Expression profiling of hypothetical genes in *Desulfovibrio vulgaris* leads to improved functional annotation

Dwayne A. Elias – corresponding author
Department of Biochemistry, 117 Schweitzer Hall, University of Missouri-Columbia, Columbia, MO 65211, and Virtual Institute for Microbial Stress and Survival, phone: 1 (573) 882-9771, fax: 1 (573) 882-5635, email: eliasd@missouri.edu

Aindrila Mukhopadhyay‡
Physical Biosciences, Lawrence Berkeley National Lab, 1 Cyclotron Road Mail Stop 978R4121, Berkeley, CA 94720, and Virtual Institute for Microbial Stress and Survival, phone: 1 (510) 495-2628, fax: 1 (510) 486-4252, email: AMukhopadhyay@lbl.gov

Marcin P. Joachimiak‡
Physical Biosciences, Lawrence Berkeley National Lab, 1 Cyclotron Road Mail Stop STANLEY, Berkeley, CA 94720, and Virtual Institute for Microbial Stress and Survival, phone: 1 (510) 643-5678, fax: 1 (510) 486-6219, email: MJoachimiak@lbl.gov

Elliott C. Drury
Department of Biochemistry, 117 Schweitzer Hall, University of Missouri-Columbia, Columbia, MO 65211, and Virtual Institute for Microbial Stress and Survival, phone: 1 (573) 882-9771, fax: 1 (573) 882-5635, email: elliott.drury@gmail.com

Alyssa M. Redding
Lawrence Berkeley National Lab, 1 Cyclotron Road Mail Stop 978R4121, Berkeley, CA 94720, and Virtual Institute for Microbial Stress and Survival, phone: 1 (510) 495-2420, fax: 1 (510) 486-5225, email: ARedding@lbl.gov

Huei-Che B. Yen
Department of Biochemistry, Room 117 Schweitzer Hall, University of Missouri-Columbia, Columbia, MO 65211, and Virtual Institute for Microbial Stress and Survival, phone: 1 (573) 882-9771, fax: 1 (573) 882-5635, email: yenb@missouri.edu

Matthew W. Fields
Department of Microbiology, Montana State University, 366 EPS Building, Bozeman, MT 59717, and Virtual Institute for Microbial Stress and Survival, phone: 1 (406) 994-7340, fax: 1 (406) 994-6098, email: matthew.fields@erc.montana.edu

Terry C. Hazen
Earth Sciences, Lawrence Berkeley National Lab, 1 Cyclotron Road Mail Stop 70A3317, Berkeley, CA 94720, and Virtual Institute for Microbial Stress and Survival, phone: 1 (510) 486-6223, fax: 1 (510) 486-4254, email: TCHazen@lbl.gov

Adam P. Arkin
Physical Biosciences, Lawrence Berkeley National Lab, 1 Cyclotron Road Mail Stop STANLEY, Berkeley, CA 94720, and Virtual Institute for Microbial Stress and Survival, phone: 1 (510) 643-5678, fax: 1 (510) 486-6219, email: APArkin@lbl.gov

Jay D. Keasling
Lawrence Berkeley National Lab, 1 Cyclotron Road Mail Stop 978R4121, Berkeley, CA 94720, and Virtual Institute for Microbial Stress and Survival, phone: 1 (510) 495-2620, fax: 1 (510) 495-2630, email: JDKeasling@lbl.gov

Judy D. Wall
Department of Biochemistry, 117 Schweitzer Hall, University of Missouri-Columbia, Columbia, MO 65211, and Virtual Institute for Microbial Stress and Survival, phone: 1 (573) 882-8726, fax: 1 (573) 882-5635, email: wallj@missouri.edu

‡ - both authors contributed equally to this work.

submitted to: Nucleic Acids Research;
Running title: *D. vulgaris* hypothetical gene annotation
Keywords: *Desulfovibrio*, sulfate-reducing bacteria, hypothetical, conserved hypothetical, annotation.

Abstract

Hypothetical and conserved hypothetical genes account for >30% of sequenced bacterial genomes. For the sulfate-reducing bacterium *Desulfovibrio vulgaris* Hildenborough, 347 of the 3634 genes were annotated as conserved hypothetical (9.5%) along with 887 hypothetical genes (24.4%). Given the large fraction of the genome, it is plausible that some of these genes serve critical cellular roles. The study goals were to determine which genes were expressed and provide a more functionally based annotation. To accomplish this, expression profiles of 1234 hypothetical and conserved genes were used from transcriptomic datasets of 11 environmental stresses, complemented with shotgun LC-MS/MS and AMT tag proteomic data. Genes were divided into putatively polycistronic operons and those predicted to be monocistronic, then classified by basal expression levels and grouped according to changes in expression for one or multiple stresses. 1212 of these genes were transcribed with 786 producing detectable proteins. There was no evidence for expression of 17 predicted genes. Except for the latter, monocistronic gene annotation was expanded using the above criteria along with matching Clusters of Orthologous Groups. Polycistronic genes were annotated in the same manner with inferences from their proximity to more confidently annotated genes. Two targeted deletion mutants were used as test cases to determine the relevance of the inferred functional annotations.

Introduction

The application of genome sequencing and sequence annotation to numerous bacterial species has yielded a "road map" for several avenues of research. These include the incorporation of gene expression changes at both the mRNA and protein levels (1-5) with physiological and metabolic studies, to deduce the behavior of the microbe as a whole, a field now called Systems Microbiology. Other approaches to discern function include genetic manipulations such as gene/protein tagging for the identification and visualization of protein complexes (6-8) and deletion mutagenesis (9-12) for confirming the function(s) of a given gene or protein. One aspect that has come to light through the sequencing of more than 780 completed bacterial and archeal genomes (http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi) with an additional 2179 ongoing (http://www.genomesonline.org/gold.cgi), is that approximately 1/3 of all of the genes within a given genome are typically predicted to encode hypothetical and conserved hypothetical genes (13). Hypothetical (HyP) proteins are defined as those with no significant sequence similarity (*i.e.* homology) to any characterized or uncharacterized predicted proteins, while conserved hypothetical (CHyP) proteins are those that have significant similarity to a predicted protein in another species or strain without direct evidence of the expression of the gene as defined by TIGR (http://cmr.jcvi.org/tigr-scripts/CMR/CmrHomePage.cgi).

This lack of similarity to well-characterized genes and proteins increases the interest in these groups of sequences, as they may well be important to specialized cellular physiology and metabolism, may possess unique functions, or complete tentatively assigned pathways. In the case of HyP, their potential existence is only supported by the output of gene prediction programs such as GLIMMER (14,15). More recently however, computational tools for determining plausible functions of gene products have been developed. These include methods such as genomic context analysis (16-19) and phylogenomic profiling (20,21). Such tools are still of limited application for HyP or CHyP due to the lack of sequence homology for the gene or protein of interest within a well-characterized, sequenced genome. Hence, a working knowledge of the function(s) of the gene products encoded is expected to lead to a greater mechanistic understanding of the metabolic capabilities of a target microorganism and possibly to a better physiological knowledge of other organisms through gene sequence and neighborhood association.

As a first step to the identification of gene function, it is important to determine if the genes are actually expressed, if expression results in production of a functional protein, and under which conditions the genes and proteins may be differentially regulated or influenced. Experiments to obtain expression information have applied a combination of transcriptomic and proteomic methods. In *Haemophilus influenzae*, 54 hypothetical genes of a total 1744 genomic loci were found to be expressed with both methods of analysis and were used to assign specific functions for 16 genes and general roles for another 27 genes (13). More recently, the metal-reducing bacterium *Shewanella oneidensis* MR-1 was the focus of three investigations (22-24). In this organism, 40% of the sequenced genome consists of genes predicted to encode HyP and CHyP. The first study of *S. oneidensis* MR-1 by Kolker and co-workers (24) showed the actual transcription of >500 HyP and CHyP genes with at least a general functional assignment for 240 of them. In a related study by Elias et al. (22) of only HyP, comprehensive MS-based proteomics data were queried across many culture conditions and the results confirmed the existence of 262 predicted proteins. Additionally, inferences were made for the subcellular localization and function from differential expression in these discreet culturing conditions (22). A study of the CHyP in this same organism by Elias et al. (23), confirmed protein production from 758 such

genes with improved functional assignments that were also inferred in part from the culturing conditions of expression (23).

The sulfate-reducing bacterium (SRB) *Desulfovibrio vulgaris* Hildenborough has a sequenced genome (25) and several physiological and metabolic studies have taken advantage of this information (3,26-29). This bacterium is well-known as a model SRB (30), and is known to reduce and immobilize metals and radionuclides (31-33) making it of interest for bioremediation efforts. Additionally, *D. vulgaris* Hildenborough was recently the focus of an effort to assign putative functions to predicted HyP from proteomics data obtained through LC-MS/MS analysis of cultures grown with different electron donors (34). These assignments were based not only on proteomics data but also relied on a number of computational methods and resulted in the re-annotation of 20 CHyP genes and the confirmation of gene expression for 129 HyP genes.

In the current study, we have incorporated the previously published results and expanded the analyses. Transcriptomics data derived from eleven different stresses as well as corresponding shotgun LC-MS/MS proteomics data from selected stress conditions were used to query the expression of each HyP or CHyP gene. Some of these data have already been published as part of a genome-wide transcriptional response to a particular stress condition. These published data include responses to heat shock (26), high salt (28), nitrate or nitrite exposure (35,36), stationary versus exponential growth (2), high pH (37), effect of deleting the Fur gene (10) and O_2 exposure (38). We have incorporated the HyP and CHyP microarray and proteomics data from these studies as well as currently unpublished results. The latter conditions include peroxide stress in the wild type and a mutant lacking the *perR* gene; a strain lacking pDV1, the 202kb native plasmid found in *D. vulgaris* Hildenborough, vs the wild type *D. vulgaris* that has pDV1; a co-culture with a methanogen vs *D. vulgaris* alone; acidic conditions; cold stress; chromium exposure; NaCl adaptation; and Fe(II)-limitation. The microarray datasets for all hypothetical and conserved hypothetical genes in published studies have been made publicly available (http://www.microbesonline.org).

These data were compiled and the genes categorized by basal expression rates, by the presence versus the absence of differential gene expression in response to a particular stress, by whether the response was specific to a single stress or seen in multiple stress conditions, and by the operon status of the genes, monocistronic or polycistronic. For the latter, differential expression in a given stress that coordinated with the rest of the operon was also considered. Additionally, bioinformatic information such as COG and subcellular location were used for functional annotation. This may be the first such comprehensive investigation utilizing mRNA microarrays and proteomics to infer a more robust functional annotation of HyP and CHyP genes from such a large number of stress conditions.

Materials and Methods

Biomass production. *Desulfovibrio vulgaris* Hildenborough (ATCC 29579) was grown in a defined lactate (60 mM)/sulfate (50 mM) medium, LS4D (28), under a variety of different stress conditions as have been reported (28,35,36). The chilled samples were harvested via centrifugation, flash frozen in liquid nitrogen, and stored at -80°C until analysis.

Culture Maintenance*. D. vulgaris* cultures from the American Type Culture Collection (ATCC) were grown to mid log phase in 1 L of LS4D, checked for purity by the appearance of anaerobic SRB colonies on LS4D plates as well as the absence of colonies on aerobic glucose plates, dispensed into 2 mL cryogenic vials (Nalgene) with 0.5 mL 30% (vol/vol) glycerol, and frozen at

-80°C until used as previously described (28). To minimize phenotypic drift from repetitive culturing, all experiments used cells that were started from frozen stocks and were fewer than three subcultures from the original ATCC culture. All experiments, inoculations, and transfers were performed in an anaerobic glove chamber (Coy Laboratory Products Inc., Grass Lake, MI) with an atmosphere of approximately 5% CO_2, 5% H_2 and 90% N_2.

Microarray transcriptomic and data analysis. DNA microarrays using 70-mer oligonucleotide probes covering 3,482 of the 3,534 annotated protein-coding sequences of the *D. vulgaris* genome that were constructed as previously described (39). Of the 52 genes not found on the microarrays, 14 were either HyP and CHyP (under Expression Category; Supplemental Tables 5,6). Total RNA extraction, purification, and labeling with the fluorophore Cy5-dUTP (Amersham Biosciences, Piscataway, NJ) were performed independently on each cell sample using previously described protocols (38). Genomic DNA was extracted from *D. vulgaris* cultures and hybridized as previously described (36). Microarray data analyses were performed using gene models from NCBI. All mRNA changes were assessed with total genomic DNA (gDNA) as a control for each of the experimental and control hybridizations. Log_2 ratios of mRNA to gDNA hybridized to gene oligonucleotides and z-scores were computed as previously described (9). A mean Log_2 ratio cut off across time points of $\geq |1.2|$ and an accompanying z-score $\geq |2|$, respectively, were used to identify the genes whose expression changed significantly.

$$Z = \frac{Log_2(Treatment / Control)}{\sqrt{0.25 + \sum \text{var}\,iance}}$$

Proteomics and proteomic data analyses.

Shotgun LC-MS/MS analysis. Sample preparation, chromatography, and mass spectrometry for shotgun LC-MS/MS proteomics were performed as described previously (35). Briefly, frozen cell pellets from triplicate 50 mL cultures were thawed and pooled prior to cell lysis. Cells were lysed via sonication in lysis buffer, composed of 4 M urea with 500 mM triethylammonium bicarbonate (TEAB), pH 8.5 (Sigma-Aldrich, St. Louis, MO), and the clarified lysate was used as total cellular protein. Sample denaturation, reduction, blocking, digestion, and labeling with isobaric reagents were performed according to the manufacturer's directions (Applied Biosystems, Framingham, MA). Strong cation exchange (SCX) chromatography was used to separate the iTRAQ labeled samples into 21-23 salt fractions. Fractions were desalted using C_{18} MacroSpin Columns (Nest Group, Southborough, MA), dried, and separated on a PepMap100 C_{18} reverse phase nano-LC-MS column (Dionex-LC Packings, Sunnyvale, CA) using an Ultimate HPLC with Famous Autosampler and Switchos Micro Column Switching Module coupled with an ESI-QTOF mass analyzer (QSTAR® Hybrid Quadrupole TOF, Applied Biosystems) as previously described (26). A 2 h gradient from 0-25% acetonitrile was used. Two product ion scans were collected from each cycle with a 1 s accumulation time. A threshold of 50 counts was required for ions to be selected for fragmentation. Parent ions and their isotopes were excluded from further selection for 1 min, with a mass tolerance of 100 ppm.

Collected mass spectra were analyzed using Analyst 1.1 with ProQuant 1.1, ProGroup 1.0.6 (Applied Biosystems), and MASCOT version 2.1 (Matrix Science, Inc, Boston, USA). A FASTA file containing all the putative ORF sequences of *D. vulgaris*, obtained from Microbes Online (http://www.microbesonline.org) (40), was used to form the theoretical search database along with the common impurities trypsin, keratin, cytochrome c, and bovine serum albumin.

The same search parameters were used in both programs; namely, trypsin was designated as the cleavage enzyme, a maximum of one missed cleavage was allowed, mass tolerances of 0.1 for mass spectrometry and 0.15 for tandem mass spectrometry were allowed, and charge states from +2 to +4 were searched. Only proteins identified by at least two unique peptides in at least one of the data sets at greater than 95% confidence by both ProQuant and MASCOT were considered.

AMT tag analysis. Whole cells lysis via bead beating and whole cell lysate tryptic digestion were performed as described previously (22,23). Separation of insoluble (i.e., membrane bound/associated) from soluble proteins in whole cell lysates was achieved with ultracentrifugation (356,000 x g, 10 min, 4^{o}C) as described elsewhere (22,23). The capillary LC system and controller, in-house manufactured mixer, capillary column selector, and sample loop for manual injections as well as separations are also as previously described (22,23).

All samples were analyzed as previously described (22,23,41). The collision induced dissociation (CID) tandem mass spectra from the LC-ion trap mass spectrometer measurements were analyzed with SEQUEST (42) using the protein sequences deduced from the *D. vulgaris* Hildenborough genome sequence (25). All samples were analyzed by a 9.4-T FTICR-MS (Bruker Daltonics) as described previously (43). Mass spectra were acquired with approximately 10^5 resolution.

High stringency constraints were used in the filtering of the data to maximize peptide identification confidence as described previously (22,23). All peptides were required to be fully or partially tryptic. In order to gauge the confidence of MS peak matching from the FTICR data to the SEQUEST^{TM} result, an algorithm to determine the quality of the match score, termed the "discriminant score", was employed (23). This scoring system computed a measure of confidence for each observation of each peptide via an extension of the approach described by Aebersold and co-workers (44). This incorporated the predicted central normalized elution time (NET) values instead of filtering out low-confidence peptides solely using observed and predicted NET values. It also utilized several SEQUEST^{TM} scoring parameters (peptide cleavage state, difference in observed and computed mass, difference in observed and predicted NET, and other indicators) to compute a confidence score for each peptide identification. This eliminated a fixed score threshold, e.g., SEQUEST ^{TM} Xcorr value of 2, to filter peptides for inclusion in a database. The advantage was that a discriminant based score is less likely to discard peptide identifications than a score based upon threshold criteria. Incorporation of NET data improved peptide identification confidence by ~10% compared to not using elution information (45). At least one high-confidence "unique" peptide (i.e., mapping to only one possible parent protein) and a total of two peptides was required for protein identification in each AMT tag analysis.

The FTICR data was processed using the PRISM data analysis system as described previously (22). Since the separation systems for both FTICR and LCQ analyses were identical, peptide confirmation was based on both the calculated (from the mass tag database) and measured mass (from the FTICR analysis) of the peptide matching to within 6 ppm and the elution times matching to within 5%.

Expression categorization of hypothetical and conserved hypothetical proteins. Each HyP or CHyP gene was identified and sorted as monocistronic or part of a polycistronic operon. This distinction allowed for inferences in the latter as to functional annotations by basing the expression responses to stresses and association with characterized genes in the same operon.

Each of these two groups of genes was then categorized solely on the basis of the microarray expression profiles. The first category was divided into those genes that exhibited "high expression", where the basal expression levels were within the top $1/8^{th}$ (12.5%) of all gene transcript levels, and those with "normal expression", where basal expression was below the 12.5%. The basal gene expression level was determined by calculating the mean Log_2 ratio of mRNA to gDNA hybridization intensities normalized as described above for all microarray experiments. With this method, the more negative Log_2 value (e.g. -14.9) indicated a smaller degree of absolute expression while a less negative number (e.g. -10.5) indicated a more highly expressed gene (9). Genes were categorized as 'not expressed' if their 2^{nd} highest observed mean Log_2 ratio of mRNA to gDNA hybridization intensity on any individual array was < -14.0, an arbitrary cutoff determined by visual inspection of probability density distributions for HyP+CHyP genes compared to annotated genes. Each of the basal expression groups was then further subdivided into the following differential gene expression categories: 1) expressed genes that lacked differential expression in response to any of the stress conditions, 2) those that showed differential expression to only one stress, 3) those that showed differential expression to multiple stresses "multiple stress response proteins", and 4) the category "not expressed" included those genes that showed no expression under any of the tested conditions.

Differential gene expression in response to either single or multiple stresses was determined by the observation of a minimum $|Log_2\ R|$ value ≥ 1.2 and a corresponding $|z\text{-score}| \geq 2$ as compared to the control condition. Because samples were analyzed at several time points after the induction of the stress condition, we established that this differential level of expression had to be observed in at least 20% of the time points to be further considered. If these parameters were met for only one of the stress conditions, then the gene was placed into the "single stress response" category. If the gene met these criteria in 2 or more of the 11 stress conditions, then they were placed in the "multiple stress response" category. In either case, the current annotation was retained for genes not meeting this criterion since every conceivable growth condition was not tested, making it premature to classify these predicted genes as "non-coding gene region".

Deletion mutagenesis. Specific HyP or CHyP genes were selected for targeted deletion based upon the microarray datasets. These genes included the monocistronic hypothetical gene DVUA0095 that is on the native plasmid of the organism and responded only to chromate stress. The second was a pair of genes (DVU0303 and DVU0304) currently annotated as an operon on the main chromosome. The expression of these HP genes was predicted and found to be directly or indirectly influenced by the Fur regulon (10), and exhibited differential expression in virtually all stress conditions tested.

Deletion cassette construction. Deletion cassettes were constructed by a method similar to the molecular bar-coding methods described for *Saccharomyces cerevisiae* (46,47). PCR primer sets were designed to amplify approximately 800 bp up- and downstream of the selected open reading frame (ORF) with unique barcode sequences between the common sequences and Km^r sequences (Supplementary Table 1). The PCR mixtures, marker exchange procedures, transformation and mutant selection procedures including Southern blot analyses were performed as previously described (10). The three segments; up- and down- stream of the gene of interest as well as the Kanamycin cassette were individually amplified by PCR and then ligated by a fourth SOEing PCR. This mutagenic cassette was then introduced into *D. vulgaris* via electroporation, where a double recombination event replaced the target gene with the drug resistance gene.

Physiological assessment of mutants. The deletion mutants were tested for growth compared to wild type *D. vulgaris* under the same conditions used in the original stress experiments from which the microarray and proteomic datasets were generated, along with growth in LS4D medium at pH 7.2 (28) as a control. Amendments and modifications for the stress conditions included the addition of 250 mM NaCl (salt stress), lowering the pH to 5.5 (acid stress), addition of 100 mM or 150 mM sodium nitrate (nitrate stress), 1 mM or 2 mM sodium nitrite (nitrite stress), and 0.2 mM, 0.4 mM, or 0.45 mM potassium chromate (chromate stress). Optical density (A_{600}) measurements were taken periodically up to 400 hours in duplicate cultures.

Sequence analysis. Protein sequence similarity was determined using FastBLAST (48) with an e-value threshold of 0.01 and an effective database size equal to $2.23x10^7$. *D. vulgaris* Hildenborough protein sequences obtained from RefSeq (release 28 March 2008) (49) were searched against the non-redundant protein (NR) database from NCBI (as of May 15, 2008) (ftp://ftp.ncbi.nih.gov/blast/). Operon predictions (50) and Clusters of Orthologous Groups (COG) (51) assignments were based on MicrobesOnline (40) data from the April 7, 2008 release (including the Nov. 2007 release of the NCBI Conserved Domain Database (52)).

Homology searches and putative functional assignments. Several publicly available *in-silico* tools were utilized in an attempt to assign a more detailed putative function to each of the HyP or CyHP genes that were expressed according to the microarray experiments. The first tool used was PSORTb version 2.0.4 (http://www.psort.org/psortb/) that was set for Gram negative organisms (53,54). This tool predicts the subcellular location of a given protein by estimating the presence and number of trans-membrane helices, the presence of signal pathway motifs, as well as other parameters. These tools were used along with the final localization estimate in conjunction with the microarray and proteomic data to give the most accurate functional annotation possible. The second method was TMHMM (55) (http://www.cbs.dtu.dk/services/TMHMM-2.0/) that predicts the number of transmembrane helices and determines if the protein is inside or outside the cytoplasm of the cell, and was used to corroborate the findings with PSORTb. Other bioinformatics tools such as those used previously (34) were attempted for several HyP and CHyP genes other than those already reported (34), but the results were ambiguous and so they were not pursued for this work.

Statistical comparisons of basal expression distributions. Probability density plots were created in the statistical software environment R (http://www.r-project.org/) with probability densities estimated by smoothing with a Gaussian kernel. Statistical tests for differences in expression level distributions were computed in R using the two-sided Mann-Whitney test with a continuity correction in the normal approximation for the p-value.

RT-PCR HyP and CHyP Basal Expression. Eight genes were selected for reverse transcription PCR (RT-PCR) in order to provide a biological verification of the microarray results. These genes were selected across the range of the average of the basal expression range with emphasis on the lower end so that if all these genes were expressed according to RT-PCR, then the assumption could be made that most if not all of the other genes were expressed as well. The positive control was the constitutively expressed *dsrC* gene (DVU2776) with an average basal expression rate of -9.7 which would place it above the top $1/8^{th}$ percentile cutoff of -11.8 so as to

be placed in the "highly expressed category". The test genes included the chromosomal genes DVU1127 (-17.1), DVU1721 (-16.6), DVU1723 (-16.6), DVU2456 (-7.6), and DVU2880 (-16.4) as well as the native plasmid genes DVUA0070 (-9.5), DVUA0144 (-15.0), and DVUA0146 (-11.4). Two negative test genes were also included. These were DVU1526 for which expression has yet to be detected via either microarrays or proteomics and DVUA0044 that was not on the microarrays as has also not been detected by proteomics. *D. vulgaris* cells were cultured and harvested as above. Total RNA was isolated immediately as described above and DNA removed using three treatments of the "DNA-free" DNAse removal kit (Applied Biosystems). To ensure the DNA was removed, PCR amplification of DVU0847 (adenylyl-sulphate reductase, α-subunit) and DVU2776 (dissimilatory sulfite reductase, γ-subunit) was performed and yielded no PCR product (data not shown). cDNA was produced using the ImProm II Reverse Transcription System A3800 (Promega). PCR reactions were then conducted for the eight test genes, the two negative controls and two positive controls (DVU2776 and DVU0847). The primers were designed to amplify as much of the gene sequence as possible without any upstream or downstream sequence (Supplementary Table 2).

Results and Discussion

Global detection of HyP and CHyP gene expression products

The sequenced genome of *D. vulgaris* shows an expected 887 HyP and 347 CHyP genes for a total of 1234 possible gene products. In general, mRNA was confidently detected for 1212 of these genes using microarrays, thus indicating actual expression of the gene (Table 1). Additionally, shotgun LC-MS/MS and AMT tag proteomic analysis further confirmed the expression of 786 proteins from HyP or CHyP genes. This represents the detection of gene expression for over 99% of the annotated HyP and over 95% of the CHyP genes with a complementary 471 (53%) of the HyP and 306 (88%) of the CHyP genes detected at the protein level (Table 1; Supplementary Table 3). In comparison, a recent report detailed the expression of 129 HyP and CHyP *D. vulgaris* genes via proteome analysis, with possible functional reassignments using *in-silico* approaches (34).

Reverse transcription PCR was conducted in order to corroborate the findings of the microarray and proteomic analysis. Eight HyP and CHyP genes were selected with five being at the lower end of the basal expression range. Two genes (DVU1526 and DVUA0044) were included as negative test cases since all data to date suggests that DVU1526 is not expressed, while DVUA0044 was not on the microarrays and was not detected by either proteomic method. The well annotated DVU2776 (*dsrC*) served as the positive control. The RT-PCR revealed that all eight of the test genes yielded bands at the predicted sizes along with the positive control, while the two negative test genes yielded no bands (Figure 1). These data were consistent with the microarray and proteomic data. Further, because 5 of the 8 test cases were among the lowest recorded basal expression rates, the results give an increased confidence that at least the large majority of genes deemed to be expressed according to the microarray and proteomics data are correct.

Neither proteomic approach gave evidence for protein synthesis from 448 transcribed genes. We sought to determine whether any bias existed that might explain this omission such as, was there a general difference in the expression of well-annotated genes versus the HyP and CHyP? Basal expression statistical profile comparisons were performed to answer this question. Overall, HyP and CHyP displayed lower gene expression levels than well characterized proteins ($p= 1.5x10^{-12}$, two-sided Mann-Whitney test; Figure 2A). This was not surprising since the core

metabolic genes required for survival are unlikely to be among the HyP and CHyP and so might be expected to be expressed at or above the average levels. However, on an individual basis, the HyP and CHyP genes have appeared amongst the most highly differentially expressed genes under particular stress conditions (2,10,26,28,35,36,38).

With respect to confident identification of HyP and CHyP by proteomics, those that were identified showed a higher gene expression level than those not identified ($p = 2.9x10^{-5}$, two-sided Mann-Whitney test; Figure 2B). Again, this is as expected, since the most abundant proteins should be preferentially identified. Additionally, a determination of any bias in the proteomic data revealed that in a comparison of HyP vs CHyP, monocistronic vs polycistronic, proteins above or below 100 amino acids in length, and highly expressed vs all others, both methods underrepresented proteins <100 amino acids in length (Figure 3A, 3B). Given this information, one possibility was that there was a lack of tryptic cut sites yielding no peptides for mass spectrometric detection. However, a query for the presence of either a lysine (K) or arginine (R) revealed that this was the case in only 3 of the 448 genes that did not have a protein confidently identified by either method (Supplementary Table 4). Curiously, 12 HyP and CHyP genes were confidently identified at the protein level but not with mRNA. Further investigation found that only 5 of these genes were on the microarrays (DVU0522, DVU1148, DVU1748, DVU2022, DVUA0050) while the remaining 7 genes (DVU0509, DVU0797, DVU0833, DVU1852, DVUA0052, DVUA0088, DVUA0145) were not (Supplemental Tables 5,6). This suggests that there may have been an issue with the microarray hybridization for these 5 genes or that transcription was below detection.

Given this information and that one of the goals of this study was to assign a functional annotation to as many of the HyP and CHyP genes as possible, a separation of the polycistronic from monocistronic genes was performed. Such classifications allow the evaluation of the hypothesis that more of the polycistronic HyP and CHyP genes would be expressed compared to the monocistronic ones. We assumed a greater likelihood for a more accurate functional annotation if the gene were in a predicted operon and displayed similar stress response patterns to more confidently annotated genes in that same operon. Hence, monocistronic and polycistronic genes were treated separately for the remainder of the study.

Categorization of differential expression patterns under stress conditions

The first step in characterization of the 882 HyP and 330 CHyP genes that were transcribed was to categorize them according to observed expression patterns under one or more of the cultivation/stress conditions tested. Collectively, 45% of all of the expressed HyP and CHyP genes were found to be polycistronic with the remaining 56% being monocistronic (Table 1). However, it is interesting that despite a greater number of monocistronic genes (687) compared to polycistronic genes (548), the distribution amongst the expression categories was quite similar (Figure 4). In each case, only 0.3% of the monocistronic and 0.2% of the polycistronic genes were highly expressed with no observed differential transcription. Highly expressed genes, differentially expressed or not, comprised 7.3% and 6.2% of these groups, respectively. Similarly, 10.9% of the monocistronic and 11.7% of the polycistronic genes that were not highly expressed responded to a single stress, while a considerably higher percentage (73.9% and 71.5%, respectively) responded to multiple stress conditions. This similarity in the categorizational proportions between the two groups of genes continued in the “expressed” (6.4% and 8.4%, respectively) and not expressed (1.5% and 2.2%, respectively) categories. For each of the polycistronic genes not observed to be expressed (12 cases), the other genes in the

operon were expressed under some condition. Therefore, these were not cases of an operon not being expressed, but rather particular genes not showing expression. Reasons for these patterns are not clear. In fact, a preliminary assumption was that more of the monocistronic genes would not be expressed than genes within operons, since the latter would be more likely to be co-expressed with the rest of the operon. However, this was not the case since the percentage of genes in each category was similar as detailed above.

Expression profiling and putative functional assignment to monocistronic genes

Monocistronic genes are arguably more difficult to functionally annotate because the reference point of neighboring genes is absent. However, the 81 monocistronic genes that responded to a single stress did give clues as to function, 65 encoded a HyP with the remaining 16 encoding a CHyP (Supplementary Table 5). In order to demonstrate the observed expression profiles, the stress responses for a randomly selected set of genes responding to a single stress condition are shown (Figure 5). It is interesting that even among these essentially unknown genes, several are found to be responsive to the same stress. Several exhibited differential expression when transcripts in stationary phase were compared to those in exponential cells or when acid treated culture transcripts were compared to base treated. In contrast, there were cases where multiple hypothetical genes responded specifically to one stress, as in the case of chromate exposure. Both DVUA0095 (Figure 5B) and DVU1338 were upregulated by chromate exposure while DVU2436 was down-regulated (Supplemental Table 5). No polycistronic genes were solely influenced by chromate. Based on these findings and the observation that *D. vulgaris* lacking pDV1 is less tolerant of chromate exposure (M. Fields, pers. comm.), DVUA0095, located on pDV1, was targeted for deletion to ascertain the cellular response to chromate in this mutant.

For the purposes of functional annotation, each of the Hyp genes has been renamed to reflect the stress response influencing its expression along with any other *in-silico* features as determined by the use of COGs, TMHMM and PSORTb (Supplementary Tables 5,6). For example, DVUA0095 was found to be up-regulated upon exposure to chromate. Results from *in-silico* analyses predicted that it possesses three transmembrane helices but no signal peptide motifs, with a final score of 9.46 (out of 10.0) that it is associated with the cytoplasmic membrane. Given the lack of a signal peptide or assigned COG, we infer that the protein resides in the cytoplasmic membrane with the bulk of the protein facing the cytoplasm. Hence, this protein has been re-annotated to be a “chromate-induced, cytoplasmic membrane protein”. For others where no such structural features were predicted, the genes were simply renamed e.g.; “acid-induced protein” or “heat-repressed protein”. The remaining monocistronic genes that did not display a differential response to any of the stresses, have simply been renamed as e.g.; “expressed protein in *D. vulgaris*” or “expressed cytoplasmic membrane protein in *D. vulgaris*” (e.g. DVU1006; Supplementary Table 5). The remaining monocistronic HyP and CHyP genes, representing 80% of those expressed, were differentially regulated in multiple stress conditions. Genes that responded by increasing or decreasing in three or more conditions were predicted to encode “general stress response proteins.” In a similar manner, a gene that responded to only two stresses was renamed by the observed responses such as a “NaCl induced, cold repressed protein” (DVU3354), or a “cold and co-culture induced protein” (DVU3130) (Supplementary Table 5). The 8 genes that were not expressed under any of the conditions tested have been left with their original annotation, as opposed to being designated as a “non-coding region,” since not all conceivable cultivation or stress conditions have been tested to date.

Expression profiling and putative functional assignment of polycistronic genes

The polycistronic HyP and CHyP genes often have a reference point for their plausible biochemical function based on their location within operons that include ORFs with more characterized, orthologous genes in other bacteria. One of the main criteria used to assign function to the HyP or CHyP genes was the similarity of the differential expression pattern of the gene to that of other genes within the operon. An additional criterion was the degree of nucleotide overlap with other genes within the operon that could suggest transcriptional coupling. Good examples of such scenarios were found in areas of the genome apparently containing temperate bacteriophages or their remnants, such as DVU2488-DVU2729 containing several predicted operons. Some temperate phages may be induced by catastrophic stress conditions or as cells enter the stationary phase of growth. In fact, stresses from high heat and the stationary phase of growth resulted in the differential expression of the greatest number of polycistronic genes that are likely to encode phage functions. One example is the apparent increased expression of a seven gene operon likely to be involved in temperate phage activity during stationary phase versus exponential growth (Figure 6A). It is prudent to note that other changes in culture conditions usually coincide with these events, such as sulfide and acetate accumulation with a concomitant change in pH and a decrease in the specific growth rate.

Another example of HyP and CHyP genes within temperate phage operons was the four-gene operon of DVU0192 (adenine specific DNA methyltransferase), DVU0194 (terminase) along with the HyP genes DVU0193 and DVU0195. In this case, all four genes responded coordinately to the onset of stationary phase compared to exponential growth (Figure 6B), suggesting that the two HyP genes could be involved in phage DNA metabolism.

Other cases were not as straightforward, such as with the predicted operon of DVU1639-DVU1642 (Supplementary Table 6) that contains two HyP and two CHyP genes. DVU1639 showed increased expression only to stationary phase whereas DVU1640 was not expressed. DVU1641 and DVU1642 were both up-regulated in multiple stresses including showing Fur-regulation. These results question the validity of the operon assignment. The three transcribed genes were renamed based upon the microarray stress response data. Hence, DVU1639 was reannotated as a “stationary phase induced protein”, whereas DVU1641 and DVU1642 were reannotated as a “Fur influenced, multiple stress induced protein” and a “multiple stress induced outer membrane protein”, respectively. The original annotation for DVU1640 remained.

Fur influenced regulation of HP and CHP genes.

D. vulgaris possesses three Fur regulator paralogs DVU0942 (Fur), DVU3095 (PerR), and DVU1340 (Zur) (56). While little information is available on the latter in *D. vulgaris*, the global regulation roles of Fur and PerR have been explored in more depth. Gene deletions within *D. vulgaris* are available for the putative global regulators Fur (10) and PerR (unpublished), and transcriptional analyses in a few stress conditions have been performed. In a number of bacteria, the Fur system regulates the uptake of ferric iron (57,58). Although Fur has also been shown to control the synthesis of specialized Fe(III) chelators known as siderophores (57,59,60), there is no evidence for siderophore production in *D. vulgaris*. In contrast, Fur is suggested to play a regulatory role in oxidative stress, motility, virulence, and acid tolerance (57,58). PerR, the second of the potential global regulators, has been best studied in *Bacillus subtilis* (61,62). Experimental evidence supported a role for the PerR system in the cellular response of *D. vulgaris* to oxidative stresses such as peroxide or metal ion limitation through increased

expression of rubrerythrin and rubredoxin genes (10). Bender and co-workers (2007) reported that the HyP gene DVU2681 was reported to have the greatest increase in transcription of all genes in the Fur deletion mutant, while 6 of the 21 genes showing strong increases in transcription in the absence of Fur were HyP or CHyP genes.

In the current work, a number of HyP and CHyP genes appeared to be either differentially transcribed in Fur or PerR deletion mutants compared to the wild type strain, or displayed stronger transcriptional responses in the deletion mutant. Expression changes occurring in the deletion mutants were inferred to result from the altered genetic background, and, in particular, when the PerR deletion mutant was exposed to a 0.1% O_2 stress or the Fur deletion mutant was exposed to high salt or high nitrate. A small proportion of genes apparently were influenced by both Fur and Per (Table 2). Hence in these cases the prefix of "Fur-influenced", "Per-influenced", or "Fur- and Per-influenced" was added to the annotation where appropriate (Supplementary Tables 5, 6). While such numbers of proteins are probably not directly linked to either of these global regulatory systems, it is conceivable that indirect, cascading regulators could affect this number of genes.

Validation of putative assignments with targeted deletion mutagenesis

In order to test the relevance of the stress responses recorded via microarrays and proteomics, as well as the inferred functional annotations applied to the HyP and CHyP genes in *D. vulgaris* during this study, two targeted deletions were constructed. A two-gene operon, DVU0303-DVU0304, predicted to be part of the Fur regulon (56), showed altered transcription rates in all conditions tested (Supplementary Table 6). The second targeted deletion was the monocistronic gene DVUA0095 that increased in expression only upon exposure to chromate (Figure 5B; Supplemental Table 5). Both deletion mutants and the wild type grew similarly in the control unamended medium containing 30 mM sodium lactate and 60 mM sodium sulfate at pH 7.2, pH 5.5 and when amended with 250 mM NaCl (Figures 7A-7C). These results were as predicted for ΔDVUA0095 since it responded only to chromium exposure. However the growth results of Δ(DVU0303-DVU0304) was unexpected since both genes originally increased in transcription at pH 5.5 and in 250 mM NaCl. However, such a lack of correlation between gene expression and cellular fitness in an imposed treatment is not uncommon (46).

Other stress conditions included amendments of 100 mM $NaNO_3$, or $NaNO_2$ at 1 mM or 2 mM. Expression of DVU0303 in wild type cultures increased with $NaNO_3$ or $NaNO_2$ (35,36). A slightly increased sensitivity of Δ(DVU0303-DVU0304) with these stressors (Figures 7D-7F) was observed. The interpretation of changes in mutant growth rate and extent may suggest a functional role in tolerance to the treatment or a general metabolic perturbation that impedes the production of cell material. Surprisingly, the ΔDVUA0095 mutant also grew more poorly relative to the wild type when exposed to the various nitrogen species (Figures 7E, 7F). Reasons for this phenotype are not clear, but could include oxidative stress and subsequent protein or DNA damage due to the unstable nature of nitrite.

The final stress conditions used to test the deletion mutants were three concentrations of K_2CrO_4 at 0.2 mM, 0.4 mM, and 0.45 mM. When wild type cultures were challenged with 0.45 mM K_2CrO_4, the transcriptional analysis showed large increases in expression of DVUA0095 as well as for DVU0303-DVU0304. At the lower concentration of 0.2 mM K_2CrO_4, there were no apparent effects on the growth of any of the strains (Figure 7G). At 0.4 mM K_2CrO_4, ΔDVUA0095 exhibited a lag phase of 75 h, compared to 25-30 h in Δ(DVU0303-DVU0304) and the wild type (Figure 7H), suggesting that the removal of this gene interfered with the response

to this level of K_2CrO_4. At 0.45 mM K_2CrO_4, both ΔDVUA0095 and Δ(DVU0303-DVU0304) showed extended lag phases of 190 and 175 hours, respectively, approximately twice that of the wild type (Figure 7I).

Conclusions

The main purpose of this study was to validate the expression of *D. vulgaris* genes annotated as HyP and CHyP and to infer additional functions when possible. Overall, 98% of the HyP and CHyP genes were found to be transcribed via microarrays with 63% also being translated. Among these, many displayed specific transcriptional responses to single stresses whereas others showed responses to multiple treatments. Some of these genes were also shown to be influenced by, or in regulons of, the global regulatory systems of Fur and/or PerR. The fact that these genes actually produce proteins increases the possibility that they may play some role in the responses to environmental perturbations.

Assessment of the ΔDVUA0095 and Δ(DVU0303-DVU0304) mutants confirmed a likely role for DVUA0095 in cellular responses to chromate as inferred from the microarray results. However, further complexity exists as revealed by some surprising phenotypes of the mutants as described above. Ongoing and future work is anticipated to include a detailed analysis of many HyP and CHyP genes, particularly those that displayed responses to only a single stress. This will include gene tagging with eventual protein complex elucidation, assessment of gene deletions, and additional stress treatments. Through these efforts, a better understanding of the physiological and metabolic aspects of bacteria with sequenced genomes such as *D. vulgaris* will be achieved. This type of methodical analysis of unknown genes may well be useful as a means to derive more meaningful functional annotations in other organisms where a few or many RNA microarray and/or proteomic datasets exist. Reanalysis of these data can range from simply comparing HyP and CHyP transcript or proteome abundances in different culturing conditions to the more discreet stress conditions such as were used to generate the data for the present work. In addition, decreasing the lists of HyP and CHyP genes will improve annotations through inter-organismal comparisons.

Acknowledgements

We thank Amy Kucken (UMC) for performing the RT-PCR work and Angela Norbeck (PNNL) for assistance with revision of the proteomics database. This work was part of the Environmental Stress Pathways Project and the Virtual Institute for Microbial Stress and Survival (http://vimss.lbl.gov), supported by the U.S. Department of Energy, Office of Science, Office of Biological and Environmental Research, Genomics Program: GTL through contract DE-AC02-05CH11231 between Lawrence Berkeley National Laboratory and the U.S. Department of Energy. The AMT tag proteomics portion of this research was performed using the Environmental Molecular Sciences Laboratory (EMSL), a national scientific user facility sponsored by the Department of Energy's Office of Biological and Environmental Research located at Pacific Northwest National Laboratory.

Reference:

1. Elias, D.A., Yang, F., Mottaz, H.M., Belieav, A.S. and Lipton, M.S. (2007) Enrichment of functional redox reactive proteins and identification by mass spectrometry results in several terminal Fe(III)-reducing candidate proteins in *Shewanella oneidensis* MR-1. *J. Microbiol. Meth.*, **68**, 367-375.
2. Clark, M.E., He, Q., He, Z., Huang, K.H., Alm, E.J., Wan, X.F., Hazen, T.C., Arkin, A.P., Wall, J.D., Zhou, J.Z. *et al.* (2006) Temporal transcriptomic analysis as *Desulfovibrio vulgaris* Hildenborough transitions into stationary phase during electron donor depletion. *Appl. Environ. Microbiol.*, **72**, 5578-5588.
3. Zhang, W., Gritsenko, M.A., Moore, R.J., Culley, D.E., Nie, L., Petritis, K., Strittmatter, E.F., II, D.G.C., Smith, R.D. and Brockman, F.J. (2006) A proteomic view of *Desulfovibrio vulgaris* metabolism as determined by liquid chromatography coupled with tandem mass spectrometry. *Proteomics*, **6**, 4286-4299.
4. Beliaev, A.S., Thompson, D.K., Fields, M.W., Wu, L., Lies, D.P., Nealson, K.H. and Zhou, J. (2002) Microarray transcription profiling of a *Shewanella oneidensis etrA* mutant. *J. Bacteriol.*, **184**, 4612-4616.
5. Beliaev, A.S., Thompson, D.K., Khare, T., Lim, H., Brandt, C.C., Li, G., Murray, A.E., Heidelberg, J.F., Giometti, C.S., Yates, J., 3rd *et al.* (2002) Gene and protein expression profiles of *Shewanella oneidensis* during anaerobic growth with different electron acceptors. *Omics*, **6**, 39-60.
6. Drepper, T., Eggert, T., Circolone, F., Heck, A., Kraub, U., Guterl, J., Wendorff, M., Losi, A., Gartner, W. and Jaeger, K. (2007) Reporter proteins for *in vivo* fluorescence without oxygen. *Nat. Biotechnol.*, **25**, 443-445.
7. Regoes, A. and Hehl, A.B. (2005) SNAP-tag™ mediated live cell labeling as an alternative to GFP in anaerobic organisms. *Biotech*, **39**, 809-812.
8. Butland, G., Peregrín-Alvarez, J.M., Li, J., Yang, W., Yang, X., Canadien, V., Starostine, A., Richards, D., Beattie, B., Krogan, N. *et al.* (2005) Interaction network containing conserved and essential protein complexes in *Escherichia coli*. *Nat*, **433**, 531-537.
9. Wall, J.D., Arkin, A.P., Balci, N.C. and Rapp-Giles, B.J. (2007) In Dahl, C. and Friedrich, C. G. (eds.), *Microbial sulfur metabolism*. Springer-Verlag, Berlin, pp. 1-11.
10. Bender, K.S., Yen, H.B., Hemme, C.L., Yang, Z., He, Z., He, Q., Zhou, J., Huang, K.H., Alm, E.J., Hazen, T.C. *et al.* (2007) Analysis of a ferric uptake regulator (Fur) mutant of *Desulfovibrio vulgaris* Hildenborough. *Appl. Environ. Microbiol.*, **73**, 5389-5400.
11. Beliaev, A.S. and Saffarini, D.A. (1998) *Shewanella putrefaciens mtrB* encodes an outer membrane protein required for Fe(III) and Mn(IV) reduction. *J. Bacteriol.*, **180**, 6292-6297.
12. Beliaev, A.S., Saffarini, D.A., McLaughlin, J.L. and Hunnicutt, D. (2001) MtrC, an outer membrane decahaem c cytochrome required for metal reduction in *Shewanella putrefaciens* MR-1. *Molec. Microbiol.*, **39**, 722-730.
13. Kolker, E., Makarova, K.S., Shabalina, S., Picone, A.F., Purvine, S., Holzman, T., Cherny, T., Armbruster, D., Jr, R.S.M., Kolesov, G. *et al.* (2004) Identification and functional analysis of hypothetical genes expressed in *Haemophilus influenzae*. *Nucl. Acids Res.*, **32**, 2353-2361.
14. Delcher, A.L., Harmon, D., Kasif, S., White, O. and Salzberg, S.L. (1999) Improved microbial gene identification with GLIMMER. *Nucl. Acids Res.*, **27**, 4636-4641.

15. Salzberg, S.L., Delcher, A.L., Kasif, S. and White, O. (1998) Microbial gene identification using interpolated Markov models. *Nucl. Acids Res.*, **26**, 544-548.
16. Doerks, T., vonMering, C. and Bork, P. (2004) Functional clues for hypothetical proteins based on genomic context analysis in prokaryotes. *Nucl. Acids Res.*, **32**, 6321-6326.
17. Huynen, M., Snel, B., III, W.L. and Bork, P. (2000) Predicting protein function by genomic context: quantitative evaluation and qualitative inferences. *Gen. Res.*, **10**, 1204-1210.
18. Lu, P., Szafron, D., Greiner, R., Wishart, D.S., Fyshe, A., Pearcy, B., Poulin, B., Eisner, R., Ngo, D. and Lamb, N. (2005) PA-GOSUB: a searchable database of model organism protein sequences with their predicted gene ontology molecular function and subcellular localization. *Nucl. Acids Res.*, **33**, D147-D153.
19. vonMering, C., Jensen, L.J., Snel, B., Hooper, S.D., Krupp, M., Foglierini, M., Jouffre, N., Huynen, M.A. and Bork, P. (2005) STRING: known and predicted protein–protein associations, integrated and transferred across organisms. *Nucl. Acids Res.*, **33**, D433-D437.
20. Enault, F., Suhre, K., Poirot, O., Abergel, C. and Claverie, J. (2003) Phydbac (phylogenomic display of bacterial genes): an interactive resource for the annotation of bacterial genomes. *Nucl. Acids Res.*, **31**, 3720-3722.
21. Sjölander, K. (2004) Phylogenomic inference of protein molecular function: advances and challenges. *Bioinformatics*, **20**, 170-179.
22. Elias, D.A., Monroe, M.E., Marshall, M.J., Romine, M.F., Beliaev, A.S., Fredrickson, J.K., Anderson, G.A., Smith, R.D. and Lipton, M.S. (2005) Global detection and characterization of hypothetical proteins in *Shewanella oneidensis* MR-1 using LC-MS based proteomics. *Proteomics*, **5**, 3120-3130.
23. Elias, D.A., Monroe, M.E., Smith, R.D., Fredrickson, J.K. and Lipton, M.S. (2006) Confirmation of the expression of a large set of conserved hypothetical proteins in *Shewanella oneidensis* MR-1. *J. Microbiol. Meth.*, **66**, 223-233.
24. Kolker, E., Picone, A.F., Galperin, M.Y., Romine, M.F., Higdon, R., Makarova, K.S., Kolker, N., Anderson, G.A., Qiu, X., Auberry, K.J. *et al.* (2005) Global profiling of *Shewanella oneidensis* MR-1: expression of hypothetical genes and improved functional annotations. *Proc. Natl. Acad. Sci. USA*, **102**, 2099-2104.
25. Heidelberg, J.F., Seshadri, R., Haveman, S.A., Hemme, C.L., Paulsen, I.T., Kolonay, J.F., Eisen, J.A., Ward, N., Methe, B., Brinkac, L.M. *et al.* (2004) The genome sequence of the anaerobic, sulfate-reducing bacterium *Desulfovibrio vulgaris* Hildenborough. *Nat. Biotechnol.*, **22**, 554-559.
26. Chhabra, S.R., He, Q., Huang, K.H., Gaucher, S.P., Alm, E.J., He, Z., Hadi, M.Z., Hazen, T.C., Wall, J.D., Zhou, J. *et al.* (2006) Global analysis of heat shock response in *Desulfovibrio vulgaris* Hildenborough. *J. Bacteriol.*, **188**, 1817-1828.
27. Fournier, M., Aubert, C., Dermoun, Z., Durand, M., Moinier, D. and Dolla, A. (2006) Response of the anaerobe *Desulfovibrio vulgaris* Hildenborough to oxidative conditions: proteome and transcript analysis. *Biochimica*, **88**, 85-94.
28. Mukhopadhyay, A., He, Z., Alm, E.J., Arkin, A.P., Baidoo, E.E., Borglin, S.C., Chen, W., Hazen, T.C., He, Q., Holman, H. *et al.* (2006) Salt stress in *Desulfovibrio vulgaris* Hildenborough: an integrated genomics approach. *J. Bacteriol.*, **188**, 4068-4078.

29. Zhang, W., Culley, D.E., Scholten, J.C.M., Hogan, M., Vitiritti, L. and Brockman, F.J. (2006) Global transcriptomic analysis of *Desulfovibrio vulgaris* on different electron donors. *Anton. van Leeuwen.*, **89**, 221-237.
30. Postgate, J.R. (1984) *The sulphate-reducing bacteria*. 2nd ed. Cambridge University Press, Cambridge, MA, USA.
31. Elias, D.A., Krumholz, L.R., Wong, D., Long, P.E. and Suflita, J.M. (2003) Characterization of microbial activities and U reduction in a shallow aquifer contaminated by uranium mill tailings. *Microb. Ecol.*, **46**, 83-91.
32. Elias, D.A., Suflita, J.M., McInerney, M.J. and Krumholz, L.R. (2004) Periplasmic cytochrome c_3 of *Desulfovibrio vulgaris* is directly involved in H_2-mediated metal but not sulfate reduction. *Appl. Environ. Microbiol.*, **70**, 413-420.
33. Lovley, D.R. and Phillips, E.J.P. (1994) Reduction of chromate by *Desulfovibrio vulgaris* and its c_3 cytochrome. *Appl. Environ. Microbiol.*, **60**, 726-728.
34. Zhang, W., Culley, D.E., Gritsenko, M.A., Moore, R.J., Nie, L., Scholten, J.C.M., Petritis, K., Strittmatter, E.F., II, D.G.C., Smith, R.D. *et al.* (2006) LC–MS/MS based proteomic analysis and functional inference of hypothetical proteins in *Desulfovibrio vulgaris*. *Biochem. Biophys. Res. Commun.*, **349**, 1412-1419.
35. Redding, A.M., Mukhopadhyay, A., Joyner, D.C., Hazen, T.C. and Keasling, J.D. (2006) Study of nitrate stress in *Desulfovibrio vulgaris* Hildenborough using iTRAQ proteomics. *Brief. Funct. Gen. Prot.*, 1-11.
36. He, Q., Huang, K.H., He, Z., Alm, E.J., Fields, M.W., Hazen, T.C., Arkin, A.P., Wall, J.D. and Zhou, J. (2006) Energetic consequences of nitrite stress in *Desulfovibrio vulgaris* Hildenborough inferred from global transcriptional analysis. *Appl. Environ. Microbiol.*, **72**, 4370-4381.
37. Stolyar, S., He, Q., Joachimiak, M.P., He, Z., Yang, Z.K., Borglin, S.E., Joyner, D.C., Huang, K., Alm, E., Hazen, T.C. *et al.* (2007) Response of *Desulfovibrio vulgaris* to alkaline stress. *J. Bacteriol.*, **189**, 8944-8952.
38. Mukhopadhyay, A., Redding, A.M., Joachimiak, M.P., Arkin, A.P., Borglin, S.E., Dehal, P.S., Chakraborty, R., Geller, J.T., Hazen, T.C., He, Q. *et al.* (2007) Cell-wide responses to low-oxygen exposure in *Desulfovibrio vulgaris* Hildenborough. *J. Bacteriol.*, **189**, 5996-6010.
39. He, Z., Wu, L., Fields, M.W. and Zhou, J. (2005) Use of microarrays with different probe sizes for monitoring gene expression. *Appl. Environ. Microbiol.*, **71**, 5154-5162.
40. Alm, E.J., Huang, K.H., Price, M.N., Koche, R.P., Keller, K., Dubchak, I.L. and Arkin, A.P. (2005) The MicrobesOnline web site for comparative genomics. *Gen. Res.*, **15**, 1015-1022.
41. 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. *et al.* (2002) Global analysis of the *Deinococcus radiodurans* proteome by using accurate mass tags. *Proc Natl Acad Sci U S A*, **99**, 11049-11054.
42. Eng, J.K., McCormack, A.L. and III, J.R.Y. (1994) An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database *J Amer Soc Mass Spec*, **5**, 976-989.
43. Harkewicz, R., Belov, M.E., Anderson, G.A., Pasa-Tolic, L., Masselon, C.D., Prior, D.C., Udseth, H.R. and Smith, R.D. (2002) ESI-FTICR mass spectrometry employing data-

dependent external ion selection and accumulation *J. Amer. Soc. Mass Spec.*, **13**, 144-154.

44. Keller, A., Nesvizhskii, A.I., Kolker, E. and Aebersold, R. (2002) Empirical Statistical Model To Estimate the Accuracy of Peptide Identifications Made by MS/MS and Database Search. *Anal Chem*, **74**, 5383 - 5392.
45. Strittmatter, E.F., Kangas, L.J., Petritis, K., Mottaz, H.M., Anderson, G.A., Shen, Y., Jacobs, J.M., II, D.G.C. and Smith, R.D. (2004) Application of Peptide LC Retention Time Information in a Discriminant Function for Peptide Identification by Tandem Mass Spectrometry. *J. Prot. Res.*, **3**, 760 - 769.
46. Giaever, G., Chu, A.M., Ni, L., Connelly, C., Riles, L., Veronneau, S., Dow, S., Lucau-Danila, A., Anderson, K., Andre, B. *et al.* (2002) Functional profiling of the *Saccharomyces cerevisiae* genome. *Nat*, **418**, 387-391.
47. Shoemaker, D., Lashkari, D.A., Morris, D., Mittmann, M. and Davis., R.W. (1996) Quantitative phenotypic analysis of yeast deletion mutants using a highly parallel molecular bar-coding strategy. *Nat. Gen.*, **14**, 450-456.
48. Price, M.N., Dehal, P.S. and Arkin, A.P. (2008) FastBLAST: homology relationships for millions of proteins. *PLoS One*, **3**, e3589-.
49. Pruitt, K.D. and Maglott, D.R. (2001) RefSeq and Locus Link: NCBI gene-centered resources. *Nucleic Acids Res*, **29**, 137-140.
50. Price, M.N., Huang, K.H., Alm, E.J. and Arkin, A.P. (2005) A novel method for accurate operon predictions in all sequenced prokaryotes. *Nucleic Acids Res*, **33**, 880-892.
51. Tatusov, R.L., Fedorova, N.D., Jackson, J.D., Jacobs, A.R., Kiryutin, B., Koonin, E.V., Krylov, D.M., Mazumder, R., Mekhedov, S.L., Nikolskaya, A.N. *et al.* (2003) The COG database: an updated version includes eukaryotes. *BMC Bioinformatics*, **4**.
52. Marchler-Bauer, A., Panchenko, A.R., Shoemaker, B.A., Thiessen, P.A., Geer, L.Y. and Bryant, S.H. (2002) CDD: A database of conserved domain alignments with links to domain three-dimensional structure. *Nucleic Acids Res*, **30**, 281-283.
53. Gardy, J.L., Laird, M.R., Chen, F., Rey, S., Walsh, C.J., Ester, M. and Brinkman, F.S.L. (2005) PSORTb v.2.0: Expanded prediction of bacterial protein subcellular localization and insights gained from comparative proteome analysis. *Bioinformatics*, **21**, 617-623.
54. Gardy, J.L., Spencer, C., Wang, K., Ester, M., Tusnady, G.E., Simon, I., Hua, S., deFays, K., Lambert, C., Nakai, K. *et al.* (2003) PSORT-B: improving protein subcellular localization prediction for Gram-negative bacteria. *Nucl. Acids Res.*, **31**, 3613-3617.
55. Krogh, A., Larsson, B., Heijne, G.v. and Sonnhammer, E.L.L. (2001) Predicting transmembrane protein topology with a hidden markov model: application to complete genomes. *J. Molec. Biol.*, **305**, 567-580.
56. Rodionov, D.A., Dubchak, I., Arkin, A.P., Alm, E. and Gelfand, M.S. (2004) Reconstruction of regulatory and metabolic pathways in metal-reducing δ-proteobacteria. *Gen. Biol.*, **5**, R90.91-R90.27.
57. Escolar, L.A., Perez-Martin, J. and Lorenzo, V.D. (1999) Opening the iron box: transcriptional metalloregulation by the Fur protein. *J. Bacteriol.*, **181**, 6223-6229.
58. Hantke, K. (2001) Iron and metal regulation in bacteria. *Curr. Opin. Microbiol.*, **4**, 172-177.
59. Rowland, B.M. and Taber, H.W. (1996) Duplicate isochorismate synthase genes of *Bacillus subtilis*: regulation and involvement in the biosyntheses of menaquinone and 2,3-dihydroxybenzoate. *Molec. Microbiol.*, **178**, 854-861.

60. Schneider, R. and Hantke, K. (1993) Iron-hydroxamate uptake systems in *Bacillus subtilis*: identification of a lipoprotein as part of a binding proteindependent transport system. *Molec. Microbiol.*, **8**, 111-121.
61. Bsat, N., Herbig, A., Casillas-Martınez, L., Setlow, P. and Helmann, J.D. (1998) *Bacillus subtilis* contains multiple Fur homologues; identification of the iron uptake (Fur) and peroxide regulon (PerR) repressors. *Molec. Microbiol.*, **29**, 189-198.
62. Lee, J.W. and Helmann, J.D. (2006) The PerR transcription factor senses H_2O_2 by metal-catalysed histidine oxidation. *Nat*, **440**, 363-367.
63. Heidelberg, J.F., Paulsen, I.T., Nelson, K.E., Gaidos, E.J., Nelson, W.C., Read, T.D., Eisen, J.A., Seshadri, R., Ward, N., Methe, B. *et al.* (2002) Genome sequence of the dissimilatory metal ion-reducing bacterium *Shewanella oneidensis*. *Nat. Biotechnol.*, **20**, 1118-1123.
64. Luo, Q., Hixson, K.K., Callister, S.J., Lipton, M.S., Morris, B.E.L. and Krumholz, L.R. (2007) Proteome analysis of *Desulfovibrio desulfuricans* G20 mutants using the accurate mass and time (AMT) tag approach. *J. Prot. Res.*, **6**, 3042-3053.
65. Romine, M.F., Elias, D.A., Monroe, M.E., Auberry, K., Fang, R., Fredrickson, J.K., Anderson, G.A., Smith, R.D. and Lipton, M.S. (2004) Validation of *Shewanella oneidensis* MR-1 Small Proteins by AMT Tag-Based Proteome Analysis. *OMICS*, **8**, 239-254.

Figure 1: RT-PCR confirmation of microarray expression data. Agarose gel showing the results of RT-PCR reactions in order to confirm the expression or lack of expression of various HyP and CHyP genes in *D. vulgaris*. Circled areas indicate the expected molecular weight band in each case. Eight such genes were selected over a range of average basal expression rates (expression cateogory in brackets) while two genes that showed no expression to date were also selected (broken circles). In both cases, PCR with gDNA ensured that the primers performed as expected. The well- annotated DsrC (DVU2776) served as the cDNA and gDNA control.

Figure 2: Comparison of the average basal gene expression levels as quantified by microarray analysis. (A) A comparison of the 1212 expressed HyP and CHyP genes (solid line) versus the 2278 better annotated genes (broken line) showed that the former displayed significantly lower expression levels overall; $p= 1.5x10^{-12}$, two-sided Mann-Whitney test. (B) The 774 HyP and CHyP gene expression products detected at the protein level (broken line) were significantly more abundant than the 438 proteins not confidently identified by either proteomics method (solid line); $p= 3.0x10^{-5}$, two-sided Mann-Whitney test.

Figure 3: Comparison of HyP and CHyP detection at the protein level using the (A) AMT tag and (B) shotgun LC-MS/MS approaches. The bars are additive between the number of proteins detected (solid bar) and those not detected (open bar). A comparison of each of the subgroups with both methods showed a bias against the detection of proteins under 100 amino acids in length.

Figure 4: Pie charts showing the stress response distribution of all (A) monocistronic HP and CHP genes (680) and (B) HP and CHP genes (557) predicted to be in polycistronic operons. In each case the genes that were categorized as having a high basal expression are grey while the rest are black. Those not showing expression are not colored. Genes displaying no stress response are solid colors while a single stress response is denoted by a striped pie slice and multiple stress responses are checkered. Those displaying a single stress response accounted for ~10% of the genes while >80% were differentially expressed in more than one stress.

Figure 5: Microarray expression profiling of several monocistronic (A) CHP and (B) HP genes from various stresses that displayed differential expression ($Log_2R \geq 1.2$) in a single stress. Stat vs Exp= stationary phase compared to exponential growth while all others are the listed stress condition compared to normal growth on lactate sulfate medium.

Figure 6: Microarray expression profiling of hypothetical genes within operons allows for a putative functional assignment by using the profile and gene association. The condition shown is stationary phase vs exponential growth. (A) Up-regulation of a 7 gene operon containing DVU2710 (♦; prophage protein), DVU2711 (■; major head subunit), DVU2712 (▲; hypothetical protein), DVU2713 (□; prophage protein), DVU2714 (○; prophage protein), DVU2715 (◊; conserved hypothetical), and DVU2716 (Δ; tail sheath protein). (B) Up-regulation of a 4 gene operon of DVU0192 (♦; adenine specific DNA methyltransferases) and DVU0194 (▲; terminase) that are well-conserved while DVU0193 (■) and DVU0195 (●) were annotated as hypothetical proteins.

Figure 7: Growth curves of the two targeted deletion mutants (genes DVU0303-0304 that were differentially expressed in 13 and 9 stresses, respectively (o), and a deletion of DVUA0095 that was only up-regulated in the presence of chromate (Δ) ($Log_2R = 4.2$; see also Figure 2A) and wild type *D. vulgaris* (♦) under various stress conditions. (A) baseline lactate/sulfate growth, (B) pH 5.5, (C) 250 mM NaCl, (D) 100 mM $NaNO_3$, (E) 1 mM $NaNO_2$, (F) 2 mM $NaNO_2$, (G) 0.2 mM K_2CrO_4, (H) 0.4 mM K_2CrO_4, (I) 0.45 mM K_2CrO_4. Each stress experiment was conducted twice.

Table 1: HP and CHP genes with evidence of expressions

Current Annotation	Number of Possible Genes[A]	Transcript Identified[B]	Protein Identified[C]
Polycistronic			
Hypothetical	327	324	227
Conserved Hypothetical	220	211	194
Monocistronic			
Hypothetical	560	557	247
Conserved Hypothetical	127	120	113

A. Computational identification of putative open reading frames were as previously described (25,63).
B. Transcript evidence obtained from microarray experiments reported by VIMSS/ESPP efforts (2,10,26,28,36,38).
C. Proteins identified by shotgun LC-MS/MS and/or AMT tag proteomics as previously described (35,64,65).

Table 2: Apparent HyP and CHyP Influence by the Global Regulators Fur and Per

	Fur	Per	Fur & Per
Polycistronic (535 total transcripts detected)			
Highly expressed, single stress response	0	0	0
Highly expressed, multiple stress response	8	3	1
Single stress response	0	2	0
Multiple stress response	108	23	9
Monocistronic (677 total transcripts detected)			
Highly expressed, single stress response	0	0	0
Highly expressed, multiple stress response	7	7	2
Single stress response	3	3	2
Multiple stress response	152	33	23

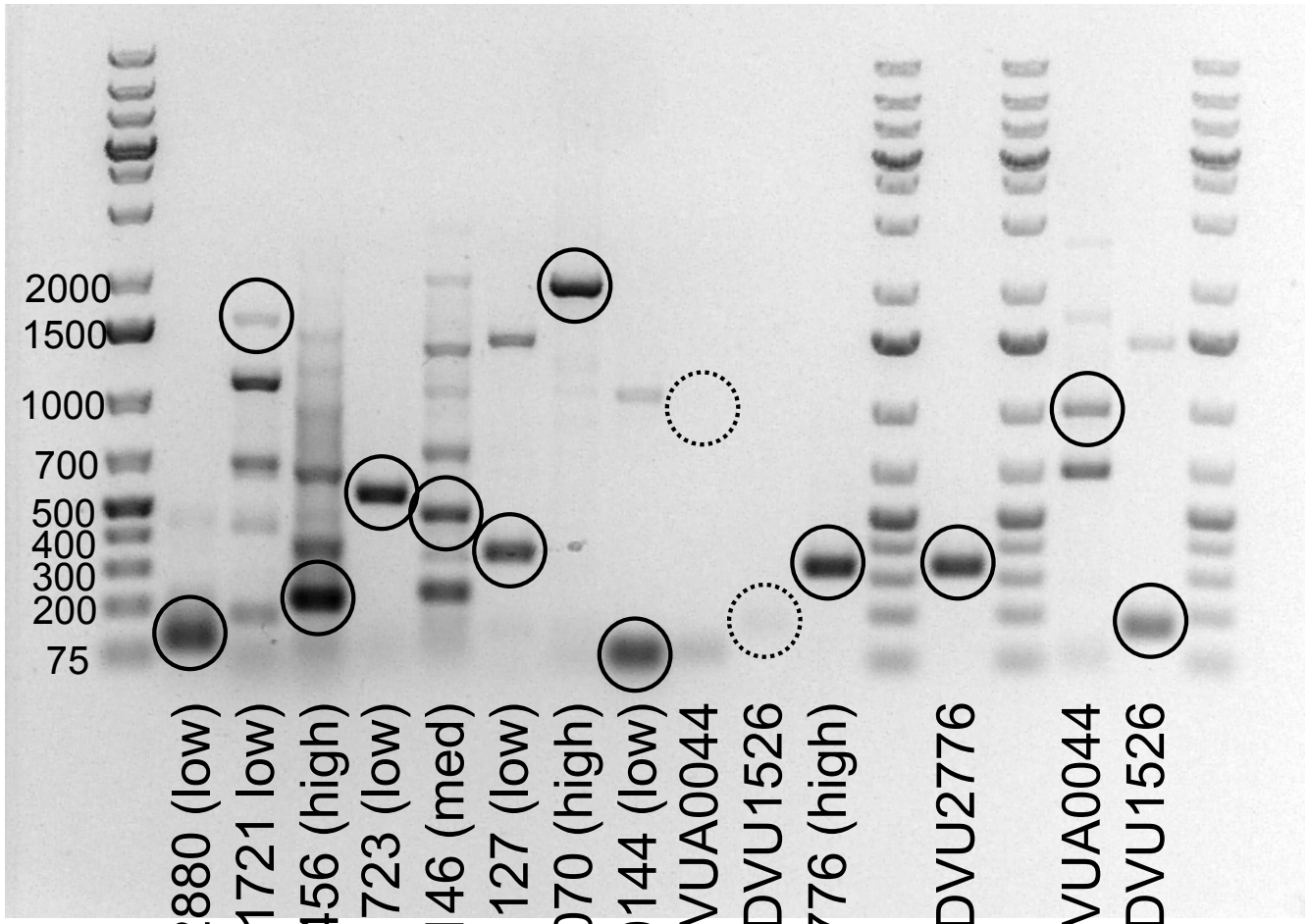
2000
1500
1000
700
500
400
300
200
75
DVU2880 (low)
DVU1721 low)
DVU2456 (high)
DVU1723 (low)
DVUA0146 (med)
DVU1127 (low)
DVUA0070 (high)
DVUA0144 (low)
DVUA0044
DVU1526
DVU2776 (high)
gDNA DVU2776
gDNA, DVUA0044
gDNA, DVU1526

Figure 1

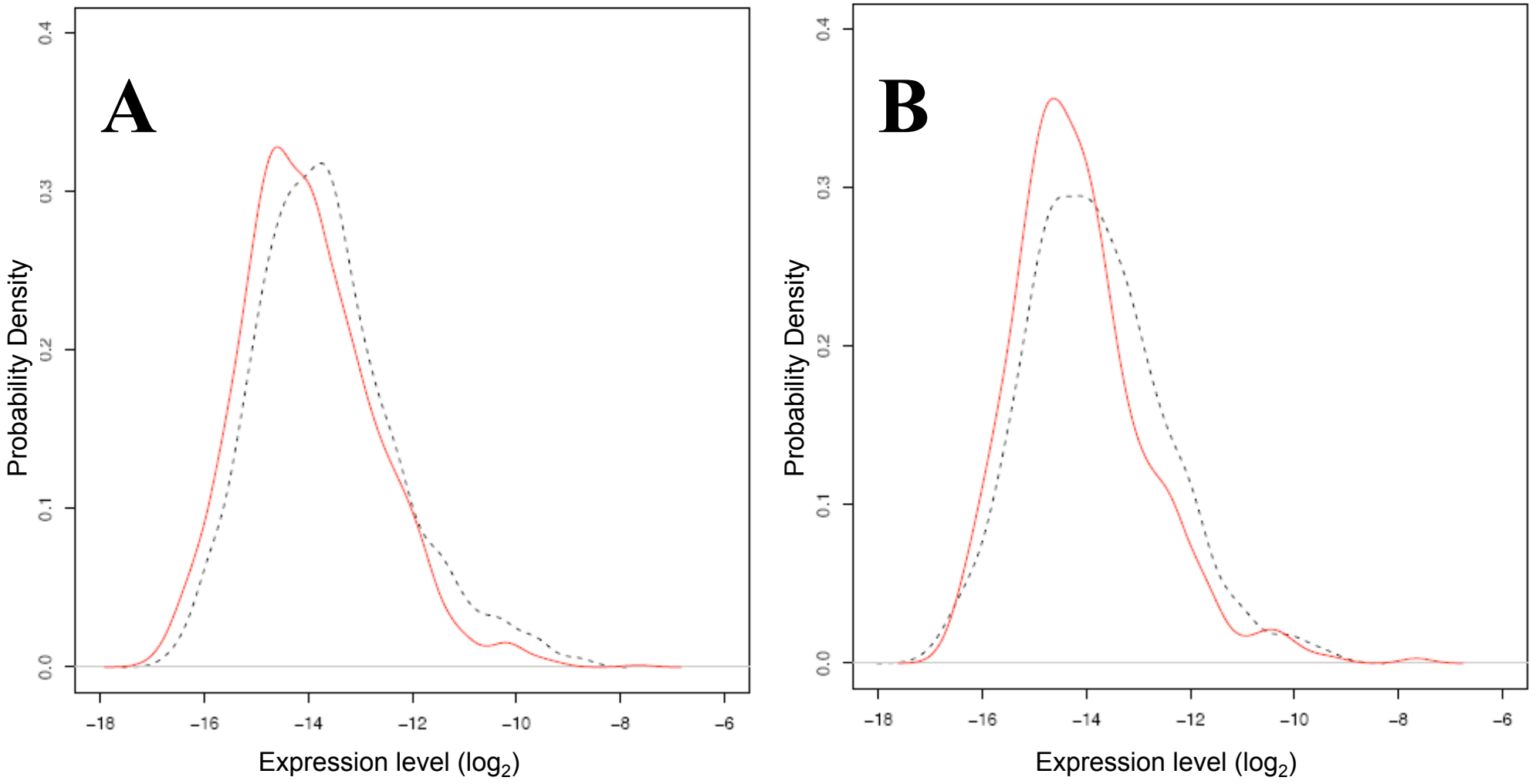
A
B
Probability Density
0.0
0.1
0.2
0.3
0.4
−18
−16
−14
−12
−10
−8
−6
Expression level (log2)

Figure 2

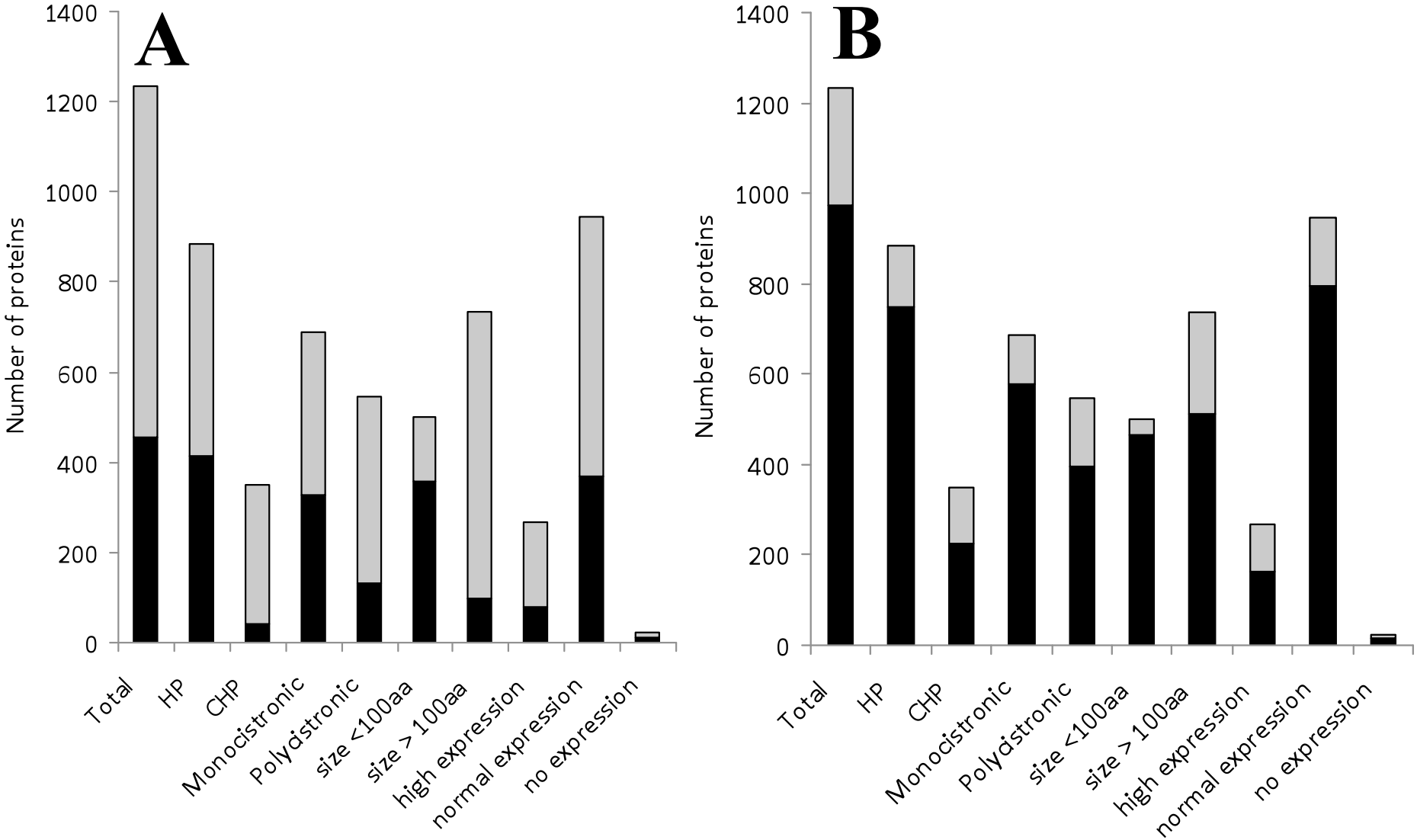
A
B
Number of proteins
0
200
400
600
800
1000
1200
1400
Total
HP
CHP
Monocistronic
Polycistronic
size <100aa
size > 100aa
high expression
normal expression
no expression

Figure 3

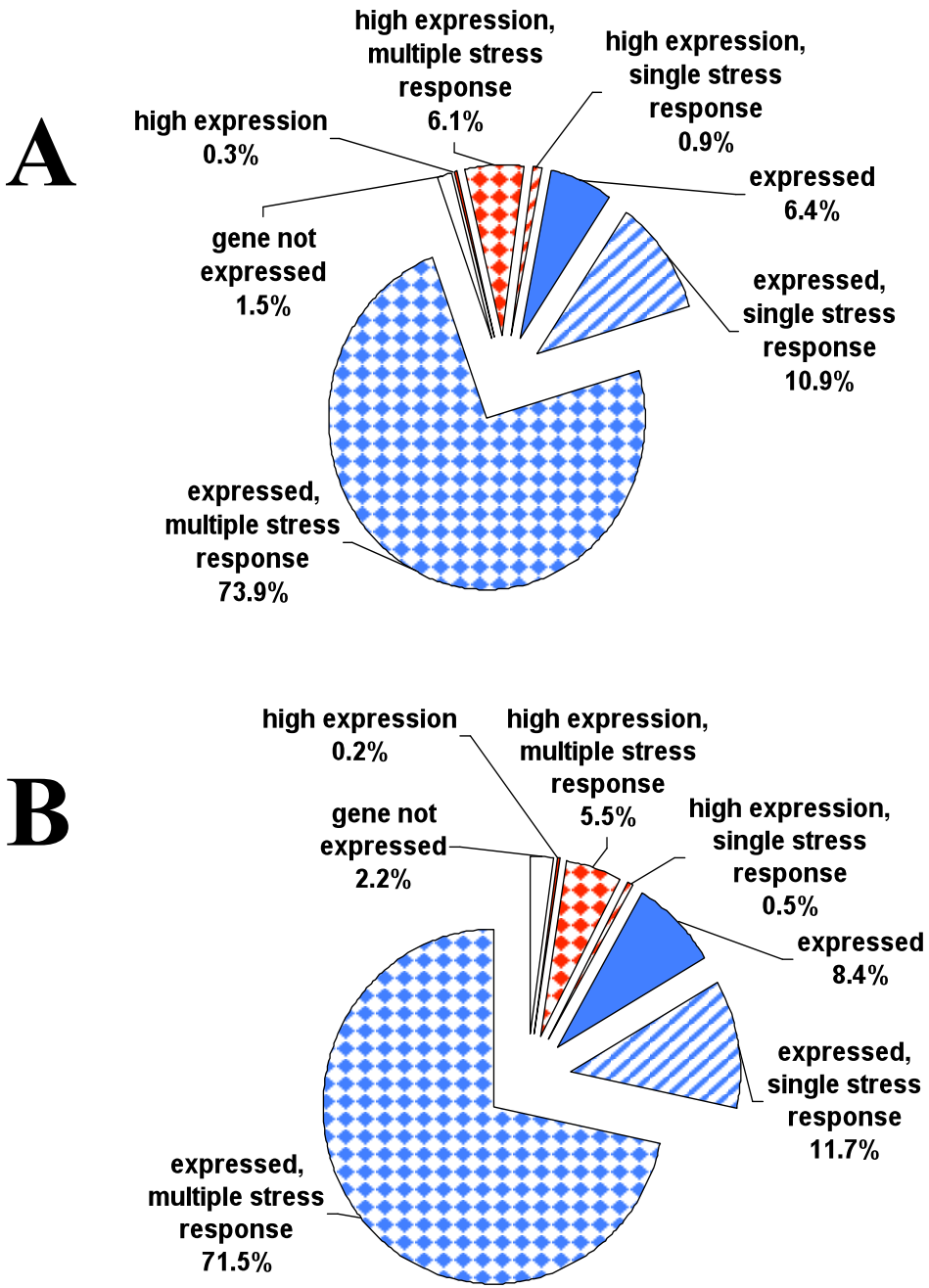
A
high expression, multiple stress response 6.1%
high expression, single stress response 0.9%
high expression 0.3%
expressed 6.4%
gene not expressed 1.5%
expressed, single stress response 10.9%
expressed, multiple stress response 73.9%
B
high expression 0.2%
high expression, multiple stress response 5.5%
gene not expressed 2.2%
high expression, single stress response 0.5%
expressed 8.4%
expressed, single stress response 11.7%
expressed, multiple stress response 71.5%

Figure 4

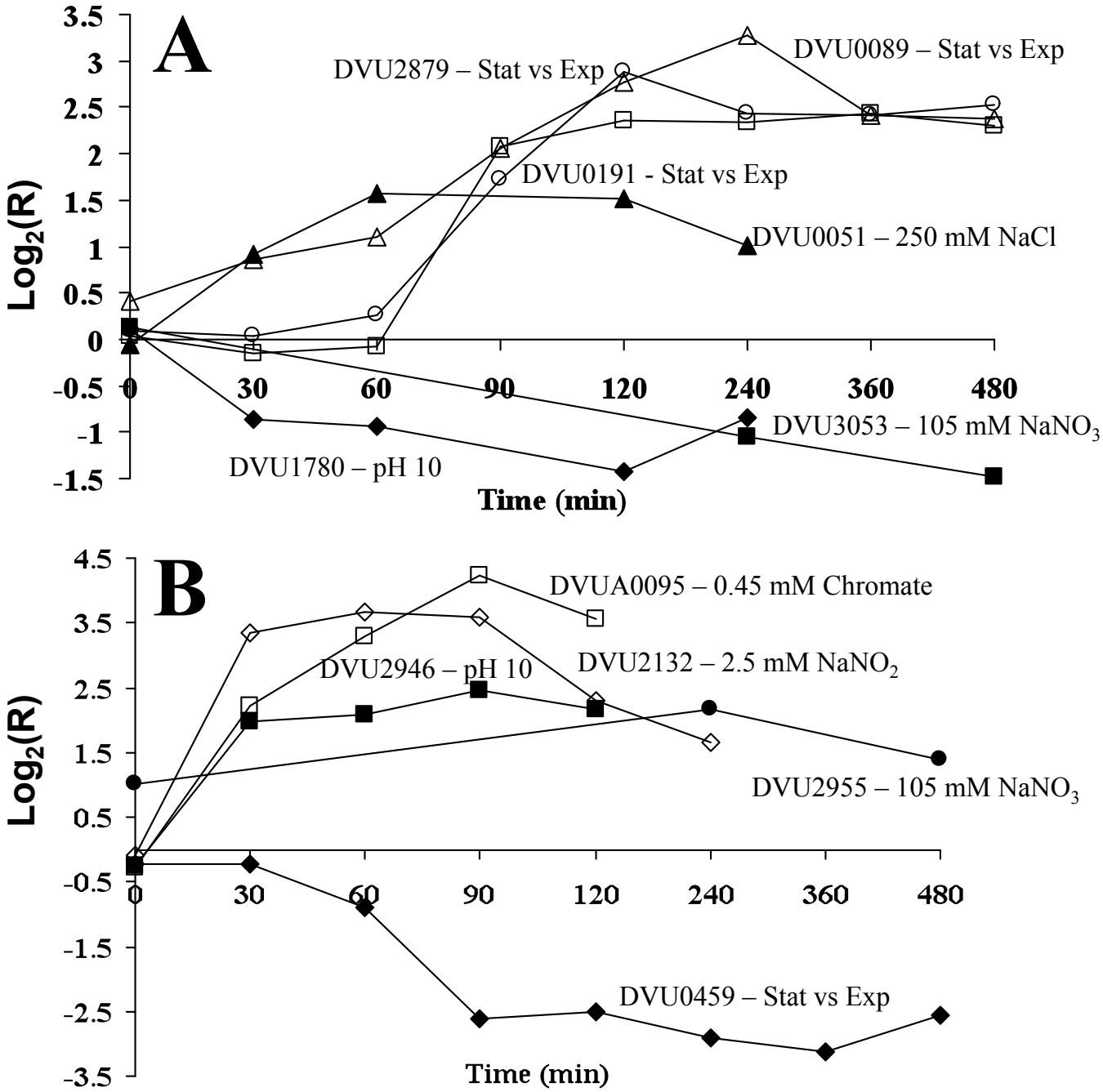
A
DVU2879 – Stat vs Exp
DVU0089 – Stat vs Exp
DVU0191 - Stat vs Exp
DVU0051 – 250 mM NaCl
DVU3053 – 105 mM NaNO3
DVU1780 – pH 10
Log2(R)
Time (min)
B
DVUA0095 – 0.45 mM Chromate
DVU2946 – pH 10
DVU2132 – 2.5 mM NaNO2
DVU2955 – 105 mM NaNO3
DVU0459 – Stat vs Exp
Log2(R)
Time (min)

Figure 5

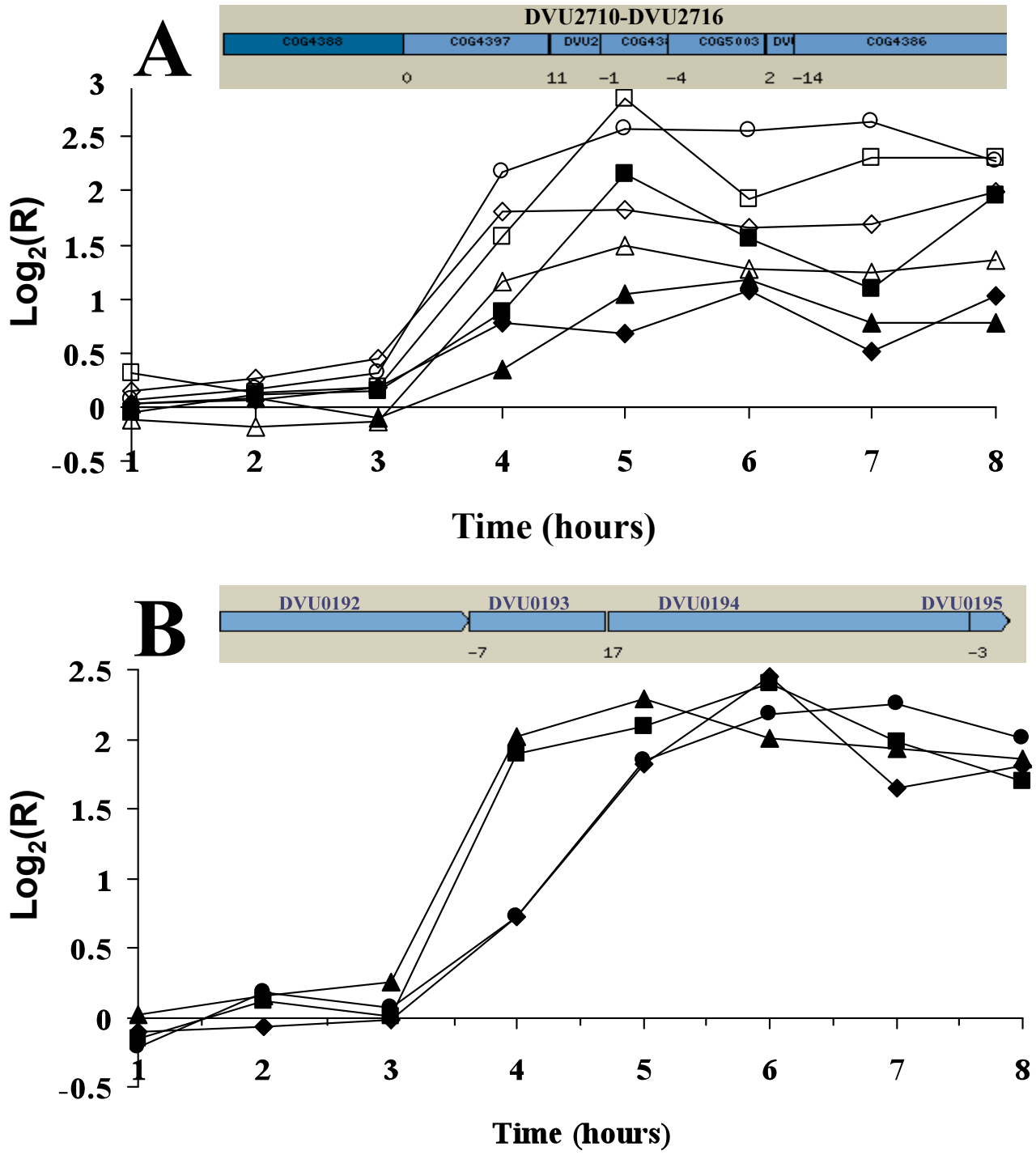
A
DVU2710-DVU2716
COG4388
COG4397
DVU2
COG43
COG5003
DV
COG4386
0
11
-1
-4
2
-14
Log2(R)
Time (hours)
B
DVU0192
DVU0193
DVU0194
DVU0195
-7
17
-3
Log2(R)
Time (hours)

Figure 6

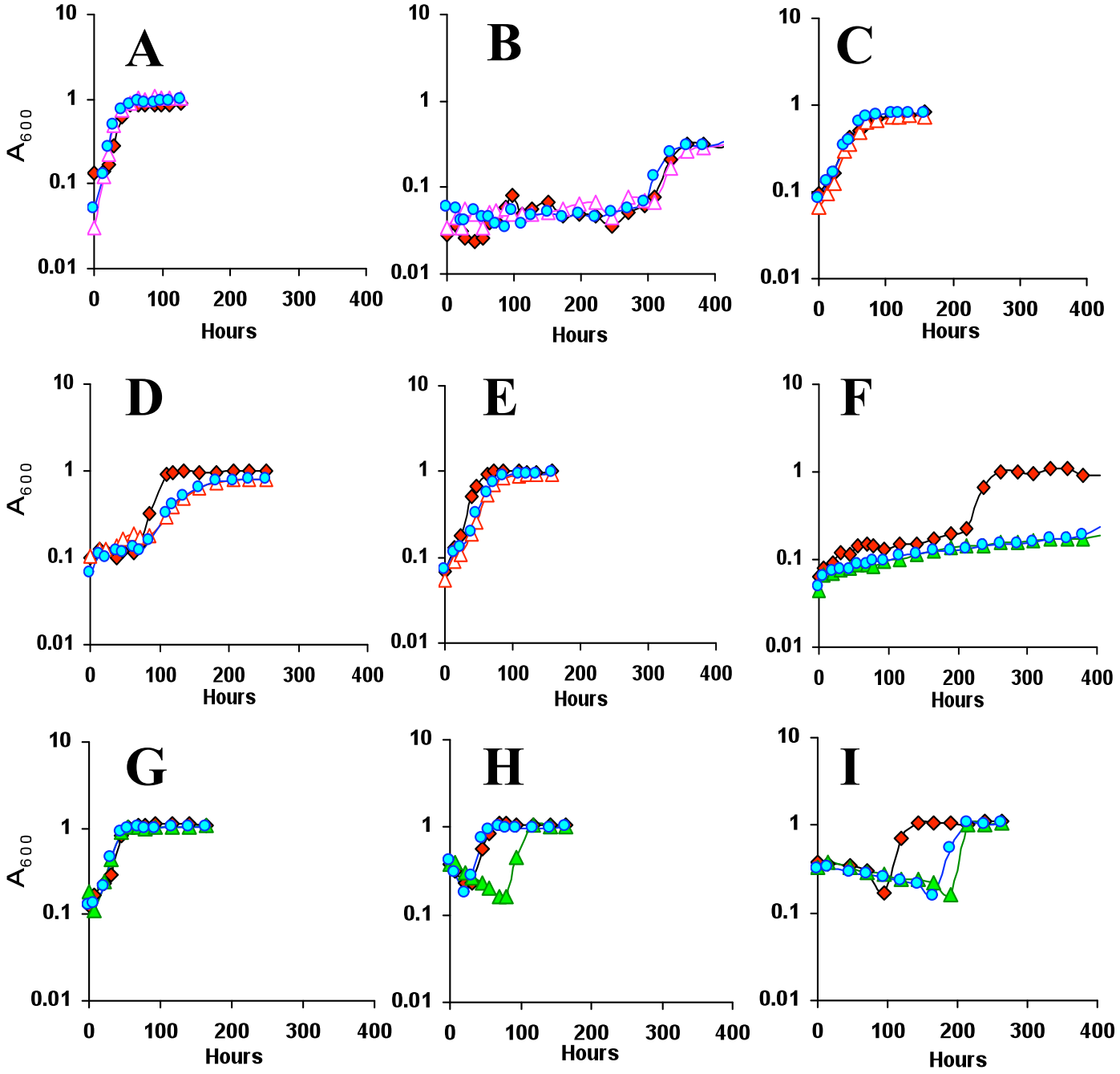

Figure 7