research articles Journal of proteome research

Halobacterium salinarum NRC-1 PeptideAtlas: Toward Strategies for Targeted Proteomics and Improved Proteome Coverage

Phu T. Van,[†,‡] Amy K. Schmid,[†] Nichole L. King,[†] Amardeep Kaur,[†] Min Pan,[†] Kenia Whitehead,[†] Tie Koide,[†] Marc T. Facciotti,[†,§] Young Ah Goo,[†,∥] Eric W. Deutsch,[†] David J. Reiss,[†] Parag Mallick,[⊥] and Nitin S. Baliga*[,†,#]

Institute for Systems Biology, 1441 North 34th Street, Seattle, Washington 98103, Departments of Biology, Microbiology, and Medicinal Chemistry, University of Washington, Seattle, Washington 98195, and Spielberg Family Center for Applied Proteomics, Cedars-Sinai Medical Center, 8750 West Beverly Boulevard, Los Angeles, California 90048

Received January 15, 2008

The relatively small numbers of proteins and fewer possible post-translational modifications in microbes provide a unique opportunity to comprehensively characterize their dynamic proteomes. We have constructed a PeptideAtlas (PA) covering 62.7% of the predicted proteome of the extremely halophilic archaeon *Halobacterium salinarum NRC-1* by compiling approximately 636 000 tandem mass spectra from 497 mass spectrometry runs in 88 experiments. Analysis of the PA with respect to biophysical properties of constituent peptides, functional properties of parent proteins of detected peptides, and performance of different mass spectrometry approaches has highlighted plausible strategies for improving proteome coverage and selecting signature peptides for targeted proteomics. Notably, discovery of a significant correlation between absolute abundances of mRNAs and proteins has helped identify low abundance of proteins as the major limitation in peptide detection. Furthermore, we have discovered that iTRAQ labeling for quantitative proteomic analysis introduces a significant bias in peptide detection by mass spectrometry. Therefore, despite identifying at least one proteotypic peptide for almost all proteins in the PA, a context-dependent selection of proteotypic peptides appears to be the most effective approach for targeted proteomics.

Keywords: PeptideAtlas • Halobacterium • iTRAQ • bioinformatics • archaea • proteomics

Introduction

A complete genome sequence presents a one-dimensional perspective of the physiological potential of an organism. It is the temporally and spatially coordinated expression of genes into functional protein networks that yields emergent behavior that is unique to each species. Therefore, to fully understand how cells function at a systems level, it is imperative to measure, assimilate and simultaneously analyze changes that occur at all levels of genetic information processing.[1] The transcriptome is dynamic and relatively easy to monitor comprehensively using whole genome microarrays and high throughput next generation sequencing, providing insight into which genes respond and assist in adaptation of the organism to a particular environment.[2] However, much more information on regulatory processes remains locked within the proteome. There exist important differences between the transcriptome and the proteome that stem from a variety of post-transcriptional processes, such as regulated degradation and post-translational modifications, thus elevating the importance of comprehensive analysis of dynamic changes in the proteome in response to various environmental perturbations.[3,4]

However, comprehensive detection of the proteome is fraught with technical challenges, especially with regard to proteins that are present in low abundance, integral to the membrane, or uniquely expressed in an environment-specific manner. Even within the same protein, some peptides are more tractable than others using mass spectrometry-based approaches. Although there are several existing hypotheses regarding the underlying reasons that make some peptides more tractable than others (*i.e.* biophysical properties such as isoelectric point (p*I*), hydrophobicity, and length),[5] certain properties such as protease accessibility, protein structure, and protein modifications complicate any attempt to make accurate predictions of peptide tractability by mass spectrometry (MS) using purely theoretical approaches.

The PeptideAtlas (PA) project was initiated to map the proteome of a given organism, cell type or tissue as experimentally detected by a mass spectrometer.[6] The PA is technology agnostic and can make use of data from a variety of MS

* To whom correspondence should be addressed. Nitin S. Baliga, Telephone, 206-732-1266; Fax, 206-732-1299; E-mail, nbaliga@systemsbiology.org.
[†] Institute for Systems Biology.
[‡] Department of Biology, University of Washington.
[§] Current affiliation: Department of Biomedical Engineering and Genome Center, University of California, Davis, One Shields Drive, Davis, California 95616.
[∥] Department of Medicinal Chemistry, University of Washington.
[⊥] Cedars-Sinai Medical Center.
[#] Department of Microbiology, University of Washington.

10.1021/pr800031f CCC: $40.75

Published on Web 07/25/2008

proteomic approaches such as qualitative proteomics surveys, tandem MS of immunoprecipitated complexes, and label-based quantitative proteomics (e.g., ICAT and iTRAQ). Once constructed, the PA can be used as a reference for designing targeted proteomic strategies such as multiple reaction monitoring (MRM) as well as absolute protein quantification.[7] PeptideAtlas databases have been constructed for the human,[8] human plasma,[9] fly,[10] and yeast[11] proteomes.

Here we report a PA for *Halobacterium salinarum NRC-1*, an obligate halophilic archaeon that evolved unique adaptations, such as increased surface negative charges on folded proteins for survival in its extreme environment of 4.5 M salt.[12] *H. salinarum NRC-1* has a completely sequenced and easily manipulable genome and as such has been used as a model system for constructing a predictive model of cellular responses[13] to a diverse array of routine and stressful environmental changes.[2,4,14–17] The PA represents the product of integration and reprocessing of data from a wide array of proteomics experiments (surveys of fractionated proteomes, enrichment of complexes by immunoprecipitation, and ICAT- and iTRAQ-based quantitative analysis of proteomic changes) in these environmental response studies. This exercise has verified the expression of 63% of the predicted proteome of *H. salinarum NRC-1* including previously undetected and potentially new members of a diverse array of physiological processes. Through extensive analysis of peptides in the PA in context of function, biophysical properties, and abundance, we have identified several factors that might have contributed to our inability to detect 37% of the proteome. Notably, by demonstrating a significant correlation between absolute abundance of proteins and transcripts, we have identified low abundance of proteins as the main limiting factor in peptide detection by mass spectrometry. We have also conducted a comparative analysis of all the various proteomics approaches that contributed data to the construction of the PA to craft strategies for improving coverage and using proteotypic reference peptides for targeted proteomics.

Materials and Methods

Cell Culture Conditions, Protein Preparation, and Mass Spectrometric Conditions. All details regarding cell culturing, protein preparation, and mass spectrometry conditions are discussed in the corresponding publications on *H. salinarum NRC-1* for each of the mass spectrometry proteomics methods included in the PA.[4,14,16–18] These methods include iTRAQ, ICAT, cell fractionation, enrichment by immunoprecipitation, and gel band extracted proteins. However, to aid clarity in the present study, we have delineated pertinent details regarding these procedures in Table 1.

PeptideAtlas Build Summary. A PA is created by identifying the peptides in MS/MS spectra, calculating the genomic coordinates of the peptides, and storing the data sets and derived information in a database for subsequent data mining.[10] The *H. salinarum NRC-1* PA was constructed from 88 experiments [immunoprecipitation (IP), quantitative proteomic analysis using isotopic reagents (ICAT and iTRAQ), and proteome surveys *via* fractionation into soluble and membrane fractions] composed of a total of 636 000 MS/MS from multiple spectrometer vendors [Sciex QStar (Applied Biosystems, Foster City, CA), Micromass QTOF (Waters, Milford, MA) and LCQ (ThermoFinnigan, Waltham, MA)] (Table 2). For each experiment, the vendor format MS/MS spectra were converted to mzXML format[19] and assigned to peptides using SEQUEST[20] and the complete set of *H. salinarum NRC-1* protein sequences derived from the original genome annotation[12] and the National Center for Biotechnology Information (NCBI) and SwissProt sequence databases. The peptide identifications were scored using PeptideProphet[21] and filtered to retain only those with $P > 0.9$, which corresponds to a spectrum identification false discovery rate of 1.1%. After all experiments were processed, the peptides were aligned to the reference proteome. The chromosomal coordinates of peptides from this analysis were verified against NCBI's Generic Features File (GFF) files and manually curated data maintained at ISB (http://baliga.systemsbiology.net/halobacterium) in Systems Biology Experimental Analysis Management System (SBEAMS), a Relational Database Management System (RDBMS) (http://www.sbeams.org), which was also used to archive all of the PA results. We generated a complete library of tryptic peptides by performing an *in silico* digest of the entire *H. salinarum NRC-1* predicted proteome, allowing for one missed cleavage. A measure of peptide observability, the Empirical Observability Score (EOS) (E. Deutsch, personal communication) was calculated for each peptide using the following equation: N_{samples} (peptide)/N_{samples} (protein). For example, if a protein was seen in 10 different samples, and one of its constituent peptides was seen in 5 of those samples, EOS of that peptide would be 0.5.

Calculation of mRNA/Protein Abundance Correlation. To calculate transcript abundance for each of the 1,646 genes whose cognate proteins were detected in the PA (Table 2), we computed the arithmetic mean intensity for that gene across 215 microarray conditions (Supplementary Table ST-1, Supporting Information). These intensities were then log10 transformed. Cultures prepared for these microarray experiments were treated identically to those used for the proteomics experiments included in the PA (Table 1, Supplementary Table ST-1, Supporting Information). Conditions included γ radiation stress,[4] UV radiation,[15] oxygen transitions,[16] and genetic knockouts.[14,15,17] To calculate sequence coverage per protein (Figure 3A), the number of amino acids in each peptide corresponding to a given protein were summed, then divided by the total number of amino acids in that protein. If peptides were detected with partially overlapping amino acid sequences, each of the bases in the overlapping region was only counted once. To calculate the spectral counts (Figure 3B), we computed the arithmetic mean of the number of spectra counted per protein which corresponded to peptides with a confidence value of $P > 0.9$. To calculate the concordance between transcript abundance and cumulative proteome coverage, the average mRNA signal intensities for each gene were organized into 100 bins with 100 intensities per bin (i.e., bin 1 = intensity 0−99; bin 2 = 100−199; etc.) (Figure 3C). The total range of intensities for this analysis was 0 to 50 000. Cumulative proteome coverage was calculated by adding the total number of proteins detected per transcript intensity bin as each successively higher bin was added to the analysis. The *p*-value of the correlation between mRNA and protein abundance was computed by counting the number of times that a set of randomly permuted mRNA and protein levels had a correlation coefficient that was greater than or equal to the reported (unpermuted) correlation.

Results and Discussion

Construction of the *H. salinarum NRC-1* PeptideAtlas. A total of 636 000 tandem mass (MS/MS) spectra from 88 proteomic experiments in 497 individual runs representing at least three

Table 1. Culturing Conditions for All Proteomics Experiments Included in the *H. salinarum* PeptideAtlas[a]

proteomics method	experiment biological purpose	data source	strains	culture conditions and perturbation	protein extraction	fractionation method	protein digestion	labeling	mass spectrometry	peptide to protein matching	data analysis
Fractionation	To detect which proteins are enriched in the membrane vs the cytoplasmic fractions	(19)	*H. salinarum NRC-1*	Standard conditions[b]	Lysed by osmotic shock. Lysates treated with nuclease and PMSF, clarified by centri-fugation.	2 × ultracentrifugation at 53 000 × *g* over 30% sucrose gradient cushion	100 μg protein digested with trypsin at 37 °C overnight	N/A	μLC−ESI−MS/MS on an LCQ-DECA.	SEQUEST[f]	INTERACT[e], Peptide-Prophet, Protein[g]
Gel bands	To identify proteins which form complexes with bacteriorhodopsin regulator	Facciotti and Vuthoori, unpublished data	*H. salinarum NRC-1* with Bat tagged on N terminus with myc	Standard conditions, cross-linked with 1.2% formald-ehyde	Lysed by sonication, treated with protease inhibitors (Roche, Switzerland), clarified by centri-fugation	Sample simplified on Ni2+-NTA resin, run on 10% polyacrylamide gel, gel bands extracted	in-gel digestion with trypsin	N/A	μLC−ESI−MS/MS [d]	SEQUEST	PeptideProphet and ProteinProphet
IP	To delineate protein−protein interaction network specified by the 13 general transcription factors	(17)	*H. salinarum NRC-1* with each of the 13 general transcription factors tagged on C terminus with myc	Standard conditions	Lysed using a micro fluidizer, treated with nuclease and PMSF, clarified by centrifugation	Immunoprecipitation by sepharose-bound IgG and mouse antimyc antibody	trypsin	N/A	μLC−ESI−MS/MS	SEQUEST	PeptideProphet and ProteinProphet
ICAT	To quantify proteome differences between cells with or without phototrophic ability	(18)	*H. Salinarum NRC-1* with overexpressed (bat+; S9) or absent (bat-; SD23) bacteriorhodopsin	Genetic perturbation, standard culture conditions	As described.[c]	cation exchange	trypsin (AFTER labeling)	As described[e] with modifications. Total protein (2.5 mg) was denatured with 6 M urea and 0.05% SDS and immediately reduced with 5 mM tributylphosphine. Cysteine residues were select-ively labeled with a 2-fold molar excess of either light (d0) (*bat*-) or heavy (d8) ICAT (*bat*+) (ABI). d0- and d8-ICAT labeled proteins were mixed in a 1:1 ratio	μLC−ESI−MS/MS	SEQUEST	EXPRESS software[e]

Table 1. Continued

proteomics method	experiment biological purpose	data source	strains	culture conditions and perturbation	protein extraction	fractionation method	protein qadigestion	labeling	mass spectrometry	peptide to protein matching	data analysis
iTRAQ	(1) To quantify protein expression changes occurring over time in cells exposed to gamma irradiation	(4)	*H. salinarum NRC-1*	Cell cultures (OD600nm = 0.4) were exposed to 2500 Gy of gamma-ray. Irradiated and control cultures recovered at 42 °C and 220 rpm shaking, during which samples were removed at various time points to extract proteins.	Lysed by osmotic shock. Lysates treated with nuclease and PMSF, clarified by centrifugation. Proteins were acetone precipitated.	cation exchange desalting, then HPLC	trypsin (BEFORE labeling)	iTRAQ labeling was conducted as per manufacturer's instructions (ABI). Reference samples were labeled with 114 Da reagent, whereas irradiated time point samples were labeled with each of 115, 116, or 117 Da reagent.	LC−MS/MS using an Applied Biosystems API QSTAR Pulsar I, in-house nanospray device	COMET, SEQUEST	PeptideProphet, ProteinProphet, Libra[h]
	(2) To quantify protein expression changes occurring over time in cells exposed to aerobic vs anaerobic conditions	(16)	*H. salinarum NRC-1*	Dissolved oxygen levels were varied between low (0−0.5% saturation) and high (80−100% saturation) in a turbidostat culture (OD 0.6) over the course of 14 h. Samples were removed at various time points to extract proteins.	Lysed by osmotic shock. Lysates treated with nuclease and PMSF, clarified by centrifugation. Proteins were acetone precipitated.	cation exchange desalting, then HPLC	trypsin (BEFORE labeling)	Reference samples were labeled with 114 Da reagent, whereas oxygen-treated time point samples were labeled with each of 115, 116, or 117 Da reagent.	LC−MS/MS using an Applied Biosystems API QSTAR Pulsar I, in-house nanospray device	COMET, SEQUEST	PeptideProphet, ProteinProphet, Libra

[a] See also refs 4, 14, 15, 16, 17, 19. [b] 37 °C, 225 rpm shaking, complex medium (CM; 250 NaCl, 20 g/L MgSO4.7H2O, 3 g/L sodium citrate, 2 g/L KCl, 10 g/L peptone), broad spectrum white light, grown to stationary phase (OD600∼1.0−2.0). Yang, C. F.; Kim, J. M.; Molinari, E.; DasSarma, S. *J. Bacteriol.* **1996**, *178*, 840−845. [c] DasSarma, S.; Fleischmann, E. M. *Halophiles*; Cold Spring Harbor Laboratory Press: Plainview, NY, 1995. [d] Yi, E. C.; Lee, H.; Aebersold, R.; Goodlett, D. R. *Rapid Commun. Mass Spectrom.* **2003**, *17*, 2093−2098. [e] Han, D. K.; Eng, J.; Zhou, H.; Aebersold, R. *Nat. Biotechnol.* **2001**, *19*, 946−951. [f] Eng, J. K.; McCormack, A. L.; Yates, J. R. 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. [g] Keller, A.; Nesvizhskii, A. I.; Kolker, E.; Aebersold, R. *Anal. Chem.* **2002**, *74*, 5383−5392; Nesvizhskii, A. I.; Keller, A.; Kolker, E.; Aebersold, R. *Anal. Chem.* **2003**, *75*, 4646−4658. [h] Pedrioli, P. G.; et al. *Nat. Biotechnol.* **2004**, *22*, 1459−1466.

Table 2. Summary of MS/MS Spectra, Proteins, and Peptides Included in the *H. salinarum NRC-1* PeptideAtlas

Number of experiments[a]	88
Number of MS runs	497
Number of MS/MS spectra	636 300
Number of MS/MS spectra searched[b]	539 950
Number of MS/MS spectra above $P = 0.9$ threshold	76 212
Number of distinct peptides in $P > 0.9$ PeptideAtlas	12 316
Number of distinct peptides aligned to reference genome	11 960
Number of proteins or ORFs in reference genome	2627
Number of proteins or ORFs detected in $P >$ 0.9 PeptideAtlas	1646 (62.7%)

[a] Includes experiments of all types as listed in Figure 2. [b] About 96 000 poor-quality spectra were excluded from SEQUEST search.

types of approaches and three types of mass spectrometers (Materials and Methods) were converted to a common file format (mzXML) (Table 2). Using SEQUEST and PeptideProphet, 76 212 MS/MS spectra or ~12% of all MS/MS spectra had significant matches ($P > 0.9$) to peptides from 1646 predicted proteins in *H. salinarum NRC-1* (Table 2), resulting in a false discovery rate of 1.1%. This represents 1646 proteins or 63% of the predicted proteome, thus improving coverage by 1.7-fold over a previous report that made use of a two-dimensional separation approach for protein cataloguing.[22] To facilitate further analysis, the PA module has been integrated with the *H. salinarum NRC-1* protein annotation module in SBEAMS, a relational database system for managing systems biology data (http://baliga.systemsbiology.net/halobacterium/).

Physiological Functions Represented in the *H. salinarum NRC-1* PA. Although gene finding algorithms such as GLIMMER can identify protein-coding genes with relatively low error rates,[23] until they are verified experimentally, these genes are considered putative. This is an especially important concept considering that over a third of all genes predicted from almost all completely sequenced genomes do not match experimentally characterized orthologs. The identification of a peptide verifies the expression of the parent protein predicted from the genome sequence. In addition, it is important to note that some proteins may only be expressed under specific cellular conditions such as thermal shock or nutritional stress.[24] As such we have verified the expression of 1646 proteins predicted in the *H. salinarum* genome under a variety of environmental conditions. Of these 1029 proteins (9330 peptides) had significant matches to PFAM signatures (e-value < 0.001);[25] 1157 proteins (10 490 peptides) had significant matches to Clusters of Orthologous Groups (COGs) (e-value < 0.001);[26] 902 proteins (9012 peptides) matched manually curated functional annotations;[12,27] and 838 proteins (12,410 peptides) mapped to distinct enzymatic steps within 77 metabolic pathways in Kyoto Encyclopedia of Genes and Genomes (KEGG).[28] In summary, we have experimentally verified the expression of at least 989 proteins (37.6% of the *H. salinarum NRC-1* proteome) with putative functional annotation from all databases combined, which represents a 2.3-fold improvement over the 16.2% verification in a previous proteomic survey.[22] More importantly, we have verified the expression of at least 300 proteins with no orthologs in other organisms, suggesting that many novel functions remain to be discovered in *H. salinarum*. Below we provide some highlights from this analysis.

Detection of Essential Cellular Functions. Consistent with previous *H. salinarum NRC-1* high-throughput proteomics studies, we observed a high degree of coverage for proteins involved in essential cellular functions (Table 3, Supplemental Table ST-2; Supplemental Figure SF-1, Supporting Information). For example, with regard to genetic information processing, unique peptides were detected from five of the six predicted DNA polymerase proteins or subunits (PolA1 was not detected). We have detected unique peptides from 10 of the 12 predicted putative RNA polymerase (RNAP) subunits. In addition, we detected a putative 7 KDa RNAP subunit, Rpc10 (COG1996), which is not cotranscribed with any of the other known RNA polymerase subunits and was not detected in any previous *H. salinarum NRC-1* proteomic surveys. With regard to protein synthesis, secretion and degradation, unique peptides from all 55 ribosomal proteins, all 20 amino acid-tRNA synthetases, elongation factors EF-1α and β, and EF-2, 11 translation initiation factors, 6 putative sec-dependent secretion proteins, 5 putative twin-arginine translocation proteins, and five proteases were detected. Also as expected, proteins involved with cellular motility and relocation including 9 chemotaxis proteins, 8 flagellar proteins, and 11 gas vesicle biogenesis proteins were detected. At least 53 out of the 75 predicted membrane ABC transport system subunits have been detected.

Table 3. Functions of *H. salinarum NRC-1* PA Parent Proteins as Classified by KEGG

category[a]	proteins[b]	%[c]
Energy metabolism	89	100.0%
Carbohydrate metabolism	120	92.3%
Lipid metabolism	50	84.7%
Amino acid metabolism	175	85.4%
Nucleotide metabolism	85	88.5%
Metabolism of cofactors and vitamins	62	72.1%
Transcription	19	65.5%
Translation	76	95.0%
Protein processing	9	100.0%
Replication and repair	4	80.0%
Membrane transport	54	78.3%
Secondary metabolite biosynthesis	24	70.6%
Xenobiotics biodegradation and metabolism	33	75.0%
Signal transduction	10	52.6%
Cell motility	9	90.0%
Metabolism of other amino acids	16	88.9%
Unknown[d]	811	49.3%
Total detected	1646	

[a] Categories represent first-level annotations from the KEGG database. Hand-curated annotations were excluded for simplicity. [b] Total number of PA parent proteins per KEGG category. [c] Percentages indicate the fraction of peptides detected out of the total predicted by the *H. salinarum NRC1* genome sequence in each KEGG category. [d] Proteins marked "unknown" had no annotations and were not represented in the KEGG database.

Identification of Specialized Components of *H. salinarum* Physiology. We have detected 20 of the 26 components of the four unique modes of energy production in *H. salinarum NRC-1*, including oxidative phosphorylation, arginine fermentation, phototrophy, and dimethyl sulfoxide (DMSO) respiration (Table 3, Supplemental Table ST-2, Supporting Information). Although all proteins from the arginine deiminase pathway, including ArcRABC, were previously detected,[22] here we have newly detected previously elusive components of the DMSO and phototrophic respiratory pathways DmsC and the regulator Bat (Table 3). Specifically, we were able to detect 183 unique peptides from 13 proteins involved in energy production via bacteriorhodopsin-mediated phototrophy by enriching for the purple membrane (Table 1: fractionation, ICAT, and bacteri-

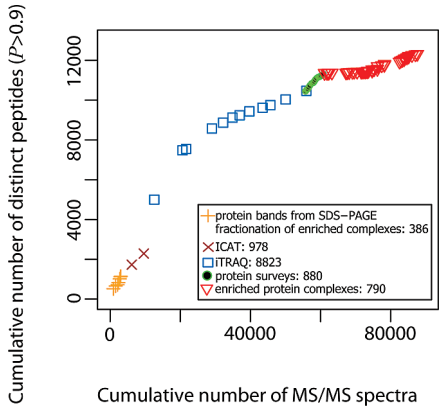

Figure 1. Threshold of peptide detections has been achieved in the *H. salinarum NRC-1* PA. The cumulative number of peptides detected with $P > 0.9$ is plotted as a function of the cumulative number of MS/MS spectra. We observe an increase in new unique peptides with an increase in the numbers of MS/MS spectra added to the PA. However, it appears that we have reached a threshold given the five different approaches that we have used thus far. Peptides detected within each type of proteomic experimental approach are color-coded (see legend for details). Numbers in legend indicate total numbers of peptides detected uniquely in each experiment type. For example, 978 peptides were detected only in ICAT experiments and not in any of the other proteomic experiments represented in the PA.

orhodopsin IP gel band extraction experiments).[29] Interestingly, in cells which overexpress the purple membrane (Table 1: ICAT experiments), we also detected VNG1459H, a protein of unknown function. *VNG1459H* colocalizes in the genome and is significantly coexpressed under relevant environmental conditions with other known phototrophy genes.[13,30,31] Although the exact function remains to be tested, these data support the prediction that this protein may be involved in phototrophy, an extension of a process that was considered well understood.

Fractionation and subsequent solubilization with detergents also improved the detection of other membrane-associated proteins, allowing detection of 188 out of 550 proteins with predicted transmembrane domains[18,32] (Figure 2D). We have also verified the expression of a large number of transcription factors despite their supposed low abundance in the cell: 68% of predicted transcription factors (88 out of 130) are included in the PA, which represents a significant improvement over previous proteomic surveys of *H. salinarum NRC-1* and other organisms, which detected at most 44% of all predicted TFs.[24] This is probably because some experiments in our study were designed to enrich transcription complexes *via* immunoprecipitation.[17] For example, several general transcription factors (GTFs: TFBa, TFBd, TFBe, TBPc, and TBPd) were identified exclusively in these enrichment experiments (Figure 1, Table 1).

Strategies for Improving Proteome Coverage. Despite the unprecedented proteome coverage of the *H. salinarum NRC-1* PA, it is significant that nearly 37% of all predicted proteins were not detected. In fact we observe from a cumulative plot of numbers of distinct peptides detected as a function of individual experiment type that we have reached an apparent threshold that was previously predicted[6,8,11] but not observed until now (Figure 1). To explore the possibility of improving coverage, we examined the influences of several parameters on proteome detection. Using these metrics as a guide, we discuss possible solutions below for improving detections in high-throughput analysis of the proteome.

Influence of Protein Molecular Weight on Detection. As expected, we observed better detection of peptides with an increase in molecular weight up to 1500 Da (Figure 2A). This is explained by a combination of peptide sequence uniqueness with an increase in length and the detection limits of the mass spectrometer. Protein size is also an important factor with regard to proteome coverage considering that lower molecular weight proteins tend to be underrepresented in total protein surveys.[33] However, it is noteworthy that we have detected at least one peptide from each of 406 (~42%) out of the total 963 predicted proteins with calculated molecular weights less than 20 KDa, which is a slight improvement over the 380 proteins detected in a recent study specifically designed to enrich these proteins.[33]

Isoelectric Point. The isoelectric point of a peptide influences its enrichment depending on the type of fractionation columns used for enriching peptides (or proteins) during sample preparation. Most proteins in *H. salinarum NRC-1* have a relatively higher number of acidic residues and the resulting negative surface charge is believed to help circumvent protein aggregation and precipitation in a hypersaline cytoplasm.[34] Consequently, most peptides in *H. salinarum NRC-1* PA are also acidic with a median isoelectric point of 4.4 (Figure 2B). It is interesting that despite the predominant use of cation exchange chromatography for sample processing in most of the experiments within the *H. salinarum NRC-1* PA, a significant fraction of basic peptides were not detected.

Hydrophobicity. Peptides of very low hydrophobicity were poorly detected. This is expected because of the property of the LC column used in most of our experiments.[32] Low hydrophobicity peptides are washed off from these columns before the mass spectrometer has a chance to analyze them. Also, as expected, peptides with hydrophobicity greater than ~30 were detected at a relatively lower frequency, perhaps due to their low solubility (Figure 2C).

Solubility. Despite enrichment of membrane proteins in some experiments, this fraction of the proteome is poorly represented in the PA (Figure 1). This was evident in the observation that over 90% of all detected peptides originated from proteins predicted to be soluble (Figure 2D). This bias in detection has been discussed previously.[4]

Influence of Protein Abundance on Peptide Detection. Abundance of proteins in a population can significantly influence the time a mass spectrometer spends analyzing each unique peptide species.[35] We evaluated the use of mRNA signal intensity from microarray-based transcription profiling experiments as a quantitative proxy for protein abundance, since no currently available methods measure absolute abundance of proteins independently on a systems scale. First we investigated whether mRNA and protein abundance were indeed proportional. The dynamic quantitative relationships between transcription and translation can be assessed at the level of absolute or relative abundance changes in the outputs of mRNA and protein. Although comparisons of relative changes across protein and mRNA concentrations have yielded variable relationships in other studies,[3,36–40] we have previously demonstrated in *H. salinarum* that a significant time-lagged correlation exists between relative changes in transcript and protein abundance provided sufficient temporal measurements for both RNA and protein level changes for most genes over time scales of minutes.[2,16]

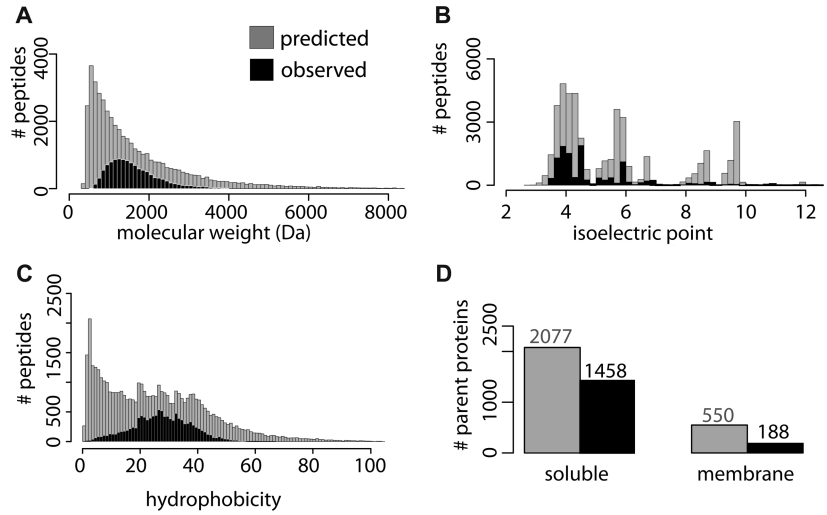

Figure 2. Characteristics of peptides detected in *H. salinarum NRC-1* proteomics experiments included in the PeptideAtlas. (A) Influence of molecular weight (MW) on peptide detection. A comparison of total predicted (gray bars) vs total detected (black bars) tryptic peptides indicates the expected peptide detection range of mass spectrometry as a function of peptide size. (B) Influence of charge on detection of peptides. A similar plot of total expected vs detected peptides as in (A) but as a function of isoelectric point (pI) shows a bias against the detection of basic peptides. (C) Peptide detection as a function of relative hydrophobicity. Comparison of predicted vs observed peptides as a function of hydrophobicity (using the SSRCalc hydrophobicity scale)[43] shows optimal detection was for peptides of hydrophobicity of ~30. (D) Detection of membrane vs soluble proteins. The membrane association predictions are based on hydropathy plots as in (C). This plot shows that despite enrichment of membrane proteins in some experiments, there is a significant bias in detection of membrane proteins (~34%) relative to soluble proteins (~70%).

To further assess this relationship using absolute RNA abundance, we compared mRNA signal intensities from 215 microarray experiments[4,15,16] (Supplementary Table ST-1, Supporting Information) to average spectral counts (over all peptides) per protein from 497 mass spectrometric runs. This comparison yielded significant correlation across the two data sets (Spearman correlation ~0.5; $P < 10^{-6}$) indicating that, irrespective of length (Supplemental Figure SF-2, Supporting Information), the abundance of most proteins is proportional to the abundance of their corresponding transcripts (Figure 3C). Notably, proteome coverage improves dramatically with small increases in mRNA abundance at the lower end of the spectrum (Figure 3C). This may be attributable to the observation that a significant fraction of the transcriptome (>60%) in *H. salinarum NRC-1* appears to be present in low abundance (300−1500 intensity units) (Figure 3C). Regardless, we find that more peptides tend to be detected from proteins whose transcripts are present in higher abundance, as reflected in better sequence coverage and spectral counts per individual protein with an increase in mRNA abundance (Figure 3A, B). Similar trends have been found in *Escherichia coli*;[41] however, an mRNA/protein correlation value *per se* was not reported. In contrast, only a small portion of the genome exhibited significant absolute mRNA/protein concordance in higher eukaryotes,[37] possibly because of the more extensive posttranscriptional and posttranslational modifications that exist in these organisms. Nonetheless, we conclude from our analysis that although targeted enrichment will help detect peptides with certain biophysical properties, approaches to enrich low abundance proteins and higher sensitivity mass spectrometers will yield higher proteome coverage.

Strategies for Targeted Proteomics. As the PA becomes increasingly comprehensive, we can use it to design strategies for high throughput approaches which rapidly characterize the proteome both qualitatively and quantitatively. A tangible approach to accomplish this is *via* the use of "proteotypic peptides",[5,42] which are peptides that map uniquely to one protein and are likely to be observed with LC-MS/MS if the protein is present. We selected proteotypic peptides as those that (i) received a PeptideProphet score of $P > 0.9$, (ii) were detected in more than one experiment, and (iii) had an Empirical Observability Score (EOS) > 0.3 (Materials and Methods). These peptides can be used as beacons for tracking specific proteins in high-throughput experiments for the targeted analysis of the proteome, greatly reducing mass-spectrometer time and improving proteome coverage at the same time. Proteotypic peptides can also aid in the AQUA approach,[7] in which known quantities of labeled synthetic peptides are spiked in and used as reference for absolute quantification of proteins.

Using the criteria listed above, we have identified proteotypic peptides for 1505 proteins or 57.3% of the proteome (Table 4). In other words, we can now, in principle, tune the mass spectrometer to specifically search for 1505 mass spectra instead of a possible ~76 212 (Table 4), which represents a 50-fold reduction in mass spectrometer time to get information on the same number of proteins. However, the selection of proteotypic peptides for practical applications is more complicated since the PA represents a diverse array of experimental techniques (ICAT, iTRAQ, immunoprecipitation, etc.) designed to address different scientific problems. Each of these approaches could have an inherent bias in peptide detection, as suggested by the observation that a majority of proteins were observed in a small number of runs. For example, 633 (43%) of the 1646 detected proteins were observed in 5 or fewer of the 497 runs (Figure 4). We also conducted a comparative analysis to determine possible biases and efficiencies of peptide detection by each individual approach. We caution that such a comparison can be confounded by the significant biases in protein and peptide populations that were intentionally enriched in several of these approaches. For example, there was very little overlap between cysteine-specific ICAT labeling

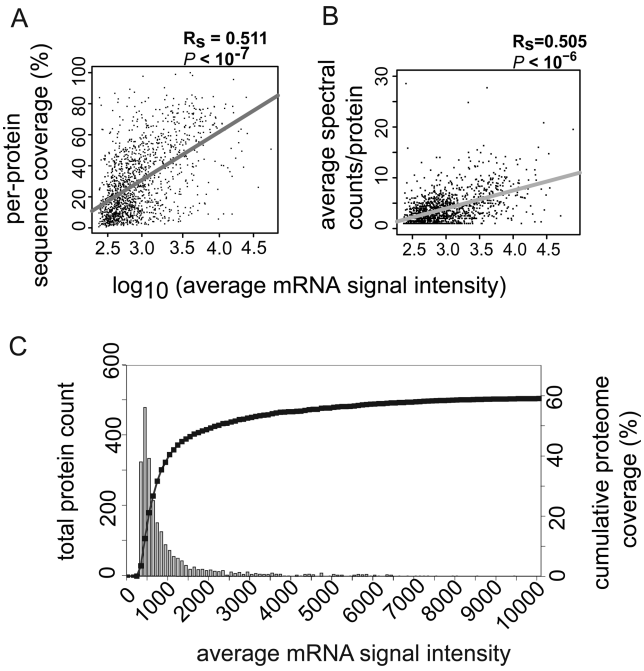

Figure 3. Significant correlation between absolute mRNA and protein abundance. (A) Concordance between transcript abundance and per-protein sequence coverage. Comparison of peptide coverage per protein (Y-axis) and transcript abundance (X-axis), each of which was calculated as described in Materials and Methods. Each point on the scatterplot corresponds to one of the 1646 genes whose proteins were detected in the PA. The Spearman correlation coefficient between the two data sets is shown on the graph ($R_s = 0.511$; $P > 10^{-7}$), and the bold gray line represents the correlation squared (R^2). (B) Concordance between transcript abundance and per-protein spectral counts. Arithmetic mean of the number of spectra counted per protein (*y*-axis; Materials and Methods) was plotted as a function of each protein's cognate transcript abundance (*x*-axis). (C) Concordance between transcript abundance and proteome coverage. Average mRNA signal intensities for each gene are organized into 100 bins with 100 intensities per bin (i.e., bin 1 = intensity 0−99; bin 2 = 100−199; etc). Note that the first three bins are empty (i.e., transcripts with low intensities were detected neither at the mRNA nor at the protein level). Cumulative proteome coverage (black connected squares; right-hand *y*-axis) was calculated by adding the total number of proteins detected per transcript intensity bin as each successively higher bin was added to the analysis. Note that although the analysis was carried out to intensities of 50 000, for brevity we have terminated the graph at 10 000 on the *x*-axis and 60% cumulative coverage on the right-hand *y*-axis, because we observed an increase of new detections of only ~3% between intensities of 10 000 and 50 000. Total protein count (gray vertical bars; left-hand *y*-axis) is represented by the height of each bar, denoting the total number of proteins detected per bin.

approach and any other proteomics method (Figure 5). While the genetic and environmental perturbations could partly explain the reason for the bias, this is more likely an outcome of the poor per-protein cysteine content in *H. salinarum NRC-1* (<65%).[14]

Despite the potential for these inherent biases, we were able to make a fair and statistically significant comparison of performance across two of the most information rich data sets that shared significant numbers of detected proteins: iTRAQ vs. all other shotgun proteomics methods (Figure 5, Table 5). Specifically, we considered proteotypic peptides for 25 proteins

Table 4. Detection of *H. salinarum NRC-1* PA Proteins by Proteotypic Peptides

distinct peptides per protein	distinct proteins represented
1	570
2	405
3	202
4	117
5	71
6	34
7	37
8	24
9	13
10	7
>10[a]	25[b]

[a] Includes all proteins with 11 or more distinct peptides per protein. [b] Sum of distinct proteins represented by 11−56 distinct peptides per protein.

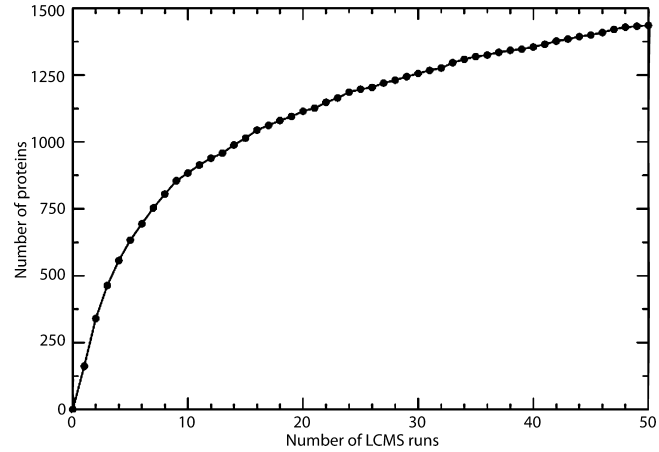

Figure 4. The majority of proteins detected by the *H. salinarum NRC-1* PA are observed in a small number of μLC−ESI−MS/MS runs. The number of μLC−ESI−MS/MS runs is plotted on the X-axis. The cumulative number of proteins detected with the addition of each successive MS run is plotted on the *y*-axis.

that were observed in the largest number of LC−MS runs in iTRAQ experiments vs all other approaches (data for two proteins are shown in Table 5, the remaining 25 proteins is given in Supplemental Table ST-3, Supporting Information). Notably, of the 180 proteotypic peptides for these 25 proteins, only 40 were detected reliably by both approaches. A likely explanation is that iTRAQ introduces a significant bias in the types of peptides that are detected. Therefore, the choice of an appropriate proteotypic peptide will clearly vary depending on the application. Until we have a clearer understanding for the reasons for these observed biases, an empirical approach remains the best option for proteotypic peptide selection. For example, searching for the glutamate dehydrogenase protein GdhB in the *H. salinarum NRC-1* PA yields 26 distinct peptides, which were detected a total of 208 times. Of these, the peptide CAVMDLPFGGAK (PA accession: PAp00363211) has an Empirical Observability Score (EOS) score of 0.59, indicating it was detected reliably. Further investigation into this peptide reveals that it was detected in both iTRAQ and ICAT experiments a total of 57 times and is therefore a reliable candidate for future use as a proteotypic peptide for both of these experimental approaches. However, this peptide was not detected in any of the experiments for targeted enrichment of transcription factors. A second peptide, VVQVSVPVER (PAp00368363) with a lower EOS score of 0.41, on the other hand, was detected

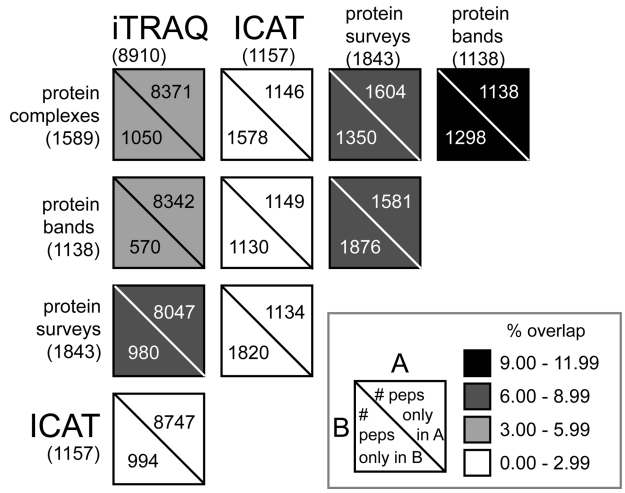

Figure 5. Comparison of all experiments included in the *H. salinarum* PA. Numbers in parentheses indicate peptide detections by each approach, numbers in boxes indicate unique detections, and shading indicates the percentage of peptides detected by both approaches. Experiment names refer to those introduced in Table 1.

Table 5. Comparison between iTRAQ and Non-iTRAQ Proteotypic Peptides for Two Proteins

method[a]	protein ID	peptide[b]
iTRAQ	VNG0414G	DIPTVVVER
iTRAQ	VNG0414G	EFDDGPAAAVIK
iTRAQ	VNG0414G	YGENPHQDAAVYR
non-iTRAQ	VNG0414G	EVVVAPGYTDDAVDVLTAK
non-iTRAQ	VNG0414G	DNTHAAASVVHADQLNPDAK
non-iTRAQ	VNG0414G	HTNPAGCATADTLADAYSDALSTDAK
non-iTRAQ	VNG0414G	VLDVGTLDGTPAPVTETPLVGGR
iTRAQ	VNG0620G	QSFLDVVMNER
iTRAQ	VNG0620G	GPAASGGYYTIAPTDK
iTRAQ	VNG0620G	AQIYAGNK
iTRAQ	VNG0620G	ASVQNYGVVYR
iTRAQ	VNG0620G	VSSPGGAVSGSEVQYR
non-iTRAQ	VNG0620G	LAQEKPVVTSVR
non-iTRAQ	VNG0620G	AVEIGLADEIGGLDAAIADAADR
non-iTRAQ	VNG0620G	VSSPGGAVSGSEVQYR
non-iTRAQ	VNG0620G	GPAASGGYYTIAPTDK

[a] Two proteins VNG0414G and VNG620G were identified in both iTRAQ experiments and in non-iTRAQ experiments in 71 and 51 LCMS runs, respectively. [b] We selected the peptides most frequently observed in each experimental method and note that they are distinct.

only 12 times but observed in at least three of the targeted enrichment experiments. Clearly, this second peptide would make for a better proteotypic peptide for these targeted enrichment applications. Using this approach, one can compile a custom list of proteotypic peptides for a specific application of interest. Further explorations of this type are possible at the *H. salinarum NRC-1* proteome annotation webpage (http://baliga.systemsbiology.net/halobacterium).

Conclusion

Analysis of the PA from *H. salinarum NRC-1* has provided two fundamental insights into approaches to improve proteome coverage and strategies for targeted proteomics. First, our findings suggest that expression level and abundance supersede certain inherent biophysical properties as the determining factors in protein detectability. Second, empirically observed proteotypic peptides will undoubtedly help design targeted proteomics approaches; however, the choice of specific peptides must be context-dependent. For a given protein, the proteotypic peptide with the best score may not always be the best candidate to track its expression in certain types of proteomics studies.

Acknowledgment. We thank Christopher Bare and David Campbell for help with programming, database construction and false discovery rate calculation, and Ning Zhang for help with generating proteotypic peptide scores. This work was supported by grants from NIH (P50GM076547 and 1R01GM077398-01A2), DoE (MAGGIE: DE-FG02-07ER64327), NSF (EF-0313754, EIA-0220153, MCB-0425825, DBI-0640950) and NASA (NNG05GN58G) and Institute for Systems Biology institutional support to N.S.B., postdoctoral fellowships are acknowledged from NSF to M.T.F. (DBI 0400598) and K.W. (0443746), and from NIH (5F32GM078980-02) to A.K.S.

Supporting Information Available: Comparative analysis revealed possible similarities between spectral counts and genome organization, a finding discussed in the Supporting Information. Supporting figures (Figure SF-1, Figure SF-2) and Supplementary Tables (ST1 and ST2).This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Facciotti, M. T.; Bonneau, R.; Hood, L.; Baliga, N. S. Systems biology experimental design-considerations for building predictive gene regulatory network models for prokaryotic systems. *Current Genomics* **2004**, *5*, 527–544.
(2) Kaur, A.; Pan, M.; Meislin, M.; Facciotti, M. T.; El-Geweley, R.; Baliga, N. S. A systems view of haloarchaeal strategies to withstand stress from transition metals. *Genome Res.* **2006**, *16*, 841–854.
(3) Gygi, S. P.; Rochon, Y.; Franza, B. R.; Aebersold, R. Correlation between protein and mRNA abundance in yeast. *Mol. Cell. Biol.* **1999**, *19*, 1720–1730.
(4) Whitehead, K.; Kish, A.; Pan, M.; Kaur, A.; Reiss, D. J.; King, N.; Hohmann, L.; Diruggiero, J.; Baliga, N. S. An integrated systems approach for understanding cellular responses to γ radiation. *Mol. Syst. Biol.* **2006**, *2*, 47.
(5) Mallick, P.; Schirle, M.; Chen, S. S.; Flory, M. R.; Lee, H.; Martin, D.; Ranish, J.; Raught, B.; Schmitt, R.; Werner, T.; Kuster, B.; Aebersold, R. Computational prediction of proteotypic peptides for quantitative proteomics. *Nat. Biotechnol.* **2007**, *25*, 125–131.
(6) Desiere, F.; Deutsch, E. W.; King, N. L.; Nesvizhskii, A. I.; Mallick, P.; Eng, J.; Chen, S.; Eddes, J.; Loevenich, S. N.; Aebersold, R. The PeptideAtlas project. *Nucleic Acids Res.* **2006**, *34*, D655–658.
(7) Beynon, R. J.; Doherty, M. K.; Pratt, J. M.; Gaskell, S. J. Multiplexed absolute quantification in proteomics using artificial QCAT proteins of concatenated signature peptides. *Nat. Methods* **2005**, *2*, 587–589.
(8) Desiere, F.; Deutsch, E. W.; Nesvizhskii, A. I.; Mallick, P.; King, N. L.; Eng, J. K.; Aderem, A.; Boyle, R.; Brunner, E.; Donohoe, S.; Fausto, N.; Hafen, E.; Hood, L.; Katze, M. G.; Kennedy, K. A.; Kregenow, F.; Lee, H.; Lin, B.; Martin, D.; Ranish, J. A.; Rawlings, D. J.; Samelson, L. E.; Shiio, Y.; Watts, J. D.; Wollscheid, B.; Wright, M. E.; Yan, W.; Yang, L.; Yi, E. C.; Zhang, H.; Aebersold, R. Integration with the human genome of peptide sequences obtained by high-throughput mass spectrometry. *Genome Biol.* **2005**, *6*, R9.
(9) Deutsch, E. W.; Eng, J. K.; Zhang, H.; King, N. L.; Nesvizhskii, A. I.; Lin, B.; Lee, H.; Yi, E. C.; Ossola, R.; Aebersold, R. Human Plasma PeptideAtlas. *Proteomics* **2005**, *5*, 3497–3500.
(10) Desiere, F.; Deutsch, E. W.; Nesvizhskii, A. I.; Mallick, P.; Eng, J. K.; Aderem, A.; Boyle, R.; Brunner, E.; Donohoe, S.; Fausto, N.; Hafen, E.; Hood, L.; Katze, M. G.; Kennedy, K.; Kregenow, F.; Lee, H.; Lin, B.; Martin, D.; Ranish, J.; Rawlings, D. J.; Samelson, L. E.; Shiio, Y.; Watts, J.; Wollscheid, B.; Wright, M. E.; Yan, W.; Yang, L.; Yi, E.; Zhang, H.; Aebersold, R. Integration of Peptide Sequences Obtained by High-Throughput Mass Spectrometry with the Human Genome. *Genome Biol.* **2004**, *5*, R9.
(11) King, N. L.; Deutsch, E. W.; Ranish, J. A.; Nesvizhskii, A. I.; Eddes, J. S.; Mallick, P.; Eng, J.; Desiere, F.; Flory, M.; Martin, D. B.; Kim,

B.; Lee, H.; Raught, B.; Aebersold, R. Analysis of the Saccharomyces cerevisiae proteome with PeptideAtlas. *Genome Biol.* **2006**, *7*, R106.
(12) Ng, W. V.; Kennedy, S. P.; Mahairas, G. G.; Berquist, B.; Pan, M.; Shukla, H. D.; Lasky, S. R.; Baliga, N. S.; Thorsson, V.; Sbrogna, J.; Swartzell, S.; Weir, D.; Hall, J.; Dahl, T. A.; Welti, R.; Goo, Y. A.; Leithauser, B.; Keller, K.; Cruz, R.; Danson, M. J.; Hough, D. W.; Maddocks, D. G.; Jablonski, P. E.; Krebs, M. P.; Angevine, C. M.; Dale, H.; Isenbarger, T. A.; Peck, R. F.; Pohlschroder, M.; Spudich, J. L.; Jung, K. W.; Alam, M.; Freitas, T.; Hou, S.; Daniels, C. J.; Dennis, P. P.; Omer, A. D.; Ebhardt, H.; Lowe, T. M.; Liang, P.; Riley, M.; Hood, L.; DasSarma, S. Genome sequence of Halobacterium species NRC-1. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 12176–12181.
(13) Bonneau, R.; Facciotti, M. T.; Reiss, D. J.; Schmid, A. K.; Pan, M.; Kaur, A.; Thorsson, V.; Shannon, P.; Johnson, M. H.; Bare, J. C.; Longabaugh, W.; Vuthoori, M.; Whitehead, K.; Madar, A.; Suzuki, L.; Mori, T.; Chang, D. E.; Diruggiero, J.; Johnson, C. H.; Hood, L.; Baliga, N. S. A predictive model for transcriptional control of physiology in a free living cell. *Cell* **2007**, *131*, 1354–1365.
(14) Baliga, N. S.; Pan, M.; Goo, Y. A.; Yi, E. C.; Goodlett, D. R.; Dimitrov, K.; Shannon, P.; Aebersold, R.; Ng, W. V.; Hood, L. Coordinate regulation of energy transduction modules in Halobacterium sp. analyzed by a global systems approach. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 14913–14918.
(15) Baliga, N. S.; Bjork, S. J.; Bonneau, R.; Pan, M.; Iloanusi, C.; Kottemann, M. C. H.; Hood, L.; DiRuggiero, J. Systems Level Insights Into the Stress Response to UV Radiation in the Halophilic Archaeon Halobacterium NRC-1. *Genome Res.* **2004**, *14*, 1025–1035.
(16) Schmid, A. K.; Reiss, D. J.; Kaur, A.; Pan, M.; King, N.; Van, P. T.; Hohmann, L.; Martin, D. B.; Baliga, N. S. The anatomy of microbial cell state transitions in response to oxygen. *Genome Res.* **2007**, *17*, 1399–1413.
(17) Facciotti, M. T.; Reiss, D. J.; Pan, M.; Kaur, A.; Vuthoori, M.; Bonneau, R.; Shannon, P.; Srivastava, A.; Donohoe, S. M.; Hood, L. E.; Baliga, N. S. General transcription factor specified global gene regulation in archaea. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 4630–4635.
(18) Goo, Y. A.; Yi, E. C.; Baliga, N. S.; Tao, W. A.; Pan, M.; Aebersold, R.; Goodlett, D. R.; Hood, L.; Ng, W. V. Proteomic Analysis of an Extreme Halophilic Archaeon, Halobacterium sp. NRC-1. *Mol. Cell Proteomics* **2003**, *2*, 506–524.
(19) Pedrioli, P. G.; Eng, J. K.; Hubley, R.; Vogelzang, M.; Deutsch, E. W.; Raught, B.; Pratt, B.; Nilsson, E.; Angeletti, R. H.; Apweiler, R.; Cheung, K.; Costello, C. E.; Hermjakob, H.; Huang, S.; Julian, R. K.; Kapp, E.; McComb, M. E.; Oliver, S. G.; Omenn, G.; Paton, N. W.; Simpson, R.; Smith, R.; Taylor, C. F.; Zhu, W.; Aebersold, R. A common open representation of mass spectrometry data and its application to proteomics research. *Nat. Biotechnol.* **2004**, *22*, 1459–1466.
(20) Eng, J. K.; McCormack, A. L.; Yates, J. R. I. An approach to correlate tandem mass spectral data of pepties with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom.* **1994**, *5*, 976–989.
(21) Keller, A.; Eng, J.; Zhang, N.; Li, X. J.; Aebersold, R. A uniform proteomics MS/MS analysis platform utilizing open XML file formats. *Mol. Syst. Biol.* **2005**, *1*, 2005–0017.
(22) Gan, R. R.; Yi, E. C.; Chiu, Y.; Lee, H.; Kao, Y. C.; Wu, T. H.; Aebersold, R.; Goodlett, D. R.; Ng, W. V. Proteome analysis of Halobacterium sp. NRC-1 facilitated by the biomodule analysis tool BMSorter. *Mol. Cell Proteomics* **2006**, *5*, 987–997.
(23) Delcher, A. L.; Bratke, K. A.; Powers, E. C.; Salzberg, S. L. Identifying bacterial genes and endosymbiont DNA with Glimmer. *Bioinformatics* **2007**, *23*, 673–679.
(24) Lipton, M. S.; Pasa-Tolic, L.; Anderson, G. A.; Anderson, D. J.; Auberry, D. L.; Battista, J. R.; Daly, M. J.; Fredrickson, J.; Hixson, K. K.; Kostandarithes, H.; Masselon, C.; Markillie, L. M.; Moore, R. J.; Romine, M. F.; Shen, Y.; Stritmatter, E.; Tolic, N.; Udseth, H. R.; Venkateswaran, A.; Wong, K. K.; Zhao, R.; Smith, R. D. Global analysis of the Deinococcus radiodurans proteome by using accurate mass tags. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 11049–11054.
(25) Finn, R. D.; Mistry, J.; Schuster-Bockler, B.; Griffiths-Jones, S.; Hollich, V.; Lassmann, T.; Moxon, S.; Marshall, M.; Khanna, A.; Durbin, R.; Eddy, S. R.; Sonnhammer, E. L.; Bateman, A. Pfam: clans, web tools and services. *Nucleic Acids Res.* **2006**, *34*, D247–251.
(26) Tatusov, R. L.; Galperin, M. Y.; Natale, D. A.; Koonin, E. V. The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res.* **2000**, *28*, 33–36.
(27) Bonneau, R.; Baliga, N. S.; Deutsch, E. W.; Shannon, P.; Hood, L. Comprehensive de novo structure prediction in a systems-biology context for the archaea Halobacterium sp. NRC-1. *Genome Biol.* **2004**, *5*, R52.
(28) Kanehisa, M. The KEGG database. *Novartis Found. Symp.* **2002**, *247*, 91–101, and discussion 101−103, 119−128, 244−152.
(29) Hartmann, R.; Sickinger, H. D.; Oesterhelt, D. Anaerobic growth of halobacteria. *Proc. Natl. Acad. Sci. U.S.A.* **1980**, *77*, 3821–3825.
(30) Bonneau, R.; Reiss, D. J.; Shannon, P.; Facciotti, M.; Hood, L.; Baliga, N. S.; Thorsson, V. The Inferelator: an algorithm for learning parsimonious regulatory networks from systems-biology data sets de novo. *Genome Biol.* **2006**, *7*, R36.
(31) Reiss, D. J.; Baliga, N. S.; Bonneau, R. Integrated biclustering of heterogeneous genome-wide datasets for the inference of global regulatory networks. *BMC Bioinformatics* **2006**, *7*, 280.
(32) Schindler, P. A.; Van Dorsselaer, A.; Falick, A. M. Analysis of hydrophobic proteins and peptides by electrospray ionization mass spectrometry. *Anal. Biochem.* **1993**, *213*, 256–263.
(33) Klein, C.; Aivaliotis, M.; Olsen, J. V.; Falb, M.; Besir, H.; Scheffer, B.; Bisle, B.; Tebbe, A.; Konstantinidis, K.; Siedler, F.; Pfeiffer, F.; Mann, M.; Oesterhelt, D. The Low Molecular Weight Proteome of Halobacterium salinarum. *J. Proteome Res.* **2007**, *6*, 1510–1518.
(34) Kennedy, S. P.; Ng, W. V.; Salzberg, S. L.; Hood, L.; DasSarma, S. Understanding the Adaptation of Halobacterium Species NRC-1 to Its Extreme Environment through Computational Analysis of Its Genome Sequence. *Genome Res.* **2001**, *11*, 1641–1650.
(35) Zhang, H.; Li, X. J.; Martin, D. B.; Aebersold, R. Identification and quantification of N-linked glycoproteins using hydrazide chemistry, stable isotope labeling and mass spectrometry. *Nat. Biotechnol.* **2003**, *21*, 660–666.
(36) Bitton, D. A.; Okoniewski, M. J.; Connolly, Y.; Miller, C. J. Exon level integration of proteomics and microarray data. *BMC Bioinformatics* **2008**, *9*, 118.
(37) Cox, B.; Kislinger, T.; Wigle, D. A.; Kannan, A.; Brown, K.; Okubo, T.; Hogan, B.; Jurisica, I.; Frey, B.; Rossant, J.; Emili, A. Integrated proteomic and transcriptomic profiling of mouse lung development and Nmyc target genes. *Mol. Syst. Biol.* **2007**, *3*, 109.
(38) Flory, M. R.; Lee, H.; Bonneau, R.; Mallick, P.; Serikawa, K.; Morris, D. R.; Aebersold, R. Quantitative proteomic analysis of the budding yeast cell cycle using acid-cleavable isotope-coded affinity tag reagents. *Proteomics* **2006**, *6*, 6146–6157.
(39) Washburn, M. P.; Koller, A.; Oshiro, G.; Ulaszek, R. R.; Plouffe, D.; Deciu, C.; Winzeler, E.; Yates, J. R. 3rd Protein pathway and complex clustering of correlated mRNA and protein expression analyses in Saccharomyces cerevisiae. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 3107–3112.
(40) Schmidt, M. W.; Houseman, A.; Ivanov, A. R.; Wolf, D. A. Comparative proteomic and transcriptomic profiling of the fission yeast Schizosaccharomyces pombe. *Mol. Syst. Biol.* **2007**, *3*, 79.
(41) Corbin, R. W.; Paliy, O.; Yang, F.; Shabanowitz, J.; Platt, M.; Lyons, C. E.; Root, K.; McAuliffe, J.; Jordan, M. I.; Kustu, S.; Soupene, E.; Hunt, D. F. Toward a protein profile of Escherichia coli: comparison to its transcription profile. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 9232–9237.
(42) Craig, R.; Cortens, J. P.; Beavis, R. C. The use of proteotypic peptide libraries for protein identification. *Rapid Commun. Mass Spectrom.* **2005**, *19*, 1844–1850.
(43) Krokhin, O. V. Sequence-specific retention calculator. Algorithm for peptide retention prediction in ion-pair RP-HPLC: application to 300- and 100-A pore size C18 sorbents. *Anal. Chem.* **2006**, *78*, 7785–7795.

PR800031F