Mass spectrometry of the *M. smegmatis* proteome: Protein expression levels correlate with function, operons, and codon bias

Rong Wang, John T. Prince, and Edward M. Marcotte[1]

Center for Systems and Synthetic Biology, Institute for Cellular and Molecular Biology, and Department of Chemistry and Biochemistry, University of Texas at Austin, Austin, Texas 78712, USA

The fast-growing bacterium *Mycobacterium smegmatis* is a model mycobacterial system, a nonpathogenic soil bacterium that nonetheless shares many features with the pathogenic *Mycobacterium tuberculosis*, the causative agent of tuberculosis. The study of *M. smegmatis* is expected to shed light on mechanisms of mycobacterial growth and complex lipid metabolism, and provides a tractable system for antimycobacterial drug development. Although the *M. smegmatis* genome sequence is not yet completed, we used multidimensional chromatography and tandem mass spectrometry, in combination with the partially completed genome sequence, to detect and identify a total of 901 distinct proteins from *M. smegmatis* over the course of 25 growth conditions, providing experimental annotation for many predicted genes with an ~5% false-positive identification rate. We observed numerous proteins involved in energy production (9.8% of expressed proteins), protein translation (8.7%), and lipid biosynthesis (5.4%); 33% of the 901 proteins are of unknown function. Protein expression levels were estimated from the number of observations of each protein, allowing measurement of differential expression of complete operons, and the comparison of the stationary and exponential phase proteomes. Expression levels are correlated with proteins' codon biases and mRNA expression levels, as measured by comparison with codon adaptation indices, principle component analysis of codon frequencies, and DNA microarray data. This observation is consistent with notions that either (1) prokaryotic protein expression levels are largely preset by codon choice, or (2) codon choice is optimized for consistency with average expression levels regardless of the mechanism of regulating expression.

[Supplemental material is available online at www.genome.org. The mass spectrometry raw data from this study have been deposited in the Open Proteomics Database http://bioinformatics.icmb.utexas.edu/OPD, under accession nos. opd00007_MYCSM–opd00031_MYCSM. The following individuals kindly provided reagents, samples, or unpublished information as indicated in the paper: D. Graham.]

The fast-growing nonpathogenic bacterium *Mycobacterium smegmatis* is particularly useful in studying basic cellular processes of relevance to pathogenic mycobacteria, such as the related species *Mycobacterium tuberculosis*, the causative agent of tuberculosis. Although the genome sequencing of *M. smegmatis* is nearly complete (http://www.tigr.org; Brosch et al. 2001), much is unknown about the mechanisms controlling growth in mycobacterial species. The large-scale study of the proteins expressed by *M. smegmatis* in different growth states has the potential to generate information about the mechanisms of cell growth, division, and adaptation, as well as inform about mycobacterial proteomes in general.

Until recently, the method of choice for profiling a complete proteome was two-dimensional gel electrophoresis coupled with mass spectrometry (2DE-MS). For example, using this approach, a total of 263 proteins were identified in *M. tuberculosis* and *Mycobacterium bovis* BCG strains, the proteome of *M. tuberculosis* H37Rv was compared with that of *M. bovis* BCG Chicago, and 25 proteins differing in position or intensity were identified (Jungblut et al. 1999). Similarly, 137 proteins were detected in *M. tuberculosis* H37Rv culture supernatant, and 27 unique proteins were identified in *M. tuberculosis* H37Rv by comparing proteins in the culture supernatant of virulent *M. tuberculosis* H37Rv to that of attenuated *M. bovis* BCG Copenhagen (Mattow et al. 2003). However, recent advances in multidimensional liquid chromatography coupled with tandem mass spectrometry (LC/LC/MS/MS) (Washburn et al. 2001) have produced a technology capable of direct analysis of the composition of protein mixtures as complex as cell lysates (Aebersold and Mann 2003). In this method, protein mixtures are digested with proteases, and the resulting peptides are separated by multidimensional liquid chromatography, bypassing potential limitations of gel electrophoresis and protein insolubility; then the separated peptides are analyzed sequentially by MS/MS. Interpretation of the MS/MS peptide spectra, for example, by using algorithms such as SEQUEST (Eng et al. 1994) or Mascot (Perkins et al. 1999), leads to identification of the proteins in the mixture.

Using this method, ~1500 *Saccharomyces cerevisiae* proteins were detected (Washburn et al. 2001; Peng et al. 2003). Similarly, in mycoplasma, Jaffe et al. (2004) detected the expression of 557 open reading frames (ORFs) in *Mycoplasma pneumoniae* strain M129 by using proteogenomic mapping, the mapping of peptides detected in the cell lysate onto the uninterpreted genome. Here, we apply LC/LC/MS/MS to characterize the expressed proteome of *M. smegmatis*, and we report observation of 901 distinct proteins under differing growth conditions, estimate relative abundance of each protein, and demonstrate that the relative

[1] Corresponding author.
E-mail marcotte@icmb.utexas.edu; fax (512) 232-3432.

Article and publication are at http://www.genome.org/cgi/doi/10.1101/gr.3994105.

15:1118–1126 ©2005 by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/05; www.genome.org

abundance correlates with codon choice and mRNA expression levels.

Results

Experimental observation of 901 proteins in the *M. smegmatis* proteome and functions of the observed proteome

Approximately 825,000 MS/MS peptide fragmentation spectra were collected and analyzed over the course of 25 LC/LC/MS/MS experiments, characterizing the proteins expressed in each of 25 samples drawn from time courses of *M. smegmatis* growing in three different media. At an estimated false-positive identification rate <5%, we identified a total of 901 *M. smegmatis* proteins (Fig. 1). These identified proteins represent ~10% of the 8968 predicted genes identified in the unfinished *M. smegmatis* genome. Of the proteins 94% were detected in more than one experiment, with a few proteins (2%) detected in every one of the 25 experiments.

Each observed *M. smegmatis* protein was associated with a functional category by comparing the amino acid sequences (using BLASTP) to a database of 350,111 protein sequences from 89 fully sequenced genomes and transferring the broad-level Clusters of Orthologous Groups (COG) annotation (Tatusov et al. 2001) from the top-scoring homologs, where significant, to the *M. smegmatis* proteins. The broad functions of the set of 901 detected proteins are plotted in Figure 1. Major functions represented include energy production and conversion, amino acid transport and metabolism, translation, ribosomal structure and biogenesis, and lipid transport and metabolism. Roughly 33% of the proteins could not be assigned functions in this manner (Fig. 1).

The 901 proteins were ranked by the number of observations of each protein across the 25 experiments. The set of proteins observed in 21–25 experiments, likely corresponding to highly expressed proteins, was significantly enriched for proteins involved in translation (30%), energy production (15%), small molecule transport and metabolism (18%), as well as a large fraction of uncharacterized proteins (15%). Examples of proteins in this set include peptidyl-prolyl *cis–trans* isomerase A (PpiA), glyceraldehyde 3-phosphate dehydrogenase (Gap), ATP synthase β chain (AtpD), elongation factor Tu (Tuf), 60 kDa chaperonin (GroEL1), and 23 ribosomal proteins (e.g., RpsA, RplR). Several of these highly expressed proteins (3.6%) are involved in the metabolism of lipids, important components of the mycobacterial cell wall, accounting for ~60% of the cell wall weight (Ratledge and Dale 1999). Highly expressed lipid metabolism proteins included acyl carrier protein (involved in meromycolate extension), acetyl-/propionyl-coenzyme A carboxylase (AccA3), and propionyl-CoA carboxylase β chain 5 (AccD5), enzymes responsible for creating lipid structures in the cell wall. In addition to proteins involved in lipid metabolism, the enzymes necessary for glycolysis, the tricarboxylic acid cycle, and a large number of ribosomal proteins are also highly expressed, consistent with expectation.

Focusing instead on proteins observed only once among the 25 experiments reveals a very different trend. These proteins, probably representing low-expression proteins, are substantially enriched for uncharacterized proteins, especially proteins whose broad function can be approximately assigned by homology (e.g., "dehydrogenase"), but whose specific function in *M. smegmatis* is unknown. The proteins involved in translation are substantially underrepresented in this set. The complete set of protein identifications and growth-phase expression levels are available as Supplemental Table 1.

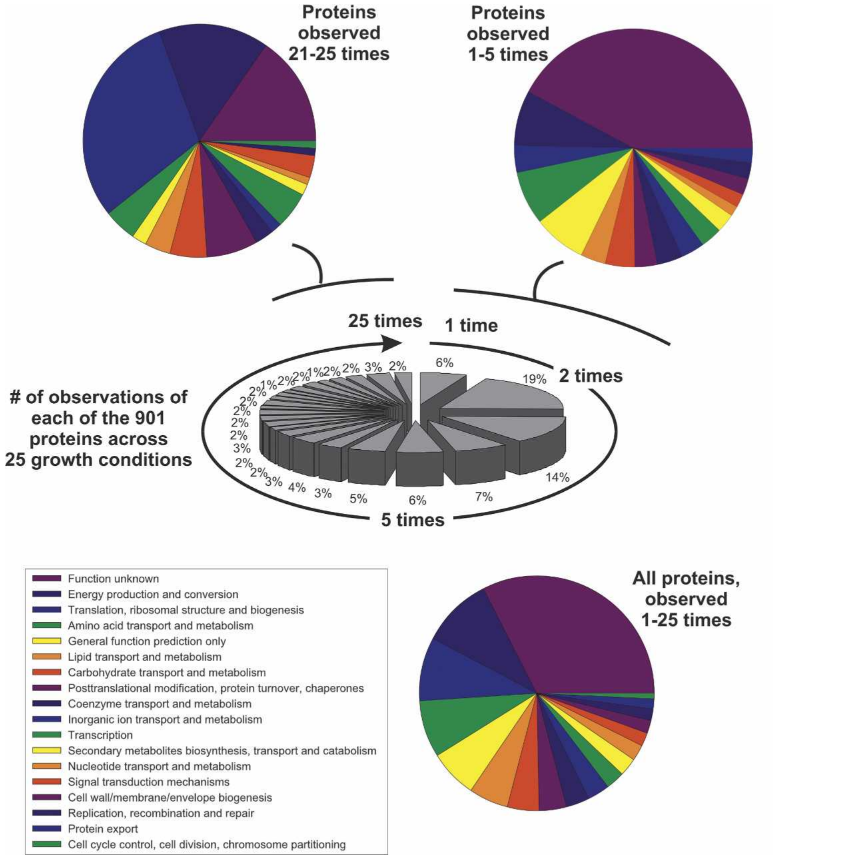

Figure 1. The distribution of observations of each of the 901 proteins (*central* chart) identified across 25 LC/LC/MS/MS experiments and the associated protein functions for the complete set of proteins (*bottom* chart), proteins detected in only one to five of the 25 experiments (*top right* chart), and the high-abundance proteins detected in 20–25 of the 25 experiments (*top left* chart).

Proteomic observation of coexpression of operon-encoded proteins

Given that operon-encoded proteins are cotranscribed and are generally cotranslated, we expect that proteins in the same operon should be coexpressed across the proteomic profiling experiments. Figure 2 shows four such examples of proteins in the same operon coordinately expressed across the 25 experiments. In each case, proteins encoded within the operon are observed, while proteins encoded by flanking genes and those on the opposite strand

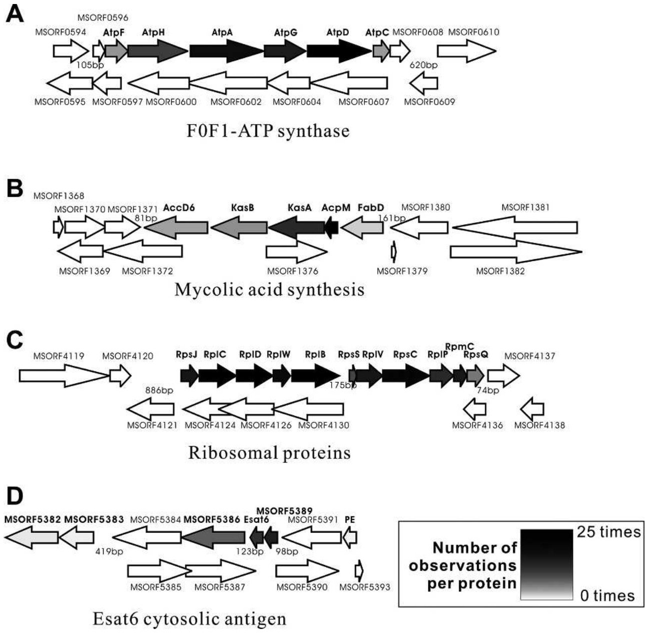

Figure 2. Proteins encoded in the same operon were observed to be coexpressed across the 25 experiments. All proteins observed in LC/LC/MS/MS experiments are labeled in bold, with arrows shaded according to the number of observations. Four distinct operons (*A–D*) are indicated.

are not observed. For example, Figure 2A shows AtpF, AtpH, AtpA, AtpG, AtpD, and AtpC, subunits of the F_0F_1-type ATP synthase. Coexpression of these proteins reveals they are in the same operon while the neighboring genes and those encoded on the opposite strand, such as MSORF0600, 0602, and 0604, are not detected.

We see similar behavior for a uniquely mycobacterial system (Fig. 2B): the genes that synthesize mycolic acid, the main component of the cell wall in mycobacteria and the end product in the metabolic pathway of InhA, which is the primary target of the mycobacterial drug isoniazid (Banerjee et al. 1994). KasB (3-oxoacyl-[acyl-carrier-protein] synthase2) is in the same operon as KasA (3-oxoacyl-[acyl-carrier-protein] synthase1), AcpM (acyl carrier protein), FabD (malonyl CoA-acyl carrier protein transacylase), and AccD6 (acetyl-/propionyl-CoA carboxylase). Both FabD and KasB are known components of the mycolic acid synthesis pathway (Wilson et al. 1999). Figure 2B shows that proteins encoded by the KasB operon are coexpressed, while proteins encoded by flanking genes are not.

Ribosomal proteins are often encoded in large operons to regulate ribosome synthesis (Allen et al. 1999). Figure 2C shows that the RpsJ and RpsS operons, which are components of the 11-gene S10 ribosomal protein operon, produce large and small ribosomal subunit proteins; the proteins are strongly expressed in a coordinate fashion.

Similarly, we observe an operon formed of antigen proteins (Fig. 2D). MSORF5388 (ESAT6) and MSORF5389 (a homolog of the Esat-6 protein family), identified in 17 and 19 experiments, respectively, are secreted antigens (Ratledge and Dale 1999). These two proteins are coexpressed with MSORF5386, an uncharacterized protein, suggesting that these three proteins form an operon, and by this association, may be functionally linked. Membership in the same operon argues that MSORF5386 may function with ESAT6 and MSORF5389.

Correlation between protein expression levels and mRNA expression levels of orthologous genes

We reason that the number of experiments in which a protein is observed (sampled) should roughly correlate with the protein's expression level. This assumption is based on the idea that in an LC/LC/MS/MS analysis of a cell lysate, which may consist of several hundred thousand tryptic peptides, only a subset of MS peaks is sampled for further MS/MS analysis, in total collecting ~30,000–40,000 MS/MS spectra per experiment. As the sampled peptides are typically chosen according to peak height in the MS parent spectra (here, choosing the three tallest peaks per MS spectrum), this introduces an intrinsic element of stochastic sampling and a bias toward high-abundance proteins. Therefore, over repeated analyses, we expect to sample highly abundant proteins more often and lower-abundance proteins less often (e.g., as recently described in Liu et al. 2004). In the analyses presented here, rather than repeated analysis of identical samples, we have analyzed related samples; thus this trend will be conflated with the tendency for a protein to be broadly expressed, rather than just highly expressed. Nonetheless, the general trend should hold.

We wished to test our assumption that observation frequency will correlate with relative protein abundance. As a first-order test of this idea, we compared the *M. smegmatis* protein observation frequency data with mRNA expression levels, under the assumption that protein and mRNA levels should be reasonably consistent. For this test, we required an absolute measurement of mRNA expression levels, such as provided by Affymetrix-style DNA microarrays. As no such data are available for *M. smegmatis* prior to completing the genome sequence, we instead used data for the orthologous genes of *Escherichia coli* grown under roughly equivalent conditions (exponential phase growth) (Covert et al. 2004). The number of observations of *M. smegmatis* proteins correlates moderately well with the absolute abundance of their *E. coli* orthologs' mRNAs ($R^2 = 0.72$) (Fig. 3). *E. coli* orthologs of proteins observed in all 25 experiments showed high mRNA expression levels (2313 ± 1126, arbitrary units from Affymetrix array data), while *E. coli* orthologs of proteins observed in only a single experiment were significantly lower (455 ± 557; p-value < 0.001 by t-test). An ANOVA F-test of the simple linear regression between the mRNA abundance and number of mass spectrometry observations is significant ($p < 0.001$). These sup-

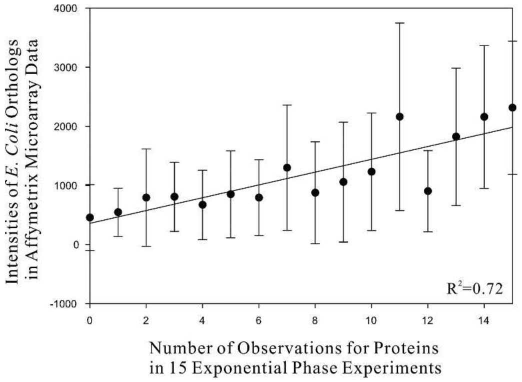

Figure 3. The number of experiments in which each protein was observed in exponential phase correlates ($R^2 = 0.72$) with the corresponding mRNA expression levels, estimated from *E. coli* orthologs' measured expression (Covert et al. 2004), suggesting that the depth of mass spectrometry sampling provides a rough estimate of protein abundance.

port the notion that the depth of sampling of a protein by mass spectrometry does, indeed, roughly correspond to protein abundance.

Correlation between protein expression levels and codon bias

Because the number of observations of each protein correlates at least roughly with protein expression levels, we tested if these approximate expression levels were predictable from protein amino acid sequence properties. First, we compared proteins' expression levels to their codon biases, via the protein codon adaptation indices (CAI) (Sharp and Li 1987), which often correlate with protein expression level (Bennetzen and Hall 1982; Sharp and Li 1987; Futcher et al. 1999; Gygi et al. 2000; Jansen et al. 2003). As seen in Figure 4A, the proteins' average CAI values are positively correlated ($R^2 = 0.74$) with the number of observations of each protein, suggesting the expression levels are, indeed, consistent with codon choice. Proteins identified in all of 25 experiments typically showed high CAI values (CAI = 0.75 ± 0.06)—examples include proteins involved in general metabolic pathways (e.g., PpiA, Gap, AtpD, Tuf, GroEL1, and RpsA) and specific metabolic pathways, especially lipid metabolism, such as acyl carrier protein, AccA3 and AccD5. In contrast, proteins observed only in a single experiment showed considerably lower CAI values (CAI = 0.63 ± 0.06), a significant difference in CAI under a *t*-test (p-value < 0.001).

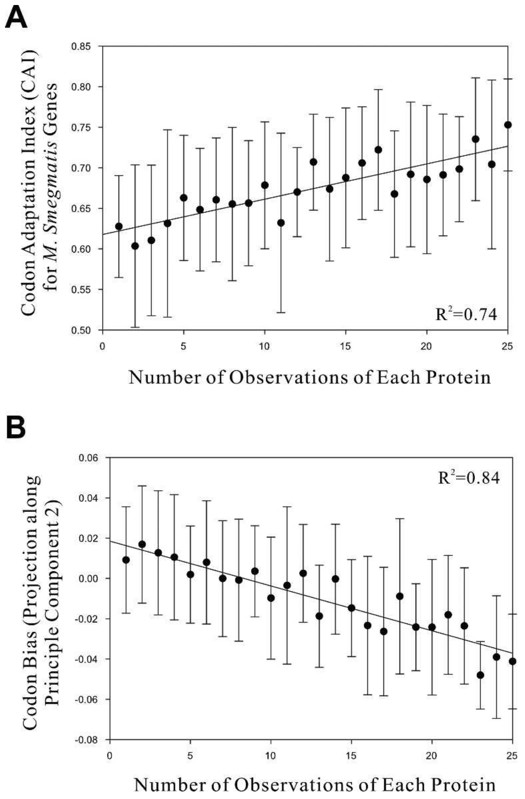

Figure 4. The approximate expression levels of each protein are predicted by two different measures of the codon bias. (*A*) The number of experiments in which each protein was observed (estimating protein abundance) correlates with the proteins' codon adaptation indices (CAI; $R^2 = 0.74$), indicating that the proteins' average expression levels can be partially predicted from the choice of codons used for each protein. (*B*) The number of observations of each protein also correlates well with codon bias calculated by principle component analysis of the proteins' codon frequencies. Here, we plot the number of observations of each protein (*x*-axis) versus the projection of the gene's codon frequencies onto the second principle component (*y*-axis) (PC2; $R^2 = 0.84$), which best captures variation in codon choice between proteins with high and low expression levels. In this case, PC2 is negatively correlated with protein abundance.

To further investigate the predictability of protein expression levels, we performed principle component analysis (PCA) of *M. smegmatis* proteins' codon frequencies. PCA is a technique for summarizing the major variation in data. In PCA, the dimensionality of a data set is reduced by projecting the original data along new coordinate axes (the principle components) that capture the major trends in the data. Principal components (PCs) are linear transformations of the original sets of variables, uncorrelated and ordered, with the first few PCs containing most of the variation in the original data set (Yeung and Ruzzo 2001; Jolliffe 2002). Performing PCA on the codon frequency vectors associated with each protein in our experiments therefore reveals the major trends in codon choice among observed *M. smegmatis* genes. While codon adaptation indices capture one aspect of codon bias, PCA should return the major trends in codon bias, regardless of whether these correlate with CAI. We tested the major PCs for correlation with the number of observations of each protein. The dominant PC (PC1) captured the background frequency of codons used by *M. smegmatis* and did not correlate with protein expression levels (data not shown). However, the second largest PC (PC2) correlated well with protein expression levels (Fig. 4B) ($R^2 = 0.84$), supporting the notion that protein expression levels are largely predictable by codon choice. PC3 and PC4 are no longer correlated with expression level. The association of both PC2 and CAI-derived codon bias with protein abundance are significant under an ANOVA *F*-test ($p < 0.001$). Thus, codon bias, as captured by CAI or PCA, can be used to predict approximate expression levels of *M. smegmatis* proteins.

Growth-phase-specific expression

The physical and metabolic changes in mycobacteria during exponential and stationary phase are largely unknown. As comparative proteomics has the potential to detect changes in the proteome, for example, changes in abundance as well as post-translational modifications such as phosphorylation and glycosylation (Link et al. 1999), we compared the proteins expressed in exponential and stationary growth phases (Fig. 5A). Exponential-phase cultures showed a higher fraction of proteins associated with active growth, such as DNA replication, recombination and repair, transcription, and translation. In contrast, stationary-phase cells show higher fractions of proteins involved in competition for limited nutrients, including energy production and conversion, inorganic ion and carbohydrate transport, and metabolism and post-translational modification.

We were able to investigate more subtle differences in protein abundance between exponential and stationary phase by comparing the number of experiments in which each protein was observed in the two growth phases (Fig. 5B,C). In this analysis, we expected to find sets of proteins over represented in stationary phase, presumably in order to make cells more competitive with stresses in nutrient-starved culture, and we did, indeed, find many such proteins. Among the proteins enriched in stationary-phase cells was the transcriptional activator MtrA, which is also induced upon bacterial entry into macrophages in *M. bovis* BCG (Zahrt and Deretic 2000). Other stationary-phase-enriched

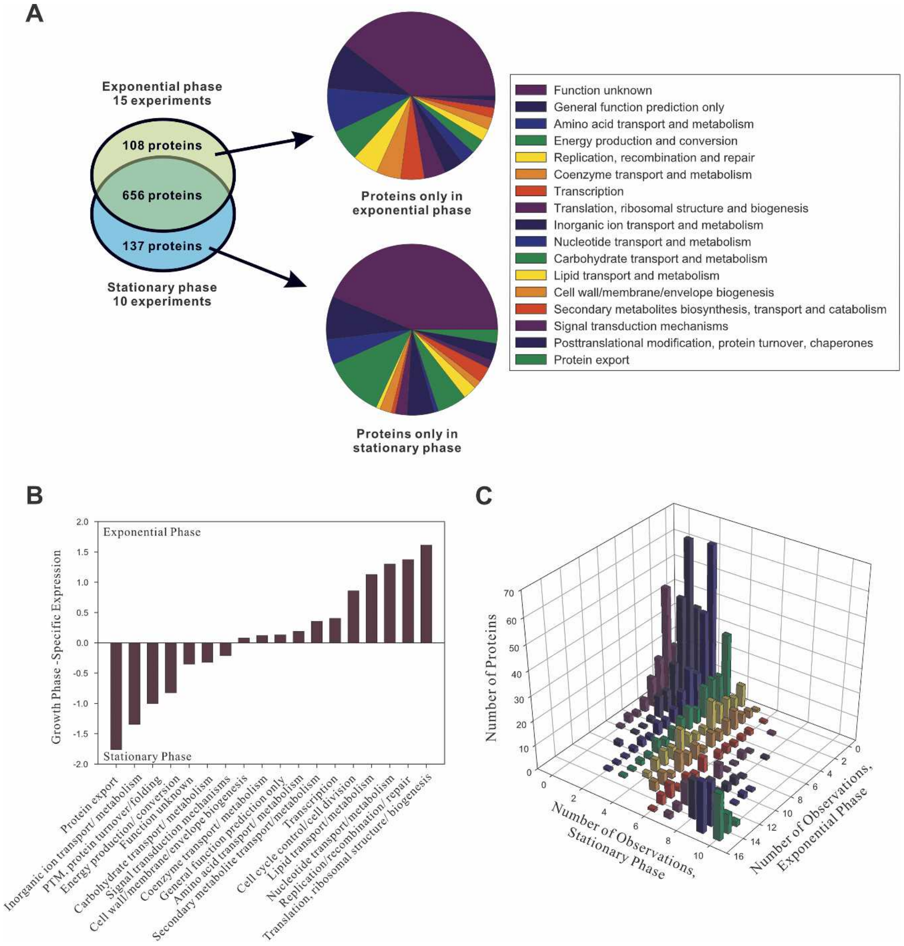

Figure 5. A comparison of the proteomes of exponential and stationary-phase cells. (*A*) The overlap of identified proteins in exponential and stationary phases (*left* chart), showing the functions of proteins specific to exponential phase (*top right* chart) or stationary phase (*bottom right* chart). For clarity, proteins with uncharacterized functions are excluded from the pie charts. (*B*) COG functions for the proteins differentially expressed in exponential and stationary phases. For the proteins in each COG category, we plot the mean difference in expression level, calculated as mean($N_{\mathrm{exp},i} - \frac{3}{2} N_{\mathrm{stat},i}$), where $N_{\mathrm{exp},i}$ and $N_{\mathrm{stat},i}$ are the number of observations of protein i in exponential and stationary phase, respectively. The factor of $\frac{3}{2}$ is introduced to scale the number of stationary-phase experiments to match the number of exponential phase experiments. Proteins of cell growth (translation, replication, etc.) are highly induced in exponential phase, while proteins of transport and energy production are highly induced in stationary phase. (PTM) Post-translational modification. (*C*) A 2D histogram plots the distribution of protein abundances and differential expression between stationary- and exponential-phase cells. The bulk of proteins lie on the diagonal (no differential expression); off-diagonal proteins are differentially expressed to varying degrees.

proteins were involved in carbohydrate or fatty acid metabolism (GlgC, PrpE, LpqG, LpqY), energy metabolism (catalase, QcrC, PdhAB, and Ndh), and sugar transport (ribose ABC transporter, GntP, β-glucanase), and amino acid transport (ProV, 3-dehydroquinase). Regulatory proteins (PhoY2) were also identified, as well as proteins of unknown function.

Among the proteins we identified as up-regulated in exponentially growing cells, several share the same COG functions with proteins enriched in stationary phase but have different specific functions. For example, the σ factor SigA is only detected in exponential phase—consistent with the result that SigA mediates enhanced growth of *M. tuberculosis* (Wu et al. 2004)—while other σ factors are enriched in stationary phase, such as SigH. As σ subunits of *E. coli* are differentially expressed according to the functions of the proteins they regulate (Ishihama 2000), it seems likely that SigA may be regulating housekeeping genes, while SigH regulates stationary-phase-specific genes, such as stress response or transport genes (Kormanec et al. 2000; Thackray and Moir 2003).

In many cases the differentially expressed proteins were components of operons that were themselves differently expressed. For example, the RpsJ, RpsS, and KasB operons are up-regulated in exponential phase, while the PdhABC, Ndh, and FadE operons are up-regulated in stationary phase (Fig. 6).

Discussion

The ability to rapidly annotate genome sequences is becoming increasingly important given the current pace of genome sequencing; equally important are methods for characterizing protein dynamics on a genome-wide scale. In this study we demonstrate the suitability of coupled liquid chromatography–mass spectrometry to address both problems. Earlier analysis of mycobacterial proteomes by 2DE-MS identified roughly 300 proteins (Jungblut et al. 2001; Mattow et al. 2003). In contrast, LC/LC/MS/MS methods have proved capable of identifying ~1500 proteins from yeast lysates (Washburn et al. 2001; Peng et al. 2003). As applied here, tandem mass spectrometry (MS/MS) is biased in single experiments toward detection of abundant proteins (e.g., ribosomal proteins and heat-shock proteins), but stochastically samples lower-abundance proteins. Repeated analysis of related samples therefore leads to detection of lower-expression-level proteins, such as cell division proteins and transcription factors, and allows us to identify and estimate relative expression levels of 901 proteins from *M. smegmatis*. The data provide experimental annotation of predicted genes, provide measurement of relative protein abundance changes across conditions, and demonstrate the correlation of protein abundance with mRNA expression levels and codon choice.

One goal of this mass spectrometry-based analysis was to provide experimental annotation of abundantly expressed proteins. In particular, a painstaking manual evaluation of each predicted gene was required for other mycobacterial genomes (e.g., *M. tuberculosis* H37Rv) (Cole et al. 1998) in order to eliminate false-positive gene predictions. We expect, given the density of predicted genes and previous observations of the performance of Glimmer 2.0 on default settings (Delcher et al. 1999), that a significant fraction of the predicted genes are false positives. Assuming similar gene density to *M. tuberculosis* or *M. bovis* BCG, we expect ~6300 genes in *M. smegmatis* (David Graham, pers.

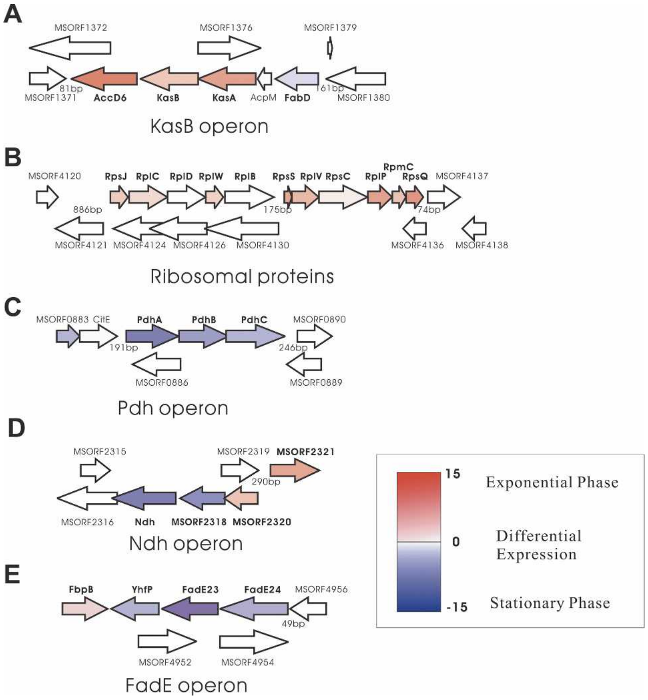

Figure 6. Proteins in the same operon coexpress in a growth-phase-specific manner. The color scale represents the relative expression of the protein between exponential phase (red) and stationary phase (blue), calculated as (number of observations in exponential phase) $-\frac{3}{2}$(number of observations in stationary phase). The KasB operon (*A*) and RpsJ, RpsS operons (*B*) are up-regulated in exponential phase; the PdhABC (*C*), Ndh operon (*D*), and FadE operon (*E*) are up-regulated in stationary phase.

comm.), suggesting that ~30% of the 8968 predicted genes are false positives. For example, genes predicted on the opposing strand to expressed proteins within the operons of Figure 2, such as MSORF1376 in the KasB operon, may be candidate false-positive-predicted genes.

Note that unlike other applications, for the purpose of constructing a mass spectrometry reference database, we desire to minimize the false-negative gene identification rate, even at the expense of increasing the false-positive gene identification rate, to ensure that all representative sequences are present. These gene prediction errors are clearly undesirable for other purposes, but for mass spectrometry serve to minimize the false-negative protein identification rate in the mass spectrometry experiment. The extreme case of this strategy is to compare the MS/MS spectral data against raw, uninterpreted genomic sequence data (Arthur and Wilkins 2003), even without predicting genes (Jaffe et al. 2004). In this mode, the false-positive gene prediction rate is maximal; however, this minimizes the false-negative protein identification rate, allowing previously unrecognized proteins to be identified. Proteomics has previously helped in this manner: for example, of 263 proteins detected in *M. tuberculosis* H37Rv by 2DE-MS, six were not previously predicted (Jungblut et al. 2001). To aid in such future annotation studies, our raw mass spectrometry data have been deposited into the public domain in the Open Proteomics Database (Prince et al. 2004). However, via this mass spectrometric approach, we are able to rapidly validate a significant subset of the predicted genes. The fraction of the proteome we identified also provides a framework for understanding the activity of pathogenic *M. tuberculosis*: In total, 709 (78.7%) of the 901 proteins we detected have homologs in *M. tuberculosis* H37Rv, compared to only 47.5% for the entire set of 8968 predicted *M. smegmatis* proteins, or to 68% if we assume a more likely total of ~6300 genes. Thus, the expressed *M. smegmatis* proteome is enriched for proteins of relevance to *M. tuberculosis*.

Beyond experimentally annotating expressed proteins, the frequency of observation of each protein gives an estimate of protein abundance. We took advantage of these data to estimate the correlation between codon bias and protein abundance. Such a correlation has previously been observed for high-abundance proteins, for example, in two-dimensional gel electrophoresis analysis of the yeast proteome (Futcher et al. 1999; Gygi et al. 2000). In our hands, the *M. smegmatis* protein expression levels were predictable from the genes' codon biases as calculated by codon adaptation indices, as well as by principle component analysis. Our data support two models of prokaryotic protein regulation. In the first, the condition under which a protein would be expressed is set by the action of transcription factors and regulatory proteins, but the quantity of the expressed protein is "preset" by its codon bias. Under this view, regulation of a prokaryotic protein is simplified to deciding when to express it, with the actual amount of the protein synthesized being optimized over evolutionary time. A perhaps more plausible, equally supported explanation is that codon choice has been optimized to be nonlimiting. That is, protein abundance is set in some fashion (e.g., translation initiation, promoter strength) and may be dynamically regulated, but rarely exceeds a characteristic expression level. Codons might then face selective evolutionary pressure to optimize such that the codon choice is nonlimiting for the normal range of each protein's expression. This scenario would still produce a correlation between codon bias and average protein abundance. We note that these expression measurements could in principle be made more accurate through the use of isotope labeling approaches, such as the use of metabolic labeling of stable-isotope tags or ICAT (Gygi et al. 1999) for quantitative protein expression profiling (Flory et al. 2002), as used to quantitate 280 *M. tuberculosis* proteins (Schmidt et al. 2004).

Numerous physiological and metabolic changes happen during the transition from exponential phase to stationary phase, and we capture some of these changes in the differential expression analysis. Proteins involved in the pathways of DNA replication, recombination and repair, transcription, and translation show a higher fraction of proteins associated with active growth. In accord with our results, UmaA1 (lipid biosynthesis) and ParA (functions in chromosome partitioning), which we observe enriched in exponential phase, have been reported to be down-regulated in *M. tuberculosis* in nutrient-starved media (Betts et al. 2002).

When cells enter stationary phase, many growth-related genes are down-regulated; instead, the stationary-phase-specific genes are expressed, such as proteins involved in energy production and conversion, post-translation modification, and carbohydrate transport and metabolism (Fig. 5A). The PdhABC proteins are up-regulated in stationary phase, consistent with their up-regulation in *M. tuberculosis* starved for nutrients (Betts et al. 2002). We speculate Mpt53, an immunogenic protein induced only in stationary phase, the transcription factor MtrA, and several proteases (ClpP, MSORF2945, and MSORF8009, a zinc metalloprotease) up-regulated in stationary phase, may be important for stress response and therefore might also be important to the virulence of pathogenic mycobacteria. Finally, we noticed the enrichment of numerous uncharacterized proteins in stationary phase. Although their exact function remains unknown,

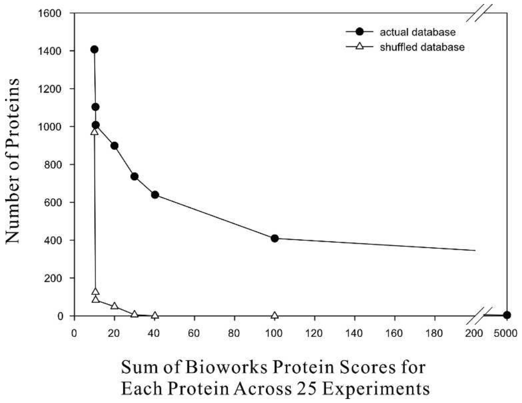

Figure 7. Estimating the protein identification error rate. The false-positive protein identification rate (triangles; estimated by comparing MS/MS spectra against a mass spectrometry reference database of shuffled *M. smegmatis* proteins) is plotted as a function of total BioWorks score for each protein. At a false-positive identification rate of ~5% (corresponding to proteins with total scores >20 across the 25 experiments), 899 proteins are identified (circles).

these proteins are likely to be stress-response proteins, heat-shock proteins, or other proteins involved in coping with the nutrient limitation and high cell density in this phase of growth, suggesting that future experiments directly monitoring the stress response of mycobacteria may be useful in more precisely determining their functions.

Methods

Initial identification of *M. smegmatis* protein-coding sequences

The partially complete genomic sequence for *M. smegmatis* mc^{2}155 was obtained from The Institute for Genomic Research (TIGR) through the Web site at http://www.tigr.org in the form of 15 sequence contigs, and protein-encoding genes were predicted by searching with the gene-finding program Glimmer 2.0 using default settings, which minimize the false-negative identification rate of coding sequences but increase the false-positive rate (Delcher et al. 1999). A total of 8968 genes were predicted and assembled into a searchable database for interpreting mass spectrometry peptide fragmentation spectra data using the program BioWorks 3.1 (ThermoFinnigan). The predicted genes were functionally annotated by comparing the predicted amino acid sequences against those of 350,111 protein sequences from 89 fully sequenced genomes (downloaded from the Entrez genome database) using the program BLASTP and default parameters, and selecting the protein functional annotation of the top BLAST match when it surpassed a BLAST expectation value threshold of 1e-6. A broad functional category was also assigned to each protein by selecting the COG database annotation (Tatusov et al. 2001) associated with the top BLAST hit, when significant. For clarity in all discussion, COG category N ("cell mobility") was labeled "protein export", as *M. smegmatis* is nonmobile and the *M. smegmatis* proteins in this category are predominantly involved in protein export and stress adaptation.

Growth of *M. smegmatis* mc^{2}155 and preparation of cell lysates

M. smegmatis mc^{2}155 bacteria were grown with agitation at 37°C in minimal medium (Brosch et al. 2001; Chacon et al. 2002) and in rich media (Middlebrook7H9 medium and Luria broth [LB] medium) (Jacobs Jr. et al. 1991), collecting samples for 16, six, and three time points from Middlebrook7H9 medium, LB, and minimal medium, respectively. Samples were centrifuged at 12,000*g* for 30 min, suspended in ice-cold lysis buffer (25 mM Tris-HCl at pH 7.5, 2.5 mM DTT, 1.0 mM EDTA, 0.02% [w/v] Brij35, 1 × Calbiochem Protease Inhibitor Cocktail Set I [CPICSI]) (1 mL/g cell pellet), and disrupted by bead-beating with 1-mm glass beads (Primm et al. 2000; Parish and Stoker 2001). Cell lysates were clarified by centrifugation at 20,000*g* for 30 min, with typical protein concentrations of 10 mg/mL.

Preparation and LC/LC/MS/MS analysis of *M. smegmatis* peptides

M. smegmatis soluble protein extracts were diluted in digestion buffer (50 mM Tris-HCl at pH 8.0, 1.0 M Urea, 2.0 mM $CaCl_2$), denatured at 95°C for 15 min, and digested with sequencing grade trypsin (Sigma) at 37°C for 20 h. Tryptic peptide mixtures were separated by automated two-dimensional high-performance liquid chromatography. Chromatography was performed at 2 µL/min with all buffers acidified with 0.1% formic acid. Chromatography salt step fractions were eluted from a strong cation exchange column (SCX) with a continuous 5% acetonitrile (ACN) background and 10-min salt bumps of 0, 20, 40, 60, 80, 100, 150, 200, 300, 500, and 900 mM ammonium chloride. Each salt bump was eluted directly onto a reverse-phase C18 column and washed free of salt. Reverse-phase chromatography was run in a 60-min gradient from 5% to 45% ACN, then purged at 95% ACN. Peptides were analyzed online with electrospray ionization (ESI) ion trap mass spectrometry (MS) (Link et al. 1999; Washburn et al. 2001) using a ThermoFinnigan Surveyor/DecaXP+ instrument. In each MS spectrum, the three tallest individual peaks, corresponding to peptides, were fragmented by collision-induced dissociation with helium gas to produce MS/MS spectra.

Estimating the mRNA expression levels

E. coli orthologs of *M. smegmatis* proteins were identified by using BLASTP under default settings and the bidirectional best hit method (Overbeek et al. 1999), only accepting hits with BLAST *E*-values less than 1e-10. By this approach, 1250 *M. smegmatis* predicted genes have *E. coli* K-12 orthologs. The mRNA abundance of each of these genes was calculated as the average abundance across three replicate Affymetrix DNA microarrays from Covert et al. (2004), representing the mRNA abundance of *E. coli* K-12 MG1655 cells growing in M9 media in exponential phase.

Calculation of codon bias

The codon adaptation index (CAI) of each *M. smegmatis* open reading frame was calculated as in Sharp and Li (1987). A second estimate of codon bias was generated by performing principle component analysis (Yeung and Ruzzo 2001; Jolliffe 2002) of the codon frequencies associated with each of the 901 observed *M. smegmatis* proteins. Using a Perl program, a vector was created for each gene composed of the usage frequencies of each of the 61 possible amino-acid-encoding codons. Principle component analysis was performed on the resulting frequency vectors using the program Cluster (Eisen et al. 1998). Projection of each gene's frequency vector onto each of the principle components associated a set of numerical indices with each gene describing the major trends in the gene's codon frequencies.

An error model for mass spectrometry proteomics data

Proteins were identified from the resulting peptide MS/MS fragmentation spectra by searching against the custom *M. smegmatis* predicted protein database using the program BioWorks 3.1

(Xcorr ≥ 2.5) and filtering the results with selection criteria that minimize the false-positive identification rate to <5%. Calculation of the false-positive identification rate was based on the following procedure: The amino acid sequences of each predicted protein encoded in the *M. smegmatis* genome were shuffled, thereby preserving the length and amino acid frequency distribution of each protein but not the amino acid order, and fragmentation spectra from the 25 LC/LC/MS/MS experiments were analyzed against the shuffled database by using the program BioWorks. For each protein identified, we calculated the sum of the BioWorks protein scores across the 25 experiments using either the correct or shuffled version of the database (Fig. 7). At a protein score threshold corresponding to a 5% false-positive identification rate in the shuffled database, a total of 899 proteins were found. A similar analysis using DTASelect (Tabb et al. 2002) identified 426 proteins versus 2 in the shuffled database, at a false-positive rate of 0.5%. By this analysis, DTASelect therefore gives a lower false-positive identification rate, with a higher false-negative identification rate. Using either criterion gave a total of 901 *M. smegmatis* proteins identified at a false-positive rate of ~5%.

Acknowledgments

The authors acknowledge funding from the Welch Foundation (F-1515), Packard Foundation, NSF, and NIH (GM067779-01). We thank Shailesh Date for his help in *M. smegmatis* sequence comparisons, David Graham for performing the Glimmer analysis and his critical comments, and Ian Molineux for helpful discussion. We thank The Institute for Genomic Research (TIGR) for access to the preliminary *M. smegmatis* genome sequence data, obtained from the Web site at http://www.tigr.org and used by permission here. Sequencing of *M. smegmatis* was accomplished with support from the National Institute of Allergy and Infections Diseases (NIAID).

References

Aebersold, R. and Mann, M. 2003. Mass spectrometry-based proteomics. *Nature* **422:** 198–207.

Allen, T., Shen, P., Samsel, L., Liu, R., Lindahl, L., and Zengel, J.M. 1999. Phylogenetic analysis of L4-mediated autogenous control of the S10 ribosomal protein operon. *J. Bacteriol.* **181:** 6124–6132.

Arthur, J.W. and Wilkins, M.R. 2003. Using proteomics to mine genome sequences. *J. Proteome Res.* **3:** 393–402.

Banerjee, A., Dubnau, E., Quemard, A., Balasubramanian, V., Um, K.S., Wilson, T., Collins, D., de Lisle, G., and Jacobs Jr., W.R. 1994. InhA, a gene encoding a target for isoniazid and ethionamide in *Mycobacterium tuberculosis*. *Science* **263:** 227–230.

Bennetzen, J.L. and Hall, B.D. 1982. Codon selection in yeast. *J. Biol. Chem.* **257:** 3026–3031.

Betts, J.C., Lukey, P.T., Robb, L.C., McAdam, R.A., and Duncan, K. 2002. Evaluation of a nutrient starvation model of *Mycobacterium tuberculosis* persistence by gene and protein expression profiling. *Mol. Microbiol.* **43:** 717–731.

Brosch, R., Pym, A.S., Gordon, S.V., and Cole, S.T. 2001. The evolution of mycobacterial pathogenicity: Clues from comparative genomics. *Trends Microbiol.* **9:** 452–458.

Chacon, O., Feng, Z., Harris, N.B., Caceres, N.E., Adams, L.G., and Barletta, R.G. 2002. *Mycobacterium smegmatis* D-alanine racemase mutants are not dependent on D-alanine for growth. *Antimicrob. Agents Chemother.* **46:** 47–54.

Cole, S.T., Brosch, R., Parkhill, J., Garnier, T., Churcher, C., Harris, D., Gordon, S.V., Eiglmeier, K., Gas, S., Barry III, C.E., et al. 1998. Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature* **393:** 537–544.

Covert, M.W., Knight, E.M., Reed, J.L., Herrgard, M.J., and Palsson, B.O. 2004. Integrating high-throughput and computational data elucidates bacterial networks. *Nature* **429:** 92–96.

Delcher, A.L., Harmon, D., Kasif, S., White, O., and Salzberg, S.L. 1999. Improved microbial gene identification with GLIMMER. *Nucleic Acids Res.* **27:** 4636–4641.

Eisen, M.B., Spellman, P.T., Brown, P.O., and Botstein, D. 1998. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci.* **95:** 14863–14868.

Eng, J.K., McCormack, A.L., and Yates III, J.R. 1994. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom.* **5:** 976–989.

Flory, M.R., Griffin, T.J., Martin, D., and Aebersold, R. 2002. Advances in quantitative proteomics using stable isotope tags. *Trends Biotechnol.* **20:** S23–S29.

Futcher, B., Latter, G.I., Monardo, P., McLaughlin, C.S., and Garrels, J.I. 1999. A sampling of the yeast proteome. *Mol. Cell. Biol.* **19:** 7357–7368.

Gygi, S.P., Rist, B., Gerber, S.A., Turecek, F., Gelb, M.H., and Aebersold, R. 1999. Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nat. Biotechnol.* **17:** 994–999.

Gygi, S.P., Corthals, G.L., Zhang, Y., Rochon, Y., and Aebersold, R. 2000. Evaluation of two-dimensional gel electrophoresis-based proteome analysis technology. *Proc. Natl. Acad. Sci.* **97:** 9390–9395.

Ishihama, A. 2000. Functional modulation of *Escherichia coli* RNA polymerase. *Annu. Rev. Microbiol.* **54:** 499–518.

Jacobs Jr., W.R., Kalpana, G.V., Cirillo, J.D., Pascopella, L., Snapper, S.B., Udani, R.A., Jones, W., Barletta, R.G., and Bloom, B.R. 1991. Genetic systems for mycobacteria. *Methods Enzymol.* **204:** 537–555.

Jaffe, J.D., Berg, H.C., and Church, G.M. 2004. Proteogenomic mapping as a complementary method to perform genome annotation. *Proteomics* **4:** 59–77.

Jansen, R., Bussemaker, H.J., and Gerstein, M. 2003. Revisiting the codon adaptation index from a whole-genome perspective: Analyzing the relationship between gene expression and codon occurrence in yeast using a variety of models. *Nucleic Acids Res.* **31:** 2242–2251.

Jolliffe, I.T. 2002. *Principal component analysis*. Springer, New York.

Jungblut, P.R., Schaible, U.E., Mollenkopf, H.J., Zimny-Arndt, U., Raupach, B., Mattow, J., Halada, P., Lamer, S., Hagens, K., and Kaufmann, S.H. 1999. Comparative proteome analysis of *Mycobacterium tuberculosis* and *Mycobacterium bovis* BCG strains: Towards functional genomics of microbial pathogens. *Mol. Microbiol.* **33:** 1103–1117.

Jungblut, P.R., Muller, E.C., Mattow, J., and Kaufmann, S.H. 2001. Proteomics reveals open reading frames in *Mycobacterium tuberculosis* H37Rv not predicted by genomics. *Infect. Immun.* **69:** 5905–5907.

Kormanec, J., Sevcikova, B., Halgasova, N., Knirschova, R., and Rezuchova, B. 2000. Identification and transcriptional characterization of the gene encoding the stress-response σ factor σH in *Streptomyces coelicolor* A3(2). *FEMS Microbiol. Lett.* **189:** 31–38.

Link, A.J., Eng, J., Schieltz, D.M., Carmack, E., Mize, G.J., Morris, D.R., Garvik, B.M., and Yates III, J.R. 1999. Direct analysis of protein complexes using mass spectrometry. *Nat. Biotechnol.* **17:** 676–682.

Liu, H., Sadygov, R.G., and Yates III, J.R. 2004. A model for random sampling and estimation of relative protein abundance in shotgun proteomics. *Anal. Chem.* **76:** 4193–4201.

Mattow, J., Schaible, U.E., Schmidt, F., Hagens, K., Siejak, F., Brestrich, G., Haeselbarth, G., Muller, E.C., Jungblut, P.R., and Kaufmann, S.H. 2003. Comparative proteome analysis of culture supernatant proteins from virulent *Mycobacterium tuberculosis* H37Rv and attenuated *M. bovis* BCG Copenhagen. *Electrophoresis* **24:** 3405–3420.

Overbeek, R., Fonstein, M., D'Souza, M., Pusch, G.D., and Maltsev, N. 1999. The use of gene clusters to infer functional coupling. *Proc. Natl. Acad. Sci.* **96:** 2896–2901.

Parish, T. and Stoker, N.G., eds. 2001. *Mycobacterium tuberculosis* protocols (Methods in molecular medicine). Humana Press, Totowa, NJ.

Peng, J., Elias, J.E., Thoreen, C.C., Licklider, L.J., and Gygi, S.P. 2003. Evaluation of multidimensional chromatography coupled with tandem mass spectrometry (LC/LC-MS/MS) for large-scale protein analysis: The yeast proteome. *J. Proteome Res.* **2:** 43–50.

Perkins, D.N., Pappin, D.J., Creasy, D.M., and Cottrell, J.S. 1999. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **20:** 3551–3567.

Primm, T.P., Andersen, S.J., Mizrahi, V., Avarbock, D., Rubin, H., and Barry III, C.E. 2000. The stringent response of *Mycobacterium tuberculosis* is required for long-term survival. *J. Bacteriol.* **182:** 4889–4898.

Prince, J.T., Carlson, M.W., Wang, R., Lu, P., and Marcotte, E.M. 2004. The need for a public proteomics repository. *Nat. Biotechnol.* **22:** 471–472.

Ratledge, C. and Dale, J., eds. 1999. *Mycobacteria: Molecular biology and virulence.* Blackwell Science, London.

Schmidt, F., Donahoe, S., Hagens, K., Mattow, J., Schaible, U.E., Kaufmann, S.H., Aebersold, R., and Jungblut, P.R. 2004. Complementary analysis of the *Mycobacterium tuberculosis* proteome by two-dimensional electrophoresis and isotope-coded affinity tag technology. *Mol. Cell Proteomics* **3:** 24–42.

Sharp, P.M. and Li, W.H. 1987. The codon Adaptation Index—A measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res.* **15:** 1281–1295.

Tabb, D.L., McDonald, W.H., and Yates III, J.R. 2002. DTASelect and Contrast: Tools for assembling and comparing protein identifications from shotgun proteomics. *J. Proteome Res.* **1:** 21–26.

Tatusov, R.L., Natale, D.A., Garkavtsev, I.V., Tatusova, T.A., Shankavaram, U.T., Rao, B.S., Kiryutin, B., Galperin, M.Y., Fedorova, N.D., and Koonin, E.V. 2001. The COG database: New developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res.* **29:** 22–28.

Thackray, P.D. and Moir, A. 2003. SigM, an extracytoplasmic function σ factor of *Bacillus subtilis*, is activated in response to cell wall antibiotics, ethanol, heat, acid, and superoxide stress. *J. Bacteriol.* **185:** 3491–3498.

Washburn, M.P., Wolters, D., and Yates III, J.R. 2001. Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nat. Biotechnol.* **19:** 242–247.

Wilson, M., DeRisi, J., Kristensen, H.H., Imboden, P., Rane, S., Brown, P.O., and Schoolnik, G.K. 1999. Exploring drug-induced alterations in gene expression in *Mycobacterium tuberculosis* by microarray hybridization. *Proc. Natl. Acad. Sci.* **96:** 12833–12838.

Wu, S., Howard, S.T., Lakey, D.L., Kipnis, A., Samten, B., Safi, H., Gruppo, V., Wizel, B., Shams, H., Basaraba, R.J., et al. 2004. The principal σ factor sigA mediates enhanced growth of *Mycobacterium tuberculosis* in vivo. *Mol. Microbiol.* **51:** 1551–1562.

Yeung, K.Y. and Ruzzo, W.L. 2001. Principal component analysis for clustering gene expression data. *Bioinformatics* **17:** 763–774.

Zahrt, T.C. and Deretic, V. 2000. An essential two-component signal transduction system in *Mycobacterium tuberculosis*. *J. Bacteriol.* **182:** 3832–3838.

Web site references

http://bioinformatics.icmb.utexas.edu/OPD; Open Proteomics Database.

http://www.tigr.org; The Institute for Genomic Research (TIGR).

Received August 24, 2004; accepted in revised form May 4, 2005.