–24} to build upon previous work using empirical nulls in two ways, one experimental and the other statistical: First, we employ an control-control approach to estimate the technical variation in quantitative mass spectrometry (*i.e.* an empirical null). Second, we modify a nonparametric statistical approach to fairly evaluate heuristics by generalizing the npCI¹⁰ to multivariate data and applying it to quantitative proteomics.

Materials and Methods

Cell culture and arrest

HeLa S3 (ATCC, CCL-2.2, Manassas, VA) cells were cultured in DMEM supplemented with 10% Fetal Bovine Serum, 1% penicillin/streptomycin and l-glutamine (Gibco, Grand Island, NY) following standard cell culture protocols. At 70% confluency, cells were rinsed with PBS and harvested using a cell lifter (Corning, New York, NY) to produce asynchronous sample. A parallel culture was grown until 50% confluency. Cells were grown in media supplemented with 2mM Thymidine for 22 hrs. Cells were released by washing Thymidine for 3 hours. Following thymidine arrest, cells were treated with medium containing 100ng/ml Nocodazole for 12 hours. Prometaphase arrested synchronous cells were washed with regular medium, rinsed with PBS and harvested immediately afterward using cell lifter.

Sample preparation

Cells were resuspended in Mammalian protein prep kit (Qiagen, MD) supplemented with protease and phosphatase inhibitors and lysed using sonication. Protein concentration was determined using Pierce BCA Protein Assay kit (Thermo Fisher Scientific, Rockford, IL). Equal amount (100µg) for each sample was taken and total protein precipitation was done

using methanol/chloroform extraction. Extracted proteins were reduced with 10mM DTT, alkylated with 1% acrylamide and subjected to overnight trypsin digestion at 37°C using porcine trypsin (Promega, Madison, WI) with a protein:trypsin ratio of 50:1.

TMT labeling and LC-MS/MS analysis

Following digestion, samples were labeled separately using isobaric TMT labels (Thermo Fisher Scientific, Rockford, IL) following manufacturers recommendations (Figure 1). Once the reactions were quenched with 5% hydroxylamine, the samples were combined into one and desalted using Oasis columns (Waters Corp., Milford, MA). Combined sample was further fractionated using offgel pH 3–10 immobiline dry strips (GE Healthcare, Pittsburgh, PA) into 24 fractions. Fractions were desalted using Nestgroup c18 tips (Southborough, MA). Fractions were run on Thermo Fisher Q Exactive coupled with Exigent LC system (AB Sciex, Framingham, MA) over a 60 minute gradient. As is normal practice for TMT-LC-MS/MS, each PSM yields 6 intensities (one per channel), and thus one TMT experiment can analyze multiple replicate experiments.

Processing with Scaffold

RAW files were converted into .mgf files using MSConvert (from ProteoWizard^{25;26}). Database search was carried out using Mascot²⁷ server (from Matrix Science). Search results from 24 runs were imported into Scaffold²⁸ (Proteome Software, Portland, OR). The complete spectrum report resulting in 32094 PSMs, 17216 peptides and 2698 proteins was exported from Scaffold for further analysis.

Postprocessing

Importantly, any PSM that could potentially come from a likely contaminant (using the cRAP proteins listed in Serang *et al.*¹⁰) was removed. This was essential for excluding data from the null when the data could possibly represent a true quantity change. Removing contaminant evidence in this way was essential to producing a clean empirical null that did not include large variation, for example large differences in keratin content found because a technician wearing a sweater walked through the lab during the preparation of exactly one control sample²⁹. Likewise, PSMs were removed where any channel had a missing value (as indicated by “No Values” or “Value Missing” in the spectrum report file exported from Scaffold). 5652 spectra were eliminated due to a missing value, and of those remaining, 590 were excluded for matching a contaminant protein.

Zero-intensity values were thresholded up using pseudocounts, small values added to small intensities to set a lower bound > 0 , which are frequently used in microarrays and sequence analysis³⁰. The pseudocount value used, 0.1, was arbitrarily chosen. Although this arbitrary parameter choice is unsatisfying, improvements in instrument sensitivity should allow no values to truly be zero. Furthermore, this constant could be chosen in a non-arbitrary manner by directly using the smallest relative intensity on a given channel in a given experiment (this should be close to the minimum intensity that gives a nonzero result).

The remaining data were scaled linearly so that all channels had the same mean intensity, thereby removing channel-specific biases in quantity. This corrects for such events as when a larger amount of the sample labeled with the +126 channel was pipetted into the final mixture (compared to the +127, +128, +129, +130, and +131 samples). Every spectrum reported by Scaffold was included in the graph exactly once (the resulting graph was a protein-to-peptide-to-spectrum tripartite graph), and proteins were grouped if they had identical peptide connectivity².

Results

A method to evaluate sets of differential proteins

Our method first estimates a control-control empirical distribution of technical variation and then uses that distribution to evaluate sets of differential proteins (Figure 1).

Control-control TMT experiment—First, we similarly construct a distribution of technical variation (*i.e.* a null distribution) by using a standard mass spectrometry approach to compare two samples of the same or very similar content; however, we use the resulting intensity-intensity heatmap as an empirical null distribution rather than model the null distribution parametrically.

Non-parametric statistical analysis—Second, we paired this two-dimensional data with a generalization of the npCI. The npCI is a nonparametric statistical method that evaluates other methods using a nonparametric Bayesian goodness of fit¹⁰. By assuming the most agnostic (*i.e.* uniform) prior on differential data, we can compute the likelihood that a set of putatively differential proteins are correct by using only the non-differential proteins and the empirical null. The likelihood is computed as proportional to the probability that the empirical null and the *non*-significant data are drawn from the same distribution (the *non*-significant data should be drawn from the null if proteins are correctly labeled as differential and non-differential).

The original, single-variate npCI used in protein identification computes the Dirichlet-based expected divergence between the null PSM data (generated by searching a decoy protein database and the remaining target PSM data. If we label the null data as α and the unidentified target data (*i.e.* the target data that remains after a given set of proteins are identified) as β , then the npCI for identification can be defined as follows:

$$\begin{aligned} \text{divergence}^{(ident)}(\alpha, \beta) &= e^{E_x \left[\log \left(\frac{\epsilon' PDF_{\alpha}(x) \epsilon' n_{\beta} PDF_{\beta}(x)}{\Gamma(1 + \epsilon' n_{\beta} PDF_{\beta}(x))} \right) \right]} \text{ (w.l.o.g)} \\ \text{npCI}^{(ident)} &= \text{divergence}^{(ident)}(\alpha, \beta) \times \text{divergence}^{(ident)}(\beta, \alpha) \end{aligned}$$

where PDF denotes the univariate probability density function of PSM scores, n denotes the number of PSMs (with zero eliminated), and ϵ' is the inverse of the size of the domain of x , the one-dimensional set of possible PSM scores. This (geometric mean) expected value is computed via one-dimensional numeric integration. The probability density functions are computed by smoothing the observed points with kernel density estimation^{31,32} (KDE).

If we denote the null control-control PSM data as α and the non-differential case-control data (*i.e.* the case-control data that remains after a given set of proteins are considered differential) as β , the multivariate npCI we introduce here can be defined as an expectation in a similar manner:

$$\begin{aligned} \text{divergence}^{(quant)}(\alpha, \beta) &= e^{E_{x,y} \left[\log \left(\frac{\epsilon' PDF_{\alpha}(x,y) \epsilon' n_{\beta} PDF_{\beta}(x,y)}{\Gamma(1 + \epsilon' n_{\beta} PDF_{\beta}(x,y))} \right) \right]} \text{ (w.l.o.g)} \\ \text{npCI}^{(quant)} &= \text{divergence}^{(quant)}(\alpha, \beta) \times \text{divergence}^{(quant)}(\beta, \alpha) \end{aligned}$$

where x is the log of the first intensity (*i.e.* the x -axis in Figure 2, y is the log of the second intensity (*i.e.* the y -axis in Figure 2, ϵ' is the inverse of the area of the domain of both x and y , and the PDF is the two-dimensional probability density function of log intensity scores of

the two compared channels. As in the original identification version, the expectation is computed using a numeric integral (although the quantification variant uses a two-dimensional numeric integral).

$npCI^{(ident)}$ uses the fact that eliminating all evidence from the correct set of present proteins should yield a remaining set of PSMs with a score distribution similar to decoy PSMs. $npCI^{(quant)}$ exploits a similar principle, but in two dimensions: correctly specifying which of the case-control proteins are differential and non-differential will result in non-differential two-dimensional log intensity vs. log intensity distribution that most closely resembling the empirical null intensity-intensity distribution. As with the single dimensional npCI, data are analyzed in a way that respects graphical dependencies: proposing a protein as differential means that any constituent peptide or peptide-spectrum match (PSM) evidence *may* be differential; the remaining peptide or PSM evidence is compared to the null distribution.

Technical replicates (either from a single MS/MS run or from different runs) can be aggregated automatically by the npCI by treating them as conditionally independent given the set of differential proteins.

The control-control empirical null distributions, as well as the similarity of the case-control distributions (with 0, 54, and 300 proteins labeled as differential) are shown in Figure 2.

Probability density functions (PDFs) are estimated via kernel density estimation^{31:32} (KDE): Each datum is replaced with a Gaussian kernel, and estimated density function is proportional the sum of these kernels (the proportionality constant is chosen so that the PDF integrates to 1). The degree of smoothing is determined by the width of the Gaussian kernels, which is sometimes referred to as the “bandwidth” parameter. A classic approach chooses the bandwidth that minimizes the mean squared error^{31:32} (MSE), but we maximize the similarity (*i.e.* the npCI likelihood) between the remaining PSM evidence and the empirical null PSM evidence (essentially this computes a projection between the remaining evidence and the empirical null evidence).

Corroborating our evaluation method

Our method for evaluating strategies for deciding what constitutes a “differential” protein was corroborated by analyzing lists of positive controls, which are known to be involved in mitosis and the anaphase promoting complex (APC) and negative controls, which are known to not substantially fluctuate in quantity during the cell cycle (Table 1).

These control proteins were determined in advance on the basis of publications and without input or analysis from our experimental data. The positive controls included a few proteins that were not identified in our experiment (*i.e.* there were no spectra matching their constituent peptides, regardless of whether they were differential or not). These proteins were re-labeled by looking for the most similar protein identified (*i.e.* the protein in the same family with any supporting spectral evidence, regardless of whether they were differential or not). A literature search revealed that all of these observed proteins were also positive controls: KIF23³³, UBE2S³⁴, CKAP5³⁵, CDC42³⁶, SKP1³⁷, and AURKA³⁸ replaced the unobserved positive controls KIF22³⁹, UBE2C⁴⁰, CKAP2⁴¹, CDC20⁴², SKP2⁴³, and AURKB⁴⁴, respectively (citations indicate the basis for involvement in mitosis or the APC). The following proteins were regarded as positive controls, but were not identified (regardless of whether they were differential or not): CCNA1⁴⁵, CCNA2⁴⁵, CCNB1⁴⁶, CDCA5⁴⁴, CENPF⁴⁷, CKS1B⁴³, CLSPN⁴⁸, FAM64A⁴⁹, GLS⁵⁰, GTSE1⁴², McI-1⁵¹, Nek9⁵², NINL⁵³, PFKFB3⁵⁴, PTTG1⁴⁰, RPL30⁵², RPS6KA4⁵², RRM2⁵⁵, SGOL1⁵⁶, and TRB3⁵⁷.

Actins and tubulins were used as negative controls because of their established quantitative stability throughout the cell cycle⁵⁸.

Discussion

Figure 2 indicates strong evidence that the number of differential proteins is between 0 (*i.e.* too few differential proteins in the ranking) and 300 (*i.e.* too many differential proteins in the ranking), and that the most likely number of differential proteins is close to 54 (near the mode of Figure 3). If follow-up experimental validation is to be performed, ultimately the limits lie with how many “significant” discoveries can reasonably be evaluated. If high-throughput follow-up is performed, the greatest utility comes from the most permissive of all plausible sets of differential proteins. In this case, our method can be used by the researcher in a subjective manner to discover where the ranking (of most differential to least differential) reaches diminishing returns. This can be interpreted as a confidence interval spanning plausible sets of differential proteins. On the other hand, if low-throughput follow-up is performed, or if follow-up experimental validation will not be performed (*e.g.* if researchers will simply query the significant proteins for particular GO terms), then the most likely protein set (*i.e.* the protein set with the greatest npCI likelihood, shown by the mode of Figure 3) is most appropriate.

Table 1 shows the most likely differential protein set contains several of the positive controls (negative controls labeled as non-differential would indicate false negatives), and contains none of the negative controls (negative controls labeled as differential would indicate false positives). Furthermore, the plausible differential region (*i.e.* sets of differential proteins from the ranking that have likelihoods substantially greater than zero) contains 13/16 of the identified positive controls and zero of the negative controls. Interestingly, the boundary of the plausible differential region is remarkably close to the commonly employed, arbitrary 1.2-fold change rule.

Figure 3 also shows that ranking from the arbitrary 3-PSM rule does not substantially improve results; the 3-PSM rule ranking gives a likelihood curve that is nearly identical to using a 1-PSM rule ranking. The principle difference between the two curves is that the 3-PSM rule ranking is compressed by the removal of subset proteins. On the surface this is a surprising result: HeLa cells are cultured human cells, and so they present a protein-to-peptide graph with substantial complexity (*i.e.* with many shared peptides), especially when searched against such a complex database. However, the 3-PSM rule ranking will bias against proteins with few detectable peptides. With such proteins low in the ranking, many false positive proteins will need to be admitted in order to eliminate the blue corners shown in Figure 2. We hypothesize that in an experiment with substantial coverage relative to the number of present proteins, including proteins without many PSMs or peptides will be detrimental, because their fold changes may simply due to shared peptides. Labeling such “hanger-on” proteins as differential will also eliminate their non-shared, non-differential PSMs, which should be drawn from the null distribution; eliminating these non-differential null PSMs will thus decrease the similarity between the leftover PSM distribution and the empirical null, lowering the npCI likelihood. In the end, we believe that a probabilistic protein ranking will eventually emerge as the superior strategy for ranking proteins from differential to non-differential. The npCI approach will make fair evaluation of such methods possible.

There are many pros and cons in deciding the ideal type of replicate for generating the empirical null comparison. In this work we have proposed a technical replicate (*i.e.* one plate of cells isolated and treated in a similar manner) strategy for making the null distribution. It may be possible that a biological replicate (*i.e.* two plates of cells grown

separately but with the same treatment) will be a better model for the empirical null; however, our approach marks a significant step toward formalizing such nulls, and can be applied to choosing the best type of empirical null distribution (it will be the strategy that maximizes the similarity between the non-differential PSMs and the null). Furthermore, from-scratch recreation of the null may not be necessary for every experiment: First, a single null may be applied to several case-control experiments if the conditions are very similar. Second, non-biologically interesting variation may be analyzed by looking at fold changes between isotope peaks. Another similarly motivated approach to using isotope peaks would be to look for the variability between PSMs from non-shared peptides from the same protein.

One important aspect of this method is that it can be applied regardless of how the data are processed (*i.e.* it can be run with other quantification software other than Scaffold). Our approach is paired trivially applicable to multiple acquisition schemes (*i.e.* data dependent as well as data independent). Distortions and biases in the intensities will apply to the control-control evidence as well as the case-control evidence, and thus will not alter the similarity between the two. Also, it is important to note a pleasing theoretical property of the method when applied to “bad data”: if there is no set of differential proteins that can be eliminated so that the leftover evidence resemble the control-control evidence or if insufficient data are provided (*i.e.* there is no reasonable differential protein set for the ranking family provided), then the smoothing will increase. This increase in smoothing will also yield a broader, more uniform likelihood curve with a less distinct mode, forcing the user to recognize the limitations on the data. Although greater coverage and greater numbers of replicates will increase the confidence in the mode, there is no minimum.

As long as the case-control and control-control data are processed in an identical manner, the non-differential case-control data should resemble the control-control data in the long run. This principle holds true even in the presence of peptide- and protein-specific biases and parameters, *e.g.* in the presence of intensity- or protein-specific variances. This is one of the key contributions of our approach: Rather than treat the data as individual points, which are used to compute a *p*-value by integrating the tails of the null, we compare the *entire* null to the remaining data (*i.e.* the data that are posited to be non-differential). By treating these data as whole populations, each of which may consist of a mixture of distinct sub-populations, these sub-populations will occur in both the non-differential case-control data and the control-control data; as a result, the non-differential case-control data should still resemble the control-control data in the long run. For example, highly abundant proteins represent an example of such a sub-population. A ranking (of proteins from differential to non-differential) that biases toward (or against) such proteins, will result in a remaining distribution depleted (or enriched) for high-intensity PSMs when compared to the control-control distribution, and thus a decreased npCI. When these sub-populations cannot be distinguished from one another based on the data used by the npCI, then such biases will not decrease the npCI, but will provide several alternative protein sets with nearly identical likelihood; as a result, the mode will be much more diffuse, indicating a lack of certainty that the mode represents the best estimate.

It is noteworthy to point out that creating multiple control-control experiments may not be necessary. Instead, they could be permuted and used to evaluate case-control samples, thus increasing the efficiency of experiments conducted with our approach. In the extreme, a single pair of control-control experiments could be used for several case-control experiments. However, aggregating replicate data sets by assuming conditional independence may no longer be valid in such strategies, because some data may be shared among the replicates (*e.g.*, using notation from Figure 1, a case-control sample from C1 vs. B1 compared to a control-control sample from B1 vs. A1 may not be conditionally

independent from a case-control sample from C2 vs. A1 compared to a control-control sample from B1 vs. A1). In practice, such a worry may prove to be pedantic, and it may be possible to aggregate data sets using a single or permuted controls.

It is also worthwhile to note that this general approach (*i.e.* using a control-control experiment to generate an empirical null, and then using the empirical null with a multivariate npCI to perform nonparametric evaluation) is not limited to TMT-LC-MS/MS; on the contrary, such an empirical null can easily be created by using SILAC or by using label-free quantification. For SILAC, an empirical null would be generated in an almost identical manner, by comparing a control with light isotopes to another control with heavy isotope labeling (of course, SILAC would use precursor MS intensities rather than the MS/MS intensities from TMT fragment ions used by TMT-LC-MS/MS analysis). As in the TMT example, the precursor MS intensities could be normalized to prevent batch effects. Likewise, an empirical null for a label-free approach would be created by separately analyzing two control samples, and using peptide-level spectral counts (normalized to prevent batch effects) as an indicator of quantity). Also, like in the TMT case, there would be no minimum or maximum number of replicates, so long as each have an empirical null.

Critical evaluation

Our method does not avoid all arbitrary parameters: notably, we use minimum values called pseudocounts to prevent the necessity of using zero values in log space. Because these zeros are the result of the instrument's precision (after all no intensity should ever truly be zero), more reasonable value for these minimum intensities could be derived using the instrument's sensitivity. Also, instruments with greater sensitivity will label such values with small nonzero values rather than zeros. Values labeled as missing because they are approximately zero could be handled similarly.

Likewise, the choice of a $(\log(intensity_1), \log(intensity_2))$ scale for these two-dimensional densities is somewhat arbitrary, and many such schemes are possible, including $(intensity_1, intensity_2)$ and the corresponding polar coordinates

$\left(\sqrt{\|intensity_1 - intensity_2\|}, \theta_{intensity_1 intensity_2} \right)$. Although these different scales will result in the same basic principle (projection of the non-differential case-control evidence to the control-control evidence), using different (*i.e.* relatively warped) spaces for this projection will increase or decrease the relative weight contributed by certain regions. This is reminiscent of the effects of nonlinear scaling in support vector machines (SVMs)⁶⁶, which may also influence the result for a similar reason.

Lastly, the smoothing performed by the npCI assumes that the leftover PSM distribution *can* resemble the null distribution when the differential proteins are eliminated. In general, this will be the case in the long run (both samples should asymptotically approach the identical distributions that they are drawn from), even in the presence of statistical artifacts such as TMT ratio distortion⁶⁷, as long as these artifacts are applied equally to the case and control populations. In short, although it is not necessary that an outlier in the case sample have a corresponding outlier from the *same* protein, peptide, or PSM in the control sample, an outlier in *some* protein, peptide, or PSM should exist (and will exist with sufficient sampling). In practice, without very deep sampling there may still be dissimilarity from extreme, infrequent outliers, which would have identical distributions in the long run, but which only appear in one data set due to too little sampling (*e.g.* contaminants). In this paper we remove these potential outliers by eliminating the contaminants; however, even when these potential outliers cannot be accounted for, and thus there exists no ranking where the distribution of leftover non-differential PSMs resemble the null distribution (*e.g.* because of

unknown contaminants or existence of an extremely rare outlier), Bayesian rules to minimize the mean squared error (MSE) can be used to smooth the data^{31;32}.

Although method is not computationally prohibitive (a single ranking processed in under 5 minutes on an Core i3 laptop), alternatives to numeric integration would make the method applicable to aggregating the information in joint distributions of many variables (as is done for aggregating peptide identification scores with Percolator²²).

Acknowledgments

This work was funded by NIH grants NS007473 and NS066973 (O.S.), W.R. Hearst Fellowship (E.C.), NS066973 (J.S.), and GM096319 and GM094844 (H.S.). This work was also supported by a grant from the Swedish Research Council (L.K.). Thanks to Manor Askenazi for his suggestion to use the isotope peaks to generate an empirical null in future work. We are also grateful to the reviewers, whose suggestions made this a better paper.

References

1. Nesvizhskii AI. A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics. *Journal of Proteomics*. 2010; 73(11):2092–2123. [PubMed: 20816881]
2. Serang O, Noble WS. A review of statistical methods for protein identification using tandem mass spectrometry. *Statistics and Its Interface*. 2012; 5(1):3–20. [PubMed: 22833779]
3. Li YF, Radivojac P. Computational approaches to protein inference in shotgun proteomics. *BMC bioinformatics*. 2012; 13(Suppl 16):S4.
4. Huang T, Wang J, Yu W, He Z. Protein inference: a review. *Briefings in bioinformatics*. 2012; 13(5):586–614. [PubMed: 22373723]
5. Sandberg R, Yasuda R, Pankratz DG, Carter TA, Del Rio JA, Wodicka L, Mayford M, Lockhart DJ, Barlow C. Regional and strain-specific gene expression mapping in the adult mouse brain. *Proceedings of the National Academy of Sciences of the United States of America*. 2000; 97:11038–11043. [PubMed: 11005875]
6. Roxas BAP, Li Q. Significance analysis of microarray for relative quantitation of lc/ms data in proteomics. *BMC bioinformatics*. 2008; 9(1):187. [PubMed: 18402702]
7. Oberg AL, Mahoney DW, Eckel-Passow JE, Malone CJ, Wolfinger RD, Hill EG, Cooper LT, Onuma OK, Spiro C, Therneau TM, et al. Statistical analysis of relative labeled mass spectrometry data from complex samples using anova. *Journal of Proteome Research*. 2008; 7(1):225–233. [PubMed: 18173221]
8. Ishihama Y, Schmidt T, Rappsilber J, Mann M, Hartl FU, Kerner MJ, Frishman D. Protein abundance profiling of the escherichia coli cytosol. *BMC genomics*. 2008; 9(1):102. [PubMed: 18304323]
9. Granholm V, Noble WS, Käll L. On using samples of known protein content to assess the statistical calibration of scores assigned to peptide-spectrum matches in shotgun proteomics. *Journal of Proteome Research*. 2011; 10(5):2671–2678. [PubMed: 21391616]
10. Serang O, Paulo J, Steen H, Steen JA. A non-parametric cutout index for robust evaluation of identified proteins. *Molecular & Cellular Proteomics*. 2013; 12(3):807–812. [PubMed: 23292186]
11. Alm H, Scholz B, Fischer C, Kulima K, Viberg H, Eriksson P, Dencker L, Stigson M. Proteomic evaluation of neonatal exposure to 2, 2, 4, 4, 5-pentabromodiphenyl ether. *Environmental health perspectives*. 2006; 114(2):254. [PubMed: 16451863]
12. Corzett TH, Fodor IK, Choi MW, Walsworth VL, Chromy BA, Turteltaub KW, McCutchen-Maloney SL. Statistical analysis of the experimental variation in the proteomic characterization of human plasma by two-dimensional difference gel electrophoresis. *Journal of proteome research*. 2006; 5(10):2611–2619. [PubMed: 17022632]
13. Nissom PM, Sanny A, Kok YJ, Hiang YT, Chuah SH, Shing TK, Lee YY, Wong TK, Hu W, Sim M, et al. Transcriptome and proteome profiling to understanding the biology of high productivity cho cells. *Molecular biotechnology*. 2006; 34(2):125–140. [PubMed: 17172658]

14. Keenan J, Murphy L, Henry M, Meleady P, Clynes M. Proteomic analysis of multidrug-resistance mechanisms in adriamycin-resistant variants of dlkp, a squamous lung cancer cell line. *Proteomics*. 2009; 9(6):1556–1566. [PubMed: 19242932]
15. Dore S, Kar S, Quirion R, et al. Rediscovering an old friend, igf-i: potential use in the treatment of neurodegenerative diseases. *Trends in neurosciences*. 1997; 20(8):326. [PubMed: 9246719]
16. Kohno K, Izumi H, Uchiumi T, Ashizuka M, Kuwano M. The pleiotropic functions of the y-box-binding protein, yb-1. *Bioessays*. 2003; 25(7):691–698. [PubMed: 12815724]
17. Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences*. 2001; 98(9):5116–5121.
18. Yang IV, Chen E, Hasseman JP, Liang W, Frank BC, Wang S, Sharov V, Saeed AI, White J, Li J, Lee NH, Yeatman TJ, Quackenbush J. Within the fold: assessing differential expression measures and reproducibility in microarray assays. *Genome Biol*. 2002; 3(11):1–0062.
19. Serang O, Moruz L, Hoopmann M, Käll RL. Recognizing uncertainty increases robustness and reproducibility of mass spectrometry-based protein inferences. *Journal of Proteome Research*. 2012; 11(12):5586–91. [PubMed: 23148905]
20. Granholm V, Navarro J, Noble WS, Käll L. Determining the calibration of confidence estimation procedures for unique peptides in shotgun proteomics. *Journal of proteomics*. 2012
21. Searle BC, Dasari S, Wilmarth PA, Turner M, Reddy AP, David LL, Nagalla SR. Identification of protein modifications using ms/ms de novo sequencing and the opensea alignment algorithm. *Journal of proteome research*. 2005; 4(2):546–554. [PubMed: 15822933]
22. Käll L, Canterbury J, Weston J, Noble WS, MacCoss MJ. A semi-supervised machine learning technique for peptide identification from shotgun proteomics datasets. *Nature Methods*. 2007; 4:923–25. [PubMed: 17952086]
23. Käll L, Storey J, Noble WS. Nonparametric estimation of posterior error probabilities associated with peptides identified by tandem mass spectrometry. *Bioinformatics*. 2008; 24(16):i42–i48. [PubMed: 18689838]
24. Choi H, Nesvizhskii AI. Semisupervised model-based validation of peptide identifications in mass spectrometry-based proteomics. *Journal of Proteome Research*. 2008; 7(1):254–265. [PubMed: 18159924]
25. Kessner D, Chambers M, Burke R, Agus D, Mallick P. Proteowizard: open source software for rapid proteomics tools development. *Bioinformatics*. 2008; 24(21):2534–2536. [PubMed: 18606607]
26. Chambers MC, Maclean B, Burke R, Amodei D, Ruderman DL, Neumann S, Gatto L, Fischer B, Pratt P, Egertson J, et al. A cross-platform toolkit for mass spectrometry and proteomics. *Nature biotechnology*. 2012; 30(10):918–920.
27. Perkins DN, Pappin DJC, Creasy DM, Cottrell JS. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis*. 1999; 20:3551–3567. [PubMed: 10612281]
28. Searle BC. Scaffold: A bioinformatic tool for validating ms/ms-based proteomic studies. *PROTEOMICS*. 2010; 10(6):1265–1269. [PubMed: 20077414]
29. Keller BO, Sui J, Young AB, Whittall RM. Interferences and contaminants encountered in modern mass spectrometry. *Analytica chimica acta*. 2008; 627(1):71–81. [PubMed: 18790129]
30. Cui X, Churchill GA, et al. Statistical tests for differential expression in cDNA microarray experiments. *Genome Biol*. 2003; 4(4):210. [PubMed: 12702200]
31. Parzen E. On estimation of a probability density function and mode. *The annals of mathematical statistics*. 1962; 33(3):1065–1076.
32. Rosenblatt M. Remarks on some nonparametric estimates of a density function. *The Annals of Mathematical Statistics*. 1956:832–837.
33. Nislow C, Lombillo VA, Kuriyama R, McIntosh JR. A plus-end-directed motor enzyme that moves antiparallel microtubules in vitro localizes to the interzone of mitotic spindles. 1992
34. Garnett MJ, Mansfeld J, Godwin C, Matsusaka T, Wu J, Russell P, Pines J, Venkitaraman AR. Ube2s elongates ubiquitin chains on APC/C substrates to promote mitotic exit. *Nature Cell Biology*. 2009; 11(11):1363–1369.

35. van der Vaart B, Manatschal C, Grigoriev I, Olieric V, Gouveia SM, Bjeli S, Demmers J, Vorobjev I, Hoogenraad CC, Michel OS, et al. Slain2 links microtubule plus end– tracking proteins and controls microtubule growth in interphase. *The Journal of cell biology*. 2011; 193(6): 1083–1099. [PubMed: 21646404]
36. Ocegüera-Yanez F, Kimura K, Yasuda S, Higashida C, Kitamura T, Hiraoka Y, Haraguchi T, Narumiya S. Ect2 and mgcracgap regulate the activation and function of cdc42 in mitosis. *The Journal of cell biology*. 2005; 168(2):221–232. [PubMed: 15642749]
37. Olsen JV, Vermeulen M, Santamaria A, Kumar C, Miller ML, Jensen LJ, Gnad F, Cox J, Jensen TS, Nigg EA, et al. Quantitative phosphoproteomics reveals widespread full phosphorylation site occupancy during mitosis. *Science signaling*. 2010; 3(104):ra3. [PubMed: 20068231]
38. Honda K, Mihara H, Kato Y, Yamaguchi A, Tanaka H, Yasuda H, Furukawa K, Urano T, et al. Degradation of human aurora2 protein kinase by the anaphase-promoting complexubiquitin-proteasome pathway. *Oncogene*. 2000; 19(24):2812. [PubMed: 10851084]
39. Feine O, Zur A, Mahbubani H, Brandeis M. Human kid is degraded by the apc/ccdh1 but not by the apc/ccdc20. *Cell Cycle*. 2007; 6(20):2516–2523. [PubMed: 17726374]
40. Rape M, Kirschner MW. Autonomous regulation of the anaphase-promoting complex couples mitosis to s-phase entry. *Nature*. 2004; 432(7017):588–595. [PubMed: 15558010]
41. Seki A, Fang G. Ckap2 is a spindle-associated protein degraded by apc/c-cdh1 during mitotic exit. *Journal of Biological Chemistry*. 2007; 282(20):15103–15113. [PubMed: 17376772]
42. Pffeger CM, Kirschner MW. The ken box: an apc recognition signal distinct from the d box targeted by cdh1. *Genes & development*. 2000; 14(6):655–665. [PubMed: 10733526]
43. Bashir T, Dorrello NV, Amador V, Guardavaccaro D, Pagano M. Control of the scf/skp2–cks1 ubiquitin ligase by the apc/ccdh1 ubiquitin ligase. *Nature*. 2004; 428(6979):190–193. [PubMed: 15014502]
44. Rankin S, Ayad NG, Kirschner MW. Sororin, a substrate of the anaphase-promoting complex, is required for sister chromatid cohesion in vertebrates. *Molecular Cell*. 2005; 18(2):185–200. [PubMed: 15837422]
45. Huang X, Summers MK, Pham V, Lill JR, Liu J, Lee G, Kirkpatrick DS, Jackson PK, Fang G, Dixit VM. Deubiquitinase usp37 is activated by cdk2 to antagonize apc; sup; cdh1/sup; and promote s phase entry. *Molecular Cell*. 2011; 42(4):511–523. [PubMed: 21596315]
46. Casaletto JB, Nutt LK, Wu Q, Moore JD, Etkin LD, Jackson PK, Hunt T, Kornbluth S. Inhibition of the anaphase-promoting complex by the Xnf7 ubiquitin ligase. *The Journal of Cell Biology*. 2005; 169(1):61–71. [PubMed: 15824132]
47. Gurden MDJ, Holland AJ, van Zon W, Tighe A, Vergnolle MA, Andres DA, Spielmann HP, Malumbres M, MF Wolthuis R, Cleveland DW, et al. Cdc20 is required for the post-anaphase, ken-dependent degradation of centromere protein f. *Journal of Cell Science*. 2010; 123(3):321–330. [PubMed: 20053638]
48. Bassermann F, Frescas D, Guardavaccaro D, Busino L, Peschiaroli A, Pagano M. The cdc14b-cdh1-plk1 axis controls the g2 dna-damage-response checkpoint. *Cell*. 2008; 134(2):256–267. [PubMed: 18662541]
49. Zhao W, Fang G. Anillin is a substrate of anaphase-promoting complex/cyclosome (apc/c) that controls spatial contractility of myosin during late cytokinesis. *Journal of Biological Chemistry*. 2005; 280(39):33516–33524. [PubMed: 16040610]
50. Colombo SL, Palacios-Callender M, Frakich N, De Leon J, Schmitt CA, Boorn L, Davis N, Moncada S. Anaphase-promoting complex/cyclosome-cdh1 coordinates glycolysis and glutaminolysis with transition to s phase in human t lymphocytes. *Proceedings of the National Academy of Sciences*. 2010; 107(44):18868–18873.
51. Harley ME, Allan LA, Sanderson HS, Clarke PR. Phosphorylation of mcl-1 by cdk1–cyclin b1 initiates its cdc20-dependent destruction during mitotic arrest. *The EMBO Journal*. 2010; 29(14): 2407–2420. [PubMed: 20526282]
52. Merbl Y, Kirschner MW. Large-scale detection of ubiquitination substrates using cell extracts and protein microarrays. *Proceedings of the National Academy of Sciences*. 2009; 106(8):2543–2548.

53. Wang Y, Zhan Q. Cell cycle-dependent expression of centrosomal ninein-like protein in human cells is regulated by the anaphase-promoting complex. *Journal of Biological Chemistry*. 2007; 282(24):17712–17719. [PubMed: 17403670]
54. Tudzarova S, Colombo SL, Stoeber K, Carcamo S, Williams GH, Moncada S. Two ubiquitin ligases, apc/c-cdh1 and skp1-cull1-f (scf)-trcp, sequentially regulate glycolysis during the cell cycle. *Proceedings of the National Academy of Sciences*. 2011; 108(13):5278–5283.
55. Cotto-Rios XM, Jones MJK, Busino L, Pagano M, Huang TT. Apc/ccdh1 dependent proteolysis of usp1 regulates the response to uv-mediated dna damage. *The Journal of Cell Biology*. 2011; 194(2):177–186. [PubMed: 21768287]
56. Karamysheva Z, Diaz-Martinez LA, Crow SE, Li B, Yu H. Multiple anaphase-promoting complex/cyclosome degrons mediate the degradation of human sgo1. *Journal of Biological Chemistry*. 2009; 284(3):1772–1780. [PubMed: 19015261]
57. Ohoka N, Sakai S, Onozaki K, Nakanishi M, Hayashi H. Anaphase-promoting complex/cyclosome-cdh1 mediates the ubiquitination and degradation of trb3. *Biochemical and Biophysical Research Communications*. 2010; 392(3):289–294. [PubMed: 20064487]
58. Thellin O, Zorzi W, Lakaye B, De Borman B, Coumans B, Hennen G, Grisar T, Igout A, Heinen E. Housekeeping genes as internal standards: use and limits. *Journal of Biotechnology*. 1999; 75(2):291–295. [PubMed: 10617337]
59. Ke P, Kuo Y, Hu C, Chang Z. Control of dtp pool size by anaphase promoting complex/cyclosome is essential for the maintenance of genetic stability. *Genes & development*. 2005; 19(16):1920–1933. [PubMed: 16103219]
60. Lindon C, Pines J. Ordered proteolysis in anaphase inactivates plk1 to contribute to proper mitotic exit in human cells. *The Journal of cell biology*. 2004; 164(2):233–241. [PubMed: 14734534]
61. Stewart S, Fang G. Anaphase-promoting complex/cyclosome controls the stability of tpx2 during mitotic exit. *Molecular and cellular biology*. 2005; 25(23):10516–10527. [PubMed: 16287863]
62. Zhu C, Lau E, Schwarzenbacher R, Bossy-Wetzel E, Jiang W. Spatiotemporal control of spindle midzone formation by prc1 in human cells. *Proceedings of the National Academy of Sciences*. 2006; 103(16):6196–6201.
63. Li L, Zhou Y, Sun L, Xing G, Tian C, Sun J, Zhang L, He F. Nusap is degraded by apc/c-cdh1 and its overexpression results in mitotic arrest dependent of its microtubules' affinity. *Cellular Signalling*. 2007; 19(10):2046–2055. [PubMed: 17618083]
64. Qi W, Yu H. Ken-box-dependent degradation of the bub1 spindle checkpoint kinase by the anaphase-promoting complex/cyclosome. *Journal of Biological Chemistry*. 2007; 282(6):3672–3679. [PubMed: 17158872]
65. Ghoshal K, Datta J, Majumder S, Bai S, Kutay H, Motiwala T, Jacob ST. 5-aza-deoxycytidine induces selective degradation of dna methyltransferase 1 by a proteasomal pathway that requires the ken box, bromo-adjacent homology domain, and nuclear localization signal. *Molecular and Cellular Biology*. 2005; 25(11):4727–4741. [PubMed: 15899874]
66. Noble, WS. Support vector machine applications in computational biology. In: Schoelkopf, B.; Tsuda, K.; Vert, J-P., editors. *Kernel methods in computational biology*. Cambridge, MA: MIT Press; 2004. p. 71-92.
67. Ting L, Rad R, Gygi SP, Haas W. Ms3 eliminates ratio distortion in isobaric multiplexed quantitative proteomics. *Nature methods*. 2011; 8(11):937–940. [PubMed: 21963607]

Use npCI to evaluate strategies for identifying differential proteins:
Case-control PSM evidence that *cannot* come from differential proteins is drawn from the same distribution as the control-control PSM evidence.

Therefore, the similarity between the distributions of non-differential case-control PSM evidence and control-control PSM evidence approximates the quality of the differential protein set.

Figure 1. Experimental setup for TMT-LC-MS/MS

An empirical null distribution is created by comparing two untreated samples (A and B), which should not exhibit biologically interesting quantitative variation. Cells in experimental sample C are synchronized by arresting them in prometaphase, yielding differential quantities in proteins involved in mitosis (when compared to the asynchronous cells in samples A and B). Using TMT, two technical replicates can be generated from one experiment: the first technical replicate labels A, B, and C with +126, +127, and +128 Da tags (yielding A1, B1, and C1), and the second labels A, B, and C with +129, +130, and +131 Da tags (yielding A2, B2, C2). The npCI is used to aggregate the two technical replicates, and provide a single probabilistic evaluation for any given set of putatively differential proteins or a ranking of proteins (from more to less differential).

Figure 2. Empirical null distributions and fold changes relative to null

(a) The empirical null distribution of PSM intensities from B1 compared to A1 (from replicate 1) and B2 compared to A2 (from replicate 2). Most PSMs lie along the diagonal (with the low-intensity PSMs dominating the mass of the distribution); however, off-diagonal technical variation is apparent. (b) For varying numbers of proteins considered “differential” in the ranking, we show the local log fold change of the case-control PSM densities relative to the control-control figure from (a)). Columns of (b) eliminating the top 0, 54, and 300 differential proteins. The off-diagonal (*i.e.* the top-left and bottom-right corners) points indicate differential PSMs, which have a greater intensity in either the treated (*i.e.* case) or untreated (*i.e.* control) samples. A blue point indicates a local enrichment of PSMs relative to the null and a red point indicates a local depletion relative to the null. When no proteins are considered as differential, the case-control off-diagonal contains many more (hence blue) PSMs than are found in the off-diagonal of the null (*i.e.* when 0 proteins are considered as differential, there are too many remaining PSMs with high intensity and high fold change). When 54 proteins are considered differential (close to the npCI maximum likelihood estimate) and their constituent PSMs are eliminated, the distribution of remaining PSMs looks nearly identical to the empirical null, and thus the distribution is nearly entirely white (a local log fold change of 0 or a fold change of 1). When too many (*i.e.* 300) proteins are considered differential and their constituent PSMs are eliminated, the blue off-diagonal enrichment has been successfully removed, but too few PSMs remain in the high-intensity region along the diagonal. For this reason, identifying 300 differential proteins results in a depletion (hence red) in the top-right corner. The npCI score is not only a function of the most extreme changes (*i.e.* the darkest blue and red), it is a function of the *overall deviation*, and is also influenced by the size of the off-white area.

Figure 3. The npCI for different numbers of differential proteins

The npCI curves for three rankings (proteins matching at least 1, at least 2, and at least 3 PSMs) of proteins by mean absolute log fold change are shown. The npCI is the likelihood that the putative differential protein set is correct, as established using the similarity between the empirical null evidence and the non-differential evidence. The three rankings all give very similar results and the ≥ 1 PSM ranking is very slightly superior, despite the convention that at least three PSMs or peptides are necessary to estimate protein fold change. Both technical replicates closely agree for all rankings.

Table 1
Known positive controls (mitotic / APC) and negative controls (actins and tubulins)

Positive and negative controls are shown, along with their rank from the 1-PSM ranking. The largest npCI containing each protein is also shown. This value is shown in place of the raw npCI because the top hit (TK1) alone still only achieves a fairly low npCI (see Figure 3); however, the top hit is included in the most likely protein set (the mode of Figure 3). Each row also shows the exponentiated average log absolute fold change (*i.e.* a protein where the average case/control intensity ratio is 0.833 will be labeled as 1.2-fold change). The most permissive of the plausible fold change thresholds is between 1.20 and 1.28. Using a 1.28-fold change threshold will yield a likelihood just under $\frac{1}{20}$ times the maximum likelihood, and using a 1.20-fold change (the next threshold for the positive control proteins) will yield a likelihood less than $\frac{1}{300,000}$ times as likely as the maximum likelihood.

Proteins with known role in mitosis / APC (positive controls):			
Gene	Rank	Highest npCI containing	Fold change
TK1 ⁵⁹	1	1.0	3.42
PLK1 ⁶⁰	2	1.0	3.21
KIF23 ³³	3	1.0	2.95
TPX2 ⁶¹	4	1.0	2.65
AURKA ³⁸	7	1.0	2.32
ANLN ⁴⁹	8	1.0	2.24
PRC1 ⁶²	14	1.0	2.02
UBE2S ³⁴	22	1.0	1.89
NUSAP1 ⁶³	55	0.775	1.70
BUB1B ⁶⁴	62	0.0539	1.59
CKAP5 ³⁵	68	0.0484	1.55
CALM2 ⁵²	96	0.0484	1.47
CDC42 ³⁶	189	0.0484	1.28
SKP1 ³⁷	537	3.13e-6	1.20
DNMT1 ⁶⁵	2176	1.11e-6	1.04
CCNB2 ⁴⁶	2531	2.31e-23	1.02
Actins and tubulins (negative controls):			
Gene	Rank	Highest npCI containing	Fold change
Actins:			
ACTZ	830	1.11e-6	1.15
ACTN4	2162	1.11e-6	1.04
ACTY	2574	1.91e-30	1.02
ACTB	5237	0	1.01
ACTG	5237	0	1.01
Tubulins:			
TBB5	1650	1.11e-6	1.07

Proteins with known role in mitosis / APC (positive controls):			
Gene	Rank	Highest npCI containing	Fold change
TBB6	2189	1.11e-6	1.04
TBB2C	2691	0	1.01