Published in final edited form as:
Analyst. 2006 December ; 131(12): 1335–1341. doi:10.1039/b610957h.

Comparison of spectral counting and metabolic stable isotope labeling for use with quantitative microbial proteomics

Erik L. Hendrickson[a], **Qiangwei Xia**[a,b], **Tiansong Wang**[a,b], **John A. Leigh**[a], and **Murray Hackett**[b,*]

a *Department of Microbiology, University of Washington, Seattle, Washington 98195 USA*

b *Department of Chemical Engineering, University of Washington, Seattle, Washington 98195 USA*

Summary

Spectral counting, a promising method for quantifying relative changes in protein abundance in mass spectrometry-based proteomic analysis, was compared to metabolic stable isotope labeling using $^{15}N/^{14}N$ "heavy/light" peptide pairs. The data were drawn primarily from a *Methanococcus maripaludis* experiment comparing a wild-type strain with a mutant deficient in a key enzyme relevant to energy metabolism. The dataset contained both proteome and transcriptome measurements. The normalization technique used previously for the isotopic measurements was inappropriate for spectral counting, but a simple adjustment for sampling frequency was sufficient for normalization. This adjustment was satisfactory both for *M. maripaludis*, an organism that showed relatively little expression change between the wild-type and mutant strains, and *Porphyromonas gingivalis*, an intracellular pathogen that has demonstrated widespread changes between intracellular and extracellular conditions. Spectral counting showed lower overall sensitivity defined in terms of detecting a two-fold change in protein expression, and in order to achieve the same level of quantitative proteome coverage as the stable isotope method, it would have required approximately doubling the number of mass spectra collected.

Introduction

An important goal of proteomic analysis is to globally determine differences in protein levels between different biological states, such as mutant versus wild-type microbial strains or growth under different nutrient conditions. In recent years the "gold standard" for such global measurements of relative protein expression has been multidimensional capillary HPLC coupled with tandem mass spectrometry using differential stable isotope labeling.[1,2] Unfortunately, isotopic labeling is not always straightforward, or even possible. Chemical labeling strategies employed after cell harvesting often tend to yield poor coverage of the proteome, and it is not always possible to metabolically label at the cellular or tissue level, especially in the case of human proteomics in a clinical setting. In a laboratory setting, prokaryotic organisms can often be grown on minimal media in which isotopically enriched salt or gas can be used to introduce ^{13}C or ^{15}N as the sole carbon or nitrogen source, respectively. However not all microorganisms of research interest can be so cultured. For these cases, label-free methods for quantifying expression differences are also needed. A promising label-free quantitation method is spectral counting, where the number of mass spectra identified for a protein is used as a measure of the protein's abundance.[3,4] Old *et al.* have shown the spectral counting method compares favorably with label-free peak area quantitation, although

*Correspondence to Department of Chemical Engineering, Box 355014, University of Washington, Seattle, Washington USA. Voice: (206) 616-8071; Fax: (206) 616-5721; e-mail: mhackett@u.washington.edu.

they suggested that the method is likely less sensitive than isotopic labeling.[5] The same group applied spectral counting to analyzing proteins extracted from leukemia cell membranes.[6] Zybailov and coworkers recently compared quantitative MudPIT results using spectral counting and stable isotope peak intensity to derive protein expression ratios from *S. cerevisiae* grown under minimal and rich media conditions, demonstrating moderately positive correlations between the two approaches to quantitation.[7] In our studies of the proteome and transcriptome of the archaeal methanogen *Methanococcus maripaludis* we have obtained an extensive, metabolically labeled proteomic dataset[8] comparing a mutant,[9] called S40, and a wild-type strain, S2. For this organism, transcription measurements tend to parallel the direction, but not the magnitude, of expression change as measured by proteomics for most protein encoding ORFs, subject to certain caveats regarding growth phase and timing with respect to sample collection for mRNA and protein.[8] In order to see how well spectral counting compared to isotopic labeling and the transcriptome, we have re-examined the proteomic dataset using spectral counting and compared these results with both the isotopic labeling and transcriptome results. We also reference results for the invasive intracellular oral pathogen *Porphyromonas gingivalis*, an organism quite different from *M. maripaludis*. These two organisms can serve as models to represent the extremes in the continuum of prokaryotic biology with respect to two qualities: ease of growth on fully defined growth medium with a ^{15}N-labeled nitrogen source and the degree of change observed under common experimental conditions in a two-state differential expression comparison. *M. maripaludis* is easy to label metabolically with ^{15}N, and showed only modest expression differences between strains S40 and S2.[8,9] *P. gingivalis* is difficult to grow economically on a single labeled nitrogen source,[10] and in two-state experiments contrasting *P. gingivalis* grown under extracellular reference conditions and internalized within model human host cells, the proteome changes are dramatic and widespread.[11,12] The primary difference observed between protein expression ratios determined using ^{14}N and ^{15}N peptide MS^1 signal intensity measurements and spectral counting in MS^1 was in sensitivity to changes in protein expression determined from portions of the raw data that were either low signal-to-noise or low in spectral counts relative to the dataset as a whole.

Experimental

Culture conditions, mass spectrometry and transcription microarrays

For information regarding the *M. maripaludis* strains, growth conditions, isotope labeling, mass spectrometry, proteomic data collection, mRNA extraction, cDNA preparation, labeling, and hybridization see Porat *et al.*[9] and Xia *et al.*[8] The latter reference contains a detailed explanation of how the *M. maripaludis* protein expression ratios were originally calculated using stable isotope labeling. Briefly, after tryptic digestion of the entire proteome extracted from 10^9 to 10^{10} cells per preparation following standard procedures for shotgun proteomics, for *M. maripaludis* an LCQ ion trap mass spectrometer (Thermo Electron Corp., San Jose, CA, USA) was interfaced to an in-house modified Michrom Magic 2002 HPLC system (Michrom BioResources, Auburn, CA, USA) and used for data dependent scanning[13,14] of proteolytic digests using a variant of MudPIT (multidimensional protein identification technology)[1,2] that was optimized for organisms with approximately 2,000 protein encoding ORFs. Raw data collection of approximately 700,000 mass spectra for *M. maripaludis* was followed by matching the peptide mass spectra using SEQUEST[15] with a database consisting of all known ORFs from *M. maripaludis* concatenated with the human subset of the nrdb (non-redundant database).[16] Proteins were reassembled and quantified *in silico* by using DTASelect[17] to globally filter the raw data for quality and to group the filtered SEQUEST output files for each peptide according to the protein from which they were derived. The data were converted into text format using routines contained in the Xcalibur data system developer's kit (Thermo) and stored in a Filemaker Pro database. Subsequent data processing was carried out in either

Filemaker Pro or a Microsoft Excel spread sheet. Redundant spectral counts as defined below were summed for each ORF and normalized protein level expression ratios were calculated as described below. The *P. gingivalis* data also shown in Fig. 1 were acquired similarly, the significant differences being the use of a *P. gingivalis* ORF database from TIGR[18] and an LTQ (Thermo) rather than an LCQ ion trap. The details of the growth of *P. gingivalis* ATCC strain 33277, protein extraction, prefractionation, and MudPIT chromatography are given in Zhang *et al.*[11] The most current details regarding data acquisition parameters used specific to the LTQ mass spectrometer and software routines used to generate protein expression ratios for *P. gingivalis* (Fig. 1) using spectral counting can be found in Xia, Wang *et al.*[12]

The entire *M. maripaludis* dataset is available as an electronic supplement.[8] The same dataset can also be downloaded from the GEO depository[19] in a somewhat different form using GEO Series Accession Numbers GSE2744 for the proteomics and GSE2745 for the spotted cDNA microarrays.

Protein expression ratio and *p*-value calculations

For information regarding the post acquisition data analysis of the transcriptome or the isotopic ratio calculations for the *M. maripaludis* proteome, see Xia *et al.*[8]

For spectral counting, the S40/S2 protein expression ratios were calculated as shown:

$$R_{sc}=log_2[(n_{s40}+1)(t_{s2}/t_{s40})/(n_{s2}+1)] \quad (1)$$

Where for each protein, R_{sc} is the $\log_2$ ratio of abundance between S40 and S2; n_{s40} and n_{s2} are the redundant spectral counts for the protein in S40 and S2 respectively; t_{s40} and t_{s2} are the sum of n_{s40} + 1 and n_{s2} + 1, respectively, over all proteins. Redundant spectral counts are defined as those acquired at all stages of the MudPIT analysis, including instances of the same peptide fragment being detected in different fractions. The ratio t_{s2}/t_{s40} was used as a normalization factor. The Differences in spectral counts were identified by applying a likelihood ratio test (G test)[5,20] for independence using normalized values and a null hypothesis of even distribution between the two strains as shown:

$$G=2[c_{s2}\,ln(c_{s2}/t_{cs2})+c_{s40}\,ln(c_{s40}/t_{cs2})] \quad (2)$$

Where for each protein, G is the G test statistic; c_{s2} is n_{s2} + 1; c_{s40} is $(n_{s40} + 1)(t_{s2}/t_{s40})$; and t_{cs2} is $(c_{s2} + c_{s40})/2$. The G statistic is approximately distributed as χ^2 with 1 degree of freedom, allowing the calculation of *p*-values for identifying differential expression. The *p*-value calculations were performed as described previously for proteomics data.[5,12] Briefly, after generating a G test statistic for each protein, a *p*-value was calculated as the probability that a χ^2 distribution with 1 degree of freedom was more extreme than our G statistic for that protein.

Results and discussion

Here we present the results from a global, quantitative, "two state" protein expression ratio analysis comparing two strains of *M. maripaludis*, using spectral counting (SC) to determine the ratios, which were then tested statistically for significant change in expression level between the two strains. In the previous paper cited in the introduction, we compared the genome wide differential expression results from proteome and transcriptome analyses of a mutant, S40, and a wild-type strain, S2, of *M. maripaludis*. In the earlier paper,[8] the proteome results were obtained using *m/z* peak height calculations from isotopically labeled samples. Here we

compare the results calculated using SC with the results for the same raw data previously calculated using conventional stable isotope methods.

Data reduction and normalization

When using SC, some proteins will have spectra in only one of the conditions, *i.e.* $n = 0$ in one state but not the other (see Eq. 1), which presents problems in the calculations. Such proteins may be very important in terms of gene regulation, so in order to handle these values, all the spectral counts were increased by one before beginning the calculations. Adding the additional count does have the undesired effect of smoothing out differences between samples at the low end, but the effect decreases with increasing counts, becoming increasingly less significant beyond the quantized region shown in Figs. 1, 2. The term "quantized region" refers to the left side, near zero on the x-axis, where total spectral counts are low and assume a narrow range of allowed values, an artifact that is intrinsic to the method. How far to the right the quantized region extends in relation to the complete dataset is a function of the number of peptides recovered that map *in silico* to each protein encoding ORF predicted by the genome annotation. The better the coverage for each ORF, the smaller the quantized region as a fraction of all the protein level expression ratios in the dataset. In absolute terms, this quantized region is a function of the allowable ratios that can be calculated using discrete values, a number that rapidly diminishes as the total spectral counts approach zero. For example, in Fig. 1, panels B and C, essentially all the values between 0 and 4 on the x-axes are predictors of a true value of zero, despite the wide spread of a few allowed values on the y-axes. Very few proteins show a change in expression in this dataset regardless of total spectral count value, so most of the points shown across the x-axes beyond a value of 4 also reflect random scatter about a net expression change of zero, and random scatter rapidly diminishes as total spectral counts increase. The artifacts potentially introduced by the addition of one spectral count (Eq. 1) are most likely to have their greatest impact in the quantized region. The quantized region at low counts is of little practical use for generating biologically meaningful expression ratios, regardless of which method is chosen to avoid a divide-by-zero condition and other problems. Smoothing algorithms that hide the discontinuities in the low counts region only serve to help disguise the fact that such data consists mostly of random scatter about zero, quantized into a small number of allowed values, see Eq. 1 and Fig. 2. Only in extreme cases is it possible to call an ORF as alternative, i.e. non-zero expression change on a $\log_2$ scale, in the quantized low counts portion of an SC dataset.

Like transcriptome datasets, proteomic datasets need to be normalized to account for differences in overall signal between samples from the two biological conditions under study. However, the degree of normalization employed with proteomics data is usually less than that used routinely with microarrays dependent on the use of fluorescent dyes, when the adjustments may be as much as an order of magnitude or greater due to differences in fluorescence quantum yield, among other factors. For the peak height calculations in Xia *et al.*,[8] the ratios were normalized to an average $\log_2$ ratio of zero by plotting a frequency distribution histogram of the ratios and applying a correction factor to center the distribution at zero. Given the highly quantized nature of the SC data (Fig. 1A), the above mentioned approach was impractical for SC. However, a plot of ratios against the total counts (Fig. 1B), both $\log_2$ transformed, showed a skew towards the S2 sample. This observation was consistent with the larger overall number of counts for the S2 sample. To normalize for the difference in overall counts, the S40 counts were multiplied by the ratio of overall S2 counts to overall S40 counts making the total counts for each strain equal, as shown in Eq. 1. This is a similar normalization scheme to that applied by Old *et al.*[5] The normalized values were then used to calculate expression ratios as well as the total spectral counts for each protein. A plot of the normalized ratios versus total counts (Fig. 1C) showed the $\log_2$ transformed ratios centered around zero, indicating that further normalization was unwarranted. Experience with *M. maripaludis* shows that expression

changes in this organism tend to be few under many experimental conditions, so it is a good example of a system with minimal changes. Many microorganisms of interest, especially pathogens, are more likely to have widespread changes between conditions. In order to see if the same considerations and normalization would work with such an organism, a dataset from the oral pathogen *Porphyromonas gingivalis* was analyzed using SC in the same manner as the *M. maripaludis* data (Fig. 1D, E, F). The *P. gingivalis* data appears to have significantly more scatter than the *M. maripaludis* data (Fig. 1D). This observation is all the more of interest because the *P. gingivalis* dataset was collected using an LTQ mass spectrometer with superior analytical figures-of-merit in terms of signal-to-noise, mass accuracy and run-to-run repeatability, relative to the LCQ used to acquire the *M. maripaludis* dataset.[21] Another difference between the two datasets is that for *M. maripaludis*, the proteomes from both strains were mixed and analyzed at the same time, as is commonly done for a stable isotope experiment. For *P. gingivalis*, the two biological states referenced in Fig. 1 were run separately. The frequency distribution histogram was not as strongly influenced by the quantized nature of the data, compared to *M. maripaludis*, but it was still not ideal for centering (Fig. 1D). Figs. 1E and 1F show that while a significantly larger number of ORFs show expression changes in the *P. gingivalis* dataset, the normalization is still effective.

Detection of differentially expressed protein-encoding ORFs

Differential expression of protein between the S2 and S40 strains was identified using a G test for significance (Eq. 2, Fig. 2). Discussion of counting statistics as applied to SC in proteomics and SAGE (serial analysis of gene expression) can be found in Old *et al.*,[5] Xia, Wang *et al.* [12] and the references contained therein. SAGE is a global transcription analysis technique with similar data analysis requirements[22] that has inspired to some degree the use of G tests and related counting statistics with SC in proteomics. The G test calculations were conducted using the normalized dataset with a null hypothesis of equal distribution between S2 and S40, *i.e.* all $\log_2$ transformed protein expression ratios were equal to zero. Significance levels were determined by comparing the G value to a χ^2 statistic with one degree of freedom for the selected level of significance. As shown previously,[12] it is important to note that in this application the G statistic is conservative, tending on average to generate more false negatives than false positives, but not as conservative as the Bonferroni correction[20] for multiple hypothesis testing. Old *et al.*[5] found that a 95% critical value might be more than 95% accurate when applied to a replicate test case; that is their cut-off value for determining significant change was also conservative. For our case the 95% critical value line (Fig. 2) does seem to fit the scatter in our data appropriately (p-value = 0.05 or less for all ORFs called as alternative). The critical value line is to be understood as predictive of the null hypothesis being true within its boundaries, and false for ratios that fall outside. Validation using other methods[8,9] tends to support the predictions based on p-value for *M. maripaludis* protein expression.

Using the 95% significance cutoff, 33 *M. maripaludis* proteins were found to be differentially expressed between S2 and S40 (Table 1 and Fig. 2). Of these, only 12 were found to be regulated by all three methods: arrays (mRNA), SC (protein) and stable isotope (protein). All of these were cases of greater expression in the S40 mutant strain (Fig. 2). As seen in Table 1, there was a greater inconsistency among all methods in identifying reduced expression in S40. For reduced expression, the large differences between the two proteomics methods imply that there is unlikely to be a biologically significant difference between the protein and mRNA datasets, but rather a situation where most of the true significant changes occur only in the up direction as a result of the underlying biology.[8,9] We found that in the isotope analysis[8] a two-fold change, 1 or −1 on the $\log_2$ scale, could be detected reliably if the number of heavy/light $^{15}N/^{14}N$ peptide pairs exceeded ~10; 417 ORFs met the two-fold change detection criterion. As seen in Fig. 2, SC could detect a two-fold change in expression if the peptide counts, after adding one to the counts and normalizing, exceeded ~34; 277 ORFs fell into this

two-fold range for SC. The number of peptide pairs and number of SC peptide counts cannot be directly compared, as the SC counts have been modified and normalized while the peptide pairs for the isotopic analysis have been filtered to exclude outliers. Nonetheless, looking at the number of ORFs that fall into the two-fold detection range for each method, the isotopic analysis is clearly more sensitive.

With large datasets, such as the *M. maripaludis* proteome, a correction is often applied to the *p*-values to account for the large number of significance tests. The most common is the Bonferroni correction noted above, which controls for error over the entire set. However, for very large datasets like those under consideration here, this correction is often too conservative and eliminates all results from consideration as alternatives to the null hypothesis (no expression change). A newer, alternative correction method is to calculate *q*-values that control for the proportion of false positives among those ORFs identified as significant.[23] However, no such correction was applied to the array or stable isotope measurements described here. Such a correction would make the significance calls slightly more conservative, eliminating some calls close to the confidence line cutoff shown in Fig. 2. The G test and *p*-values (Eq. 2) were adequate for the *M. maripaludis* dataset in terms of striking a proper balance between false positives and false negative calls without further adjustment.

Fig. 3 shows plots, centered about a net expression change of zero on a $\log_2$ scale, comparing the results from the two protein expression ratio methods against the transcription microarray results[8] and against each other. In the isotope analysis there was a weak overall correlation between the proteomic and transcriptome results (R = 0.27, Fig. 3A). The weak correlation was mostly driven by the large number of results that were unchanged between the two strains, those near zero where random noise would cause significant scatter between the two measurements. When restricted to the regulated ORFs the correlation improved significantly (R = 0.58, Fig. 3A). The SC analysis produced very similar results (Fig. 3B), with a slightly lower correlation for the overall comparison (R = 0.26) and slightly higher when comparing only the regulated ORFs (R = 0.66). While neither proteomics analysis produced a better correlation with the transcriptome, there were significant differences between the results of the two protein methods (Fig. 3C). Overall correlation between the two methods is better than that with the transcriptome (R = 0.57 overall and 0.89 for the regulated ORFs), but still far from unity despite being applied to the same dataset. However, most of the differences fall into two categories that have relatively little effect on the most important results. First, as with the transcriptome comparison, unchanged ORFs do not correlate well. Since the goal is to identify changes in expression between the two strains, unchanged ORFs are not of great interest. Second, given the level of quantitative uncertainties in the whole cell proteomics data, the direction of change in expression is normally viewed as of greater practical interest than the magnitude of change. For most of the ORFs where a change in expression is seen, both protein methods show that change in the same direction, see Fig. 3C. Other, better methods exist to determine the magnitude of expression change, where it is usually more efficient to work at the transcription level using an approach like quantitative real-time RT-PCR (reverse-transcriptase polymerase chain reaction)[24] for the small number of genes that are typically of sufficient interest to justify further study. The exception to this last statement would be cases of known or suspected post-transcriptional gene regulation, where a more precise direct measure of protein abundance would be of greater interest. More in-depth discussion of transcription and translated protein expression correlations for *M. maripaludis* can be found in Porat *et al*.[9] and Xia *et al*.[8]

Conclusions

The most significant difference observed between metabolic isotope labeling and SC was overall sensitivity. Stable isotope labeling yielded approximately 50% more ORFs that fell into

a range where a two-fold expression difference could be detected. Making the simplifying assumption that sampling remained somewhat consistent across the proteome, under our conditions it would thus take roughly double the number of mass spectra under the two-fold change criterion to obtain the same number of statistically significant expression ratios using SC. Despite the many problems with such an assumption regarding uniform sampling and the overall complexity of the analytical scheme, such an estimate is still useful as a guide to the depth and breadth of sampling required to achieve a desired level of confidence in the expression ratios generated in a shotgun proteomics experiment. At high signal-to-noise and (or) counts, the two methods converge on average, within the dynamic range of the isotope ratio measurements, which is not as great as that associated with SC.[4] It is in the more problematic portions of the dataset, where signal-to-noise is relatively poor and (or) total counts are low that metabolic labeling with stable isotopes appears to allow generation of a larger number of biologically meaningful and statistically significant expression ratios. We agree with the conclusions of Zybailov *et al.*[7] and others that SC approaches appear to have a wider dynamic range relative to measurements based on signal intensity, and have so noted in two prior publications, as well as noting the excellent reproducibility typical of SC combined with multidimensional capillary HPLC separations.[10,12] However, dynamic range was not systematically investigated as part of the present study. Our overall impression of spectral counting to date is that it performs poorly when counts are low (see Figs. 1 and 2), but performs quite well when counts and signal-to-noise are high, a conclusion that is not inconsistent with the data presented by Zybailov *et al.* in their study of yeast membrane proteins.[7] The problem with SC in the present context is that for biological questions driving ongoing research with *M. maripaludis* and *P. gingivalis*, the low counts and (or) low signal-to-noise portion of the data is often of the greatest experimental interest, where traditional stable isotope methods still offer clear advantages for quantitation. The issue of under-sampling is always a concern in a shotgun proteomics study, even with organisms expressing relatively small numbers of protein encoding ORFs, such as the model systems used in this report. While the current state-of-the-art in mass spectrometry instrumentation and pre-fractionation methods could allow for doubling the amount of data collected as a practical strategy for organisms such as *M. maripaludis* or *P. gingivalis*, it would be impractical at present for organisms with larger, more complex proteomes. Microbiologists usually have the option of metabolic labeling, and thus avoiding the lower quantitative proteome coverage we have observed with SC. However, many scientists studying human proteomics and others already dealing with serious under-sampling issues do not have this option, and for them SC may still be the best available quantitative approach, despite the limitations described above. Improved mass spectrometry instrumentation, that can maintain unit resolution or better for precursor ion selection while increasing the number of scans that can be acquired per unit time, will facilitate the application of spectral counting methods to a wider range of biological investigations by reducing the impact of under-sampling.

Acknowledgements

We thank Fred Taub, Kevin Wheeler, Jim Shofstahl, Richard J. Lamont, and Yoonsuk Park for their assistance; Michrom Bioresources for equipment support; supported by the United States Department of Energy under Microbial Cell Project DE-FG03-01ER15252 (J. A. L.), and by the Department of Health and Human Services under NIH grants DE014372 (M. H.) and GM60403 (J. A. L.).

References

1. Washburn MP, Wolters D, Yates JR III. Nature Biotech 2001;19:242–247.
2. Washburn MP, Ulaszek R, Deciu C, Schieltz DM, Yates JR III. Anal Chem 2002;74:1650–1657. [PubMed: 12043600]
3. Gao J, Opiteck GJ, Friedrichs MS, Dongre AR, Hefta SA. J Proteome Res 2003;2:643–649. [PubMed: 14692458]

4. Liu H, Sadygov RG, Yates JR 3rd. Anal Chem 2004;76:4193–4201. [PubMed: 15253663]
5. Old WM, Meyer-Arendt K, Aveline-Wolf L, Pierce KG, Mendoza A, Sevinsky JR, Resing KA, Ahn NG. Mol Cell Proteomics 2005;4:1487–1502. [PubMed: 15979981]
6. Ruth MC, Old WM, Emrick MA, Meyer-Arendt K, Aveline-Wolf LD, Pierce KG, Mendoza AM, Sevinsky JR, Hamady M, Knight RD, Resing KA, Ahn NG. J Proteome Res 2006;5:709–719. [PubMed: 16512687]
7. Zybailov B, Coleman MK, Florens L, Washburn MP. Anal Chem 2005;77:6218–6224. [PubMed: 16194081]
8. Xia Q, Hendrickson EL, Zhang Y, Wang T, Taub F, Moore BC, Porat I, Whitman WB, Hackett M, Leigh JA. Mol Cell Proteomics 2006;5:868–881. [PubMed: 16489187]
9. Porat I, Kim W, Hendrickson EL, Xia Q, Zhang Y, Wang T, Taub F, Moore BC, Anderson IJ, Hackett M, Leigh JA, Whitman WB. J Bacteriol 2006;188:1373–1380. [PubMed: 16452419]
10. Lamont RJ, Meila M, Xia Q, Hackett M. Infect Disorders Drug Targ 2006;6:311–325.
11. Zhang Y, Wang T, Chen W, Yilmaz O, Park Y, Jung IL, Lamont RJ, Hackett M. Proteomics 2005;5:198–211. [PubMed: 15619293]
12. Xia Q, Wang T, Park Y, Lamont RJ, Hackett M. Int Jour Mass Spectrom. 2006in press
13. Ducret A, Oostveen IV, Eng JK, Yates JR III, Aebersold R. Prot Science 1998;7:706–719.
14. Washburn MP, Yates JR III. Curr Opin Microbiol 2000;3:292–297. [PubMed: 10851159]
15. Eng JK, McCormack AL, Yates JR III. J Am Soc Mass Spectrom 1994;5:976–989.
16. ftp://ftp.ncbi.nih.gov/blast/db/FASTA/
17. Tabb DL, McDonald WH, Yates JR III. J Proteome Res 2002;1:21–26. [PubMed: 12643522]
18. http://cmr.tigr.org/tigr-scripts/CMR/CmrHomePage.cgi
19. See the GEO URL, http://www.ncbi.nlm.nih.gov/projects/geo/query/acc.cgi.
20. Sokal, R.; Rohlf, F. Biometry: the Principles and Practice of Statistics in Biological Research. W. H. Freeman; New York: 1995.
21. Blackler AR, Klammer AA, MacCoss MJ, Wu CC. Anal Chem 2006;78:1337–1344. [PubMed: 16478131]
22. Man MZ, Wang X, Wang Y. Bioinformatics 2000;16:953–959. [PubMed: 11159306]
23. Storey JD, Tibshirani R. Proc Natl Acad Sci U S A 2003;100:9440–9445. [PubMed: 12883005]
24. Bustin SA, Benes V, Nolan T, Pfaffl MW. Jour Molec Endocrinol 2005;34:597–601. [PubMed: 15956331]
25. Hendrickson EL, Kaul R, Zhou Y, Bovee D, Chapman P, Chung J, Conway de Macario E, Dodsworth JA, Gillett W, Graham DE, Hackett M, Haydock AK, Kang A, Land ML, Levy R, Lie TJ, Major TA, Moore BC, Porat I, Palmeiri A, Rouse G, Saenphimmachak C, Soll D, Van Dien S, Wang T, Whitman WB, Xia Q, Zhang Y, Larimer FW, Olson MV, Leigh JA. J Bacteriol 2004;186:6956–6969. [PubMed: 15466049]

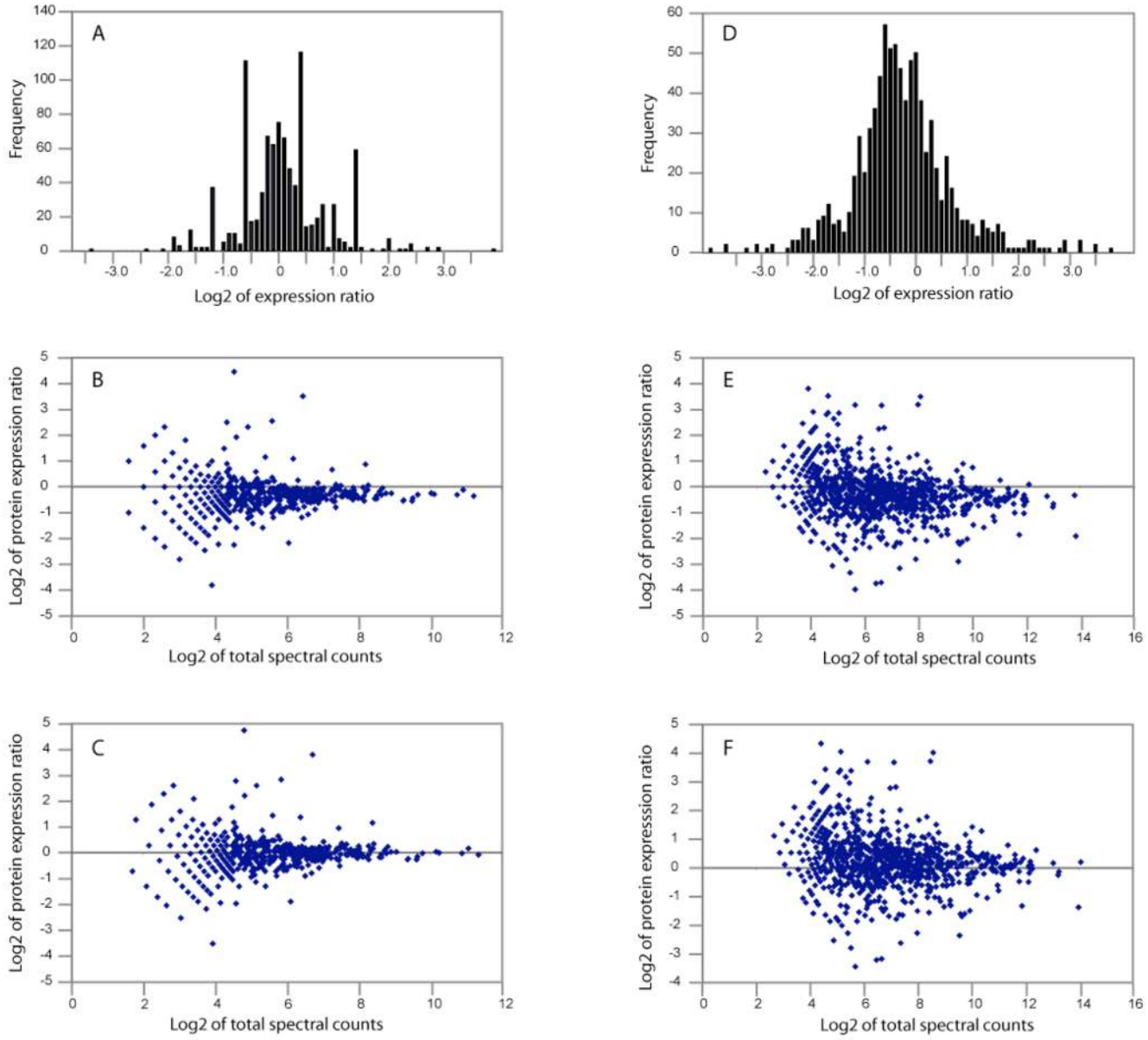

Fig. 1.

Normalization of proteomic expression ratios for spectral counting. (A) Histogram of the $\log_2$ transformed S40/S2 ratios for the proteomics data calculated by spectral counting, prior to normalization. (B) Plot of the $\log_2$ transformed S40/S2 expression ratios against the $\log_2$ transformed total number of spectral counts used to calculate the ratio. (C) Plot of the $\log_2$ transformed S40/S2 expression ratios against the $\log_2$ transformed total number of spectral counts used to calculate the ratios, after normalization. The data were normalized by multiplying the number of S40 counts used in the calculation by the total number of S2 counts divided by the total number of S40 counts for the entire dataset, see text. (D) Histogram of the $\log_2$ transformed ratios from a *Porphyromonas gingivalis* dataset contrasting two different growth conditions, prior to normalization (E) Plot of the $\log_2$ transformed *P. gingivalis* expression ratios against the $\log_2$ transformed total number of spectral counts used to calculate the ratio. (F) Plot of the $\log_2$ transformed *P. gingivalis* expression ratios against the $\log_2$

transformed total number of spectral counts used to calculate the ratio, after normalization. The data were normalized by multiplying the number of growth state 1 counts used in the calculation by the total number of growth state 2 counts divided by the total number of growth state 1 counts for the entire dataset, similarly to the *M. maripaludis* data.

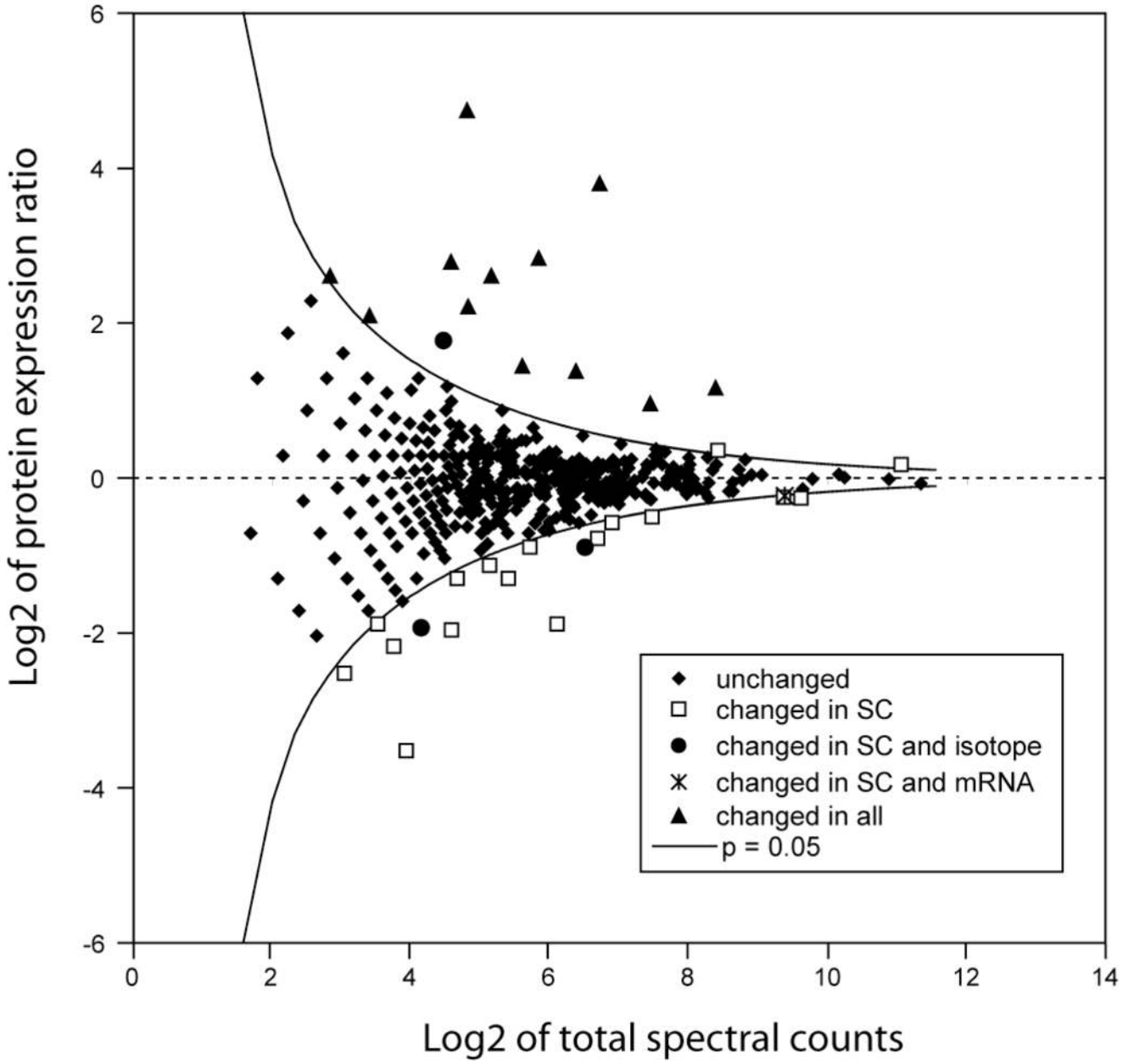

Fig. 2.

Plot of normalized log2 transformed S40/S2 expression ratios against the total number of spectral counts used to calculate the ratios. Curved lines show the 95% critical value thresholds ($p = 0.05$) for determining differential expression from spectral counting calculated using the G test statistic, see text. The dashed line indicates a $\log_2$ transformed ratio of zero, indicating no expression change between the two strains. Diamonds: unchanged expression between S40 and S2; Open squares: ORFs called differentially expressed only by the SC method; Circles: ORFs called differentially expressed by both SC and isotopic labeling analysis of the proteome: Crosses: ORFs called differentially expressed by both SC and mRNA measurements; Triangles: ORFs called differentially expressed by all three methods.

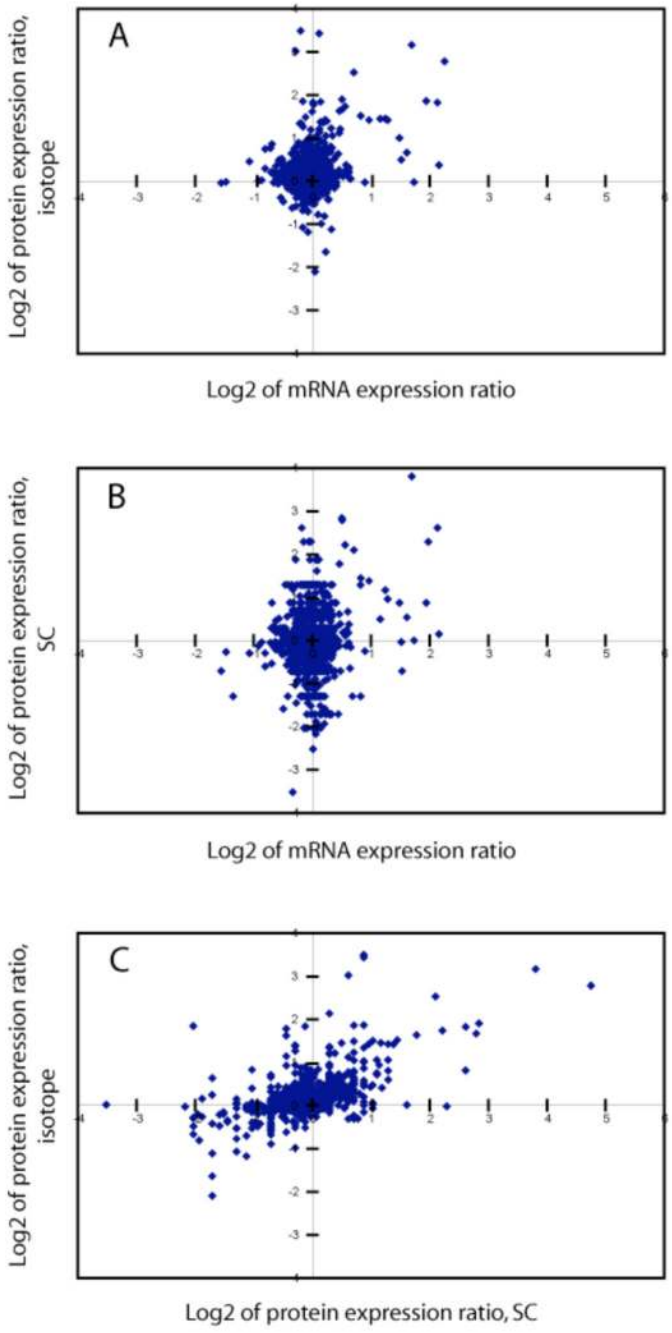

Fig. 3.

Scatter plots showing correlation between mRNA expression ratios and protein expression ratios calculated by both spectral counting and isotopic peak height methods. (A) Plot of $\log_2$ transformed differential mRNA and protein expression ratios generated by isotopic peak height measurements.[8] Product-moment correlation coefficients[20] were: 0.27 for all data points and 0.59 for the data that showed significant changes in both mRNA and by the isotopic peak height calculations. (B) Plot of $\log_2$ transformed differential mRNA and protein expression ratios generated by SC. Product-moment correlation coefficients were: 0.26 for all data points and 0.66 for the data that showed significant changes in both mRNA and by SC. (C) Plot of $\log_2$ transformed differential protein expression ratios generated by isotopic peak

height measurements and SC respectively. Product-moment correlation coefficients were: 0.57 for all data points and 0.89 for the data that showed significant changes in protein expression ratios by both methods.

Table 1

Combined summary of transcriptome and both isotopic peak height analysis and spectral counting analysis for the proteome for S40/S2 expression ratios. The number of ORFs in each category are shown: Up = significantly higher expression in S40 compared to S2; Down = significantly lower expression in S40 compared to S2.; No change = no significant difference; NP = not probed in microarrays; ND = not detected using proteomics. The total number of ORFs in the annotated *M. maripaludis* genome[25] is 1722. Portions of this table have been reproduced from Xia *et al.*[8] with permission of the ASBMB.

			Transcriptome				Isotope			
			Up	**Down**	**No Change**	**NP**	**Up**	**Down**	**No Change**	**ND**
		Total	**50**	**45**	**1502**	**125**	**60**	**34**	**845**	**783**
Spectral counting	Up	15	12	0	3	0	13	0	2	0
	Down	18	0	1	16	1	0	2	16	0
	No Change	917	26	33	809	49	46	32	828	11
	ND	772	11	11	675	75	0	0	0	772
Isotope	Up	60	15	2	41	2				
	Down	34	0	0	31	3				
	No Change	845	24	32	744	45				
	ND	783	11	11	686	75				