Published in final edited form as:

Proteomics. 2007 August ; 7(16): 2904–2919. doi:10.1002/pmic.200700267.

Protein abundance ratios for global studies of prokaryotes

Qiangwei Xia^{1,2}, Erik L. Hendrickson², Tiansong Wang^{1,2}, Richard J. Lamont³, John A. Leigh², and Murray Hackett¹

¹ Department of Chemical Engineering, University of Washington, Seattle, WA, USA

² Department of Microbiology, University of Washington, Seattle, WA, USA

³ Department of Oral Biology, University of Florida, Gainesville, FL, USA

Abstract

The use of multidimensional capillary HPLC combined with tandem mass spectrometry has allowed high qualitative and quantitative proteome coverage of prokaryotic organisms. The determination of protein abundance change between two or more conditions has matured to the point that false discovery rates can be very low and for smaller proteomes coverage is sufficiently high to explicitly consider false negative error. Selected aspects of using these methods for global protein abundance assessments are reviewed. These include instrumental issues that influence the reliability of abundance ratios; a comparison of sources of non-linearity, errors, and data compression in proteomics and spotted cDNA arrays; strengths and weaknesses of spectral counting versus stable isotope metabolic labeling; and a survey of microbiological applications of global abundance analysis at the protein level. Proteomic results for two organisms that have been studied extensively using these methods are reviewed in greater detail. Spectral counting and metabolic labeling data are compared and the utility of proteomics for global gene regulation studies are discussed for the methanogenic Archaeon *Methanococcus maripaludis*. The oral pathogen *Porphyromonas gingivalis* is discussed as an example of an organism where a large percentage of the proteome differs in relative abundance between the intracellular and extracellular phenotype.

Keywords

Differential protein abundance; Tandem mass spectra; Quantitative analysis; Multidimensional liquid chromatography; Prokaryote; False negative; False positive

1 Introduction

The concept of using global abundance measurements to study cellular function in microbial cells has its roots in genomics and the transcription microarray community, and has more recently been applied at the level of expressed protein using the tools of proteomics. Proteomics is both more complex experimentally and less developed as a science relative to microarray technology for global transcription measurements, therefore the number of published studies that have achieved sufficiently high qualitative and quantitative coverage to be truly global at a whole cell level are few. In terms of cost and complexity, such studies are more comparable to SAGE or genome sequencing than they are to microarray experiments. The reasons for persisting in such a costly endeavor are several, but can be distilled down to one central issue: measurements at the protein level are required to fully understand certain aspects of gene

regulation and functional genomics. One of the key assumptions behind systems biology is that, given a data matrix of transcription, expressed protein, PTM status and metabolite measurements, it should be possible to construct predictive models of cell function [1–3].

The need for more rigor and application of statistical principles to proteomics is now widely recognized, as has been discussed in recent reviews of statistical methods and data analysis in proteomics [4,5]. The review by Urfer *et al.* [6] emphasizes methods for distinguishing significant from insignificant abundance change. Zhang *et al.* [7] have discussed the comparative merits of different approaches to multiple hypothesis testing for large sets of abundance ratios determined using spectral counting, including Fisher's exact test and the *G*-test. Old and coworkers [8] and Xia *et al.* [9] have described *G* statistics applied to spectral counting. Xia *et al.* applied the *q*-value statistic [10,11], a measure of FDR (false discovery rate), to abundance ratios determined using signal intensities for unlabeled samples and spectral counting. The *q*-statistic served as a correction for multiple hypotheses applied to the *p*-values generated using a *t*-test for intensity data and a *G*-test for spectral counting [9].

Simply stated, FDR in this review refers to the fraction of the proteins judged to show a change in abundance between two states, where in fact there is no change; i.e. false positives divided by all positives at some predetermined level of significance. For a more rigorous definition, see Storey and Tibshirani [10] and the references therein. While strategies have evolved to assess qualitative and quantitative false positive error (Type 1 error), the issue of controlling false negative (Type 2) quantitative errors in proteomics has not been explored. Microarray analysis has reached a level of maturity and cost effectiveness such that formal assessments of both types of quantitative error can be accomplished at the transcriptional level [12,13], but Type 2 error is often ignored in practice. It is not practical at the present time to repeat a high coverage global proteomics experiment with many biological replicates to adequately control Type 2 error, noting that the actual number of such replicates is best decided by a power analysis [14,15], informed by the difference in protein abundance that needs to be reliably detected in order to satisfy the goals of the study. Despite the low number of replicates usually employed, there are features of "bottom up" proteomics data [4,16] that can allow biologically meaningful global assessments of quantitative error, while restricting the amount of raw data collection required to a level that is at least achievable with instrumentation and methods that exist today. Each protein will be associated with a number of proteolytic fragments, each one of which can potentially yield a separate estimate of protein abundance.

The present review is restricted to studies of prokaryotes, a reflection of the research interests of the authors. An implicit assumption of protein identifications based on searching databases against mass spectra acquired from proteolytic fragments is that the old ideas of "one gene, one polypeptide" still hold true for purposes of unambiguous qualitative identification. Even for microbial systems, this is no longer accepted dogma without many exceptions, but it is a much closer approximation to reality than would be the case with any higher organism. How well the database search paradigm works in practice for whole cell studies is determined largely by the degree to which a one-to-one correspondence exists between the amino acid sequence inferred from the annotated genomic database and the amino acid sequence of the experimentally observed protein [17].

The choice of topics discussed here is highly selective and concerns key issues related to quantitative reproducibility that should be of general interest, especially among microbiologists that work with both proteomic and transcription data. In addition, we discuss examples of biologically driven studies where the newer high coverage and mostly gel-free quantitative proteomics methods have been applied to prokaryotic systems on a cellular scale. Yeast proteomic studies are not discussed, except peripherally. Their relevance to the present discussion is that these unicellular eukaryotes have been widely used as model systems for

proteomic technology development to the extent that the most cited high coverage shotgun proteomics methods were originally developed using the proteome of *Saccharomyces cerevisiae* as a model system [18–20]. The application of proteomic methods to *S. cerevisiae* biology has been reviewed [21].

Lastly, by way of introductory comments, the authors acknowledge differing opinions exist regarding the terms “protein expression ratio” and “protein abundance ratio.” “Abundance ratio” emphasizes that what is being measured is really a balance between biosynthetic and degradative processes. A more subtle point is that regulation of a metabolic pathway may involve other factors that are independent mRNA or protein expression. The term “expression ratio” is solidly entrenched in the transcription array literature. We therefore use the term “expression ratio” to refer to transcription data, and “abundance ratio” for protein measurements. When referring specifically to the biosynthesis of protein, independent of degradation, the term “expression” is used.

2 Instrumental aspects, proteome coverage and reproducibility issues

2.1 Strengths and Limitations of multidimensional HPLC coupled with tandem mass spectrometry for global protein abundance studies

Since the development of multidimensional capillary HPLC for large-scale proteomic studies in the 1990s [22–24], and its subsequent refinement and popularization, e.g. as MudPIT (multidimensional protein identification technology) [18–20], such approaches to large-scale proteomics now find application in studies that are primarily of a biological character, rather than focused on analytical methods development. Prokaryotic organisms represent good model systems for applications of this technology, given their relatively small genomes, uncomplicated gene regulation relative to higher organisms, and what is generally a more straightforward relationship between the detected proteolytic fragment and the whole intact protein than what can be easily established for proteins extracted from the cells of eukaryotes. Advantages have included vastly improved proteome coverage, and more reliable quantitation relative to 2-D PAGE or 2-D DIGE. The state-of-the-art is changing rapidly with respect to quantitation, as illustrated recently by high resolution orbitrap data for known abundance ratios in *S. cerevisiae* [25]. The time and expense required for such global proteomic analyses is lessening while the amount of biologically useful information gained is increasing proportionately. The mass spectrometers acquire data faster, the separations are more reproducible, inexpensive multiprocessor computing clusters allow more rapid database searches and the cost of storing Terabyte quantities of data has dropped to a level well within the budgets of individual investigators, about \$500 US per Terabyte at the time this review is being written. New computational approaches allow the mining of data for information that, up until now, has been latent in the dataset but not accessible. An example of the last point would be the global assessment of PTM status [26].

Disadvantages of the multidimensional HPLC/tandem MS approach have included experimental cost and complexity, the relatively high degree of skill required, the loss of information about the intact, whole protein that is intrinsic to “bottom up” methods, the requirement for metabolic stable isotope labeling in order to achieve the best analytical figures-of-merit for quantitation [27–29], and the absence of available software to facilitate data handling on a cellular scale. Various solutions have been proposed to these limitations, including the use of spectral counting methods as an alternative approach to metabolic labeling with stable isotopes [30–33]. There is a sense amongst experienced users that when this approach is applied to whole cell studies more separation power is still needed, even when combined with extensive pre-fractionation and higher resolution [25] mass-to-charge measurements.

More fundamental problems remain, however, not the least of which lies in the nature of the CID spectrum itself. The better the proteome coverage in a qualitative sense, the better the quantitative coverage, and the more reliable the abundance ratios derived from the mass spectral data. Of necessity, such global experiments are typically performed using one set of instrumental tuning conditions that have been optimized to yield the largest number of usable precursor (MS^1) and product ($MS^2 \dots MS^n$) ions over the duration of the experiment. Under such conditions, only a small percentage of the total data collected leads to a database match of acceptable quality [34]. With the LCQ 3-D (three-dimensional) ion trap formerly used in the senior author's laboratory until 2003, we typically retained about 8% of the total data collected. With the LTQ based-system used at present [35], that number is now about 20% for the global studies of *P. gingivalis* and *M. maripaludis* cited in this review. The instrumental characteristics that stand behind this improvement in overall performance have been described [36,37]. Despite recent improvements, the number of CID spectra with an easily interpreted and unambiguous pattern of a, b and y ions are relatively few, a situation that can lead to false positive results in the hands of investigators who fail to fully appreciate the natural tendency of database searching algorithms to force the best fit of spectrum to database. Such issues surrounding database searching have been reviewed [34]. There are additional reasons why CID spectra yield poor matches, including microbial strain differences (genome sequence strain is different than experimental strain), mutations, PTMs, contamination from growth media containing protein components and background contamination in the general laboratory environment. The use of electron transfer dissociation [38–40] as an alternative to conventional CID shows promise as a means of generating fragmentation spectra that are easier to interpret, more informative of amino acid sequence generally, more likely to retain features indicative of labile PTMs and more informative for longer peptide sequences, well beyond the limits imposed by the energetics of conventional CID. However, in its present state of development it is not a high-throughput method. Too much experimental optimization is required to apply ETD to global studies at the present time, and quantitation using ETD has received little attention thus far in the peer reviewed literature.

Another complicating factor is that capillary separations are quite unforgiving of minor errors that would often not cause problems with separations performed using larger columns at higher flow rates. One example is the extreme sensitivity to air bubbles and leaks that often cannot be detected visually or with conventional leak detection methods. Particularly at a time when peptide identifications are now often based on a single scan mass spectrum acquired over a few milliseconds [41], slight, short-duration instabilities in the electrospray used to ionize the peptides as they emerge from the final stage of separation actually have a more dramatic effect on the composition of the final, processed dataset than they did a few years ago, when signal averaging over several scans of LC/MS data was the norm [36,42]. The electrospray, as with the separation itself, is affected adversely by subtle leaks. Such instabilities also have their basis in part due to changes in solvent and dissolved gas composition, analyte concentration and solution ionic strength that take place during gradient elution, as well as the constantly changing physical condition of the spray tip and capillary HPLC column. Such subtle experimental issues can have a dramatic and adverse effect on quantitative proteome coverage and measures of statistical reliability assigned to the abundance ratios in the fully processed data. While not as significant as issues related to electrospray instabilities during gradient elution, about 2% of the single scan CID (MS^2) spectra collected using an LTQ instrument under carefully controlled isocratic conditions showed indications of spectral distortion due to mass spectrometer related centroider errors [42].

Another issue is that background contamination becomes critical in a global analysis for reasons beyond those that are obvious and common to smaller scale experiments. At present, all global experiments under-sample, even in the most optimal circumstances [43]. This is currently a fundamental limitation of mass spectrometry based proteomics for global studies.

When one is operating under a regimen where the amount of data collected is nearly always inadequate with respect to the biological goals of the study, contamination robs precious duty cycle time. The more mass spectral scans acquired for contaminants, the fewer precious scans are available for the proteolytic fragments of primary interest. This indirectly affects quantitation by reducing the number of stable isotope pairs, spectral counts or intensity measurements at natural isotopic abundance available for quantitation. The sampling relationships between analyte and contaminant in such large-scale experiments are complex, but nonetheless demonstrate patterns that are reproducible and predictable to some degree. Detection and quantitation of proteins that are large, with many enzymatic digestion sites, and abundant in terms of copy number, is proportionately less affected by sampling competition from contaminants than proteins that are small and (or) less abundant. Within a given protein, certain proteolytic fragments are often favored dramatically in terms of overall detectability, often dominating the total ion current to the point of suppressing signals from other peptide species present in solution. This has led to the concept of “proteotypic peptides” [44], peptides that can be viewed as markers for the presence of a given protein if the sequence is unique. These are all critical considerations for quantitation, because the number of qualitative peptide “hits” must in general be maximized in order to shift the number of observations that exceed a certain threshold for quantitative reliability as far to the right as possible in plots of abundance ratios versus the total number of observations in the pool used to calculate the ratio. Examples of such plots are given in Figures 1–5. This relationship will be discussed in greater detail in Section 2.3. In the study of internalized *P. gingivalis* discussed in section 3.3, background contamination from proteins in the growth media and human protein derived from gingival epithelial cells more than doubled the time and effort required to complete the study, despite the relative ease with which the SEQUEST search algorithm [45] could distinguish tryptic fragments derived from *P. gingivalis* and those with other origins described above. That fortuitous state of affairs has its basis in the relative lack of sequence similarity between *P. gingivalis* and higher organisms. From a proteomic viewpoint, finding unique and unambiguous proteolytic fragments that distinguish a bacterial protein from that of a human or a cow is usually an easier proposition than distinguishing, for example, a human gingival cell protein from a growth medium component of bovine origin.

2.2 Sources of non-linearity and error: transcription microarrays and MudPIT proteomics compared

In comparing uncertainty in transcriptome arrays to MudPIT proteomics we will limit the discussion to two color printed PCR product arrays. However, many of the concerns with these arrays are also applicable to other transcriptome array techniques. The number of independent samples assayed is determined by the amount of biological noise in the system under investigation, noise in the measurement process, and the cost in time and resources of the assay and sample production. While not inexpensive, array experiments generally have reasonable costs. If the cost of generating the samples is also reasonable, several biological replicates can be assayed, consistent with control of both Type 1 and Type 2 error at an acceptable level [12,13,15]. Presently the cost for proteomic analysis is significantly higher than that of transcriptome arrays, control for biological variation is consequently more limited, and control of Type 2 error is commonly ignored.

In the array experiment the number of replicates is higher, but the numbers of specific sequences used to probe for a given transcript are few, often just one 0.5 kbp fragment in the case of spotted PCR product arrays. In contrast, the proteomic experiments described in Sections 3.2 and 3.3 used anywhere from two to several thousand peptides to identify and quantitate protein expression by a given gene. Such a high degree of redundancy can be one of the advantages of the proteomics approach. More advanced, higher density array methods also incorporate some level of redundancy [12,13].

As each gene will be subject to different efficiencies of RNA degradation, reverse transcription, dye labeling and hybridization, comparisons cannot be made between genes. Higher signal for gene 1 than gene 2 on an array may simply be a result of these biases, rather than a higher level of expression for gene 1. Thus, transcription is most reliably measured in relative terms with each gene compared to its expression level in a different sample [46]. This is the rationale behind the two-dye experiment, where the two samples to be compared are each labeled with a different dye and hybridized to the same array. Unfortunately, the dyes usually show significant differences in incorporation efficiency and fluorescence characteristics, introducing their own systematic bias [46,47]. This bias is typically controlled by a dye swap experiment. Similar considerations apply to proteomics. Isotope label swapping becomes important when samples are differentially labeled, and such label swaps, e.g. $^{14}\text{N}/^{15}\text{N}$, are a convenient source of biological replicates. Primary isotope effects are a problem more in theory than practicality, because the proteomic methods employed in global studies are normally not sensitive enough to differentiate such subtle effects from random noise. However, the label swap replicate can detect problems with uniform incorporation of the label, e.g. unexpected sources of ^{14}N in growth medium that in theory should contain only ^{15}N . The overall detectability of proteolytic fragments [48] within a given whole proteome enzymatic digest may vary by several orders of magnitude, depending on such factors as intact protein size, abundance, number of tryptic sites, location within the lipid bilayer, instrumental response factor, etc., leading to the need to compare protein abundance by the same ORF under different conditions, rather than trying to compare protein abundances within a given sample. Thus, the basic rationale for the comparison among two or more biological states is similar for both transcription and protein level abundance analysis. Nonetheless, that there is a relationship among spectral counts, signal intensity and protein abundance within a sample has been demonstrated [43], and it is plausible that this complex relationship can be modeled [44,48], such that one day it will be possible to input spectral count or intensity data into such an algorithm that will output a reasonably accurate estimate of absolute protein abundance within the sample that can in turn be used as an input into higher order models of cell function [49].

Variation in the production of the array slides contributes to the overall variation of the experiment. This is another reason for using two dyes, so that samples are always compared for hybridization to the same spot. Print tip and more general spatial biases can be detected by a visual representation of the array readings, for example by representing the log ratios as a color map and plotting them according to their spot location on the array [50–52]. The closest analogies in the MudPIT proteomics world are with the integrated HPLC mass spectrometer, including temporal variations in instrument performance. In microarray studies, variance occurs primarily in the spatial domain. In mass spectrometry-based proteomics, variance occurs in the domain of time. It has been shown that the S/N of mass spectra are correlated with the variability and bias of abundance ratio estimation in quantitative proteomics, as in microarray assays [53]. For proteomic comparisons of two or more states, the same instrumental conditions need to be maintained for all samples and instrument performance needs to be closely monitored. The weakest link in the system is, interestingly enough, a spray tip with an inner diameter of a few micrometers, followed closely by degradation of the capillary HPLC column over time. The fused-silica electrospray tips used in the MudPIT proteomics experiment, with HPLC flow rates on the order of a few hundred nl/min, have a finite lifetime and must be changed periodically. Their useful lifetime can be extended somewhat by washing the tips between runs, a process that has been automated in the senior author's laboratory (Wang *et al.*, manuscript in preparation). Using this automated washing system a single tip will now last for approximately four weeks (24/7) of operation, although this number is highly sample dependent. Experimental design approaches such as randomization and blocking [14] can also be used to minimize the effects of tip and column degradation.

Incorrect hybridization can also be a problem on the array side. PCR products for the array should be chosen to minimize potential cross hybridization with other RNA species in the sample. However, the products are usually relatively long, hundreds of base pairs, and adequately specific regions may be difficult to find for some genes. Additionally, cross hybridizations could occur that are not obvious from the sequence data [54]. There are no easily implemented methods to access cross hybridization for every spot on a PCR array and so these errors will mostly go undetected unless selected for post array validation by a method not subject to the cross hybridization effect. For proteomics the analogous problem is less severe. Using only peptides that can be unequivocally assigned to specific protein-encoding ORFs may reduce the number of peptides available for analysis of proteins sharing sequence similarity with other proteins, but it eliminates the proteomic equivalent of cross hybridization errors. If proteome coverage is high, such unique peptides are present in sufficient quantities that it is possible and desirable to base detection and subsequent quantitation on these proteolytic fragments that are unique to the ORF in question. Only in relatively few cases is the degree of sequence similarity so high that several proteins within a given prokaryotic proteome must be grouped as a class, rather than treated individually. In general, the more data collected, the higher the qualitative proteome coverage, and less need there is to group proteins by class as opposed to treating them individually. This is a key difference between studies of single prokaryotes and studies of microbial communities or higher organisms. In the latter types of studies, such unique sequences are not as common, and one must therefore resort to reporting groups of proteins based on sequence similarity more often.

For array experiments using fluorescent dyes, dye biases can have a systematic dependence on fluorescence intensity [12,51]. Fluorescence measurements also have a somewhat limited dynamic range. Complicating matters, spot intensities are often not linear with the amount of labeled cDNA, resulting in compression of the expression ratios [55]. Microarray measurements show a dynamic range of slightly over three orders of magnitude in the best of cases [56,57]. For transcripts with very low levels of expression or limited samples, RNA amplification has been employed to produce measurable samples [58–60]. The MudPIT proteomics studies discussed in Sections 3.2 and 3.3 generally show a similar dynamic range compared to fluorescence-based arrays, or slightly less. Spectral counting, an analysis method based on sampling frequencies, may increase the dynamic range [27], and the inherent instrumental dynamic range of the newer 2-D linear ion traps has improved markedly over that achievable with older, 3-D designs [36,37]. Compression of abundance ratios observed using proteomic methods is common, the details varying to some extent with the type of quantitation approach used. This topic will be discussed further in Section 2.3.

Finally, on the array side the normalization procedures commonly used to remove biases in the data can be complex, potentially introducing biases and flattening signals [50]. This is especially true if the underlying assumptions of the normalization prove incorrect. For example, many normalization methods assume that few transcripts are differentially expressed and that differential expression will occur in both directions. In this area proteomics has an advantage. The results reviewed in Sections 3.2 and 3.3 have suggested that simple normalization schemes are adequate for analysis of global proteomic results [9,28,61], typically either centering the ratio distribution maxima at zero abundance change, or normalizing based on total observed signal in each biological state. The magnitude of normalization required for the proteome comparison is typically small compared to an array experiment, varying between zero [62] and a factor of 3.3 [9] in the most extreme case we have encountered thus far.

2.3 Strengths and limitations of spectral counting and stable isotope metabolic labeling for quantitative global relative abundance studies

As noted in Section 2.1, the highest possible qualitative coverage is a practical requirement for quantitative proteomics on a cellular scale. While higher resolution, higher mass accuracy data can help reduce the number of required measurements to achieve a desired level of quantitative detectability and precision [25], such data is still subject to this requirement. It is difficult to achieve reproducible and biologically validated protein level abundance quantitation without acquiring a great deal of redundant data. This fundamental relationship applies regardless of the choice of quantitative approach: ICAT [63–66], iTRAQ [67], SILAC [68], ^{15}N metabolic stable isotope labeling [19,20], ^{18}O exchange [69] or spectral counting [30–33], to name the most popular methods that are compatible with multidimensional capillary HPLC/tandem MS. The SILAC (stable isotope labeling with amino acids in cell culture) method was developed primarily for eukaryotic cellular and tissue level investigations, and has thus far not been widely applied as an alternative to conventional metabolic labeling strategies for prokaryotic systems. SILAC differs from conventional metabolic labeling in that the label is introduced into cell culture as amino acid. Isotopic labeling by chemical modification “post harvest” has in general not performed well in whole microbial cell studies, relative to traditional metabolic labeling and spectral counting methods. Thus we restrict further discussion to quantitative methods that do not require the use of chemical modification procedures. The primary reason for this state of affairs can be seen in Fig. 1. In Fig. 1A, the RSD of the abundance ratios shows a very clear relationship with the size of the sample pool from which the five best measurements were selected, as defined below. Simply stated, the more relative abundance estimates in the dataset for a given protein, in this case heavy/light isotopic pairs, the better, even if only a small percentage of the total data collected is used in the final ratio calculation that is reported. Any quantitative approach that limits the pool of potentially useful measurements for the sake of reducing dataset size and complexity is going to run afoul of this relationship by shifting the number of peptide pairs that potentially can be used for the abundance calculation to the left, into a region where the variance is so high that the abundance ratio is almost meaningless in terms of either precision or accuracy. A relative abundance ratio based on a small number of measurements is not likely to be reproducible. In a large dataset, a significant number of relative abundance ratios based on such a small number of measurements will show acceptable RSD values purely by chance, and an uncorrected p -value based on variance as well as the magnitude of deviation from an expected null value will not necessarily be a true gauge of significant change. The measurements selected for computing ratios shown in Fig. 1 were the five heavy/light pairs that grouped most closely about a measure of central tendency based on the median of the absolute deviation from the mean, as determined using a MAD (median of the absolute deviation) z -score [61,70,71]. Stated in non-technical language, we favored the MAD approach because it minimizes the influence of outliers. The frequency distribution of MAD z -scores for these data are shown in supplementary Fig. S5. For the data shown in Fig. 1, the true values of the ratios were known because they were based on a known 16-fold dilution of the *M. maripaludis* proteome measured against the undiluted proteome. These data were originally acquired to calibrate q -value cut-offs, see Table 1, for nutrient limitation studies described in Section 3.2 (Xia *et al.*, manuscript in preparation). The data acquisition and database searching procedures for these data were much the same as for the *P. gingivalis* data discussed below [9]. In most cases, the isotopic pairs selected were correlated with the highest S/N measurements [53].

A similar relationship holds for spectral counting: the higher the total counts used in the ratio calculation, the more reproducible the abundance ratio and the greater the likelihood that the abundance ratio will be validated by orthogonal means. Spectral counting measurements are very stable, precise and reproducible [27,43]. However, there is evidence that this same invariance can lead, at least for certain proteins, to a lack of change in counts as a function of

concentration, resulting in false negative trends when simulated abundance ratios are measured under conditions when the true ratios are known, as discussed below. For spectral counting, the relationship between total counts and the quality of the calculated abundance ratio can be seen in Figures 2A–5A. The relationship between spectral counts and S/N is more complex than that described above for MAD z -score and S/N for ratios based on heavy/light isotopic pairs or non-label signal intensity measurements, although the precision observed for spectral counting tends on average to be better [27].

Accuracy for spectral counting is another matter entirely, and in general shows greater dependencies on protein size, number of enzymatic cleavage sites, abundance, and HPLC gradient elution conditions relative to quantitation based on relative signal intensities of heavy/light stable isotope pairs. Figures 1, 3, 5, Tables 1–3, and the color ORF plots shown in supplementary Figures S1–S4 illustrate on a whole proteome level the results when different known concentrations of *P. gingivalis* and *M. maripaludis* were compared in a two-state experiment without normalization. Note that while precision was not as good with the intensity data (Figures 3B, 5B), accuracy in terms of calling the proper direction of change was significantly better with *P. gingivalis* and the five-fold dilution of *M. maripaludis*. Spectral counting largely failed to detect a five-fold change in relative concentration, see Fig. 5A and Table 2. As noted above, these data were originally acquired for purposes of calibrating q -value cut-offs [10] used to differentiate significant from insignificant abundance change [9], (Xia *et al.*, manuscript in preparation).

If the spectral count method had worked perfectly, all *P. gingivalis* ORFs in Fig. S1 (see supplement) would be color-coded either white for non-detects or green for ratios significantly higher than zero on a $\log_2$ scale. The true ratio in each case was 16 on a linear scale, 4 on a $\log_2$ scale. Experimentally, 333 of the detected proteins showed significant change in the correct direction (green), while 2 showed change in the wrong direction (red) and 750 showed no significant change (yellow), based on the q -value criteria established by Xia *et al.* (submitted for publication), see Table 3. For *M. maripaludis*, comparison of the undiluted and 16-fold diluted proteome in supplemental Fig. S3 illustrates the situation that occurs when spectral counts are very high, many times higher than in the *P. gingivalis* 16-fold dilution example, on the order of hundreds to thousands of counts per protein, and the magnitude of relative concentration change is also high, 16-fold. In this scenario, spectral counting (Fig. S3) called 767 proteins in the correct direction (green), 8 in the wrong direction (red) and 498 as ambiguous (yellow). By comparison, matched $^{14}\text{N}/^{15}\text{N}$ peptide pairs (Fig. S2) in the same data generated 1075 correct calls (green), 5 reds and 188 yellows. All but three of the correct calls (green) in the spectral counting data were included as greens in the isotope data, see the Venn diagram, supplemental Fig. S6.

The q -value cut-offs shown in Tables 1–3 also tend to support the observation that methods employed for *M. maripaludis* (see Section 3.2) and *P. gingivalis* (see Section 3.3) proteomics control false positive risk well, but do not control for false negative risk (Xia *et al.*, submitted for publication). The FNR (false negative rate) was particularly high for the spectral counting data shown in Fig. 5A for a five-fold change, 92% at a q -value of 0.01, see Table 2. Based on observations with *M. maripaludis*, spectral counting was not as sensitive as stable isotope labeling for detecting small changes in protein abundance, on the order of two-fold [28].

To summarize, for spectral counting precision is high, but quantitative false positive risk is greater (see supplemental Figures S1 and S3, Tables 1–3) and statistical power for determining subtle abundance changes is low. For metabolic stable isotope labeling, precision is moderate to high (see Fig. 1), false positive risk is lower (Fig. S2), and two-fold changes in abundance or less can be detected. As a tool for probing gene regulation in prokaryotes, metabolic stable isotope labeling in general requires less data to achieve a predefined level of statistical power

[28], but given enough data and a sufficiently large change in abundance, spectral counting can achieve similar goals. Signal intensity measured in samples run separately (Table 2, Figures 3B, 5B, supplemental Fig. S4) was noisier than spectral counting and classic metabolic stable isotope labeling, where samples are normally mixed prior to mass spectral analysis. However, the number of true positives called correctly (green dots) was the highest with this method, see supplemental Fig. S4. The number of false positive calls in wrong direction, 22, was also high, consistent with the lack of precision relative to spectral counting or the stable isotope data. All approaches yielded high rates of quantitative false negatives in the examples discussed above, see Tables 1–3, with spectral counting showing the highest FNRs. The best overall results were achieved with the isotope ratio measurements, based on FNR, FPR and FDR.

Proteomic data will often yield abundance ratios that are compressed compared to the true value, such as the data shown in Fig. 1. The sources for this compression are several and differ somewhat for spectral counting and metabolic labeling. In spectral counting, many proteins show a discontinuous and complex relationship between observed counts and concentration, leading to a restriction in the range of observed ratios. As protein molecular weight and the number of enzymatic digestion sites increase, the relationship between counts and concentration becomes more linear. For signal intensity measurements such as those based on isotopic pairs, dynamic range and sample loading restrictions imposed by the capillary chromatography and the mass spectrometer become significant factors, which also can impact spectral counting data. Nonetheless, the direction of abundance change, rather than the magnitude of change, is normally of greater interest in most global studies. Evidence to date suggests that protein abundance ratios determined using MudPIT methods accurately predict the direction of change in a two-state comparison [20,27,28,61] with low FDRs, keeping in mind the evidence for false negative bias described above.

3 Microbiological applications

3.1 An overview of recent global quantitative studies of protein abundance in prokaryotes

As noted in the introduction, the numbers of biologically driven quantitative proteomic studies that are truly global in scope at the cellular level are few. Most of the literature dealing with protein level global abundance measurements is of a technical or methodological nature, rather than biological. This is to be expected, given that the science of such measurements is still in an early stage of development. “Truly global” is defined here somewhat loosely as sufficient proteome coverage to compare two or more biological conditions for a large percentage of the proteins expressed by the organism. Knowing how many proteins are really being expressed is a complex issue, and in the studies described below this was not known with any certainty. Transcription microarray measurements have been used for such estimates, but the absence of easily accessible absolute detection information for a given transcript limits their utility for addressing this question (see Section 2.2).

Pacific Northwest Laboratory in the United States has played a leading role in such studies, including ongoing investigations of several organisms of environmental interest, including *Deinococcus radiodurans* [72], *Rhodobacter sphaeroides* [73,74], *Shewanella oneidensis* [75], *Desulfovibrio vulgaris* [76,77], and the pathogens *Yersinia pestis* [78] and *Salmonella enterica* [79]. These studies have utilized both unit resolution data from quadrupole ion traps as well as high accuracy mass tags [72] acquired with FTMS instrumentation. The last mentioned paper is particularly noteworthy in that the *Salmonella enterica* serovar Typhimurium proteome was measured over a time course for bacteria internalized within macrophages, describing this facultative intracellular pathogen’s response to a model human host environment. Groups associated with ORNL (Oak Ridge National Laboratory) have published studies of *Shewanella oneidensis* response to chromium exposure [80,81]. Li and coworkers [82] have used microarrays and quantitative proteomics with metabolic stable

isotope labeling to study acetate versus methanol metabolism in *Methanosarcina acetivorans*. Kao *et al.* [83] have applied quantitative proteomics to the study of copper ion regulation in *Methylococcus capsulatus* metabolism. Scherl, Francois *et al.* [84] used a combined microarray and proteomics approach to study glycopeptide resistance in *Staphylococcus aureus*. In one of the most successful attempts at global ICAT studies, Guina *et al.* [65,66] have studied the induction of virulence factors under conditions of magnesium limitation in the opportunistic pathogen *Pseudomonas aeruginosa*. Stensjo *et al.* [85] have used iTRAQ quantitative proteomics to study the cyanobacterium *Nostoc sp.* PCC 7120 proteome response to nitrogen fixing and non-fixing conditions. Bisle, Schmidt *et al.* [86] used a combination of 2-D DIGE and free amino group isotopic labeling of intact proteins, followed by tryptic digestion, capillary HPLC and MALDI-TOF MS, to quantitatively study the membrane proteome of the halophilic Archaeon *Halobacterium salinarum* under aerobic and anaerobic conditions.

3.2 Global protein abundance studies of the methanogenic Archaeon *Methanococcus maripaludis*

The MudPIT approach to proteomics has proven useful for *Methanococcus maripaludis*, a hydrogenotrophic methanogen belonging to the domain Archaea. These microorganisms inhabit a variety of anaerobic habitats, where they obtain energy by using molecular hydrogen as the electron donor to reduce carbon dioxide to methane. Due in part to its relatively rapid and reliable growth in the laboratory, and the availability of tools for genetic manipulation, *M. maripaludis* is one of a few species of methanogens that has emerged as a useful laboratory model. Work is also aided by the small genome (containing 1,722 ORFs, [87]) and its relative freedom from repeats and other features that would complicate annotation. *M. maripaludis* is also reliably grown in chemostats under defined nutrient conditions [88], the ideal culturing method for studies of global gene expression.

MudPIT proteomics of metabolically labeled samples has yielded protein abundance ratios with reliability approaching that of spotted PCR product mRNA expression arrays, see Section 2.2. *M. maripaludis* is easily labeled with ^{15}N , and measurements of mass spectral signal intensity ratios ($^{15}\text{N}/^{14}\text{N}$) allowed us to compare protein abundances between the wild type strain and a mutant deficient in a hydrogenase that supplies electrons for biosynthesis [61]. Proteomics and transcriptome arrays were in substantial agreement and both demonstrated increased abundance in the mutant of carbon assimilation genes that encode basic biosynthetic functions [88]. In a study completed but not yet published (Xia *et al.*, manuscript in preparation) MudPIT proteomics on metabolically labeled samples from chemostats enabled us to identify sets of proteins that increase markedly in abundance with specific nutrient limitation by hydrogen, nitrogen, or phosphate. We have not yet detected clear differential abundance at the proteome level where there is none at the transcriptome level. However, quantitative proteomics holds promise for the detection of post-transcriptional regulation in *M. maripaludis*.

In addition to abundance ratios of individual proteins, spectral counting has proven useful in the estimation of relative abundance among proteins. As noted in Section 2.2, making such inferences regarding protein abundance within a given sample or biological state is controversial, but for this organism specifically the MudPIT approach reliably yielded high counts for proteins expected to be abundant (e.g. ribosomal proteins) and agreed with predictions of protein abundance based on codon bias (Karlin indices) [61]. Similar agreement was noted based on spectral counts for the related Archaeon *Methanocaldococcus jannaschii* [89]. This same type of comparison with theoretic estimates of protein abundance has failed with other prokaryotes, including *P. gingivalis* (see Section 3.3). Protein abundance information can be especially useful when estimating the relative importance of paralogs. For

example, *M. maripaludis* encodes two S-layer proteins, but one of them appears to be much more abundant than the other and probably makes up the bulk of the cell wall.

Proteomics also holds promise for discovery of post-translational modifications, detection of translational recoding, and genomic annotation. Many instances of post-translational modification are known in Archaea, both by addition of organic moieties and proteolytic processing [90], but no doubt many remain to be discovered. Instances of stop codon read-through and frame-shifting are also known [91]. Also, a certain amount of inaccurate annotation (e.g. incorrect start codon assignment) is inevitable. Any of these events should lead to peptides whose spectra differ from predictions made from the annotated genome database. Our coverage of *M. maripaludis* is now essentially quantitative, ~92% of the predicted proteome in our most recent work, so comprehensive search strategies are being employed for anomalous peptides indicative of the various conditions described above. Many proteins are now detected at a level of less than one copy per cell, suggesting that the analytical methods have matured to a point that one must address the issue of distinguishing protein expression due to cellular machinery that may never shut down completely, versus expression that plays a functional role in the organism. When collecting on the order of 10^9 to 10^{10} cells from the chemostat, only routine sensitivity (high attomole in absolute terms) on the part of the mass spectrometer is required to measure such putative baseline expression, when protein extraction efficiencies are high, also taking into account losses during each stage of separation.

3.3 Global protein abundance studies when many genes are regulated: the intracellular pathogen *Porphyromonas gingivalis*

P. gingivalis is a gram-negative obligate anaerobe that is a major pathogen in severe cases of periodontal disease. In the oral cavity, *P. gingivalis* inhabits the multi-species biofilm known as dental plaque, that accumulates on tooth surfaces, and also resides within the epithelial cells of the gingival crevice, the area of the gum in which the root of the tooth is located [92]. *In vitro* culture models have demonstrated that *P. gingivalis* remain viable within gingival epithelial cells (GECs) and that the epithelial cells do not undergo apoptotic or necrotic cell death when infected with the organism [93–95]. The characteristics of *P. gingivalis* that facilitate colonization of the oral cavity, and can lead to destruction of the periodontal tissue, have been established through the cumulative outcomes of a large number of conventional microbiological and molecular biological investigations [92,96]. Whole cell protein profiling of intracellular *P. gingivalis* can provide insight in to this system of extreme complexity.

MudPIT proteomics with spectral counting and non-label signal intensity quantitation was used to compare the proteome of *P. gingivalis* recovered from inside GEC to *P. gingivalis* in GEC medium [9] (Xia *et al.*, submitted for publication). The analysis provided several insights in to the adaptation of *P. gingivalis* to an intracellular lifestyle. The major adhesin for gingival epithelial cells is the long fimbriae comprised of the FimA subunit protein [97]. FimA was under-expressed intracellularly, indicating that the organism dispenses with production of this molecule after entry has been accomplished. Loss of a major surface protein such as FimA may also have implications for recognition of the organism by intracellular microbial detection, systems such as the Nod-like receptors [98]. Abundance of the major proteases of *P. gingivalis*, RgpA, RgpB and Kgp, was downregulated within GEC, which will likely limit the amount of damage that can be caused by intracellular organisms. In all, 28% of the predicted proteome of *P. gingivalis* was shown to be differentially regulated in intracellular organisms (Xia *et al.*, submitted for publication) indicating that intracellular *P. gingivalis* becomes essentially a different organism on a phenotypic level compared to its extracellular counterpart. Unlike the case with *M. maripaludis* discussed in the previous section, the strain W83 database used for searching the mass spectra was not a near perfect match with the ATCC 33277 strain used experimentally, leading to a reduction in overall proteome coverage on the order of 10%.

The proteomics approach worked reasonably well under conditions where a large percentage of the proteome was changing, in violation of the assumption of minimal change that was implicit in the methods used to process the data [9]. We base this last observation on the overall consistency of the proteomic results with other studies of *P. gingivalis* invasion and internalization (Xia *et al.*, submitted for publication).

4 Concluding remarks

What can quantitative proteomic methods based on multidimensional capillary HPLC and tandem mass spectrometry tell us about protein abundance in lower organisms? Rather a great deal, even though it is obvious much is potentially being missed. For organisms that have been sequenced, annotated and contain on the order of 2,000 protein encoding ORFs, the MudPIT type of proteomic approach is competitive with microarray methods in terms of coverage and quantitative reliability, as judged by FDR. FDRs in the work cited in this review were low, generally less than 5% in the work we know well enough to assign a number. However, the FDR approach as implemented, for example, by the use of *q*-values [9,10], only tells us about bad quantitative calls among those proteins in a given dataset that are being called as significantly changed. It tells us little about how much abundance change is being missed completely. The experiments designed to address the false negative issue described in Section 2.3 suggest an uncomfortably high rate of quantitative false negative assignments, particularly for spectral counting. This shared shortcoming of quantitative proteomics and many published microarray studies is intuitively understood by many, but seldom evaluated and discussed due to the time, cost and other feasibility issues associated with measuring and controlling quantitative false negative error. To put the false negative issue into perspective, most working microbiologists who use the data are more concerned with wasting their time and resources with false leads, i.e. with FDR, than they are about leads for further research that were never discovered to begin with. However, in order for the field to progress FNRs need to be lowered. For example, in gene regulation studies of *M. maripaludis*, mRNA expression and protein abundance changes as confirmed by real time RT-PCR have been subtle [61,88], usually less than five-fold, so any gain in statistical power is of practical importance.

For organisms with smaller proteomes where qualitative and quantitative coverage is saturated in terms of amino acid sequence, the next great challenge beyond limiting quantitative false negative error is measuring global changes in PTM status [25,26,99,100,101]. To accomplish this task in a way that is sufficiently complete, reproducible, and quantitative as to allow the use of such data for the elucidation of regulatory networks is a worthy long-term goal. However, this process will require improvements in a number of areas, both experimentally and computationally [25,26,101,102].

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

Supported by the NIH through DE14372 (M. H.), DE11111 (R. J. L.) and GM074783 (J. A. L.). Additional support provided by the United States DOE through DE-FG02-05ER15709 (J. A. L.). We thank R. D. Smith and J. W. Jorgenson for their assistance.

Abbreviations

2-D

two dimensional

3-D	three dimensional
ETD	electron transfer dissociation
FDR	false discovery rate
FNR	false negative rate
FPR	false positive rate
FTMS	Fourier transform mass spectrometry
GEC	gingival epithelial cell
LOWESS	locally weighted scatter plot smoothing
MAD	median of the absolute deviation
MudPIT	multidimensional protein identification technology
SILAC	stable isotope labeling with amino acids in cell culture
TIGR	The Institute for Genomic Research

5 References

1. Ishii N, Robert M, Nakayama Y, Kanai A, Tomita M. Toward large-scale modeling of the microbial cell for computer simulation. *J Biotechnol* 2004;113:281–294. [PubMed: 15380661]
2. Souchelnytskyi S. Bridging proteomics and systems biology: what are the roads to be traveled? *Proteomics* 2005;5:4123–4137. [PubMed: 16196099]
3. Mayya V, Han KD. Proteomic applications of protein quantification by isotope-dilution mass spectrometry. *Expert Rev Proteomics* 2006;3:597–610. [PubMed: 17181474]
4. Domon B, Aebersold R. Mass spectrometry and protein analysis. *Science* 2006;312:212–217. [PubMed: 16614208]
5. Quackenbush, J. *Animal Genet.* Vol. 37 Suppl 1. 2006. From 'omes to biology; p. 48-56.
6. Urfer W, Grzegorzczak M, Jung K. Statistics for proteomics: A review of tools for analyzing experimental data. *Proteomics* 2006;48–55. [PubMed: 17031797]
7. Zhang B, VerBerkmoes NC, Langston MA, Uberbacher E, et al. Detecting differential and correlated protein expression in label-free shotgun proteomics. *J Proteome Res* 2006;5:2909–2918. [PubMed: 17081042]
8. Old WM, Meyer-Arendt K, Aveline-Wolf L, Pierce KG, et al. *Proteomics* 2005;4:1487–1502.
9. Xia QW, Wang TS, Park Y, Lamont RJ, Hackett M. Differential quantitative proteomics of *Porphyromonas gingivalis* by linear ion trap mass spectrometry: Non-label methods comparison, *q*-values and LOWESS curve fitting. *Int Jour Mass Spectrom* 2007;259:105–116.

10. Storey JD, Tibshirani R. Statistical significance for genome wide studies. *Proc Natl Acad Sci USA* 2003;100:9440–9445. [PubMed: 12883005]
11. Storey JD. The positive false discovery rate: a Bayesian interpretation and the q-value. *Ann Stats* 2003;31:2013–2035.
12. Quackenbush J. Microarray data normalization and transformation. *Nat Genet* 2002;32 Suppl:496–501. [PubMed: 12454644]
13. Quackenbush J. Computational analysis of microarray data. *Nat Rev Genet* 2001;2:418–427. [PubMed: 11389458]
14. Sokal, RR.; Rohlf, FJ. *Biometry: the principles and practice of statistics in biological research*. Vol. 3. Freeman W. H.; New York: 1995.
15. Zakharkin SO, Kim K, Bartolucci AA, Page GP, Allison DB. Optimal allocation of replicates for measurement evaluation studies. *Genom Proteom Bioinform* 2006;4:196–202.
16. Delahunty C, Yates JR III. Protein identification using 2D-LC-MS/MS. *Methods* 2005;35:248–255. [PubMed: 15722221]
17. Godovac-Zimmermann J, Kleiner O, Brown LR, Drukier AK. Perspectives in splicing up proteomics with splicing. *Proteomics* 2005;5:699–709. [PubMed: 15693068]
18. Washburn MP, Wolters D, Yates JR III. Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nat Biotechnol* 2001;19:242–247. [PubMed: 11231557]
19. Washburn MP, Ulaszek R, Deciu C, Schieltz DM, Yates JR III. Analysis of quantitative proteomic data generated via multidimensional protein identification technology. *Anal Chem* 2002;74:1650–1657. [PubMed: 12043600]
20. Washburn MP, Ulaszek RR, Yates JR III. Reproducibility of quantitative proteomic analyses of complex biological mixtures by multidimensional protein identification technology. *Anal Chem* 2003;75:5054–5061. [PubMed: 14708778]
21. Kolkman A, Slijper M, Heck AJ. Development and application of proteomics technologies in *Saccharomyces cerevisiae*. *Trends Biotechnol* 2005;23:598–604. [PubMed: 16202464]
22. Opiteck GJ, Jorgenson JW, Moseley MA, Anderegg RJ. Two-dimensional microcolumn HPLC coupled to a single-quadrupole mass spectrometer for the elucidation of sequence tags and peptide mapping. *Jour Microcol Separ* 1998;10:365–375.
23. Opiteck GJ, Lewis KC, Jorgenson JW, Anderegg RJ. Comprehensive on-line LC/LC/MS of proteins. *Anal Chem* 1997;69:1518–1524. [PubMed: 9109352]
24. Holland LA, Jorgenson JW. Characterization of a comprehensive two-dimensional anion exchange-perfusive reversed phase liquid chromatography system for improved separations of peptides. *Jour Microcol Separ* 2000;12:371–377.
25. Venable JD, Wohlschlegel J, McClatchy DB, Park SK, Yates JR III. Relative quantification of stable isotope labeled peptides using a linear ion trap-orbitrap hybrid mass spectrometer. *Anal Chem*. 2007in press
26. Tsur D, Tanner S, Zandi E, Bafna V, Pevzner PA. Identification of post-translational modifications by blind search of mass spectra. *Nat Biotechnol* 2005;23:1562–1567. [PubMed: 16311586]
27. Zybaylov B, Coleman MK, Florens L, Washburn MP. Correlation of relative abundance ratios derived from peptide ion chromatograms and spectrum counting for quantitative proteomic analysis using stable isotope labeling. *Anal Chem* 2005;77:6218–6224. [PubMed: 16194081]
28. Hendrickson EL, Xia Q, Wang T, Leigh JA, Hackett M. Comparison of spectral counting and metabolic stable isotope labeling for use with quantitative microbial proteomics. *Analyst* 2006;131:1335–1341. [PubMed: 17124542]
29. Ong SE, Mann M. Mass spectrometry-based proteomics turns quantitative. *Nat Chem Biol* 2005;1:252–262. [PubMed: 16408053]
30. Gao J, Opiteck GJ, Friedrichs MS, Dongre AR, Hefta SA. Changes in the protein expression of yeast as a function of carbon source. *J Proteome Res* 2003;2:643–649. [PubMed: 14692458]
31. Gao J, Friedrichs MS, Dongre AR, Opiteck GJ. Guidelines for the routine application of the peptide hits technique. *J Am Soc Mass Spectrom* 2005;16:1231–1238. [PubMed: 15978832]

32. Florens L, Carozza MJ, Swanson SK, Fournier M, et al. Analyzing chromatin remodeling complexes using shotgun proteomics and normalized spectral abundance factors. *Methods* 2006;40:303–311. [PubMed: 17101441]
33. Pang JX, Ginanni N, Dongre AR, Hefta SA, Opitek GJ. Biomarker discovery in urine by proteomics. *J Proteome Res* 2002;1:161–169. [PubMed: 12643536]
34. Sadygov RG, Cociorva D, Yates JR III. Large-scale database searching using tandem mass spectra: looking up the answer in the back of the book. *Nat Methods* 2004;1:195–202. [PubMed: 15789030]
35. Schwartz JC, Senko MW, Syka JE. A two-dimensional quadrupole ion trap mass spectrometer. *J Am Soc Mass Spectrom* 2002;13:659–669. [PubMed: 12056566]
36. Blackler AR, Klammer AA, MacCoss MJ, Wu CC. Quantitative comparison of proteomic data quality between a 2D and 3D quadrupole ion trap. *Anal Chem* 2006;78:1337–1344. [PubMed: 16478131]
37. Douglas DJ, Frank AJ, Mao D. Linear ion traps in mass spectrometry. *Mass Spectrom Rev* 2005;24:1–29. [PubMed: 15389865]
38. Coon JJ, Syka JE, Shabanowitz J, Hunt DF. Tandem mass spectrometry for peptide and protein sequence analysis. *Biotechniques* 2005;38:519–521. [PubMed: 15884666]
39. Syka JE, Coon JJ, Schroeder MJ, Shabanowitz J, Hunt DF. Peptide and protein sequence analysis by electron transfer dissociation mass spectrometry. *Proc Natl Acad Sci USA* 2004;101:9528–9533. [PubMed: 15210983]
40. Zubarev RA, Haselmann KF, Budnik B, Kjeldsen F, Jensen F. Towards an understanding of the mechanism of electron-capture dissociation: a historical perspective and modern ideas. *Eur Jour Mass Spectrom* 2002;8:337–349.
41. Venable JD, Yates JR III. Impact of ion trap tandem mass spectra variability on the identification of peptides. *Anal Chem* 2004;76:2928–2937. [PubMed: 15144207]
42. Li Q, Xia Q, Wang T, Meila M, Hackett M. Analysis of the stochastic variation in LTQ single scan mass spectra. *Rapid Commun Mass Spectrom* 2006;20:1551–1557. [PubMed: 16628562]
43. Liu H, Sadygov RG, Yates JR III. A model for random sampling and estimation of relative protein abundance in shotgun proteomics. *Anal Chem* 2004;76:4193–4201. [PubMed: 15253663]
44. Craig R, Cortens JP, Beavis RC. The use of proteotypic peptide libraries for protein identification. *Rapid Commun Mass Spectrom* 2005;19:1844–1850. [PubMed: 15945033]
45. Eng JK, McCormack AL, Yates JR III. An approach to correlate tandem mass-spectral data of peptides with amino-acid-sequences in a protein database. *J Am Soc Mass Spectrom* 1994;5:976–989.
46. Kerr MK, Churchill GA. Statistical design and the analysis of gene expression microarray data. *Genet Res* 2001;77:123–128. [PubMed: 11355567]
47. Dudoit S, Yang YH, Callow MJ, Speed TP. Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statistica Sinica* 2002;12:111–139.
48. Tang H, Arnold RJ, Alves P, Xun Z, et al. *Bioinformatics* 2006;22:e481–488. [PubMed: 16873510]
49. Lisacek F, Cohen-Boulakia S, Appel RD. Proteome informatics II: bioinformatics for comparative proteomics. *Proteomics* 2006;6:5445–5466. [PubMed: 16991192]
50. Cui X, Kerr MK, Churchill GA. Transformations for cDNA microarray data. *Stat Appl Genet Mol Biol* 2003;2Article 4
51. Yang YH, Dudoit S, Luu P, Lin DM, et al. Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res* 2002;30:e15. [PubMed: 11842121]
52. Park T, Yi SG, Kang SH, Lee S, et al. Evaluation of normalization methods for microarray. *BMC Bioinform* 2003;4:33.
53. Pan C, Kora G, Tabb DL, Pelletier DA, et al. Robust estimation of peptide abundance ratios and rigorous scoring of their variability and bias in quantitative shotgun proteomics. *Anal Chem* 2006;78:7110–7120. [PubMed: 17037910]
54. Chuaqui RF, Bonner RF, Best CJ, Gillespie JW, et al. Post-analysis follow-up and validation of microarray experiments. *Nat Genetics* 2002;32 Suppl:509–514. [PubMed: 12454646]
55. Ramdas L, Coombes KR, Baggerly K, Abruzzo L, et al. Sources of nonlinearity in cDNA microarray expression measurements. *Genome Biol* 2001;2RESEARCH0047

56. Kuhn K, Baker SC, Chudin E, Lieu MH, et al. A novel, high-performance random array platform for quantitative gene expression profiling. *Genome Res* 2004;14:2347–2356. [PubMed: 15520296]
57. Lockhart DJ, Dong H, Byrne MC, Follett MT, et al. Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat Biotechnol* 1996;14:1675–1680. [PubMed: 9634850]
58. Laurell C, Wirta V, Nilsson P, Lundberg J. Comparative analysis of a 3' end tag PCR and a linear RNA amplification approach for microarray analysis. *Jour Biotechnol* 2007;127:638–646. [PubMed: 16997411]
59. Li L, Roden J, Shapiro BE, Wold BJ, et al. Reproducibility, fidelity, and discriminant validity of mRNA amplification for microarray analysis from primary hematopoietic cells. *J Mol Diagn* 2005;7:48–56. [PubMed: 15681474]
60. Theilgaard-Monch K, Cowland J, Borregaard N. Profiling of gene expression in individual hematopoietic cells by global mRNA amplification and slot blot analysis. *J Immunol Meth* 2001;252:175–189.
61. Xia Q, Hendrickson EL, Zhang Y, Wang T, et al. Quantitative proteomics of the archaeon *Methanococcus maripaludis* validated by microarray analysis and real time PCR. *Proteomics* 2006;5:868–881.
62. Zhang Y, Wang T, Chen W, Yilmaz O, et al. *Proteomics* 2005;5:198–211. [PubMed: 15619293]
63. Gygi SP, Rist B, Gerber SA, Turecek F, et al. Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nat Biotechnol* 1999;17:994–999. [PubMed: 10504701]
64. Vaughn CP, Crockett DK, Lim MS, Elenitoba-Johnson KS. Analytical characteristics of cleavable isotope-coded affinity tag-LC-tandem mass spectrometry for quantitative proteomic studies. *J Mol Diagn* 2006;8:513–520. [PubMed: 16931593]
65. Guina T, Purvine SO, Yi EC, Eng J, et al. Quantitative proteomic analysis indicates increased synthesis of a quinolone by *Pseudomonas aeruginosa* isolates from cystic fibrosis airways. *Proc Natl Acad Sci USA* 2003;100:2771–2776. [PubMed: 12601166]
66. Guina T, Wu M, Miller SI, Purvine SO, et al. Proteomic analysis of *Pseudomonas aeruginosa* grown under magnesium limitation. *J Am Soc Mass Spectrom* 2003;14:742–751. [PubMed: 12837596]
67. Thompson A, Schafer J, Kuhn K, Kienle S, 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]
68. Mann M. Functional and quantitative proteomics using SILAC. *Nat Rev Mol Cell Biol* 2006;7:952–958. [PubMed: 17139335]
69. Miyagi M, Rao KC. Proteolytic ¹⁸O-labeling strategies for quantitative proteomics. *Mass Spectrom Rev* 2007;26:121–136. [PubMed: 17086517]
70. Hampel FR. The influence curve and its role in robust estimation. *J Amer Statist Assn* 1974;69:383–393.
71. Burke S. Missing Values, Outliers, Robust Statistics & Nonparametric Methods. *Statistics and data analysis. LC GC Europe online Suppl* 2001;59:19–24.
72. Lipton MS, Pasa-Tolic L, Anderson GA, Anderson DJ, et al. Global analysis of the *Deinococcus radiodurans* proteome by using accurate mass tags. *Proc Natl Acad Sci USA* 2002;99:11049–11054. [PubMed: 12177431]
73. Callister SJ, Nicora CD, Zeng X, Roh JH, et al. Comparison of aerobic and photosynthetic *Rhodobacter sphaeroides* 2.4.1 proteomes. *J Microbiol Methods* 2006;67:424–436. [PubMed: 16828186]
74. Callister SJ, Dominguez M, Nicora CD, Zeng X. Application of the accurate mass and time tag approach to the proteome analysis of sub-cellular fractions obtained from *Rhodobacter sphaeroides* 2.4.1. Aerobic and photosynthetic cell cultures. *Jour Proteome Res* 2006;5:1940–1947. [PubMed: 16889416]
75. Fang R, Elias DA, Monroe ME, Shen Y, et al. Differential label-free quantitative proteomic analysis of *Shewanella oneidensis* cultured under aerobic and suboxic conditions by accurate mass and time tag approach. *Mol Cell Proteomics* 2006;5:714–725. [PubMed: 16401633]
76. Nie L, Wu G, Zhang W. Correlation of mRNA expression and protein abundance affected by multiple sequence features related to translational efficiency in *Desulfovibrio vulgaris*: a quantitative analysis. *Genetics* 2006;174:2229–2243. [PubMed: 17028312]

77. Nie L, Wu G, Zhang W. Correlation between mRNA and protein abundance in *Desulfovibrio vulgaris*: a multiple regression to identify sources of variations. *Biochem Biophys Res Commun* 2006;339:603–610. [PubMed: 16310166]
78. Hixson KK, Adkins JN, Baker SE, Moore RJ, et al. Biomarker candidate identification in *Yersinia pestis* using organism-wide semi-quantitative proteomics. *J Proteome Res* 2006;5:3008–3017. [PubMed: 17081052]
79. Shi L, Adkins JN, Coleman JR, Schepmoes AA, et al. Proteomic analysis of *Salmonella enterica* Serovar Typhimurium isolated from RAW 264.7 macrophages: identification of a novel protein that contributes to the replication of Serovar Typhimurium inside macrophages. *J Biol Chem* 2006;281:29131–29140. [PubMed: 16893888]
80. Chourey K, Thompson MR, Morrell-Falvey J, Verberkmoes NC, et al. Global molecular and morphological effects of 24-hour chromium(VI) exposure on *Shewanella oneidensis* MR-1. *Appl Environ Microbiol* 2006;72:6331–6344. [PubMed: 16957260]
81. Brown SD, Thompson MR, Verberkmoes NC, Chourey K, et al. *Proteomics* 2006;5:1054–1071.
82. Li L, Li Q, Rohlin L, Kim U, et al. Quantitative proteomic and microarray analysis of the archaeon *Methanosarcina acetivorans* grown with acetate versus methanol. *Jour Proteome Res* 2007;6:759–771. [PubMed: 17269732]
83. Kao WC, Chen YR, Yi EC, Lee H, et al. Quantitative proteomic analysis of metabolic regulation by copper ions in *Methylococcus capsulatus* (Bath). *J Biol Chem* 2004;279:51554–51560. [PubMed: 15385566]
84. Scherl A, Francois P, Charbonnier Y, Deshusses JM, et al. Exploring glycopeptide-resistance in *Staphylococcus aureus*: a combined proteomics and transcriptomics approach for the identification of resistance-related markers. *BMC genomics* 2006;7:296. [PubMed: 17121677]
85. Stensjo K, Ow SY, Barrios-Llerena ME, Lindblad P, Wright PC. An iTRAQ-based quantitative analysis to elaborate the proteomic response of *Nostoc* sp. PCC 7120 under N₂ fixing conditions. *J Proteome Res* 2007;6:621–635. [PubMed: 17269719]
86. Bisle B, Schmidt A, Scheibe B, Klein C, et al. *Proteomics* 2006;5:1543–1558.
87. Hendrickson EL, Kaul R, Zhou Y, Bovee D, et al. Complete genome sequence of the genetically tractable hydrogenotrophic methanogen *Methanococcus maripaludis*. *J Bacteriol* 2004;186:6956–6969. [PubMed: 15466049]
88. Porat I, Kim W, Hendrickson EL, Xia Q, et al. Disruption of the operon encoding Ehb hydrogenase limits anabolic CO₂ assimilation in the archaeon *Methanococcus maripaludis*. *J Bacteriol* 2006;188:1373–1380. [PubMed: 16452419]
89. Zhu W, Reich CI, Olsen GJ, Giometti CS, Yates JR III. Shotgun proteomics of *Methanococcus jannaschii* and insights into methanogenesis. *J Proteome Res* 2004;3:538–548. [PubMed: 15253435]
90. Eichler J, Adams MW. Posttranslational protein modification in Archaea. *Microbiol Mol Biol Rev* 2005;69:393–425. [PubMed: 16148304]
91. Cobucci-Ponzano B, Rossi M, Moracci M. Recoding in archaea. *Mol Microbiol* 2005;55:339–348. [PubMed: 15659155]
92. Lamont RJ, Jenkinson HF. Life below the gum line: pathogenic mechanisms of *Porphyromonas gingivalis*. *Microbiol Mol Biol Rev* 1998;62:1244–1263. [PubMed: 9841671]
93. Yilmaz O, Jungas T, Verbeke P, Ojcius DM. Activation of the phosphatidylinositol 3-kinase/Akt pathway contributes to survival of primary epithelial cells infected with the periodontal pathogen *Porphyromonas gingivalis*. *Infect Immun* 2004;72:3743–3751. [PubMed: 15213114]
94. Nakhjiri SF, Park Y, Yilmaz O, Chung WO, et al. Inhibition of epithelial cell apoptosis by *Porphyromonas gingivalis*. *FEMS Microbiol Lett* 2001;200:145–149. [PubMed: 11425466]
95. Lamont RJ, Chan A, Belton CM, Izutsu KT, et al. *Porphyromonas gingivalis* invasion of gingival epithelial cells. *Infect Immun* 1995;63:3878–3885. [PubMed: 7558295]
96. Holt SC, Kesavalu L, Walker S, Genco CA. Virulence factors of *Porphyromonas gingivalis*. *Periodontol* 2000 1999;20:168–238. [PubMed: 10522227]
97. Yilmaz O, Watanabe K, Lamont RJ. Involvement of integrins in fimbriae-mediated binding and invasion by *Porphyromonas gingivalis*. *Cell Microbiol* 2002;4:305–314. [PubMed: 12027958]
98. Kufer TA, Banks DJ, Philpott DJ. Innate immune sensing of microbes by Nod proteins. *Ann N Y Acad Sci* 2006;1072:19–27. [PubMed: 17057187]

99. Cantin GT, Venable JD, Cociorva D, Yates JR III. Quantitative phosphoproteomic analysis of the tumor necrosis factor pathway. *J Proteome Res* 2006;5:127–134. [PubMed: 16396503]
100. Salih E. Phosphoproteomics by mass spectrometry and classical protein chemistry approaches. *Mass Spectrom Rev* 2005;24:828–846. [PubMed: 15538747]
101. Jensen ON. Interpreting the protein language using proteomics. *Nat Rev Mol Cell Biol* 2006;7:391–403. [PubMed: 16723975]
102. Tanner S, Shu H, Frank A, Wang LC, et al. InsPecT: identification of posttranslationally modified peptides from tandem mass spectra. *Anal Chem* 2005;77:4626–4639. [PubMed: 16013882]
103. Cleveland WS. A program for smoothing scatterplots by robust locally weighted regression. *Amer Statistician* 1981;35:54.

Figure 1.

Scatter plot of $\log_{10}$ of isotopic peptide pairs versus RSD for 1129 proteins associated with more than five isotope pairs, (A). Scatter plot of $\log_{10}$ isotopic heavy/light pairs versus $\log_2$ $^{14}\text{N}/^{15}\text{N}$ signal intensity ratios, (B). In order to simulate a proteome wide analysis in which all abundance ratios were known, we used AH030-54 and AH030-49, biological replicates of *M. maripaludis* with flipped isotope labeling, without normalization. AH030-54 was at natural abundance and AH030-49 was ^{15}N labeled. We adjusted the total amount of AH030-54 (~1.0 mg total protein in each undiluted sample) and AH030-49 according to the results of a Bradford total protein assay such that AH030-49 was diluted approximately 16-fold. The two samples were mixed and run through the MudPIT procedure together, as described previously [61], with changes as noted here. An LTQ instrument was used with acquisition parameters as described [9]. The MAD z -score test was applied to detect outliers with a value of 3.5 or greater [61]. Then, the top five ranked isotope ratios with the lowest MAD z -scores were used to determine the reported ratio and RSD. The average RSD was 19%.

Figure 2. Scatter plot of $\log_2$ total spectral counts for a control population of *P. gingivalis* versus the ratio of spectral counts for two technical replicates of the same sample, for 1074 proteins, (A). Such plots have been used to determine error boundaries and as an aid to determining FPRs and FDRs [9]. The solid line is a LOWESS curve [103] fit to the extreme values above and

below zero. The 19 values in red are false positives for abundance change as determined by a q -value cutoff of 0.01, see Table 3 and [9]. Scatter plot of $\log_2$ of sum of signal intensity measurements for the same two *P. gingivalis* technical replicates versus ratios, (B). The data handling procedures for the intensity data have been described [9]. Only one data point (red) generated a false positive result at a q -cutoff of 0.001, see Table 3. A total of 884 data points were plotted. Taken from Xia *et al.*, (submitted for publication).

Figure 3.

Scatter plot $\log_2$ of sum of spectral counts for the same *P. gingivalis* control sample as shown in Fig. 2 and a 16-fold dilution versus the ratio of spectral counts, (A). The data were used to establish an FNR under these conditions, see Table 3 and supplementary Fig. S1. A total of

975 data points were plotted, 293 proteins shown in red had q -values less than 0.01. The solid lines are LOWESS curves defining the region of random error about zero, see Fig. 2. Scatter plot of $\log_{10}$ of sum of signal intensities versus $\log_2$ ratios for the same data, (B). Data points for 694 proteins were plotted, 693 (red) were significantly greater than zero ($\log_2$ scale), with q -values less than 0.001, see Table 3.

Figure 4. Scatter plot of $\log_2$ sum of spectral counts for two replicates of *M. maripaludis* AH030-104 versus $\log_2$ of AH030-104 ratios of spectral counts for each replicate, (A). 1374 data points were plotted; 35 false positives for non-zero ratios with q -values less than 0.01 are shown in

red, see Tables 1, 2. Scatter plot of $\log_{10}$ sum of signal intensities versus ratios calculated from signal intensity [9], (B). The solid lines are LOWESS curves used to set boundaries on the region of random scatter about zero. Of 1214 proteins, only two (red) had q -values less than 0.001.

Figure 5.

Scatter plot of log₂ sum of spectral counts for *M. maripaludis* AH030-104 and a five-fold dilution versus log₂ AH030-104/dilution spectral count ratios, (A). The lack of adequate statistical power to detect a five-fold simulated abundance change is evident, see Table 2. There were 1397 data points, of which 120 proteins shown in red had *q*-values less than 0.01, yielding an FNR of 92%. The LOWESS curves show the expected region of random scatter about zero, see Fig. 4. Scatter plot of log₂ sum of signal intensities for AH030-104 and dilution versus log₂ AH030-104/dilution signal intensity ratios, (B). Note the much greater power of the signal intensity approach [9] to detect a five-fold change. There were 1212 data points, of which 923

ORFs shown in red had q -values less than 0.001, see Table 2. The LOWESS curves show the region of expected random scatter about zero.

Table 1

Relation of different q -value [10] cut-offs to FPRs (quantitative false positive rates) and FNRs (false negative rates) for *M. maripaludis*. FNR was based on a 16-fold dilution of the proteome. The numbers in boldface were chosen to select the q -value cut-offs for the same data shown in supplemental Figures S2 and S3. Experimental details were as described in the caption for Fig. 1. More detail is given in the caption to supplemental Fig. S2. FNR was calculated by dividing the number of detected proteins not showing change in the correct direction by the total number of detected proteins; FPR by dividing the number of proteins incorrectly showing change by the total number of detected proteins

q cutoff	Spectral Counting Ratio		Signal Intensity Ratio	
	AH030_104 run1/run2	AH030_54/49_dilute 16	AH030_104 run1/run2	AH030_54/49_dilute 16
	FPR (%) ¹	FNR (%) ²	FPR (%) ¹	FNR (%) ²
0.1	7.2	40	2.1	2.5
0.08	6.4	40	1.3	2.5
0.06	5.5	40	0.7	2.5
0.04	4.7	40	0.4	2.5
0.02	3.6	40	0.3	2.5
0.01	2.5	40	0.2	2.5
0.005	2.1	40	0.2	2.6
0.002	1.8	40	0.2	3.1
0.001	1.6	41	0.2	5.1
0.0005	1.4	41	0.2	8.2
0.0002	1.0	41	0.1	12.4
0.0001	0.9	41	0.1	15.7
0.00005	0.9	41	0.1	18.5
0.00002	0.8	42	0.1	21.7
0.00001	0.7	42	0.1	25.1
0.000005	0.6	42	0.1	28.0
0.000002	0.4	43	0.1	31.5
0.000001	0.4	43	0	33.5

¹ Technical replicates of AH030_104, in which all abundance ratios should represent random scatter about a true change of zero.

² AH030_104 and a 16-fold dilution of an isotopic flip biological replicate, AH030_54 (see Sections 2.2, 2.3), in which all abundance ratios should be true non-zeros.

Table 2

Relation of different q -value cut-offs to the FPRs and FNRs for *M. maripaludis*. FNR was based on a 5-fold dilution of the proteome. The numbers in boldface were chosen to select the q -value cut-offs for the same data shown in Figures 4, 5 and supplemental Fig. S4. Experimental details were as described in the captions for Figures 4, 5

q cutoff	Spectral Counting Ratio		Signal Intensity Ratio	
	AH030_104 run1/run	AH030_104_regular/104_dilute	AH030_104 run1/run	AH030_104_regular/104_dilute
	FPR (%) ¹	FNR (%) ²	FPR (%) ¹	FNR (%) ²
0.1	7.2	87	2.1	5.4
0.08	6.4	88	1.3	5.4
0.06	5.5	89	0.7	5.4
0.04	4.7	89	0.4	5.4
0.02	3.6	91	0.3	7.4
0.01	2.5	92	0.2	10.7
0.005	2.1	93	0.2	15.7
0.002	1.8	93	0.2	21.8
0.001	1.6	94	0.2	25.9
0.0005	1.4	94	0.2	30.5
0.0002	1.0	95	0.1	35.3
0.0001	0.9	95	0.1	39.9
0.00005	0.9	96	0.1	43.5
0.00002	0.8	96	0.1	47.7
0.00001	0.7	96	0.1	50.6
0.000005	0.6	96	0.1	53.3
0.000002	0.4	96	0.1	56.5
0.000001	0.4	96	0	59.5

¹ Technical replicates of AH030_104, in which all abundance ratios should represent random scatter about a true change of zero.

² AH030_104 and a five-fold dilution of the same sample (see Section 2.3), in which all abundance ratios should be true non-zeros, 5 on a linear scale.

Table 3

Relation of different q -value cut-offs to the FPRs and FNRs. The numbers in boldface were chosen to select the q -value cut-offs used in the original study of *P. gingivalis* protein abundance. (Xia *et al.*, submitted for publication). See also Figures 2, 3 and supplemental Figure S1

q cutoff	Spectral Counting Ratio		Signal Intensity Ratio	
	PG_PPC2 run1/run2	PG_PPC2/PG_PPC2_dilute 16	PG_PPC2 run1/run2	PG_PPC2/PG_PPC2_dilute 16
	FPR (%) ¹	FNR (%) ²	FPR (%) ¹	FNR (%) ²
0.1	2.8	55	5	1.4
0.08	2.7	56	4.6	1.4
0.06	2.5	58	4.2	1.4
0.04	2.5	61	4.0	1.4
0.02	2.2	65	3.3	1.4
0.01	1.6	69	3.0	1.4
0.005	1.0	71	2.5	1.5
0.002	0.8	74	2.2	2.0
0.001	0.7	76	1.8	2.8
0.0005	0.7	77	1.6	4.1
0.0002	0.7	79	1.1	7.3
0.0001	0.7	81	1.0	9.3
0.00005	0.7	82	0.8	11.5
0.00002	0.6	84	0.7	14.5
0.00001	0.5	84	0.6	17.2
0.000005	0.5	85	0.6	19.7
0.000002	0.5	86	0.6	23.0
0.000001	0.5	87	0.6	25.4

¹ Technical replicates of PG_PPC2, in which all abundance ratios should represent random scatter about a true change of zero.

² PG_PPC2 and a 16-fold dilution of the same sample, in which all abundance ratios should be true non-zeros, in this case the true ratio is 16 on a linear scale.