Genomics, Proteomics, and Bioinformatics:

Global Gene Expression Profiling in *Escherichia coli* K12: EFFECTS OF OXYGEN AVAILABILITY AND ArcA

Kirsty A. Salmon, She-pin Hung, Nicholas R. Steffen, Rebecca Krupp, Pierre Baldi, G. Wesley Hatfield and Robert P. Gunsalus

J. Biol. Chem. 2005, 280:15084-15096.
doi: 10.1074/jbc.M414030200 originally published online February 7, 2005

THE JOURNAL OF BIOLOGICAL CHEMISTRY
Vol. 280, No. 15, Issue of April 15, pp. 15084–15096, 2005

Printed in U.S.A.

Global Gene Expression Profiling in *Escherichia coli* K12

EFFECTS OF OXYGEN AVAILABILITY AND ArcA*[S]

Received for publication, December 14, 2004, and in revised form, January 18, 2005
Published, JBC Papers in Press, February 7, 2005, DOI 10.1074/jbc.M414030200

Kirsty A. Salmon,[a,b,c] She-pin Hung,[b,d,e,f] Nicholas R. Steffen,[g,h] Rebecca Krupp,[a,i] Pierre Baldi,[e,g,j] G. Wesley Hatfield,[d,e,k,l] and Robert P. Gunsalus[a,m,n]

From the [a]Department of Microbiology, Immunology, and Molecular Genetics and the [m]Molecular Biology Institute, University of California, Los Angeles, California 90095-1489 and the Departments of [d]Microbiology and Molecular Genetics, [g]Information and Computer Science, [j]Biological Chemistry, and [k]Chemical Engineering and Material Science and the [e]Institute for Genomics and Bioinformatics, University of California, Irvine, California 92697

The ArcAB two-component system of *Escherichia coli* regulates the aerobic/anaerobic expression of genes that encode respiratory proteins whose synthesis is coordinated during aerobic/anaerobic cell growth. A genomic study of *E. coli* was undertaken to identify other potential targets of oxygen and ArcA regulation. A group of 175 genes generated from this study and our previous study on oxygen regulation (Salmon, K., Hung, S. P., Mekjian, K., Baldi, P., Hatfield, G. W., and Gunsalus, R. P. (2003) *J. Biol. Chem.* 278, 29837–29855), called our gold standard gene set, have *p* values <0.00013 and a posterior probability of differential expression value of 0.99. These 175 genes clustered into eight expression patterns and represent genes involved in a large number of cell processes, including small molecule biosynthesis, macromolecular synthesis, and aerobic/anaerobic respiration and fermentation. In addition, 119 of these 175 genes were also identified in our previous study of the *fnr* allele. A MEME/weight matrix method was used to identify a new putative ArcA-binding site for all genes of the *E. coli* genome. 16 new sites were identified upstream of genes in our gold standard set. The strict statistical analyses that we have performed on our data allow us to predict that 1139 genes in the *E. coli* genome are regulated either directly or indirectly by the ArcA protein with a 99% confidence level.

* This work was supported in part by National Institutes of Health Grants GM49694 and AI21678 (to R. P. G.) and Grant GM68903 (to G. W. H.) and by the University of California Institute for Genomics and Bioinformatics (Irvine, CA). The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

[S] The on-line version of this article (available at http://www.jbc.org) contains a supplemental table.

[b] Both authors contributed equally to this work.

[c] Present address: Dept. of Microbiology and Molecular Genetics, University of California, Irvine, CA 92697.

[f] Recipient of a postdoctoral fellowship from the University of California Biotechnology Research and Education Program.

[h] Recipient of Biomedical Informatics Training Program Postdoctoral Fellowship T15 LM-07443 from the National Institutes of Health-National Library of Medicine.

[i] Supported by a traineeship from the UCLA-Integrative Graduate Education and Research Traineeship Bioinformatics Program funded by National Science Foundation Grant DGE-9987641.

[l] To whom correspondence may be addressed: Dept. of Microbiology and Molecular Genetics, University of California, Medical Science I, Campus Dr., Irvine, CA 92697. Tel.: 949-824-5344; Fax: 949-824-8595; E-mail: gwhatfie@uci.edu.

[n] To whom correspondence may be addressed: Dept. of Microbiology, Immunology, and Molecular Genetics, UCLA, 609 Charles Young Dr. East, 1602A MSB, Los Angeles, CA 90095. Tel.: 310-206-8201; Fax: 310-206-5231; E-mail: robg@microbio.ucla.edu.

Escherichia coli thrives in the gastrointestinal tract of many warm-blooded animals as a commensal or as a pathogen depending on a strain-dependent complement of genes (2). These enteric bacteria have the ability to switch between aerobic and anaerobic growth if oxygen is limiting. In response to microenvironments in the host, each individual cell adjusts its metabolic pathways to optimize energy generation via aerobic and/or anaerobic respiration or by fermentation of simple sugars (3). Many other cellular functions also are adjusted in response to oxygen availability, such as alterations in gene expression levels of membrane-associated nutrient uptake and/or excretion systems, biosynthetic pathways, and macromolecular synthesis (3).

Expression of *E. coli* genes involved in oxygen utilization is down-regulated as oxygen is depleted, and in a reciprocal fashion, expression of genes encoding alternative anaerobic electron transport pathways or genes needed for fermentation is switched on. Many of these metabolic transitions are controlled at the transcriptional level by the activities of the ferric nitrate reductase global regulatory protein FNR and/or the two-component ArcAB regulatory system (4, 5). The role of the FNR protein in the global control of *E. coli* gene expression has been profiled in response to anaerobiosis (1). Based on this analysis of whole genome transcription data, it was estimated that the expression of over one-third of the genes expressed during growth under aerobic conditions are altered when *E. coli* cells transition to an anaerobic growth state and that the expression of half of these genes is modulated either directly or indirectly by FNR. Thus, the *fnr* gene family was estimated to be ~10-fold larger than the 70 members previously recognized as members of the *fnr* gene regulatory network (6, 7).

The ArcAB (aerobic respiratory control) two-component regulatory system is recognized as a second global regulator of anaerobic gene regulation (3, 6, 8). The ArcAB system is composed of a classical OmpR-like receiver regulator, ArcA, and a membrane-associated sensor transmitter protein, ArcB (6). Together, these components have been shown to regulate expression of oxygen-requiring pathways, including the tricarboxylic acid cycle (*e.g. sdhCDAB*, *icd*, *fumA*, *mdh*, *gltA*, *acnA*, and *acnB*), and the aerobic cytochrome oxidase complexes (9–18). ArcAB is also known to be required for proper expression of certain catabolic genes for pyruvate utilization and sugar fermentation (19–21).

In this genome-based study, we have identified additional *E. coli* genes under oxygen control that are differentially expressed in response to the ArcA global regulatory protein. This was accomplished by the use of DNA microarrays to analyze gene expression profiles in *E. coli* cells cultured at steady-state growth rates under aerobic ($+O_2$) or anaerobic ($-O_2$) growth

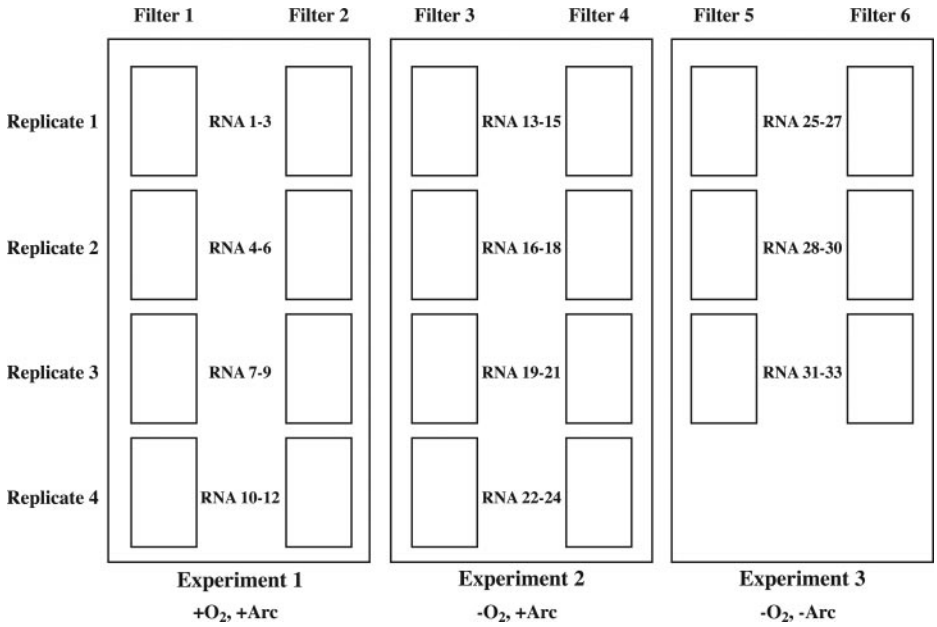

FIG. 1. **Experimental design.** See "Materials and Methods" for details.

conditions and in cells cultured under anaerobic growth conditions in the presence ($-O_2$, +ArcA) or absence ($-O_2$, −ArcA) of the ArcA protein or in otherwise *arcA*$^+$ and *arcA*$^-$ isogenic strains. These experiments show that about one-half of the genes whose expression levels are affected by aerobic to anaerobic transitions are also affected by the ArcA protein. Thus, the number of *E. coli* genes differentially regulated by the ArcA protein is much larger than the 30 (5) or 100 (22) genes/operons previously recognized. The results of the gene expression profiling experiments further show that as many as two-thirds of the genes whose expression levels are affected by the ArcA protein are also affected by the FNR protein (1).

MATERIALS AND METHODS

Chemicals and Reagents—Avian myeloblastosis virus reverse transcriptase and Sephadex G-25 Quickspin columns were obtained from Roche Applied Science. Phenol and the DNA-free kit were purchased from Ambion Inc. Ribonuclease inhibitor III was purchased from PanVera/Takara. Ultrapure deoxynucleoside triphosphates were purchased from Amersham Biosciences. Random hexamer oligonucleotides and T4 polynucleotide kinase were obtained from New England Biolabs Inc., and [α-^{33}P]dCTP (2–3000 Ci/mmol) was obtained from PerkinElmer Life Sciences. DNA filter arrays (Panorama *E. coli* gene arrays) were obtained from Sigma. SYBR Gold was purchased from Molecular Probes, Inc.. All other chemicals were obtained from Sigma. All reagents and baked glassware used in RNA manipulations were treated with diethyl pyrocarbonate prior to their use.

Bacterial Strains and Growth Conditions—*E. coli* strains MC4100 (*F*$^-$ *araD139* Δ(*argF-lac*)*U169 rpsL150 relA1 flb-5301 deoC1 ptsF25 rbsR*) (23) and PC35 (MC4100 Δ*arcA*::*kan*) (15) were used in this study. Cells were grown in MOPS[1] medium (24) containing 40 mM glucose. Aerobic cultures were grown as described previously (1) in 125-ml Erlenmeyer flasks with constant aeration. Anaerobic cultures were grown in 15-ml anaerobic tubes fitted with butyl rubber stoppers (15). The same medium was made anaerobic by flushing with O_2-free N_2 gas for 20 min and then dispensed anaerobically into N_2-flushed tubes. Cultures of the indicated strain were inoculated from overnight cultures grown under identical conditions (15). Cells were grown to A_{600} = 0.5–0.6 (mid-exponential growth phase) and harvested as described previously (1, 25).

Total RNA Isolation, cDNA Synthesis, and Target Labeling Conditions—Total RNA was isolated from 10-ml cultures; cDNA was synthesized and labeled with [α-^{33}P]dCTP; and filters were hybridized exactly as described by Hung *et al.* (25). Stripping and reusing filters four times as described here results in a <3% increase in variance (26).

Data Acquisition—The commercial software package DNA ArrayVision obtained from Research Imaging Inc. was used to grid the 16-bit image file obtained from a PhosphorImager, to record the pixel density of each of the 18,432 addresses on each filter, and to perform the background subtractions. 8580 of the addresses on each filter were spotted with duplicate copies of each of the 4290 *E. coli* open reading frames (ORFs). The remaining 9852 empty addresses were used for background measurements. Because the backgrounds were constant, a global average background measurement was subtracted from each experimental measurement, although local background calculations are possible.

Experimental Design—The experiments described here (Fig. 1) were performed at the same time as our previously reported experiments profiling gene expression levels in the presence or absence of oxygen and FNR (1). The data for strain MC4100 (ArcA$^+$) grown aerobically (Experiment 1, Filters 1 and 2) and anaerobically (Experiment 2, Filters 3 and 4) have been reported by Salmon *et al.* (1). For Experiment 3, Filters 5 and 6 were hybridized with random hexamer-generated ^{33}P-labeled cDNA fragments complementary to each of three independently prepared RNA preparations (RNA 25–27) obtained from three individual cultures of strain PC35 (*arcA*$^-$) grown under anaerobic conditions. These three ^{33}P-labeled cDNA target preparations were pooled prior to hybridization to the full-length ORF probes on the filters (Experiment 3, Replicate 1, Filters 5 and 6). Following PhosphorImager analysis, these filters were stripped and again hybridized with pooled ^{33}P-labeled cDNA target fragments complementary to each of another three independently prepared RNA preparations (RNA 28–30) from the same strain (PC35; Experiment 3, Replicate 2). This procedure was repeated one more time with Filters 5 and 6 with yet another independently prepared pool of cDNA targets (Experiment 3, Replicates 3; RNA 31–33). The data for the fourth replicate of this experiment were lost.

This experimental design produced duplicate filter data for four replicates performed with cDNA targets complementary to four independent sets of pooled RNA preparations for Experiments 1 and 2. Thus, because each filter contained duplicate spots for each ORF and duplicate filters were used for each experiment, a total of 16 measurements were obtained, four measurements for each ORF from each of four replicates. Duplicate filter data were obtained for three replicates performed with cDNA targets complementary to three independent sets of pooled RNA preparations for Experiment 3. Thus, because each filter contained duplicate spots for each ORF and duplicate filters were used for each experiment, a total of 12 measurements were obtained, four measurements for each ORF from each of three replicates.

Statistical Analyses—Data processing and statistical methods implemented in the Cyber-T software used for the analysis and interpretation of the data obtained from the DNA microarray experiments described in this study were the same as those described previously by Salmon *et al.* (1). For each target signal, a background subtracted estimate of the expression level was obtained and scaled to total counts on the membrane by dividing each individual gene expression value by the total of all target signals on the membrane. Thus, each normalized

[1] The abbreviations used are: MOPS, 4-morpholinepropanesulfonic acid; ORF, open reading frame; PPDE, posterior probability of differential expression.

gene level is expressed as a fraction of the total mRNA hybridized to each DNA array. For any given measurement, a value greater than zero (indicating an expression level) or a zero (indicating an expression level lower than background) was obtained. Only those genes exhibiting an expression level greater than zero in all replicates were used for statistical analysis. These gene expression level measurements were analyzed by a regularized *t* test based on a Bayesian statistical framework (25–29). For analysis of the data reported here, we ranked the mean gene expression levels of the replicate experiments in ascending order, used a sliding window of 101 genes, and assigned the average S.D. of the 50 genes ranked below and above each gene as the Bayesian S.D. for that gene. The *p* values for each gene measurement based on a regularized *t* test with a confidence value of 10 are reported in the Supplemental Material. A comprehensive discussion of the use of a regularized *t* test and the modifications applicable to the analysis of DNA microarray data of the type presented here is described in detail elsewhere (26).

Gene measurements containing zero expression values in one or more replicates were set aside. Among this set of genes, those with zero expression values for all replicates in one experiment and all values greater than zero for all measurements of another experiment were identified. Because these gene measurements could not be analyzed with a *t* test, the significance of these results was evaluated by ranking these genes in ascending order according to their coefficients of variance of the four greater than zero measurements of each experiment.

Cyber-T employs a mixture model-based method described by Allison *et al.* (30) for the computation of the global false positive and false negative levels inherent in a DNA microarray experiment (25, 26). With this method, described by Hung *et al.* (25), we estimated the rates of false positives and false negatives as well as true positives and true negatives at any given *p* value threshold. In other words, we obtained a posterior probability of differential expression PPDE(*p*) value for each gene measurement and a PPDE(<*p*) value at any given *p* value threshold based on the experiment-wide global false positive level and the *p* value exhibited by that gene (25, 26). In most instances, PPDE(<*p*) values are reported below and Tables I–VIII. However, both PPDE(*p*) and PPDE(<*p*) values are given for each gene in the Supplemental Material.

It is expected that for each *p* value threshold, there is a tradeoff between the rates of true and false positives. A low conservative *p* value threshold leads to few false positives, but may also reduce the true positive rate. A large *p* value threshold ultimately allows one to recover all the true positives, but at the cost of increasing the false positive rate. This fundamental tradeoff is usually captured in statistics using a receiver operating characteristic curve obtained by plotting the true positive rate (or sensitivity) defined by true positive/(true positive + false negative) *versus* the false positive rate defined by false positive/(false positive + true positive) (87). For instance, at a 77% true positive rate, we expect a 5% false positive rate when Experiment 1 ($+O_2$, +ArcA) is compared with Experiment 2 ($-O_2$, +ArcA) (Fig. 2*A*), and at a 80% true positive rate, we expect a 5% false positive rate when Experiment 2 ($-O_2$, +ArcA) is compared with Experiment 3 ($-O_2$, −ArcA) (Fig. 2*B*).

The Cyber-T software package is available at the web site for the Institute for Genomics and Bioinformatics at the University of California (Irvine, CA). The clustering methods used to determine the regulatory patterns reported below are those implemented in the GeneSpring™ software package (Silicon Genetics, Redwood City, CA).

Data Accession—All raw and processed data for the experimental results reported here are provided in tabular format as Excel files in the Supplemental Material.

RESULTS AND DISCUSSION

Differential Gene Expression in the Presence or Absence of Oxygen

In the following discussions, we often refer to the -fold change for differentially expressed genes. However, simple -fold changes are incomplete and can be misleading (26). For this reason, the mean expression levels, S.D. values, *p* values, PPDE(<*p*) values, and PPDE(*p*) values for all differentially expressed *E. coli* genes are included in the Supplemental Material. In Tables I–IX, we report only *p* values, PPDE (<*p*) values, and -fold changes.

A comparison of the wild-type *E. coli* gene expression levels between cells grown in the presence and absence of oxygen revealed 2820 genes that exhibited expression levels above the background for all replicates of Experiments 1 and 2 ($+O_2$, + ArcA *versus* $-O_2$, + ArcA) (Fig. 1) (1). The statistical analysis of these data revealed that approximately one-half of the genes expressed during aerobic growth (1445 genes) were differentially expressed following a transition from aerobic to anaerobic growth with a *p* value of 0.05 and a PPDE(<*p*) value of 0.96. Therefore, 58 of these 1445 differentially expressed genes are expected to be false positives (25).

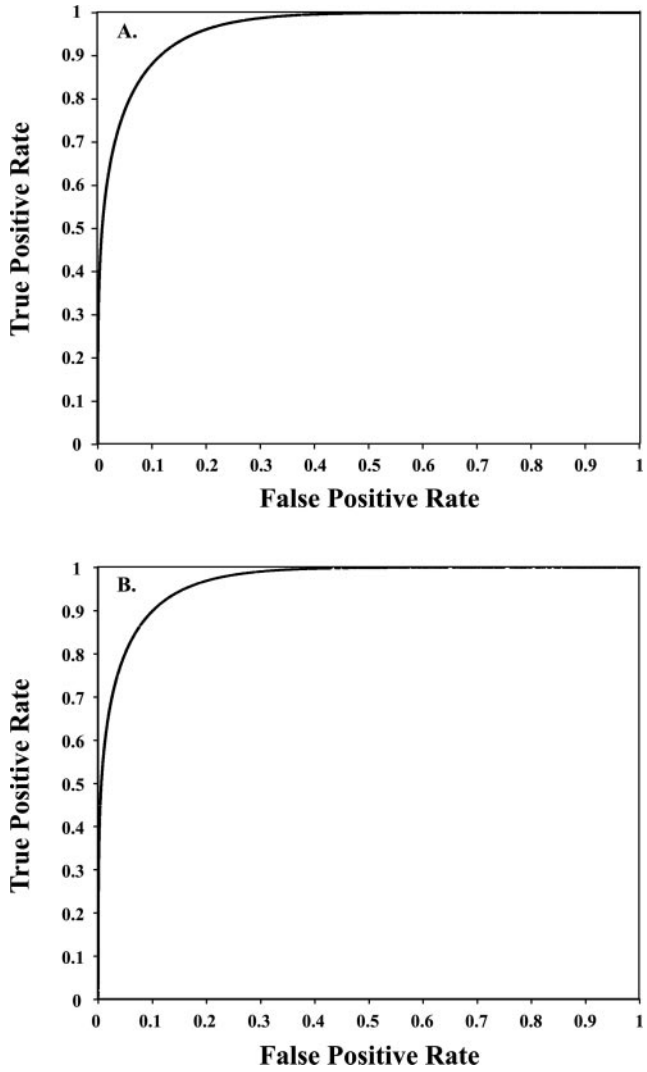

FIG. 2. **Receiver Operating Characteristic curve.** These plots correlate the fraction of correctly identified differentially expressed genes (*y axis*) with the fraction of falsely identified differentially expressed genes (*x axis*). *Panel A*: $+O_2$, +ArcA *versus* $-O_2$, +ArcA. *Panel B*: $-O_2$, +ArcA *versus* $-O_2$, −ArcA. The false positive rate is [FP/(FP+TN)]. The true positive rate is [TP/(TP+FN)], where FP is the false positive, TN is the true negative, TP is the true positive and FN is the false negative.

Differential Gene Expression in the Absence of Oxygen and in the Presence and Absence of the ArcA Global Regulatory Protein

A comparison of the gene expression levels between cells grown in the absence of oxygen and in the presence or absence of ArcA revealed 2264 genes that exhibited expression levels above the background for all replicates of Experiments 2 and 3 ($-O_2$, +ArcA *versus* $-O_2$, −ArcA) (Fig. 1). Again, about one-half of the gene expression levels were modulated by this treatment condition. An examination of the distribution of *p* values suggested that the expression levels of 1243 genes with *p* values ≤0.05 were modulated either directly or indirectly by ArcA during growth transition from aerobic to anaerobic con-

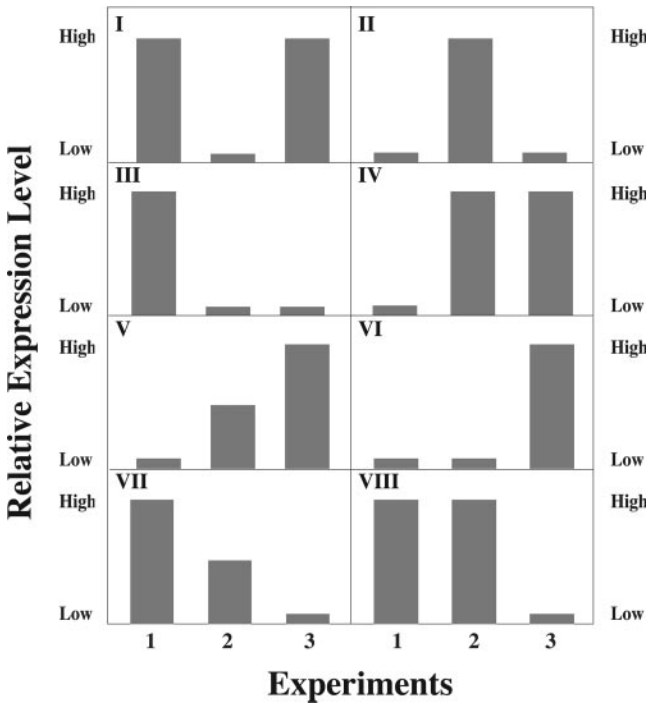

FIG. 3. **Gene expression regulatory patterns expected from the comparison of DNA array experiments with one control and two treatment conditions.** Experiment 1 (control) indicates gene expression levels during growth under aerobic conditions in an ArcA$^+$ *E. coli* strain. Experiment 2 indicates gene expression levels during growth under anaerobic conditions in an ArcA$^+$ *E. coli* strain. Experiment 3 indicates gene expression levels during growth under anaerobic conditions in an ArcA-deficient *E. coli* strain. Regulatory patterns I–VIII are indicated.

ditions. Because the PPDE($<p$) value for this group of genes is 0.97, 37 false positives are expected. The individual p values and PPDE values, as well as additional statistical data, for all genes are provided in the Supplemental Material.

Identification of Differential Gene Expression Patterns Resulting from Two-variable Perturbation Experiments

To identify the global changes and adjustments of gene expression patterns that facilitate a transition from aerobic to anaerobic growth conditions and to determine the effects of genotype on these gene expression patterns, we analyzed *E. coli* gene expression profiles obtained from cells cultured under aerobic ($+O_2$) or anaerobic ($-O_2$) growth conditions and under anaerobic growth conditions in the presence ($-O_2$, +ArcA) or absence ($-O_2$, −ArcA) of ArcA, the global regulatory protein for anaerobic metabolism. Because ArcA is presumed to be inactive under aerobic conditions (5, 6, 31), we did not perform experiments comparing *arcA* genotypes under aerobic conditions.

Only two general regulatory patterns can be observed when two experimental conditions are compared, *e.g.* growth in the presence or absence of oxygen. However, when two conditions are compared, at least eight general regulatory patterns are expected. The data in Fig. 3 diagram the eight basic regulatory patterns that could be observed among three experiments conducted in the presence and absence of oxygen in an *arcA*$^+$ strain and in the absence of oxygen in an *arcA*$^-$ strain. For simplicity, only three expression levels for each of these three experimental conditions were assumed: low, medium, and high.

To identify genes differentially expressed at a high confidence level that correspond to each of the patterns (I–VIII) diagrammed in Fig. 3, the genes differentially expressed due to the treatment conditions of Experiments 1 and 2 were sorted in ascending order according to their p values based on the regularized t test as described under "Materials and Methods." Next, the genes differentially expressed due to the treatment conditions of Experiments 2 and 3 were sorted in ascending order according to their p values. 100 genes with the lowest p values present in both lists were selected. These genes exhibited either an increased or decreased expression level between both treatment conditions (*i.e.* between Experiments 1 and 2 and between Experiments 2 and 3) (Fig. 3).

To identify those genes differentially expressed at a high level of confidence under the treatment conditions of Experiments 1 and 2 but expressed at the same or similar levels under the treatment conditions of Experiments 2 and 3 (patterns III and IV) (Fig. 3), the 500 genes of Experiments 1 and 2 with the highest probability for differential expression values were compared with the 500 genes of Experiments 2 and 3 with the lowest probability for differential expression values. This comparison identified 40 genes that were present in both lists, *i.e.* genes whose regulatory patterns fulfill this criterion. Likewise, to identify those genes differentially expressed under the treatment conditions of Experiments 2 and 3 but expressed at the same or similar levels under the treatment conditions of Experiments 1 and 2 (patterns VI and VIII) (Fig. 3), the 500 genes of Experiments 2 and 3 with the highest probability for differential expression values were compared with the 500 genes of Experiments 1 and 2 with the lowest probability for differential expression values. This comparison identified 35 genes that were present in both lists. These gene lists were combined into a single list of 175 genes differentially expressed under at least one treatment condition. All of the differentially genes of this list exhibited p values <0.00013 and a global confidence based on the experiment-wide false positive level of 99% (PPDE($<p$) = 0.99). They constitute the "gold standard" gene set for the following analyses.

Hierarchical Clustering and Principal Component Analysis

GeneSpring™ software was used to empirically determine parameters for hierarchical clustering of these 175 genes into the eight patterns of Fig. 3 as discussed by Salmon *et al.* (1) and shown in Fig. 4. Interestingly, 83 of these ArcA-regulated genes are also differentially regulated directly or indirectly by FNR (patterns I, II, and V–VIII) (1). As an independent test to corroborate the accuracy of this supervised hierarchical clustering method, we used principal component analysis to cluster and visualize the same set of 175 genes (14). The principal component analysis clustering results shown in Fig. 5 illustrate that this unsupervised method produced the same results as the supervised hierarchical clustering method.

Interpretation of Clustering Results

Although some of the genes or operons differentially expressed in the presence or absence of ArcA are expected to be affected only indirectly, others whose expression is directly regulated by ArcA should possess a DNA-binding site(s) upstream of their transcriptional start site(s). ArcA is a 28-kDa protein that contains a winged helix-turn-helix motif that interacts with a poorly conserved consensus DNA sequence (31). This ArcA-P consensus sequence, obtained from DNA footprinting experiments performed with ~15 ArcA-regulated promoters, is 5′-*W*GTTAATTAW-3′ (where *W* is A or T) (31).

Liu and De Wulf (22) used a weight matrix and a subset of 10 ArcA-footprinted promoter regions to define a slightly different consensus sequence of 5′-GTTAATTAAATGTTA-3′. This sequence resembles the previous 10-bp consensus sequence; how-

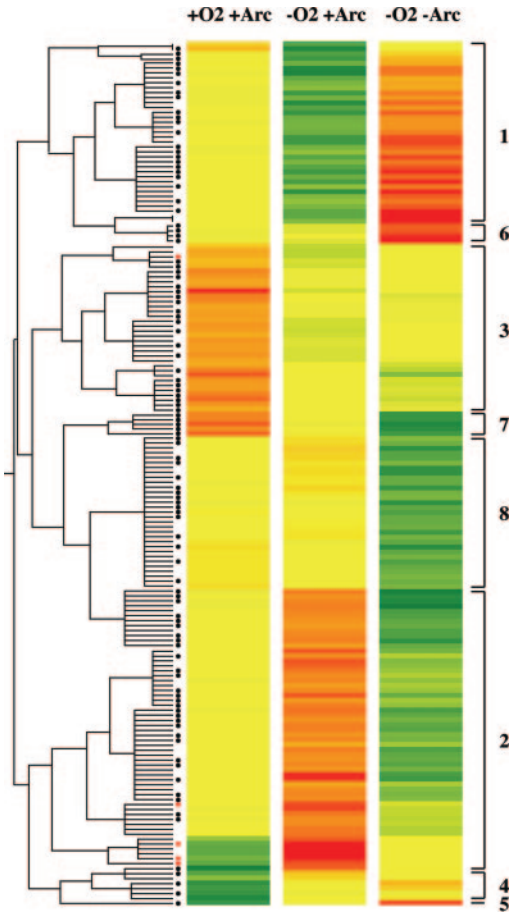

FIG. 4. **Hierarchical clustering of differentially expressed gene regulatory patterns.** The experimental cell growth conditions were as follows: wild-type *E. coli* K12 strain (ArcA$^+$) cultured under aerobic conditions (*+O2 +Arc*), wild-type *E. coli* K12 strain (ArcA$^+$) cultured under anaerobic conditions (*−O2 +Arc*), and isogenic *E. coli* K12 strain lacking the *arcA* gene cultured under anaerobic conditions (*−O2 −Arc*). Each regulatory pattern is identified by different colors on the dendrogram and by *numbers* that correspond to the regulatory patterns defined in the legend to Fig. 3. The trust parameter is directly related to the mean divided by the S.D. for each gene measurement. *Red* indicates high expression, *yellow* indicates medium expression, and *green* indicates low expression.

ever, it is extended by 5 residues at the 3′-end, and the first nucleotide of the original consensus sequence (5′-(A/T)) turned out to be poorly conserved and is not included in their motif (22). For the analyses reported here, a set of 26 known ArcA-binding sites in *E. coli*, including the 15 sites reviewed by Lynch and Lin (31) plus three newly footprinted ArcA-binding sequences,[2] was compiled (see the Supplemental Material). Analysis of these sequences with MEME Version 3.0 (32, 33) identified a partially degenerate 15-bp motif. A weight matrix was generated from the motif found by MEME. The *E. coli* K12 genome was then scanned for sequences on either strand that had a weight matrix score exceeding the threshold and that were within 300 bp of an ORF origin, as identified by Regulon_DB (34). A total of 386 such sequences were located.

When ArcA acts as an activator of gene expression, it most often binds to upstream sites centered from 60 to 120 base pairs before the transcriptional start site of the affected gene or operon. When it acts as a repressor of gene expression, it binds to other sites often located near the transcriptional start site of the affected gene or operon (31). Of the 42 genes down-regulated in the presence of ArcA (patterns I, V, and VI) (Fig. 3), 12 contain a documented ArcA-binding site or a predicted ArcA-binding site at or near the transcriptional start site using the above MEME/weight matrix method (Tables I, V, and VI). Of the 93 genes up-regulated in the presence of ArcA (patterns II, VII, and VIII) (Figs. 4 and 5), 14 contain an upstream documented or predicted ArcA-binding site (Tables II, VII, and VII). Because the expression levels of the 40 genes of patterns III and IV were not affected by the presence or absence of ArcA, they are not expected to possess binding sites for this regulatory protein. However, five of these genes are predicted to possess a putative ArcA-binding site (Tables III and IV). Of these, three genes, *cydA*, *nuoG*, and *nuoF*, are known to be ArcA-regulated; however, the expression data are not consistent with previously published data, and this is likely due to paralog issues within the *E. coli* genome. Thus, the statistical and clustering methods described here produced results consistent with biological expectations.

Functional Classes of Genes Affected by Oxygen Availability and ArcA

The following discussion is limited to the 175 genes (our gold standard set) of regulatory patterns I–VIII (Fig. 4), although ArcA control of many other genes may be deduced from the information supplied in the Supplemental Material. As in our previous study (1), they represent many genes known to be oxygen-controlled and another larger set for which no previous information is available. These genes are listed in Tables I–VIII and represent genes involved in a large number of processes, including small molecule biosynthesis, macromolecular synthesis, and aerobic/anaerobic respiration and fermentation. Regardless of their metabolic role, these genes are discussed below in the context of their expression patterns (Fig. 3).

Expression Pattern I: Decreased Expression during Anaerobiosis and Increased Expression in an ArcA Strain—Among the 175 genes displayed in the clustering procedures described above, 37 showed decreased expression under anaerobic conditions due to regulation by ArcA (Table I). Of these 37 genes, 10 have been reported to be directly regulated by ArcA (6), and 27 are newly discovered genes that are regulated either directly or indirectly by this global regulatory protein. In addition, 23 of the genes clustered into pattern I were also identified as being down-regulated by the FNR protein in our previous study (1). Previously described ArcA-regulated genes will be discussed first, followed by a discussion of the newly discovered ArcA-down-regulated genes.

Seven genes of the tricarboxylic acid cycle clustered into pattern I: *icdA*, *sdhAB*, *lpdA*, *mdh*, *sucD*, and *gltA*. Each of these seven genes has been shown previously to be anaerobically repressed by the ArcA protein (5, 6, 9, 10, 12, 13, 31, 35, 36). Regulation of *lpdA* by FNR was also observed in our previous study (1). A search for putative ArcA-binding sites using our customized MEME/weight matrix method (see "Materials and Methods") identified one or more sites upstream of each of these genes (Table I).

The *cyoA* gene is the first member of the *cyoABCDE* operon, which encodes all of the subunits of the cytochrome *o* ubiquinol oxidase. The *cyoA* gene was expressed 10-fold higher when cells were grown aerobically and 23-fold higher when cells were grown anaerobically in the ArcA-deficient strain (Table I). A previous study by our laboratory using a *cyoA*::*lacZ* fusion in the same ArcA$^+$ and ArcA$^-$ isogenic strains used in this work showed the same regulatory pattern (16). A site similar to the ArcA consensus sequence has been identified upstream of the *cyoA* promoter[3] and was also shown to be subject to regulation

[2] R. P. Gunsalus, unpublished data.

[3] J. A. Albrecht, unpublished data.

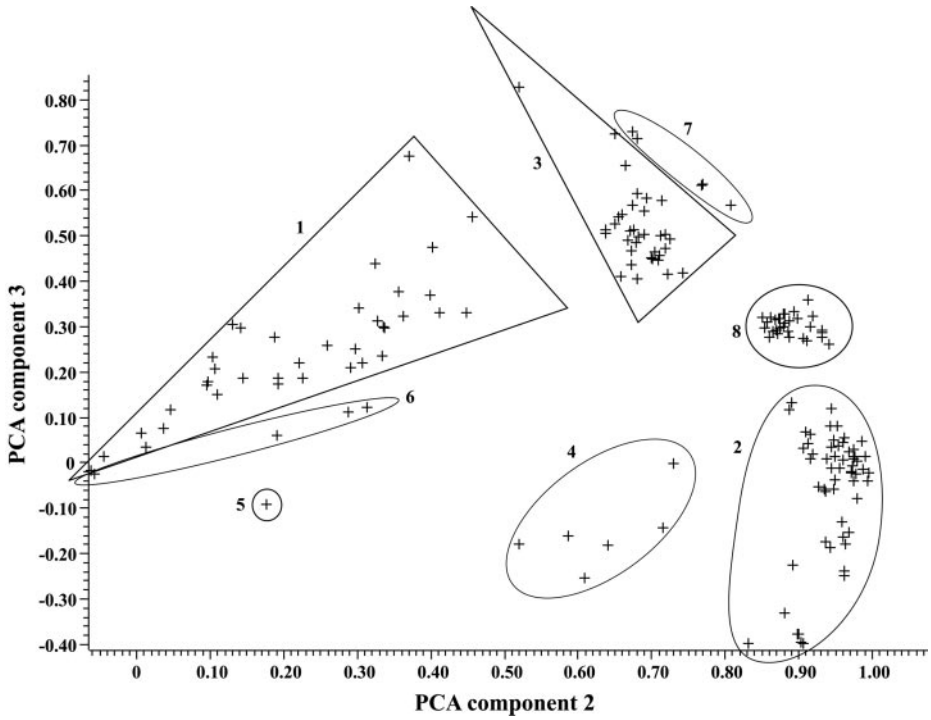

FIG. 5. **Principal component analysis clustering of differentially expressed gene regulatory patterns.** Shown is a two-dimensional projection onto a plane spanned by the second and third principal components. Each cluster is enclosed. The clusters are numbered according to the regulatory patterns indicated in the legends to Figs. 3 and 4. *PCA*, principal component analysis.

by FNR in our previous study (1), but this is likely indirect. Our MEME/weight matrix identified four putative ArcA-binding sites (Table I) upstream of the *cyoA* gene.

The *nuoB* and *nuoE* genes, which belong to the *nuoA–N* operon, encode NdhI (NADH dehydrogenase I), a membrane-associated, multisubunit, proton-translocating enzyme similar to complex I of eukaryotic mitochondria (37). Expression of both of these genes was lower under anaerobic conditions and elevated in the *arcA* mutant (Table I). A previous study using *nuo*::*lacZ* fusions established that *nuo* expression is subject to ArcA-mediated anaerobic repression (38). Two putative ArcA-binding sites were identified ~140 and 190 bp upstream of the *nuoA* gene using our MEME/weight matrix method (Table I). The *nuoE* gene also appeared to be subject to FNR regulation in our previous work (1), but the effect of FNR may be indirect as a consequence of its role in regulating ArcA expression (39).

The remaining genes in this group have not been shown previously to be subject to ArcA regulation. These newly discovered genes fall into the same functional classes as the genes regulated by the leucine-responsive regulatory protein Lrp under aerobic conditions (25) and FNR under anaerobic conditions (1). These functional classes include genes for small molecule biosynthesis and transport and macromolecule biosynthesis. More interestingly, of the remaining 27 genes of this expression group, 20 were also found to be regulated by FNR under anaerobic conditions (1).

12 genes of this cluster belong to the small molecule metabolism and transport groups. Nine of these genes were also found to be repressed in anaerobiosis due to regulation by FNR (1). These genes are *crr* (phosphocarrier protein for glucose transport); *gpmA* (phosphoglyceromutase); *gatY* (D-tagatose-1,6-bisphosphate aldolase); *talA* (transaldolase A); *trpB* (tryptophan synthase); *speD* and *speE* (biosynthesis of spermidine); *prlA* (*secY*, protein translocator of the *secYEG* operon); and *ompA*, which encodes an outer membrane protein. The remaining three genes are *rbsC* and *rbsD* (ribose high affinity transport system) and *yjcU* (*alsE*, allulose-6-phosophate 3-epimerase). Putative ArcA-binding sites were identified using the MEME/weight matrix for two of these: *gatY* and *trpB*.

11 of the remaining genes of this expression group belong to the macromolecule synthesis class. Eight of these were also observed to be regulated by FNR (1). These are *rpsA*, *rpsT*, *rpsJ*, *rplS*, *rplT*, and *rplM* (ribosomal proteins); *tufA* (elongation factor Tu); and *oppA* (oligopeptide permease). The remaining three genes are *rplX* (ribosomal protein), *pal* (essential lipoprotein), and *atpG* (ATP synthase). Putative ArcA-binding sites were identified using the MEME/weight matrix for two of these: *oppA* and *atpG*.

The functions of the remaining four genes in this list, *ycdC*, *yajG*, *yceD*, and *yfiA*, remain to be characterized. Three of these four genes, *yajG*, *yceD*, and *yfiA*, were also observed to be regulated by FNR in anaerobiosis (1).

Recently, Liu and De Wulf (22) identified 234 ORFs as being repressed by ArcA under anaerobic conditions in a microarray-based study. In our gold standard set, we identified a total of 42 genes as being up-regulated in an *arcA* mutant (patterns I, V, and VI) or 37 genes in pattern I. Only three genes, *gltA*, *icd*, and *mdh*, are conserved between the two reported data sets. However, our clustering set of 175 genes is highly restricted, with a strict PPDE($<p$) cutoff level of 0.997, and eliminates false positives and other genes for which the data are of lower statistical significance.

Expression Pattern II: Increased Expression during Anaerobiosis and Decreased Expression in an ArcA Strain—Transcription of the 57 genes of expression pattern II (Table II) was both induced in the absence of oxygen and positively regulated by ArcA. Moreover, of these 57 genes, 34 were also observed to be positively regulated by FNR in anaerobiosis (1). 19 of these genes are members of the small molecule metabolism and transport group. Among the genes for metabolism, eight were also observed to be positively regulated by FNR in anaerobiosis. These are *pyrD* (dihydro-orotate dehydrogenase), *glnD* (uridylyltransferase), *mobB* (molybdenum cofactor biosynthesis), *speC* (ornithine decarboxylase), *narY* (cryptic nitrate reductase subunit), *glnE* (glutamine synthetase/adenylyltransferase), *tdh* (threonine dehydrogenase), and *tynA* (tyramine oxidase). One of these genes, *glnD*, is predicted to have a putative ArcA-binding site (Table II).

The *gadA* and *gadB* genes, encoding two highly homologous glutamate decarboxylases, also clustered into this group. In

TABLE I

Regulatory pattern I: genes that exhibit decreased levels during anaerobic growth and increased levels in an ArcA-deficient strain

Gene name (NIH) and b no.	*p* value (+O_2, +*arcA vs.* −O_2, +*arcA*)	PPDE(<*p*) (−O_2, +*arcA vs.* −O_2, +*arcA*)	*p* value (−O_2, +*arcA vs.* −O_2, −*arcA*)	PPDE(<*p*) (−O_2, +*arcA vs.* −O_2, +*arcA*)	-Fold (+O_2, +*arcA vs.* −O_2, +*arcA*)	-Fold (−O_2, +*arcA vs.* −O_2, −*arcA*)	Predicted ArcA site[a]
ycdC (b1013)	1.58E-05	0.9999039	2.90E-04	0.9994196	−1.76	2.53	
rpsA (b0911)	7.17E-05	0.9997069	2.52E-04	0.9994788	−1.91	2.90	
rplX (b3309)	4.68E-10	1.0000000	1.26E-04	0.9996910	−3.35	2.98	
rpsT (b0023)	1.90E-05	0.9998898	2.65E-04	0.9994589	−2.26	3.20	
talA (b2464)	2.21E-04	0.9993279	2.16E-04	0.9995354	−1.64	3.46	
prlA (b3300)	2.35E-07	0.9999957	7.76E-07	0.9999935	−2.12	3.49	
crr (b2417)	2.14E-09	0.9999999	5.06E-05	0.9998458	−3.23	3.63	
rbsC (b3750)	2.89E-07	0.9999949	7.32E-05	0.9997959	−2.07	3.76	
oppA (b1243)	8.78E-05	0.9996597	3.03E-10	1.0000000	−1.66	3.79	113, 324, 472
rpsJ (b3321)	6.28E-06	0.9999512	1.09E-04	0.9997244	−2.10	3.80	
pal (b0741)	3.90E-08	0.9999988	1.45E-05	0.9999403	−2.64	3.91	301
tufA (b3339)	1.59E-07	0.9999967	6.16E-14	1.0000000	−1.66	4.01	
ompA (b0957)	1.48E-06	0.9999831	3.86E-06	0.9999781	−2.27	4.02	
yajG (b0434)	7.73E-05	0.9996902	2.24E-05	0.9999169	−1.95	4.17	
rbsD (b3748)	3.85E-08	0.9999989	7.40E-05	0.9997942	−2.81	4.83	
speD (b0120)	6.41E-06	0.9999504	5.07E-05	0.9998454	−2.30	5.11	
yjcU (b4085)	1.81E-07	0.9999964	2.81E-04	0.9994337	−1.83	5.20	
yceD (b1088)	2.00E-10	1.0000000	3.45E-05	0.9998847	−2.90	5.27	
rplS (b2606)	3.02E-05	0.9998450	4.23E-08	0.9999993	−2.23	5.81	
atpG (b3733)	2.56E-07	0.9999954	1.04E-04	0.9997345	−1.94	6.07	124, 385
speE (b0121)	7.18E-06	0.9999461	5.70E-06	0.9999705	−2.10	6.32	
yfiA (b2597)	7.16E-11	1.0000000	4.68E-06	0.9999746	−10.41	7.49	
nuoE (b2285)	2.34E-09	0.9999999	2.61E-06	0.9999837	−2.79	8.83	149, 190
rplT (b1716)	4.41E-06	0.9999624	3.48E-05	0.9998838	−2.78	9.01	
gatY (b2096)	3.49E-07	0.9999942	7.81E-06	0.9999626	−2.74	10.72	515, 520
icdA (b1136)	1.03E-11	1.0000000	1.38E-05	0.9999423	−2.74	14.04	111
sdhA (b0723)	2.06E-07	0.9999961	3.90E-05	0.9998733	−4.03	14.54	69, 267, 330
lpdA (b0116)	1.29E-11	1.0000000	2.29E-06	0.9999852	−4.75	15.29	219, 230, 232
gpmA (b0755)	1.27E-09	0.9999999	3.85E-06	0.9999781	−7.18	16.95	
rplM (b3231)	3.40E-07	0.9999943	3.07E-06	0.9999816	−2.57	17.05	
mdh (b3236)	4.00E-04	0.9989605	2.89E-04	0.9994210	−1.84	17.95	229
nuoB (b2287)	1.32E-04	0.9995414	2.12E-04	0.9995422	−6.48	19.34	149, 190
trpB (b1261)	3.12E-10	1.0000000	4.43E-05	0.9998606	−2.85	19.97	46
cyoA (b0432)	9.70E-10	0.9999999	5.06E-05	0.9998457	−9.98	23.30	62, 82, 235, 246
sdhB (b0724)	1.25E-05	0.9999188	2.09E-04	0.9995476	−2.52	27.87	69, 267, 330
sucD (b0729)	2.42E-05	0.9998684	7.28E-05	0.9997967	−5.04	86.14	69, 267, 330
gltA (b0720)	3.09E-05	0.9998421	3.80E-06	0.9999783	−2.60	107.01	348

[a] Distance is upstream from the start codon of the gene or the first gene of the operon (if internal). Sites are predicted from the gene promoter. Other putative ArcA-binding sites may also be predicted upstream of a secondary promoter, but are not mentioned here.

agreement with our previous study (1), *lacZ* fusion studies have shown that their anaerobic induction is due solely to the presence of the *arcA* gene product,[2] but only the *gadA* gene has a predicted ArcA-binding sites upstream of its start codon. The *gadX* and *gadW* genes also clustered into pattern I. These two genes encode transcription factors that control the expression of the *gadA* and *gadBC* operons (40–43). A putative ArcA-binding site(s) was identified upstream of each of these two genes (Table II). Two other genes, *rhaA* (L-rhamnose isomerase) and *glgC* (glucose-1-phosphate adenylyltransferase), have not been shown previously to be regulated by ArcA.

Six genes of this expression pattern belong to the small molecule transport functional class. Four of these genes were shown previously to be subject to FNR-mediated regulation (1). These genes are *yabM* (*setA*, glucose/lactose efflux transporter), *yadQ* (*clcA*, mammalian chloride channel protein homolog), *nanT* (sialic acid transporter), and *uraA* (transport of uracil). The remaining two genes belonging to this group are *nfrA* (an outer membrane protein) and *pnuC* (nicotinamide mononucleotide transporter).

As in our previous study (1), several genes of this expression pattern belonging to the macromolecular synthesis class are for DNA repair: *recB* and *recC* (subunits of the RecBCD enzyme complex), *dinG* (encodes a LexA-regulated DNA repair enzyme), and *sbcC* (co-suppressor of *recBC* mutations). Of the remaining five genes belonging to this functional group, only one was also observed to be regulated by FNR: *glgA* (glycogen synthesis). The other four genes are *degQ* (*hhoA*, periplasmic serine endopeptidase); *cdh* (CDP-diglyceride hydrolase); and two hydrogenase-encoding genes, *hycD* (hydrogenase-3 subunit) and *hyaB* (hydrogenase-1 subunit). Putative ArcA-binding sites were identified upstream of *recB* and *hycD* (Table II).

Of the remaining genes clustered into this expression pattern, two genes, *mrcA* (penicillin-binding protein 1A) and *rarD* (involved in chloramphenicol resistance), were also observed to be regulated by FNR (1). Two other genes, organized in an operon encoding a putative alternative cytochrome oxidase, *appCB* (*cbdAB*), were not observed previously to be regulated by FNR (1), and *xylR* (regulatory gene for the xylose operon) also clustered into this expression pattern. The 23 remaining members of this expression pattern are currently uncharacterized, 12 of which were also previously observed to be regulated by FNR (1). A putative ArcA-binding site was identified upstream of *rarD* and *xylR* and upstream of 2 of the 23 previously uncharacterized genes (*ydbA* and *yhjE*).

Only one gene in this expression pattern, *glcC*, was also identified in the study by Liu and De Wulf (22); however, their results indicated that *glcG* is repressed by ArcA (2.6-fold). Liu and De Wulf identified a total of 138 genes as being activated in the presence of ArcA. Again, in our gold standard set, we identified a total of 42 genes as being up-regulated in an *arcA* mutant (patterns II, VII, and VIII) or 57 genes in pattern II. However, our clustering set of 175 genes is highly restricted, with a strict PPDE(<*p*) cutoff level of 0.997.

Expression Pattern III: Decreased Expression during Anaerobiosis and No Change in an ArcA Strain—34 genes clustered into expression pattern III. Of these, 23 clustered into the same expression pattern in our previous study (1), indicating that

TABLE II

Regulatory pattern II: genes that exhibit increased levels during anaerobic growth and further decreased levels in an ArcA-deficient strain

Gene name (NIH) and b no.	*p* value ($+O_2$, +*arcA vs.* $-O_2$, +*arcA*)	PPDE(<*p*) ($-O_2$, +*arcA vs.* $-O_2$, −*arcA*)	*p* value ($-O_2$, +*arcA vs.* $-O_2$, −*arcA*)	PPDE(<*p*) ($-O_2$, +*arcA vs.* $-O_2$, −*arcA*)	-Fold ($+O_2$, +*arcA vs.* $-O_2$, +*arcA*)	-Fold ($-O_2$, +*arcA vs.* $-O_2$, −*arcA*)	Predicted ArcA site[a]
yhjE (b3523)	1.29E-04	0.9995485	1.79E-07	0.9999979	2.20	−50.94	202
yabM (b0070)	3.01E-04	0.9991577	7.13E-07	0.9999939	2.30	−29.69	
pyrD (b0945)	5.36E-06	0.9999565	4.37E-06	0.9999759	4.95	−18.91	
nfrA (b0568)	1.87E-05	0.9998909	4.71E-06	0.9999745	4.71	−16.63	
yadQ (b0155)	2.80E-04	0.9992013	3.23E-07	0.9999967	1.99	−16.01	
dinG (b0799)	3.20E-04	0.9991182	1.60E-07	0.999998	1.83	−11.79	
yhhT (b3474)	3.06E-04	0.9991471	5.24E-07	0.9999952	2.14	−11.26	
gadB (b1493)	1.87E-08	0.9999993	1.47E-06	0.9999895	23.98	−11.23	
gadA (b3517)	5.14E-07	0.9999923	2.50E-05	0.9999096	22.98	−9.44	567
glnD (b0167)	8.77E-05	0.9996601	2.46E-06	0.9999844	2.55	−8.58	560
mobB (b3856)	5.57E-06	0.9999553	8.68E-07	0.9999929	3.26	−7.83	
yadR (b0156)	2.03E-05	0.9998843	2.00E-06	0.9999867	3.60	−7.47	
yafZ (b0252)	2.57E-04	0.9992505	1.64E-06	0.9999886	2.04	−7.38	
aroM (b0390)	1.81E-04	0.9994207	1.72E-06	0.9999881	2.19	−7.13	
wecE (b3791)	1.34E-04	0.9995359	3.41E-06	0.99998	2.29	−6.48	
yghQ (b2983)	4.25E-04	0.9989143	2.41E-06	0.9999847	1.98	−6.38	
tra5_2 (b0541)	1.38E-04	0.9995245	5.22E-07	0.9999952	1.90	−6.31	
pnuC (b0751)	1.71E-05	0.9998978	9.66E-06	0.9999561	3.38	−6.22	
B1172 (b1172)	1.20E-07	0.9999974	1.71E-05	0.9999322	16.47	−6.14	
gadX (b3516)	2.32E-06	0.9999766	1.31E-04	0.9996818	16.09	−6.11	1,227,238,249
mrcA (b3396)	2.50E-04	0.9992655	1.86E-06	0.9999874	2.00	−6.09	
yhjJ (b3527)	2.90E-04	0.9991791	7.03E-06	0.9999655	2.08	−5.77	
sbcC (b0397)	3.89E-04	0.9989827	1.59E-05	0.9999357	2.17	−5.59	
yhjW (b3546)	1.57E-05	0.999904	2.46E-07	0.9999973	2.15	−5.58	
yhjD (b3522)	8.40E-05	0.9996706	4.68E-05	0.9998545	3.01	−5.18	
xylR (b3569)	3.55E-04	0.9990494	4.52E-06	0.9999753	1.92	−5.11	112
gadW (b3515)	2.41E-05	0.9998687	1.06E-04	0.99973	3.71	−4.63	131
recC (b2822)	2.98E-05	0.9998462	1.08E-07	0.9999985	1.78	−4.62	
yidE (b3685)	1.72E-04	0.999442	2.31E-06	0.9999851	1.82	−4.36	
yheF (b3325)	2.72E-04	0.999217	1.57E-05	0.9999366	2.01	−4.36	
B2866 (b2866)	5.29E-06	0.999957	9.06E-07	0.9999927	2.10	−4.03	
appC (b0978)	4.83E-09	0.9999998	7.60E-07	0.9999936	3.50	−4.01	
speC (b2965)	2.86E-04	0.9991885	2.02E-06	0.9999866	1.68	−2.97	
glgA (b3429)	1.74E-04	0.9994371	5.85E-06	0.99997	1.85	−2.90	
yhhJ (b3485)	3.34E-04	0.99909	8.61E-05	0.9997692	2.26	−2.79	
narY (b1467)	3.93E-05	0.9998118	1.23E-05	0.9999472	2.39	−2.72	
recB (b2820)	9.51E-06	0.9999337	1.79E-06	0.9999878	1.97	−2.68	
yjiE (b4327)	1.32E-04	0.9995418	1.13E-07	0.9999985	1.53	−2.65	
glnE (b3053)	5.30E-05	0.9997654	2.60E-06	0.9999838	1.80	−2.55	
degQ (b3234)	4.54E-04	0.9988598	1.20E-04	0.9997038	2.10	−2.47	
nanT (b3224)	7.22E-05	0.9997053	1.01E-05	0.9999546	1.87	−2.46	
appB (b0979)	3.31E-09	0.9999998	2.48E-05	0.9999101	9.26	−2.43	
rhaA (b3903)	1.48E-04	0.9994999	3.29E-06	0.9999806	1.67	−2.41	
cdh (b3918)	4.00E-04	0.9989608	3.09E-04	0.9993915	2.42	−2.28	
hycD (b2722)	1.50E-05	0.9999073	2.65E-05	0.9999055	2.40	−2.21	55
rarD (b3819)	3.61E-04	0.999036	7.45E-05	0.9997932	1.99	−2.19	395
ydbA_2 (b1405)	9.98E-08	0.9999977	1.83E-06	0.9999876	2.33	−2.94	95
hdeA (b3510)	4.82E-09	0.9999998	1.76E-04	0.9996025	40.69	−2.83	
hyaB (b0973)	7.30E-08	0.9999982	2.57E-04	0.9994709	4.50	−2.83	
tdh (b3616)	3.20E-05	0.9998381	6.46E-05	0.9998144	1.92	−2.62	
uraA (b2497)	7.91E-05	0.9996848	1.14E-04	0.999715	1.97	−2.60	
yjcS (b4083)	9.51E-05	0.9996391	3.36E-05	0.999887	1.72	−2.58	
tynA (b1386)	1.73E-04	0.9994387	2.02E-04	0.9995585	1.82	−2.52	
yhdR (b3246)	2.96E-04	0.9991674	1.30E-04	0.9996846	1.69	−2.42	
yphB (b2544)	2.71E-04	0.99922	2.88E-04	0.9994224	1.86	−2.38	
glgC (b3430)	4.69E-04	0.9988327	5.35E-05	0.9998391	1.45	−2.15	
yhgF (b3407)	3.41E-04	0.9990762	1.65E-04	0.9996211	1.50	−2.10	

[a] Distance is upstream from the start codon of the gene or the first gene of the operon (if internal). Sites are predicted from the gene promoter. Other putative ArcA-binding sites may also be predicted upstream of a secondary promoter, but are not mentioned here.

the expression of these genes, although decreased during anaerobiosis, is not regulated by either ArcA or FNR.

Two members of the *nuoA–N* operon, *nuoG* and *nuoE*, which encode NdhI, a membrane-associated, multisubunit, proton-translocating enzyme similar to complex I of eukaryotic mitochondria (37), clustered into pattern I. Expression of the *nuoE* gene (Table III) was 3.8-fold lower under anaerobic conditions and was elevated 8.8-fold in the ArcA mutant (see Table IX). A previous study using *nuo-lacZ* fusions established that *nuo* expression is subject to ArcA-mediated anaerobic repression and NarL nitrate-mediated anaerobic activation (38). Two other members of this operon clustered into pattern I (*nuoB* and *nuoE*) (Table I).

The *cydA* gene (part of the *cydAB* operon) encodes the high affinity terminal oxidase of the oxygen respiratory chain, cytochrome *d* oxidase. The data obtained here show that *cydA* was repressed ~2-fold during anaerobic growth, but was unchanged in the ArcA-deficient strain (Table III). In agreement with these findings, previous studies using *cydA*::*lacZ* fusions showed that transcription of the *cydAB* operon is ArcA-repressed when oxygen becomes limiting (16, 44, 45). Subsequent studies have shown that ArcA functions to anti-repress *cydAB* transcription when oxygen is limiting (46), whereas FNR is required for repression when the oxygen tension is decreased further (14, 17, 45). As our study was carried out in full anaerobiosis, the ArcA effect was not observed, but the FNR effect

TABLE III

Regulatory pattern III: genes that exhibit decreased levels during anaerobic growth that are unaffected in an ArcA-deficient strain

Gene name (NIH) and b no.	*p* value ($+O_2$, +*arcA vs.* $-O_2$, +*arcA*)	PPDE(<*p*) ($-O_2$, +*arcA vs.* $-O_2$, −*arcA*)	*p* value ($-O_2$, +*arcA vs.* $-O_2$, −*arcA*)	PPDE(<*p*) ($-O_2$, +*arcA vs.* $-O_2$, −*arcA*)	-Fold ($+O_2$, +*arcA vs.* $-O_2$, +*arcA*)	-Fold ($-O_2$, +*arcA vs.* $-O_2$, −*arcA*)	Predicted ArcA site[a]
ylcB (b0572)	4.14E-10	1	6.21E-01	0.7667727	−8.17	1.15	
eaeH (b0297)	4.12E-04	0.9989377	4.30E-01	0.8257776	−2.33	−1.88	
ycgC (b1198)	2.10E-08	0.9999993	4.82E-01	0.8088967	−2.68	−1.19	
rplB (b3317)	3.15E-07	0.9999946	7.33E-01	0.7359018	−2.25	−1.07	
rplC (b3320)	2.61E-06	0.9999744	5.14E-01	0.7986594	−2.15	−1.14	
rplO (b3301)	6.12E-07	0.9999912	9.23E-01	0.6887762	−2.15	1.02	
hflC (b4175)	9.53E-06	0.9999336	6.43E-01	0.760619	−2.10	1.14	
rplF (b3305)	2.06E-06	0.9999785	4.04E-01	0.8343415	−2.08	1.18	
rplQ (b3294)	1.40E-04	0.999521	3.89E-01	0.8393984	−2.05	−1.31	
rplI (b4203)	2.21E-04	0.9993285	8.49E-01	0.7062438	−2.02	−1.05	
rpsE (b3303)	1.53E-05	0.9999062	3.74E-01	0.8445234	−2.02	1.23	
yhbM (b3163)	4.07E-06	0.9999645	3.62E-01	0.8488866	−2.01	1.31	
cydA (b0733)	1.14E-05	0.9999241	4.17E-01	0.8302134	−1.98	1.32	149, 385, 404, 586
rho (b3783)	7.29E-07	0.99999	6.73E-01	0.7519987	−1.90	1.10	
prfB (b2891)	2.51E-05	0.9998648	5.42E-01	0.7901032	−1.89	1.12	
rplD (b3319)	6.91E-05	0.9997147	6.09E-01	0.7703721	−1.82	1.16	
fabG (b1093)	1.60E-04	0.9994709	9.12E-01	0.6912062	−1.80	1.03	
rpsH (b3306)	2.22E-04	0.9993253	7.09E-01	0.7423228	−1.80	1.08	
tsf (b0170)	2.86E-05	0.9998508	6.06E-01	0.7712214	−1.76	1.11	
rfbX (b2037)	6.71E-04	0.9984799	7.61E-01	0.7283532	−1.74	−1.09	
rplE (b3308)	5.65E-04	0.9986613	5.96E-01	0.7739222	−1.74	−1.13	
nuoG (b2283)	8.77E-05	0.99966	5.43E-01	0.7897395	−1.72	1.16	149, 190
yfiB (b2605)	5.20E-04	0.9987396	5.51E-01	0.7874321	−1.69	−1.15	
ykgI (b0303)	6.29E-05	0.9997338	9.85E-01	0.6746519	−1.63	−1.00	24
nfi (b3998)	3.46E-04	0.9990659	7.21E-01	0.7389198	−1.62	−1.09	
tig (b0436)	1.15E-04	0.9995855	4.12E-01	0.8316477	−1.60	1.16	
katE (b1732)	6.37E-05	0.9997312	8.83E-01	0.6980209	−1.60	1.03	
nuoF (b2284)	1.20E-04	0.9995715	8.99E-01	0.6941889	−1.59	−1.03	
lysS (b2890)	3.12E-04	0.9991345	9.00E-01	0.6939597	−1.58	1.02	
ycdT (b1025)	2.67E-04	0.9992291	9.51E-01	0.6821248	−1.56	−1.01	
B1605 (b1605)	2.74E-04	0.9992143	3.93E-01	0.8381038	−1.54	1.25	
B1808 (b1808)	3.15E-04	0.9991286	4.61E-01	0.8156317	−1.48	1.14	
yabI (b0065)	2.80E-04	0.9992007	7.45E-01	0.7327525	−1.44	−1.05	
yjiT (b4342)	2.99E-04	0.9991618	3.84E-01	0.841103	−1.42	−1.15	

[a] Distance is upstream from the start codon of the gene or the first gene of the operon (if internal). Sites are predicted from the gene promoter. Other putative ArcA-binding sites may also be predicted upstream of a secondary promoter, but are not mentioned here.

TABLE IV

Regulatory pattern IV: genes that exhibit increased levels during anaerobic growth that are unaffected in an ArcA-deficient strain

Gene name (NIH) and b no.	*p* value ($+O_2$, +*arcA vs.* $-O_2$, +*arcA*)	PPDE(<*p*) ($-O_2$, +*arcA vs.* $-O_2$, −*arcA*)	*p* value ($-O_2$, +*arcA vs.* $-O_2$, −*arcA*)	PPDE(<*p*) ($-O_2$, +*arcA vs.* $-O_2$, −*arcA*)	-Fold ($+O_2$, +*arcA vs.* $-O_2$, +*arcA*)	-Fold ($-O_2$, +*arcA vs.* $-O_2$, −*arcA*)	Predicted ArcA site[a]
htpG (b0473)	3.61E-05	0.9998229	5.51E-01	0.7873323	1.96	−1.12	
mrr (b4351)	2.30E-04	0.9993078	5.10E-01	0.8001226	3.33	−1.19	
cysK (b2414)	2.01E-04	0.9993736	6.04E-01	0.7718185	4.04	1.16	
ybeD (b0631)	6.12E-05	0.9997391	9.99E-01	0.6714978	4.36	−1.00	127,131
ygjD (b3064)	4.08E-04	0.9989455	5.51E-01	0.7874807	5.13	1.34	
cof (b0446)	3.80E-04	0.9990006	9.80E-01	0.6757591	8.65	1.01	

[a] Distance is upstream from the start codon of the gene or the first gene of the operon (if internal). Sites are predicted from the gene promoter. Other putative ArcA-binding sites may also be predicted upstream of a secondary promoter, but are not mentioned here.

TABLE V

Regulatory pattern V: genes that exhibit increased levels during anaerobic growth and further increased levels in an ArcA-deficient strain

Gene name (NIH) and b no.	*p* value ($+O_2$, +*arcA vs.* $-O_2$, +*arcA*)	PPDE(<*p*) ($-O_2$, +*arcA vs.* $-O_2$, −*arcA*)	*p* value ($-O_2$, +*arcA vs.* $-O_2$, −*arcA*)	PPDE(<*p*) ($-O_2$, +*arcA vs.* $-O_2$, −*arcA*)	-Fold ($+O_2$, +*arcA vs.* $-O_2$, +*arcA*)	-Fold ($-O_2$, +*arcA vs.* $-O_2$, −*arcA*)	Predicted ArcA site[a]
ybjX (b0877)	3.46E-05	0.9998284	5.15E-05	0.9998437	4.46	3.77	

[a] Distance is upstream from the start codon of the gene or the first gene of the operon (if internal) Sites are predicted from the gene promoter. Other putative ArcA-binding sites may also be predicted upstream of a secondary promoter, but are not mentioned here.

was observed in our previous study (1). There are three ArcA sites that have been footprinted (17, 31). The study by Liu and De Wulf (22) also identified *cydA* to be ArcA-controlled; however, their study indicated that it is ArcA-activated (5.2-fold).

The remaining 31 genes of this cluster have not been studied previously for their expression under anaerobic growth conditions; however, one contains a putative ArcA-binding site (*ykgI*) (Table III). Again, the genes of this cluster are members of the same functional classes of expression patterns I and II. Three genes (*fabG*, *rfbX*, and *katE*) are involved in small molecule metabolism. 17 genes (*rplB*, *rplC*, *rplO*, *hflC*, *rplF*, *rplQ*, *rplI*, *rpsE*, *rho*, *prfB*, *rplD*, *rpsH*, *tsf*, *rplE*, *nfi*, *tig*, and *lysS*) are involved in macromolecule synthesis or degradation. 10 genes of this cluster are of unclassified function, seven of which were also identified in our FNR study (1). The remaining gene, *eaeH* (homologous to attachment and effacement proteins), also clustered into this expression pattern.

Expression Pattern IV: Increased Expression during Anaerobiosis and No Change in an ArcA Strain—The six genes of this cluster (Table IV) showed elevated expression under anaerobic

TABLE VI

Regulatory pattern VI: genes that exhibit similar levels during aerobic and anaerobic growth but increased levels in an ArcA-deficient strain

Gene name (NIH) and b no.	*p* value ($+O_2$, +*arcA vs.* $-O_2$, +*arcA*)	PPDE(<*p*) ($-O_2$, +*arcA vs.* $-O_2$, −*arcA*)	*p* value ($-O_2$, +*arcA vs.* $-O_2$, −*arcA*)	PPDE(<*p*) ($-O_2$, +*arcA vs.* $-O_2$, −*arcA*)	-Fold ($+O_2$, +*arcA vs.* $-O_2$, +*arcA*)	-Fold ($-O_2$, +*arcA vs.* $-O_2$, −*arcA*)	Predicted ArcA site[a]
potF (b0854)	6.51E-01	0.7200042	1.74E-04	0.9996055	−1.08	3.21	
gapA (b1779)	6.54E-01	0.7190951	7.41E-07	0.9999937	−1.07	3.45	
ydcF (b1414)	9.75E-01	0.6320587	8.14E-05	0.9997787	1.01	4.73	
hisJ (b2309)	8.98E-01	0.6509644	2.04E-06	0.9999865	−1.08	59.58	

[a] Distance is upstream from the start codon of the gene or the first gene of the operon (if internal). Sites are predicted from the gene promoter. Other putative ArcA-binding sites may also be predicted upstream of a secondary promoter, but are not mentioned here.

TABLE VII

Regulatory pattern VII: genes that exhibit decreased levels during anaerobic growth and further decreased levels in an ArcA-deficient strain

Gene name (NIH) and b no.	*p* value ($+O_2$, +*arcA vs.* $-O_2$, +*arcA*)	PPDE(<*p*) ($-O_2$, +*arcA vs.* $-O_2$, −*arcA*)	*p* value ($-O_2$, +*arcA vs.* $-O_2$, −*arcA*)	PPDE(<*p*) ($-O_2$, +*arcA vs.* $-O_2$, −*arcA*)	-Fold ($+O_2$, +*arcA vs.* $-O_2$, +*arcA*)	-Fold ($-O_2$, +*arcA vs.* $-O_2$, −*arcA*)	Predicted ArcA site[a]
rpmC (b3312)	2.45E-09	0.9999998	6.70E-05	0.9998091	−2.20	−8.46	
ybdE (b0575)	2.33E-04	0.9993019	1.70E-06	0.9999882	−1.77	−6.20	
frdA (b4154)	1.41E-07	0.999997	1.44E-06	0.9999896	−2.04	−4.56	
nirB (b3365)	6.62E-07	0.9999907	3.35E-07	0.9999966	−2.04	−4.26	
ylcD (b0574)	2.00E-08	0.9999993	2.59E-05	0.9999072	−2.04	−2.79	

[a] Distance is upstream from the start codon of the gene or the first gene of the operon (if internal). Sites are predicted from the gene promoter. Other putative ArcA-binding sites may also be predicted upstream of a secondary promoter, but are not mentioned here.

growth conditions, but were not affected by deletion of the *arcA* allele. Two genes of unknown function clustered into this group (*ybeD* and *ygjD*) and also clustered into the same group in our FNR study (1). The remaining members of this cluster include *htpG* (a heat shock protein), *mrr* (involved in the restriction of methylated adenine residues; also clustered into this group in Ref. 1), *cysK* (cysteine synthase), and *cof* (complementation of *fur*). A search of the promoter regions of these six genes identified a putative ArcA-binding site upstream of one of these genes: *ybeD*. None of these genes were identified by Liu and De Wulf (22).

Expression Pattern V: Increased Expression during Anaerobiosis and Increased Expression in an ArcA Strain—This cluster contains only a single gene of unknown function: *ybjX* (Table V). A similar pattern of expression was also observed previously (1).

Expression Pattern VI: No Change during Anaerobiosis and Increased Expression in an ArcA Strain—Of the four genes of this cluster, three are involved in small molecule metabolism and transport: *gapA* (structural gene for glyceraldehyde-3-phosphate dehydrogenase A, essential for glycolysis), *potF* (member of the *potFGHI* operon involved in the transport of putrescine), and *hisJ* (member of the *hisTJQMP* operon encoding a histidine-binding protein that is part of the periplasmic permeases for the high affinity uptake of histidine). The final member, *ydcF*, is currently uncharacterized. All four members of this expression pattern clustered into the same group in our FNR study (1).

Expression Pattern VII: Decreased Expression during Anaerobiosis and Decreased Expression in an ArcA Strain—The same five genes observed in this expression pattern were also observed in our study with FNR (1). Two of the genes, *frdA* and *nirB*, have been shown previously to be FNR-regulated (47–49). As we discussed previously (1), the discrepancy in these data is likely due to paralogs in the genome with these two genes (*sdhA* to *frdA* and *nirD*, *cysI* and *cysJ* to *nirB*). The remaining genes include *rpmC* (ribosomal protein) and two uncharacterized genes, *ybdE* (*cusB*) and *ylcD* (*cusA*).

Expression Pattern VIII: No Change during Anaerobiosis and Decreased Expression in an ArcA Strain—This cluster contains 31 genes, 20 of which are of unknown function. (12 were also identified in our previous FNR study (1).) Of the 31 genes of known function (Table VIII), two are known to be regulated by oxygen and/or ArcA under anaerobic growth conditions, and four contain putative ArcA-binding sites.

The two genes reported to be regulated by oxygen and/or ArcA are *fumB* and *lysU*. The anaerobic fumarase, encoded by *fumB*, is known to be activated during anaerobic fermentative growth (50, 51), and Tseng (51) showed that both ArcA and FNR are responsible for this anaerobic activation. As stated in our previous work (1), although our microarray data indicate that *fumB* is not regulated with respect to oxygen, its presence in this expression pattern is probably a result of the high sequence identity (80%) between *fumB* and the aerobically expressed fumarase, *fumA*. The *lysU* gene encodes one of the two lysyl-tRNA synthetases (the other being *lysS*, with which it shares 79% sequence identity (52)) and was reported previously to be induced under anaerobic conditions (53).

Eight members of this expression pattern are involved in macromolecular metabolism: *cvpA* (colicin V production), *aceK* (isocitrate dehydrogenase kinase/phosphatase), *ftsY* (cell division), *dnaX* (subunit of DNA polymerase III), *umuC* (involved in mutation induction), *menD* (*o*-succinylbenzoate synthase I), *degS* (periplasmic serine endopeptidase), and *tyrB* (tyrosine aminotransferase). The final member, *fhuC* (ferric hydroxamate-dependent iron uptake), is involved in small molecule transport. The remaining 20 genes of this expression pattern have not been characterized. Putative ArcA-binding sites were identified upstream of *ycdM* and *yjhH* (Table VIII).

The functional class distribution of the 175 genes of regulatory patterns I—VIII is shown in Fig. 6. Roughly 37.7% are hypothetical or unclassified, whereas another 23.4% are involved in small molecule metabolism. Most of the previously documented oxygen-controlled genes fall into the category of carbon and energy metabolism (5%). The study by Liu and De Wulf (22) identified 58 new genes/operons that are implicated in energy metabolism, transport, survival, catabolism, and transcriptional regulation.

Genes Not Expressed in at Least One Experiment

Only those genes exhibiting an expression level greater than zero in all experiments were used for statistical analysis. To identify differentially expressed genes that were not expressed under one condition but turned on under another treatment condition (or vice versa), gene measurements containing zero expression values were set aside and are listed in Table IX. This set contains only eight genes with expression values of at

TABLE VIII

Regulatory pattern VIII: genes that exhibit similar levels during aerobic and anaerobic growth but decreased levels in an ArcA-deficient strain

Gene name (NIH) and b no.	*p* value ($+O_2$, +*arcA vs.* $-O_2$, +*arcA*)	PPDE(<*p*) ($-O_2$, +*arcA vs.* $-O_2$, −*arcA*)	*p* value ($-O_2$, +*arcA vs.* $-O_2$, −*arcA*)	PPDE(<*p*) ($-O_2$, +*arcA vs.* $-O_2$, −*arcA*)	-Fold ($+O_2$, +*arcA vs.* $-O_2$, +*arcA*)	-Fold ($-O_2$, +*arcA vs.* $-O_2$, −*arcA*)	Predicted ArcA site[a]
yihR (b3879)	6.68E-01	0.7149639	1.11E-04	0.9997199	−1.08	−6.17	
ybdL (b0600)	5.27E-01	0.760456	2.57E-04	0.9994711	1.14	−5.74	
cvpA (b2313)	5.47E-01	0.7537703	9.86E-07	0.9999922	1.08	−5.42	
yraP (b3150)	6.58E-01	0.7179875	7.04E-06	0.9999654	1.07	−5.40	
ycfT (b1115)	5.72E-01	0.745473	5.52E-05	0.9998352	1.10	−4.92	
ybhK (b0780)	9.92E-01	0.6279724	1.21E-05	0.9999478	1.00	−4.89	
yjfJ (b4182)	8.21E-01	0.6711207	2.11E-04	0.9995436	1.04	−2.85	
lysU (b4129)	9.48E-01	0.6386061	2.70E-05	0.9999043	−1.01	−2.17	201
B1012 (b1012)	7.62E-01	0.687143	3.39E-05	0.9998861	−1.04	−2.14	61
yjiL (b4334)	5.29E-01	0.7597633	4.46E-04	0.9991964	1.10	−2.14	
aceK (b4016)	5.95E-01	0.7376482	4.26E-04	0.9992244	1.09	−2.02	
ftsY (b3464)	9.83E-01	0.6300542	4.12E-05	0.9998679	−1.00	−2.64	
dnaX (b0470)	7.30E-01	0.6963859	2.27E-04	0.9995187	−1.04	−2.58	
B2512 (b2512)	6.92E-01	0.7076898	9.52E-05	0.9997508	−1.04	−2.54	
umuC (b1184)	5.60E-01	0.7493436	1.14E-04	0.9997144	1.07	−2.46	
yjhS (b4309)	7.49E-01	0.6908652	3.74E-04	0.9992973	1.04	−2.46	
yqiG (b3046)	9.91E-01	0.6283205	2.65E-04	0.9994591	1.00	−2.45	
yjcP (b4080)	7.28E-01	0.6971228	2.24E-04	0.9995225	1.04	−2.45	
menD (b2264)	8.94E-01	0.6518572	1.53E-04	0.9996433	1.01	−2.36	
degS (b3235)	8.28E-01	0.6692454	4.71E-04	0.9991623	−1.03	−2.36	
yi41 (b4278)	7.46E-01	0.6918869	9.04E-06	0.9999582	−1.03	−2.33	
yhcQ (b3241)	8.82E-01	0.6550192	4.92E-04	0.9991338	1.02	−2.27	
tyrB (b4054)	8.47E-01	0.6641473	4.20E-04	0.9992319	−1.02	−2.23	
yjiN (b4336)	7.61E-01	0.6874317	3.15E-05	0.9998923	1.03	−2.14	
fumB (b4122)	6.59E-01	0.7176352	9.68E-05	0.9997477	1.04	−2.13	242, 260, 370, 376, 381
yhjX (b3547)	7.62E-01	0.687415	4.65E-04	0.9991711	−1.04	−2.13	
fhuC (b0151)	7.66E-01	0.6861368	4.85E-04	0.9991431	1.03	−2.06	
yjgR (b4263)	8.42E-01	0.6655468	3.84E-04	0.999283	−1.02	−2.00	
yaaJ (b0007)	5.41E-01	0.7558279	2.99E-04	0.999406	1.05	−1.96	
yjhH (b4298)	6.15E-01	0.7314819	4.94E-04	0.999132	−1.05	−1.93	211, 597
yaaU (b0045)	9.99E-01	0.6262859	4.68E-04	0.9991658	−1.00	−1.91	

[a] Distance is upstream from the start codon of the gene or the first gene of the operon (if internal). Sites are predicted from the gene promoter. Other putative ArcA-binding sites may also be predicted upstream of a secondary promoter, but are not mentioned here.

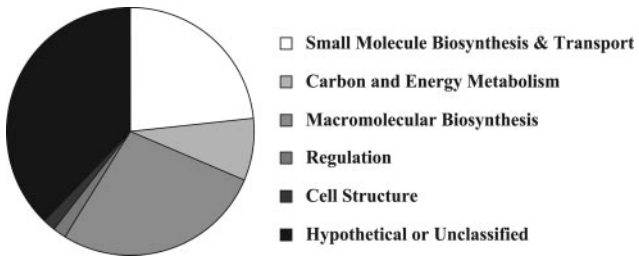

FIG. 6. **Distribution of functions for genes affected by oxygen availability and ArcA.** The distribution of the 175 genes with PPDE(<*p*) values >0.99 and *p* values <0.0013 is as follows: small molecule biosynthesis and transport, 41; carbon and energy metabolism, 14; macromolecular biosynthesis, 48; regulation, three; cell structure, three; and hypothetical or unclassified, 66.

least 1×10^{-4} of total mRNA for all measurements in at least one experiment with a coefficient of variance <0.2 (Table IX). Seven are members of pattern II (increased expression during anaerobiosis and decreased expression in an ArcA strain): *yddS*, *ftsW*, *hyaD*, *ldcC*, *ybdA*, *yhgE*, and *yrbF*. The remaining gene, *frdB*, is a member of pattern VII (decreased expression during anaerobiosis and decreased expression in a ArcA strain). Two of these genes contain putative ArcA-binding sites: *yddS* and *yhgE* (Table IX).

Venn Diagram

To better visualize the interaction between the oxygen, ArcA and FNR regulons, Venn diagrams were created (Fig. 7). The top 500 genes (sorted by *p* value) from each data set were used as in the construction of the 175-gene list described above and the 205-gene list from our previous study (1). Interestingly, 303 genes were found to be regulated by both ArcA and FNR, and 74 of these genes showed additional regulation by oxygen (Fig. 7*A*). This is in contrast to the 16 genes reported previously to be co-regulated (5, 6).

In looking at the top 500 genes from each group, 48 genes were identified as being subject solely to ArcA regulation and 57 solely to FNR regulation under anaerobic conditions. The remaining 321 genes pose an interesting question as to whether or not another global oxygen regulator that has yet to be identified exists within the *E. coli* genome. Moreover, the 378 genes in the ArcA grouping and the 369 genes in the FNR grouping that do not show regulation by oxygen, but that are regulated by each of these proteins (or co-regulated) under anaerobic conditions, suggest that these two proteins may also be important for adaptation to the anaerobic environment. It is also important to note that a large proportion of the 515 genes in this latter group are currently of unknown function. In addition to the comparisons above, a second comparison between the ArcA, FNR, and Lrp (25) regulons was also done (Fig. 7*B*), as we had indicated previously an overlap between the FNR and Lrp data sets (1). This diagram reveals 48 genes overlapping between the Lrp and FNR regulons, 43 genes overlapping between the ArcA and Lrp regulons, and 26 genes overlapping between all three (data not shown). These comparisons strongly suggest that regulatory networks are more complex than described previously.

Comparison with Other Studies

When different array formats are used, the magnitudes and sources of experimental errors are surely different. This raises the question of whether or not results obtained from experiments performed with different DNA array formats such as pre-synthesized filter arrays and *in situ* synthesized Affymetrix GeneChips can be compared with one another. We have previously addressed this question. Hung *et al.* (25) compared the results of 4-fold replicated gene expression profiles of

TABLE IX
Genes not expressed in at least one experiment

Expression pattern	Gene name (NIH) and b no.	*p* value		-Fold		CV[a]			Predicted ArcA site[b]
		$+O_2$ *vs.* $-O_2$	$-O_2$ *vs.* $-arcA$	$+O_2$ *vs.* $-O_2$	$-O_2$ *vs.* $-arcA$	$+O_2$	$-O_2$	$-arcA$	
II	*ldcC* (b0186)	8.41E-03	3.66E-03	1.95	−1374.34	0.15	0.24	NA	
II	*ftsW* (b0089)	1.93E-02	6.98E-04	1.47	−914.53	0.22	0.14	NA	
II	*ybdA* (b0591)	1.17E-02	9.79E-03	2.65	−1188.68	0.20	0.34	NA	
II	*yrbF* (b3195)	3.68E-02	1.40E-02	2.10	−823.03	0.15	0.39	NA	
II	*yhgE* (b3402)	1.78E-02	1.66E-02	3.06	−943.72	0.20	0.41	NA	172
II	*yddS* (b1487)	5.91E-04	1.45E-03	2.63	−604.07	0.18	0.18	NA	103
II	*hyaD* (b0975)	1.60E-06	2.21E-04	26.14	−1262.69	1.21	0.09	NA	
VII	*frdB* (b4153)	7.45E-03	1.28E-03	−2.35	−503.68	0.28	0.17	NA	

[a] CV, coefficient of variance; NA, not applicable.
[b] Distance is upstream from the start codon of the gene or the first gene of the operon (if internal). Sites are predicted from the gene promoter. Other putative ArcA-binding sites may also be predicted upstream of a secondary promoter, but are not mentioned here.

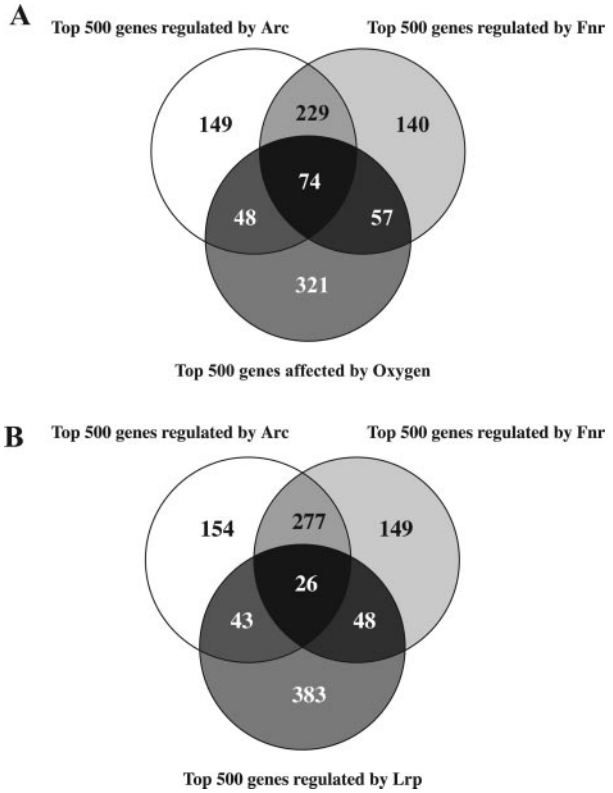

FIG. 7. **Venn diagrams.** *A*, Venn diagram for oxygen-, ArcA-, and FNR-regulated genes. The top 500 genes for ArcA (*upper left circle*), FNR (*upper right circle*), and oxygen (*lower circle*) are shown. *B*, Venn diagram for ArcA-, FNR-, and Lrp-regulated genes. The top 500 genes for ArcA (*upper left circle*), FNR (*upper right circle*), and Lrp (*lower circle*) are shown.

otherwise wild-type and *lrp* isogenic *E. coli* strains performed with these two DNA microarray formats. To emphasize variance due to format differences, the same RNA samples were used for target preparation for both formats, and the data were analyzed with Cyber-T software as described here. When the top 100 genes with the lowest *p* values obtained with each format were compared, a highly significant number of genes, 29, were in common.

Liu and De Wulf (22) have reported the transcriptional profiles of $arcA^+$ and $arcA^-$ *E. coli* cells grown under anaerobic conditions and generously provided us with their raw data. A comparison of this Affymetrix GeneChip data with our filter array data, both analyzed with Cyber-T software, does not show significant agreement. Of the top 100 genes with the lowest *p* values (<0.018) obtained with each format, only three genes were in common. Because Liu and De Wulf use a different data analysis software package (Spotfire) and defined differentially expressed genes as those with an expression level coefficient of variance <0.8 and a mutant to wild-type signal ratio of >2 with $p < 0.05$, it is not possible to directly compare their results with the results presented here. In addition, Liu and De Wulf also used a different carbon source (xylose rather than glucose). We can, however, compare conclusions. They reported 58 differentially expressed genes of operons under the direct control of ArcA as evidenced by the presence of a documented or putative DNA-binding site. In our data set, these genes exhibit *p* values ranging from 3.8×10^{-6} to 0.9 and PPDE(*p*) values ranging from 1.0 to 3.2×10^{-8}. This suggests many false negatives and false positives in the data set of Liu and De Wulf.

Implications for Genome-wide Control by ArcA and FNR

In this study, we employed statistical methods (1, 25) for the identification of differentially expressed genes based on experiment-wide false positive and false negative measurement levels. These methods previously allowed us to infer differential expression for more than one-third of the 4290 genes of *E. coli* during growth in the presence or absence of oxygen (1445 genes) (1). This study has allowed us to determine that ~1243 of these changes in expression level are mediated either directly or indirectly by ArcA (Fig. 2*B*). These results further support our previous conclusions (1) that the network of genes required for the transition of cells from aerobic to anaerobic growth conditions is as much as 10 times larger than previously suspected. A comparison of the ArcA and FNR gold standard sets showed that 303 genes were regulated by both proteins (Fig. 7*A*), 74 of which were also affected by oxygen. Previous to this study, only 16 genes had been reported to be co-regulated (5, 6). Therefore, as suggested previously by us (1) and Liu and De Wulf (22), the total number of genes directly activated or repressed by ArcA and FNR is likely to be much higher than documented previously.

Rationale of Regulatory Patterns

Regulatory pattern I (anaerobic repressed gene expression, *i.e.* decreased expression in the presence of ArcA) (Table I) and pattern II (anaerobic activated gene expression, *i.e.* increased expression in the presence of ArcA) (Table II) are most easily reconciled with previous reports. Of the 94 genes of these patterns, 24 contain known or putative ArcA-binding site motifs. These results suggest that we might expect the total number of genes directly activated or repressed by ArcA to be in the range of 290 genes. Liu and De Wulf (22) estimated 372 genes.

Regulatory pattern III (anaerobically repressed, but not affected by ArcA) and pattern IV (anaerobically activated, but not affected by ArcA) are most easily explained as genes controlled by the FNR protein or by an as yet unidentified global regulator such as Lrp, IHF, FIS, or H-NS. Only two of these genes, *nuoG* and *nuoF*, are known members of the ArcA regulon.

As in our previous work (1), it is more difficult to understand

the physiological roles that the genes of regulatory patterns V–VIII might play in anaerobic metabolism. However, these genes are still members of the same functional classes regulated by FNR (1) and Lrp (25). To illustrate the overlap between genes regulated by ArcA, FNR, and Lrp, a Venn diagram was constructed (Fig. 7*B*). The 500 genes with the highest PPDE(p) values (>0.996232) and the lowest p values (<5.26E-04) obtained from the array experiments reported here comparing *arcA* isogenic strains under anaerobic growth conditions were compared with the 500 genes with the highest PPDE(p) values (>0.991) and the lowest p values (<0.0014) obtained from the array experiments reported here comparing *fnr* isogenic strains under anaerobic growth conditions compared with the highest PPDE(p) values (>0.80) and the lowest p values (<0.027) obtained from the Lrp array experiments comparing *lrp* isogenic strains under aerobic growth conditions (25). Among these three gene sets, 26 genes are present in all three, and 43 genes overlap between the ArcA and Lrp regulons. This further supports our previous suggestion (1) that the FNR, Lrp, and now ArcA regulons reveal overlapping functions under aerobic and anaerobic conditions.

Conclusion

In this, our fourth study of global gene expression profiling in *E. coli* K12, we have again employed rigorous statistical treatment of the data to infer differential expression for 1139 genes in the presence and absence of the ArcA regulatory protein. In agreement with our previous study on the FNR protein (1) and the study of Liu and De Wulf (22), these results demonstrate that the network of genes required for the transition of cells from aerobic to anaerobic growth conditions is much larger than previously suspected (~8–10-fold).

A total of 30 genes had been documented previously as members of the ArcA regulon (5, 6). The study by Liu and De Wulf (22) suggested that 372 genes (or ~9% of the *E. coli* genome) are potential members of the ArcA regulon. The results presented here identify 135 of 175 genes with p values <0.000174 and PPDE(<p) values >0.9994 whose expression is affected by ArcA. However, if we include all genes expressed at a level above the background and examine the PPDE *versus* p value plots, we have a 63% confidence level that any gene in our oxygen-regulated set is differentially expressed (1), *i.e.* 63% of the 2820 genes or ~1700 genes. In the same manner, using the same PPDE *versus* p value plots, 67% of these 1700 genes or 1139 genes are either directly or indirectly regulated by ArcA. Thus, these results greatly expand our knowledge of genes that compose the ArcA regulatory network.

REFERENCES

1. Salmon, K., Hung, S. P., Mekjian, K., Baldi, P., Hatfield, G. W., and Gunsalus, R. P. (2003) *J. Biol. Chem.* **278,** 29837–29855
2. Blattner, F. R., Plunkett, G., III, Bloch, C. A., Perna, N. T., Burland, V., Riley, M., Collado-Vides, J., Glasner, J. D., Rode, C. K., Mayhew, G. F., Gregor, J., Davis, N. W., Kirkpatrick, H. A., Goeden, M. A., Rose, D. J., Mau, B., and Shao, Y. (1997) *Science* **277,** 1453–1474
3. Gunsalus, R. P., and Park, S. J. (1994) *Res. Microbiol.* **145,** 437–450
4. Guest, J. R., Attwood, M. M., Machado, R. S., Matqi, K. Y., Shaw, J. E., and Turner, S. L. (1997) *Microbiology* **143,** 457–466
5. Lynch, A. S., and Lin, E. C. C. (1996) in *Regulation of Gene Expression in E. coli* (Lin, E. C. C., and Lynch, A. S., eds) pp. 362–381, R. G. Landes Co., Austin, TX
6. Lynch, A. S., and Lin, E. C. C. (1996) in *Escherichia coli and Salmonella: Cellular and Molecular Biology* (Neidhart, F. C., ed) Vol. 1, pp. 1526–1549, ASM Press, Washington, D. C.
7. Guest, J. R., Green, J., Irvine, A., and Spiro, S. (1996) in *Regulation of Gene Expression in Escherichia coli* (Lin, E. C. C., and Lynch, A. S., eds) pp. 317–342, R. G. Landes Co., Austin, TX
8. Bauer, C. E., Elsen, S., and Bird, T. H. (1999) *Annu. Rev. Microbiol.* **53,** 495–523
9. Park, S. J., Chao, G., and Gunsalus, R. P. (1997) *J. Bacteriol.* **179,** 4138–4142
10. Park, S. J., Cotter, P. A., and Gunsalus, R. P. (1995) *J. Bacteriol.* **177,** 6652–6656
11. Park, S. J., and Gunsalus, R. P. (1995) *J. Bacteriol.* **177,** 6255–6262
12. Park, S. J., McCabe, J., Turna, J., and Gunsalus, R. P. (1994) *J. Bacteriol.* **176,** 5086–5092
13. Park, S. J., Tseng, C. P., and Gunsalus, R. P. (1995) *Mol. Microbiol.* **15,** 473–482
14. Cotter, P. A., Chepuri, V., Gennis, R. B., and Gunsalus, R. P. (1990) *J. Bacteriol.* **172,** 6333–6338
15. Cotter, P. A., and Gunsalus, R. P. (1989) *J. Bacteriol.* **171,** 3817–3823
16. Cotter, P. A., and Gunsalus, R. P. (1992) *FEMS Microbiol. Lett.* **70,** 31–36
17. Cotter, P. A., Melville, S. B., Albrecht, J. A., and Gunsalus, R. P. (1997) *Mol. Microbiol.* **25,** 605–615
18. Govantes, F., Albrecht, J. A., and Gunsalus, R. P. (2000) *Mol. Microbiol.* **37,** 1456–1469
19. Drapal, N., and Sawers, G. (1995) *Mol. Microbiol.* **16,** 597–607
20. Jeong, J. Y., Kim, Y. J., Cho, N., Shin, D., Nam, T. W., Ryu, S., and Seok, Y. J. (2004) *J. Biol. Chem.* **279,** 38513–38518
21. Sawers, G., and Suppmann, B. (1992) *J. Bacteriol.* **174,** 3474–3478
22. Liu, X., and De Wulf, P. (2004) *J. Biol. Chem.* **279,** 12588–12597
23. Silhavy, T. J., Berman, M. L., and Enquist, L. W. (1984) *Experiments with Gene Fusions*, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY
24. Neidhardt, F. C., Bloch, P. L., and Smith, D. F. (1974) *J. Bacteriol.* **119,** 736–747
25. Hung, S. P., Baldi, P., and Hatfield, G. W. (2002) *J. Biol. Chem.* **277,** 40309–40323
26. Baldi, P., and Hatfield, G. W. (2002) *DNA Microarrays and Gene Expression: From Experiments to Data Analysis and Modeling*, Cambridge University Press, Cambridge, UK
27. Hatfield, G. W., Hung, S. P., and Baldi, P. (2003) *Mol. Microbiol.* **47,** 871–877
28. Long, A. D., Mangalam, H. J., Chan, B. Y., Tolleri, L., Hatfield, G. W., and Baldi, P. (2001) *J. Biol. Chem.* **276,** 19937–19944
29. Baldi, P., and Long, A. D. (2001) *Bioinformatics* **17,** 509–519
30. Allison, D. B., Gadbury, G. L., Heo, M., Fernndez, J. R., Lee, C. K., Prolla, T. A., and Weindruch, R. (2002) *Comput. Stat. Data Anal.* **39,** 1–20
31. Lynch, A. S., and Lin, E. C. C. (1996) *J. Bacteriol.* **178,** 6238–6249
32. Bailey, T. L., and Elkan, C. (1994) in *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology* (Altman, R., Brutlag, D., Karp, P., Lathrop, R., and Searls, D., eds) pp. 28–36, AAAI Press, Menlo Park, CA
33. Bailey, T. L., and Gribskov, M. (1998) *Bioinformatics* **14,** 48–54
34. Salgado, H., Gama-Castro, S., Martinez-Antonio, A., Diaz-Peredo, E., Sanchez-Solano, F., Peralta-Gil, M., Garcia-Alonso, D., Jimenez-Jacinto, V., Santos-Zavaleta, A., Bonavides-Martinez, C., and Collado-Vides, J. (2004) *Nucleic Acids Res.* **32,** D303–D306
35. Chao, G., Shen, J., Tseng, C. P., Park, S. J., and Gunsalus, R. P. (1997) *J. Bacteriol.* **179,** 4299–4304
36. Shen, J., and Gunsalus, R. P. (1997) *Mol. Microbiol.* **26,** 223–236
37. Weidner, U., Geier, S., Ptock, A., Friedrich, T., Leif, H., and Weiss, H. (1993) *J. Mol. Biol.* **233,** 109–122
38. Bongaerts, J., Zoske, S., Weidner, U., and Unden, G. (1995) *Mol. Microbiol.* **16,** 521–534
39. Compan, I., and Touati, D. (1994) *Mol. Microbiol.* **11,** 955–964
40. Ma, Z., Richard, H., Tucker, D. L., Conway, T., and Foster, J. W. (2002) *J. Bacteriol.* **184,** 7001–7012
41. Ma, Z., Richard, H., and Foster, J. W. (2003) *J. Bacteriol.* **185,** 6852–6859
42. Masuda, N., and Church, G. M. (2003) *Mol. Microbiol.* **48,** 699–712
43. Tramonti, A., Visca, P., De Canio, M., Falconi, M., and De Biase, D. (2002) *J. Bacteriol.* **184,** 2603–2613
44. Iuchi, S., Chepuri, V., Fu, H. A., Gennis, R. B., and Lin, E. C. C. (1990) *J. Bacteriol.* **172,** 6020–6025
45. Tseng, C. P., Albrecht, J., and Gunsalus, R. P. (1996) *J. Bacteriol.* **178,** 1094–1098
46. Govantes, F., Orjalo, A. V., and Gunsalus, R. P. (2000) *Mol. Microbiol.* **38,** 1061–1073
47. Jones, H. M., and Gunsalus, R. P. (1987) *J. Bacteriol.* **169,** 3340–3349
48. Bell, A. I., Cole, J. A., and Busby, S. J. (1990) *Mol. Microbiol.* **4,** 1753–1763
49. Jayaraman, P. S., Cole, J. A., and Busby, S. J. (1989) *Nucleic Acids Res.* **17,** 135–145
50. Tseng, C. P., Yu, C. C., Lin, H. H., Chang, C. Y., and Kuo, J. T. (2001) *J. Bacteriol.* **183,** 461–467
51. Tseng, C. P. (1997) *FEMS Microbiol. Lett.* **157,** 67–72
52. Hirshfield, I. N., Tenreiro, R., Vanbogelen, R. A., and Neidhardt, F. C. (1984) *J. Bacteriol.* **158,** 615–620
53. Leveque, F., Gazeau, M., Fromant, M., Blanquet, S., and Plateau, P. (1991) *J. Bacteriol.* **173,** 7903–7910