

# Global Gene Expression Profiling in *Escherichia coli* K12

THE EFFECTS OF INTEGRATION HOST FACTOR\*<sup>§</sup>

Received for publication, March 17, 2000, and in revised form, June 21, 2000  
Published, JBC Papers in Press, June 27, 2000, DOI 10.1074/jbc.M002247200

Stuart M. Arfin<sup>‡</sup>, Anthony D. Long<sup>§</sup>, Elaine T. Ito<sup>¶</sup>, Lorenzo Toller<sup>¶</sup>, Michelle M. Riehle<sup>§</sup>,  
Eriks S. Paegle<sup>||</sup>, and G. Wesley Hatfield<sup>¶</sup>

From the <sup>‡</sup>Departments of Biological Chemistry and <sup>¶</sup>Microbiology and Molecular Genetics, College of Medicine,  
<sup>§</sup>Department of Ecology and Evolutionary Biology, School of Biological Sciences, and <sup>||</sup>Department of Chemical  
Engineering and Material Sciences, School of Engineering, University of California, Irvine, California, 92697

We have used nylon membranes spotted in duplicate with full-length polymerase chain reaction-generated products of each of the 4,290 predicted *Escherichia coli* K12 open reading frames (ORFs) to measure the gene expression profiles in otherwise isogenic integration host factor IHF<sup>+</sup> and IHF<sup>-</sup> strains. Our results demonstrate that random hexamer rather than 3' ORF-specific priming of cDNA probe synthesis is required for accurate measurement of gene expression levels in bacteria. This is explained by the fact that the currently available set of 4,290 unique 3' ORF-specific primers do not hybridize to each ORF with equal efficiency and by the fact that widely differing degradation rates (steady-state levels) are observed for the 25-base pair region of each message complementary to each ORF-specific primer. To evaluate the DNA microarray data reported here, we used a linear analysis of variance (ANOVA) model appropriate for our experimental design. These statistical methods allowed us to identify and appropriately correct for experimental variables that affect the reproducibility and accuracy of DNA microarray measurements and allowed us to determine the statistical significance of gene expression differences between our IHF<sup>+</sup> and IHF<sup>-</sup> strains. Our results demonstrate that small differences in gene expression levels can be accurately measured and that the significance of differential gene expression measurements cannot be assessed simply by the magnitude of the fold difference. Our statistical criteria, supported by excellent agreement between previously determined effects of IHF on gene expression and the results reported here, have allowed us to identify new genes regulated by IHF with a high degree of confidence.

It has been more than forty years since the pioneering studies of Jacob and Monod (1) on the regulation of the genes of the *lac* operon of *Escherichia coli* established the basic paradigm for protein-mediated regulation of gene expression. Since then, the molecular mechanisms responsible for the regulation of scores of operons in this organism have been elucidated. However, although a great deal has been learned about the regulation of individual operons, much less is known about the global regulatory mechanisms that coordinate the expression of these

operons with one another and with the nutritional and environmental growth state of the cell.

Much of what is known about global gene regulation has been inferred from the analysis of O'Farrell two-dimensional electrophoresis gels for the resolution of individual proteins expressed in cells grown under two experimental conditions. The heat shock- and starvation-induced proteins, for example, were originally identified by this method (2). However, the identification of each of the cellular proteins on these gels is a laborious task, and the estimation of the level of expression of each protein in the cell is not easily quantified. In fact, the expression of only about 250 proteins so far have been characterized by this method.

Classical genetic and biochemical studies have also identified global regulatory proteins that respond to small molecule co-regulators to affect the expression of large sets of genes. These proteins, such as catabolic repressor protein and leucine-responsive regulatory protein (Lrp), modulate the expression of stimulons that encompass several regulons, each of which may contain multiple operons of common function (3). Stimulons are generally regulated by nutritional (e.g. glucose starvation) or environmental (e.g. heat shock) signals. For example, it is well known that catabolic repressor protein and its co-activator, cyclic-AMP, are required for the induction of carbon utilization operons under glucose starvation conditions (4).

Another class of global regulatory proteins, which include members such as H-NS and integration host factor (IHF),<sup>1</sup> are DNA architectural proteins involved in the condensation of the bacterial nucleoid. They are abundant proteins that bind to many sequence-specific but degenerate DNA sites and affect processes that require DNA duplex destabilization such as DNA replication, recombination, and transcription (5–10). These proteins also affect the expression of many genes (operons), but unlike the global regulators of the previous class, they do not respond to small molecule metabolic co-regulators and there is no obvious metabolic coherence among the genes whose expression they affect.

IHF was initially identified as the product of a gene required for the site-specific integrative recombination of phage  $\lambda$  into the *E. coli* chromosome (11). Subsequently, it has been discovered that IHF affects many cell functions including a variety of site-specific recombination events and DNA replication (12). In addition, it was found that IHF influences the expression levels of many genes. For example, Freundlich *et al.* (13) used O'Farrell two-dimensional electrophoresis gels to demonstrate a difference in the levels of 15–20% of the proteins expressed in

\* This work was supported in part by National Institutes of Health Grant GM55073 (to G. W. H.). 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.

<sup>§</sup> The on-line version of this article (available at <http://www.jbc.org>) contains a supplemental figure.

<sup>1</sup> The abbreviations used are: IHF, integration host factor; SIDD, supercoiling-induced DNA duplex destabilized; ORF, open reading frame; MOPS, 4-morpholinepropanesulfonic acid.

IHF mutant and isogenic parent cells grown under conditions comparable with those reported here. The mechanistic role for IHF in these processes has been largely ascribed to its ability to bend DNA to bring distant sites on the bacterial chromosome together for a biological function. In the case of phage  $\lambda$ , IHF facilitates the integration reaction by bringing distant integrase binding sites into the proximity of the bacterial and phage attachment sites (14). IHF has also been shown to function as a DNA looper protein to facilitate interactions between regulatory proteins bound at upstream sites and RNA polymerase at downstream promoter sites (15, 16). Because these functions involve IHF binding to site-specific high affinity sites and because of the high intracellular concentration of this abundant chromosomal organizer protein, these IHF sites are likely saturated under all physiological conditions (17). Thus, unlike other regulatory proteins that bind small molecule effectors that affect their DNA binding properties, IHF functions as an architectural component of DNA structures that affect the constitutive or basal level expression of many promoters. This may explain the lack of any obvious metabolic coherence among the genes whose expression are affected by IHF (12).

It has recently been demonstrated that IHF can also inhibit the transition of supercoiling-induced DNA duplex destabilized (SIDD) sites from a B-form to a partially denatured duplex structure (18–20). This results in the translocation of the superhelical energy (negative twist) normally absorbed by the SIDD site to another site in a superhelically constrained DNA domain (19, 20). In the case of the *ilvGMEDA* operon, required for the biosynthesis of the branched chain amino acids in *E. coli*, IHF-mediated translocation of superhelical energy from an upstream SIDD site results in a destabilization of the DNA duplex in the  $-10$  region of the downstream *ilvP<sub>G</sub>* promoter. This supercoiling-dependent, IHF-mediated duplex destabilization in the promoter region facilitates open complex formation and an increase in transcription into the structural genes of this operon (21).

It is known that the global superhelical density of the chromosome varies over a wide range during different phases of the bacterial growth cycle and in response to various types of environmental assaults such as osmotic, temperature, and anaerobic shocks and nutritional upshifts and downshifts. Our previously published results suggest that the effect of IHF on the expression of the genes of the *ilvGMEDA* operon is to amplify basal level expression of this operon (independently of operon-specific controls) in response to small changes in the global superhelical density of the bacterial chromosome in order to coordinate the capacity for branched chain amino acid biosynthesis with the environmental and nutritional growth conditions of the cell. To determine if this represents a general control mechanism for coordinating the expression of other genes (operons) will require knowledge of the location and thermodynamic stability of each of the SIDD sites on the *E. coli* chromosome at different physiological superhelical densities encountered under different growth conditions. This information together with the location of each of the high affinity IHF sites and the effects of IHF on global gene expression profiles under these same conditions should facilitate an assessment of the generality of this mechanism. Indeed, calculations to identify the location and thermodynamic stability of all of the SIDD sites on the *E. coli* chromosome at the global chromosomal superhelical densities observed in stationary phase and aerobically and anaerobically cultured cells growing in glucose minimal MOPS medium are currently in progress (22).<sup>2</sup> We are also currently developing genomic SELEX methods to isolate *in*

*in vivo* cross-linked IHF-chromosomal DNA fragments for hybridization to DNA arrays containing probes for all of the *E. coli* inter-ORF (upstream regulatory) regions to identify each of the IHF-binding sites.<sup>3</sup>

In this report we describe the use of nylon membranes spotted in duplicate with full-length polymerase chain reaction-generated products of each of the 4,290 predicted *E. coli* K12 ORFs to measure the gene expression profiles in otherwise isogenic IHF<sup>+</sup> and IHF<sup>-</sup> strains growing in glucose minimal MOPS medium. To evaluate the data generated by these gene expression profiling experiments we used a linear analysis of variance model appropriate for the experimental design employed in this study. These statistical methods allowed us to identify and minimize experimental variables that affect the reproducibility and accuracy of DNA microarray measurements and to determine the statistical significance of observed differences between expression levels of each ORF in these two genotypes. Together with future knowledge of the location and *in vivo* occupancy of IHF at its high affinity chromosomal binding sites and the location and stability of the SIDD sites around the *E. coli* chromosome in wild-type and IHF<sup>-</sup> strains grown under different environmental and nutritional conditions, these data will allow us to assess the generality of global gene regulation by IHF-mediated translocation of superhelical energy from one site on the chromosome to another.

#### MATERIALS AND METHODS

**Chemicals and Reagents**—Avian myeloblastosis virus reverse transcriptase, RNase free DNase I, and Sephadex G-25 Quickspin columns were obtained from Roche Molecular Biochemicals. Ribonuclease inhibitor III was purchased from Panvera/Takara, ultra pure deoxynucleoside triphosphates were from Amersham Pharmacia Biotech, random hexamer oligonucleotides were from New England Biolabs, and [ $\alpha^{33}$ P] dCTP (2–3000 Ci/mmol) was from NEN Life Science Products. DNA filter arrays (Panorama *E. coli* gene arrays) and 3' ORF-specific oligonucleotides were obtained from Sigma-Genosys Biotechnologies. All other chemicals were obtained from Sigma. All reagents and baked glassware used in RNA manipulations were treated with diethylpyrocarbonate.

**Bacterial Strains and Growth Conditions**—The construction of the isogenic *E. coli* strains IH100 [*ilvP<sub>G</sub>::lacZYA*] and IH105 [*ilvP<sub>G</sub>::lacZYA,  $\Delta$ himA*] used in these experiments have been described (21). In both strains the genes of the chromosomal *lac* operon are transcribed from the *ilvP<sub>G</sub>* promoter, which is activated by the upstream binding of IHF. Cells were grown in 25 ml of MOPS medium (23) containing 0.4% glucose in 125-ml Erlenmeyer flasks at 37 °C with constant aeration.

**Isolation of Total RNA**—Total RNA was isolated from cells at an A<sub>600</sub> of 0.5–0.6. Five-ml samples of cultures of growing cells were pipetted directly into 5 ml of boiling lysis buffer (1% SDS, 0.1 M NaCl, 8 mM EDTA) and mixed at 100 °C for 2 min. These samples were transferred to 125-ml Erlenmeyer flasks, mixed with an equal volume of acid phenol (pH 4.3), and shaken vigorously for 6 min at 64 °C. After centrifugation, the aqueous phase was transferred to a fresh Erlenmeyer flask, and the hot acid phenol extraction procedure was repeated. The second aqueous phase was extracted with phenol-chloroform-isoamyl alcohol (25:24:1, pH 8) at room temperature and, finally, with chloroform-isoamyl alcohol (24:1). Total RNA was precipitated with two volumes of ethanol in 0.3 M sodium acetate (pH 5.3), washed with 70% ethanol, and redissolved in a 10 mM Tris, 1 mM EDTA solution (pH 8.0). The redissolved RNA was treated with DNase I (20 units in a 200- $\mu$ l reaction mixture containing 10 mM MgCl<sub>2</sub>, 1 mM dithiothreitol, and 5 units RNasin) for 15 min at 37 °C and re-extracted first with phenol-chloroform-isoamyl alcohol (25:24:1, pH 8) at room temperature and then with chloroform-isoamyl alcohol (24:1). To ensure that the total RNA preparation was free of genomic DNA contamination, the DNase I treatment was repeated a second time. After ethanol precipitation, the purified RNA was again washed with 70% ethanol and redissolved in a 10 mM Tris, 1 mM EDTA solution (pH 8.0). The RNA concentration was determined by absorption at 260 nm. Normally, three 5-ml samples from each culture were processed in parallel.

**cDNA Synthesis and Labeling Conditions**—For random hexamer-primed cDNA synthesis, 20  $\mu$ g of total RNA and 37.5 ng of random

<sup>2</sup> C. J. Benham, personal communication.

<sup>3</sup> J. M. Calvo and G. W. Hatfield, unpublished results.

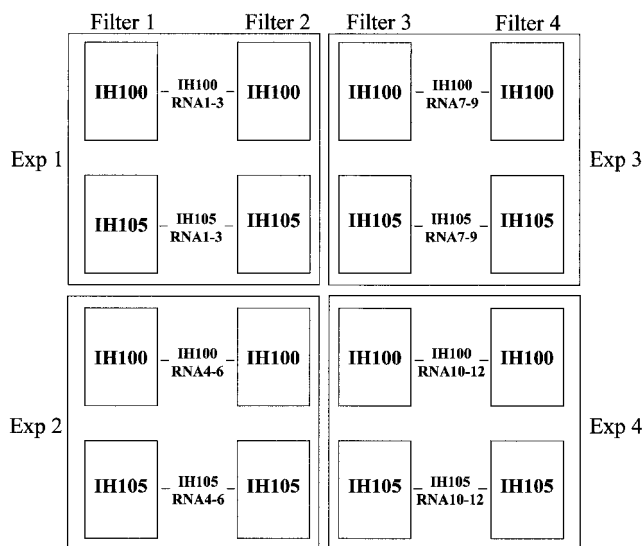


FIG. 1. **Experimental design.** See "Materials and Methods" for description.

hexamer primers were heated at 70 °C for 3 min and quick-cooled on ice. cDNA synthesis was performed at 42 °C for 3 h in a 60- $\mu$ l reaction mixture containing the RNA and primer mixture, reverse transcriptase buffer (Roche Molecular Biochemicals), 1 mM each dATP, dGTP, and dTTP, 50  $\mu$ Ci [ $\alpha^{33}$ P]dCTP, 20 units of ribonuclease inhibitor III, and 4  $\mu$ l (88 units) of avian myeloblastosis virus reverse transcriptase. Labeled cDNA was separated from unincorporated nucleotides on Sephadex G-25 spin columns.

ORF-specific oligonucleotide primed cDNA synthesis was performed in the same way except that 2  $\mu$ g of total RNA was mixed with 8  $\mu$ l of ORF-specific primers and unlabeled nucleotides, heated at 90 °C for 2 min, and slow-cooled to 42 °C. After the addition of ribonuclease inhibitor III, avian myeloblastosis virus reverse transcriptase, and [ $\alpha^{33}$ P]dCTP, the reaction mixture was incubated at 42 °C for 3 h. Approximately 50% incorporation of labeled nucleotides usually was achieved with both protocols.

**DNA Microarray Hybridization**—The nylon filters were soaked in 2 $\times$  saline/sodium phosphate/EDTA for 10 min and prehybridized in 10 ml of hybridization solution (5 $\times$  saline/sodium phosphate/EDTA, 2% SDS, 1 $\times$  Denhardt's solution containing 0.1 mg/ml sheared herring sperm DNA) for 1 h at 65 °C. 2–3  $\times 10^7$  cpm of cDNA probe in 500  $\mu$ l of the same solution was heated at 90–95 °C for 10 min, rapidly cooled on ice, and added to 5.5 ml of hybridization solution. The prehybridization solution was removed and replaced with the hybridization solution. Hybridization was carried out for 15 to 18 h at 65 °C. Following hybridization, each filter was rinsed with 50 ml of 0.5 $\times$  saline/sodium phosphate/EDTA containing 0.2% SDS at room temperature for 3 min, followed by three washes in the same solution at 65 °C for 20 min each. The filters were partially air-dried, wrapped in Saran Wrap, and exposed to a phosphor screen for 48–60 h. Filters were stripped by microwaving at half-maximal power in 500 ml of 10 mM Tris solution (pH 8.0) containing 1 mM EDTA and 1% SDS for 20 min. Stripped filters were wrapped in Saran Wrap and stored in the presence of damp paper towels in sealed plastic bags at 4 °C.

**Enzyme Assays**—Cells were harvested at an  $A_{600\text{ nm}}$  of 0.5 to 0.6 and disrupted by sonication in an appropriate buffer. Assays for cystathionase (*metC*),  $\alpha,\beta$ -dihydroxyacid dehydrase (*ilvD*),  $\beta$ -galactosidase (*lacZ*), glutamine synthetase (*glnA*), and imidazolylacetol phosphate:L-glutamate aminotransferase (*hisC*) were performed as described (18, 24–27). All assays were performed under conditions where they were linear with respect to both extract concentration and time. Protein concentration was determined by the method of Bradford (28).

**Experimental Design**—The experimental regimen for the four duplicate experiments reported here is diagrammed in Fig. 1. In Experiment 1, Filters 1 and 2 were hybridized with  $^{33}$ P-labeled, random hexamer generated cDNA fragments complementary to each of three RNA preparations (IH100 RNA1–3) obtained from the cells of three individual cultures of strain IH100 (IHF<sup>+</sup>). These three  $^{33}$ P-labeled cDNA preparations were pooled before the hybridizations. Following Phosphor-Imager analysis, these filters were stripped and hybridized with pooled,  $^{33}$ P-labeled cDNA fragments complementary to each of three RNA

preparations (IH105 RNA1–3) obtained from strain IH105 (IHF<sup>-</sup>). In Experiment 2, these same filters were again stripped, and this protocol was repeated with  $^{33}$ P-labeled cDNA fragments complementary to another set of three pooled RNA preparations obtained from strains IH100 (IH100 RNA 4–6) and IH105 (IH105 RNA 4–6) as described above. Another set of filters (Filter 3 and Filter 4) was used for Experiments 3 and 4 as described for Experiments 1 and 2. This protocol results in duplicate filter data for four experiments performed with the cDNA probes complementary to four independently prepared sets of RNA. Thus, since each filter contains duplicate spots for each ORF and duplicate filters are hybridized for each experiment, four measurements for each ORF were obtained from each of four experiments.

**Data Acquisition**—A commercial software package obtained from Research Imaging Inc. (DNA ArrayVision) was used to grid the phosphorimaging image, to record the pixel density of each of the 18,432 addresses on each filter, and to perform the background subtraction. 8,580 of the addresses on each filter are spotted with duplicate copies of each of the 4,290 *E. coli* ORFs. The remaining 9,852 empty addresses were used for background measurements. Since the backgrounds were quite constant, a global average background measurement was subtracted from each experimental measurement, although local background calculations were possible. Greater than four logs of linearity for the phosphorimaging-derived data was observed.

**Statistical Methods**—The experimental design employed in this study consisted of four independent  $^{33}$ P-labeled cDNA preparations for each of two genotypes separately hybridized to two filter pairs, with each filter containing every *E. coli* ORF spotted in duplicate. This design is depicted in Fig. 1. For each spot, a background-subtracted estimate of expression level was obtained and scaled to total counts on the membrane. For any given spot, a number greater than zero (indicating an expression level) or a zero (indicating an expression level lower than background) was obtained. A full statistical model describing this design would be both complex and over-parameterized with respect to the number of expression measures for any given ORF. For these reasons and because we are only interested in testing for differences between genotypes, we opted for a reduced model. This model consists of *t* tests, which assume that filters are not involved in two-way statistical interactions. The *t* test evaluates the difference between the means of two groups employing the variance within groups as an error term. The result is that large differences between groups for any given ORF would tend to be declared non-significant if the expression level of that ORF were unreplicable within experimental treatments. Conversely, small differences in expression could be determined to be statistically significant for a given ORF if expression levels for that ORF were replicable within treatments. In short, the test statistic employed here was constructed by scaling the difference in gene expression levels between genotypes relative to the observed variances within genotypes. *p* values based on this test statistic range from 1.0 for gene expression levels, with identical values to very small *p* values for expression level differences that are highly significant. A comprehensive discussion of the use of the *t* test and the modifications applicable to the analysis of DNA microarray data of the type presented here is available at the Genomics at the University of California, Irvine web site. To identify possible sources of experimental error we also used the *t* test to determine statistical differences among different filters hybridized with the same RNA preparation of the same genotype as well as differences among different RNA preparations of the same genotype hybridized to the same filters. Since these comparisons only involve factors that we expect to be highly replicable, an excess of hybridization signals showing significant differences in gene expression levels would indicate an experimental artifact. Depending on the magnitude of these artifacts, the detection of significant differences between genotypes requires replication over the variables (filters and RNA preparations) that lead to these false positives. The data presented here demonstrate that no experimental artifacts are contributed by filter differences and that experimental artifacts due to differences in RNA preparations of the same genotype can be eliminated by averaging over as few as four independent RNA preparations.

**Data Accession**—All of the raw and processed data for the experimental results reported here may be downloaded in tabular format through the online version of this paper.<sup>4</sup>

<sup>4</sup> The supplemental figure shows complete gene expression data for *E. coli* K12 strains IH100 (IHF<sup>+</sup>) and IH105 (IHF<sup>-</sup>). The values in each column are: *column 1*, gene name; *columns 2–5*, average of the duplicate gene measurements for each filter hybridized with cDNA pools from strain IH100 for experiments 1–4, respectively; *columns 6–9*, average

## RESULTS AND DISCUSSION

**Random Hexamer Priming of Total RNA for cDNA Probe Synthesis Is Required for Accurate Measurements of Differential Gene Expression Levels in Bacteria**—Since less than 10% of the total RNA in an *E. coli* cell is mRNA, it was feared that cDNA preparation by random hexamer priming of total RNA for hybridization to DNA microarrays might produce unacceptable backgrounds. We, therefore, used a set of 4,290 unique 25-base pair oligonucleotide primers specific for the 3' end of each *E. coli* ORF (available from Sigma-Genosys) for primer-directed synthesis of  $\alpha^{33}\text{P}$ -labeled cDNA probes. However, filter hybridizations with these cDNA probes detected only 1,760 genes, with at least 2 out of 4 or greater non-zero, background-subtracted measurements on the control (IH100) filters. Equally disturbing, it was often observed that although some genes of a given operon were detected, others were not. For example, hybridization signals above background were detected for only three of the five genes of the *ilvGMEDA* operon and two of the three genes for the *ilvP<sub>G</sub>::lacZYA* operon. We, therefore, turned to the use of random hexamers for primer-directed synthesis of  $\alpha^{33}\text{P}$ -labeled cDNA probes. Filter hybridization with these cDNA probes detected the expression of 2,592 genes with at least 2 out of 4 non-zero, background-subtracted measurements for each of the control (IH100) and experimental (IH105) data sets. Thus, the expression levels of 832 more genes including all of the genes of the *ilvGMEDA* and the *ilvP<sub>G</sub>::lacZYA* operons were detected with the random hexamer-labeled probes. Furthermore, the expression level of the genes in each operon varied less than 3-fold (see Fig. 4).

The observation that the ORF-specific probes do not detect as many mRNAs as the random hexamer-labeled probes suggested that these probes do not hybridize to about one-third of the mRNAs either because of the hybridization conditions or because they hybridize to themselves or to one another. Furthermore, the wide variation of signals obtained with the ORF-specific-labeled probes for genes of a common operon can be explained by the expectation that a variable amount of  $\alpha^{33}\text{P}$  would be incorporated into each ORF because of unequal hybridization efficiencies and different lengths of labeled cDNA fragments. On the other hand, since each mRNA (or mRNA fragment) is randomly primed with the random hexamers, the amount of  $\alpha^{33}\text{P}$  label incorporated into each probe should be largely proportional to the ORF length.

To test these interpretations of our results, random hexamers or ORF-specific primers were used for primer-directed synthesis of  $\alpha^{33}\text{P}$ -labeled cDNA probes derived from genomic DNA. In the case of the ORF-specific-labeled probes, we would expect that variable amounts of  $\alpha^{33}\text{P}$  should be incorporated into each probe because the length of the synthesized probe might significantly exceed the length of the ORF, especially for short ORFs (or might be smaller for long ORFs). Thus, a single probe might extend into adjacent ORFs or ORFs encoded on the

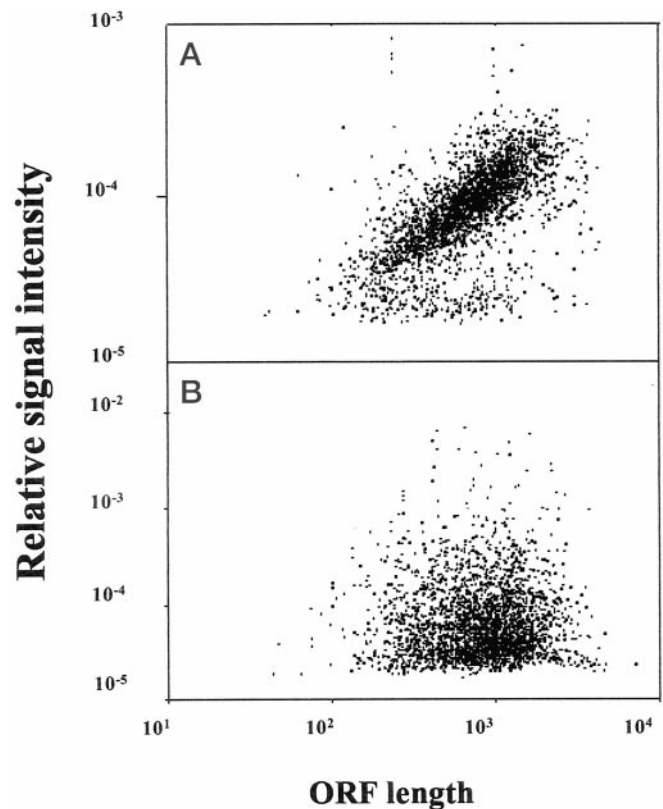


FIG. 2. Scatter plot showing the relationship between hybridization signal intensities with  $^{33}\text{P}$ -labeled cDNA probes generated from genomic DNA using random hexamer oligonucleotides (A) or 3' ORF-specific DNA primers (B).

other strand. In these cases, the hybridization signal obtained for any given ORF spot on the array should depend on a complex set of parameters including the size of the ORF, the size of the probe fragment, the position of surrounding ORFs, the placement of the labeling primers, and the hybridization conditions. In contrast, since each region of the chromosome is randomly primed with the random hexamers, probes for every ORF should be generated, and the amount of  $\alpha^{33}\text{P}$  incorporated into each probe should be largely proportional to the ORF length. The data presented in Fig. 2 confirm these expectations. Hybridization signals for each of the 4,290 ORFs on the array were observed with the random hexamer-labeled probes, whereas hybridization signals for only two-thirds of the ORFs on the array were observed with the ORF-specific primer-labeled probes (the same ratio of ORF-specific versus random hexamer-labeled probe hybridization signals observed with the cDNA probes generated from RNA). Furthermore, as predicted, the data displayed in Fig. 2A show that the hybridization signal for the random hexamer-labeled probes generated from genomic DNA was reasonably proportional to ORF length ( $r^2 = 0.41$ ), but no significant correlation between ORF length and hybridization signal was observed with the ORF-specific-labeled probes ( $r^2 = 0.004$ ; Fig. 2B).

An additional complication that might contribute to the disparate results obtained with random hexamer and ORF-specific-labeled probes could be the result of the widely differing degradation rates of the 25-base pair region of each message measured by the ORF-specific primers. It is known that rapid mRNA decay in *E. coli* is initiated by endonucleolytic cleavages followed by 3' to 5' exonucleolytic degradation (29, 30). Therefore, if the initial endonucleolytic site were adjacent to the 3' ORF-specific primer binding site, this region might be rapidly degraded, and little or no steady-state message would be ex-

of the duplicate gene measurements for each filter hybridized with cDNA pools from strain IH105 for experiments 1–4, respectively; *column 10*, number of non-zero IH100 measurements for Experiments 1–4; *column 11*, number of non-zero IH105 measurements for Experiments 1–4; *column 12*, average of the values in columns 2–5; *column 13*, average of the values in columns 6–9; *column 14*, the S.D. of the mean for the IH100 values for Experiments 1–4 (columns 2–5); *column 15*, the S.D. of the mean for the IH105 values for Experiments 1–4 (columns 6–9); *column 16*, the value of the *t* test statistic; *column 17*, the degrees of freedom associated with the *t* test; *column 18*, the ratio of the variances of the IH100 and IH105 measurements; *column 19*, the *p* values associated with the differences between the IH100 and IH105 measurements based on the *t* test distribution; *column 20*, the ratios of the means of the IH100 and IH105 data, a negative sign implies decreased expression in strain IH105. These data may be viewed and downloaded from the on-line journal (<http://www.jbc.org>).

tracted for primer extension labeling of this gene-specific transcript. In fact, we have previously demonstrated that different probe hybridization sites on the same mRNA do exhibit different half-lives (31). On the other hand, the random hexamer-labeling procedure produces RNA-DNA duplexes for primer extension from all of the partial degradation products of each message. Since the exonucleolytic clearance of mRNA degradation products to free nucleotides follows endonucleolytic message inactivation (at a presumably more constant rate), the random hexamers should detect the steady-state level of all of these intermediate degradation products. This suggests that although the functional half-lives of *E. coli* mRNA are rapid and message-specific, the "clearance" rate for message degradation intermediates must occur at a more constant rate. If this were the case, the relative expression levels of genes measured with the random hexamer-labeled probes would be more closely related to their rates of synthesis and, therefore, their relative abundance in the cell. This conclusion is supported by the fact that a positive correlation between mRNA and protein abundance is observed with the random hexamer but not with the ORF-specific primer data (see below).

Since the microarray hybridization data measured with random hexamer-labeled probes obtained from genomic DNA showed a correlation with ORF length, we considered the possibility of correcting the expression data obtained with random hexamer-labeled cDNA probes obtained from RNA for ORF length. However, because the less than 10-fold variance in ORF lengths contributes less than one percent of the four logs of variance in expression level measurements obtained from the RNA derived probes, the expression data presented here are not corrected for differences in ORF lengths.

In conclusion, these data demonstrate that random hexamer priming for cDNA probe synthesis is required for accurate measurement of gene expression levels in bacteria. This is explained largely by the fact that it is difficult to obtain a set of 4,290 unique primers that hybridize to each ORF with equal efficiency and by the fact that widely differing degradation rates (steady-state levels) can be observed for the 25-base pair region of each message complementary to each ORF-specific primer.

*Replication and Appropriate Statistical Analysis Are Required for Determining the Accuracy of DNA Microarray Measurements*—A basic problem that is encountered by all types of DNA microarray experiments stems from the fact that thousands of measurements are obtained from a single experiment. This means that if there is any source of experimental error, a Gaussian distribution of these measurements will be observed. For example, if 5,000 measurements are obtained from two DNA microarray experiments performed under identical experimental conditions, then, based on a standard *t* test distribution, 250 (5%) of the individual measurements are expected to differ sufficiently by chance alone to produce a *p* value less than 0.05. Now, if 5,000 measurements are obtained from two DNA microarray experiments performed under different experimental conditions and 500 differences are observed at a 95% confidence level ( $p < 0.05$ ), half of these differences (250) will be false positives due to chance alone. Therefore, to interpret data from DNA microarray experiments in which thousands of measurements are obtained from a single experiment, it is necessary to employ statistical methods capable of distinguishing chance occurrences from biologically meaningful data. In this respect, an advantage of nylon DNA microarray filters is that they are relatively inexpensive and can be reused several times. It is therefore economically feasible to repeat each experiment a sufficient number of times to obtain statistically reliable data. For example, in the experiments reported here

four filters spotted in duplicate with each *E. coli* ORF were used four times in four separate experiments, resulting in 16 measurements for each ORF for each of two different genotypes (Fig. 1).

To design an appropriate analysis of variance model it is necessary to know the sources of experimental errors. The principal sources of experimental error in these experiments are expected to arise from differences among filters and differences among RNA preparations. Therefore, our experimental strategy was designed to allow us to determine the reproducibility of results obtained from different filters hybridized with the same cDNA preparations or from different cDNA preparations hybridized to the same filters. To assess the reproducibility between different filters, we employed a statistical *t* test to compare data from each pair of filters hybridized with cDNA probes prepared from the same RNA preparation. In this case, the data for each duplicate ORF measurement on each filter were averaged, the IH100 or IH105 data for Filter 1 were compared with the data on Filter 2, and the IH100 or IH105 data for Filter 3 were similarly compared with the data on Filter 4. Of the 2,592 genes expressed, 13 are expected to exhibit *p* values  $< 0.005$  (0.5% of 2,592 genes). Our analyses identified an average of 13 false positives for the same filters hybridized with cDNA preparations from strain IH100 and 3 false positives for the same filters hybridized with cDNA preparations from IH105. Therefore, since no false positives beyond those expected by chance alone result from the use of replicate filters, we conclude that no significant experimental error is contributed by differences among filters.

To ascertain the experimental error contributed by differences in RNA preparations, we employed a statistical *t* test to compare the data from the same filters hybridized with cDNA probes prepared from different RNA preparations. For example, the data for each duplicate ORF measurement on Filters 1 and 2 of Experiment 1 were compared with the data from Filters 1 and 2 of Experiment 2. Likewise, the data for each duplicate ORF measurement on Filters 3 and 4 of Experiment 3 were compared with the data from Filters 3 and 4 of Experiment 4. At a *p* value = 0.005, these comparisons revealed an average of 39 false positives between filters hybridized with cDNA preparations from IH100 and 27 false positives between filters hybridized with cDNA preparations from IH105. Thus, differences among RNA preparations account for a slightly elevated false positive rate. The number of false positives due to this error is approximately 2.5 times that expected by chance alone. However, when the data are averaged across RNA preparations, the variance contributed by differences in RNA preparations is minimized, and the number of false positives is decreased. For example, inspection of the *p* values obtained from a comparison of the data from the IH100 filters of experiments 1 and 3 with the IH100 filters of experiments 2 and 4 identified only 4 false positives at a *p* value  $< 0.005$  and no false positives at a *p* value less than 0.0001. Similar results were obtained when the data from the IH105 filters were compared in this way.

To determine the differential gene expression levels between strains IH100 (IHF<sup>+</sup>) and IH105 (IHF<sup>-</sup>), the background-subtracted and normalized IH100 or IH105 measurements for each ORF from each of the four experiments were averaged. These four averaged IH100 and IH105 sets of measurements were analyzed according to the statistical methods described under "Methods and Materials." This analysis identified 23 genes that differ in expression with a *p* value less than 0.0001 (Table I; Fig. 3A). Since a comparable analysis of the filters hybridized with identical genotypes revealed no genes that differ with a *p* value  $< 0.0001$  (see above), we can be nearly certain that these

TABLE I  
Genes differentially expressed between *E. coli* K12 strains IH100 (IHF<sup>+</sup>) and IH105 (IHF<sup>-</sup>) with a *p* value less than 0.0001

The data are presented as the average (Avg) and S.D. of four independent gene expression measurements expressed as a fraction of the total hybridization signal (total mRNA) on each DNA microarray filter. The *p* values are calculated on the basis of the *t* test distribution. Positive fold differences indicate increased gene expression in strain IH105. Negative fold differences indicate decreased gene expression in strain IH105.

Gene	Avg IH100	Avg IH105	S.D. IH100	S.D. IH105	<i>p</i> value	Fold
<i>b2226</i>	1.41E - 06	1.23E - 05	1.82E - 07	1.67E - 05	1.31E - 06	8.69
<i>yeeE</i>	8.20E - 05	6.64E - 04	1.83E - 05	6.37E - 05	2.18E - 06	8.11
<i>yeeD</i>	1.61E - 04	6.26E - 04	2.38E - 05	5.05E - 05	3.00E - 06	3.90
<i>intB</i>	2.41E - 06	7.33E - 05	2.28E - 06	8.27E - 06	3.13E - 06	30.38
<i>b1667</i>	6.21E - 05	1.04E - 04	4.01E - 06	4.71E - 06	1.03E - 05	1.67
<i>dniR</i>	1.91E - 04	9.75E - 04	4.20E - 05	1.09E - 04	1.05E - 05	5.10
<i>tra8_2</i>	2.48E - 04	6.58E - 04	3.59E - 05	5.57E - 05	1.71E - 05	2.65
<i>b1285</i>	3.04E - 05	8.74E - 05	7.35E - 06	5.64E - 06	1.76E - 05	2.87
<i>b2876</i>	7.94E - 06	5.01E - 05	6.05E - 06	3.45E - 06	1.94E - 05	6.30
<i>hdeB</i>	1.09E - 03	5.51E - 06	1.80E - 04	3.47E - 06	1.96E - 05	-198.50
<i>ilvA</i>	5.06E - 04	3.42E - 04	1.86E - 05	2.26E - 05	3.02E - 05	-1.48
<i>b1585</i>	8.69E - 06	4.86E - 05	4.44E - 06	5.62E - 06	3.13E - 05	5.59
<i>b2006</i>	8.49E - 06	3.58E - 05	3.25E - 06	4.05E - 06	4.32E - 05	4.22
<i>b2556</i>	3.92E - 05	3.57E - 04	7.99E - 06	6.07E - 05	4.67E - 05	9.13
<i>rfaD</i>	1.97E - 04	1.01E - 04	1.35E - 05	1.28E - 05	4.88E - 05	-1.95
<i>tra8_3</i>	4.39E - 04	1.13E - 03	5.97E - 05	1.26E - 04	5.84E - 05	2.58
<i>b1434</i>	4.55E - 06	7.66E - 06	7.22E - 07	1.88E - 07	6.10E - 05	1.68
<i>b0281</i>	1.01E - 04	7.97E - 04	4.16E - 05	1.37E - 04	6.89E - 05	7.91
<i>sodA</i>	3.80E - 04	9.74E - 04	1.06E - 04	6.26E - 05	6.99E - 05	2.57
<i>yagP</i>	2.43E - 05	1.46E - 04	7.12E - 06	2.44E - 05	7.67E - 05	5.98
<i>b1375</i>	3.10E - 05	7.29E - 05	5.44E - 06	6.97E - 06	7.88E - 05	2.35
<i>b1511</i>	7.98E - 07	8.68E - 06	5.53E - 07	7.95E - 06	9.48E - 05	10.87
<i>lacY</i>	1.62E - 03	4.08E - 04	2.53E - 04	7.95E - 05	9.79E - 05	-3.96

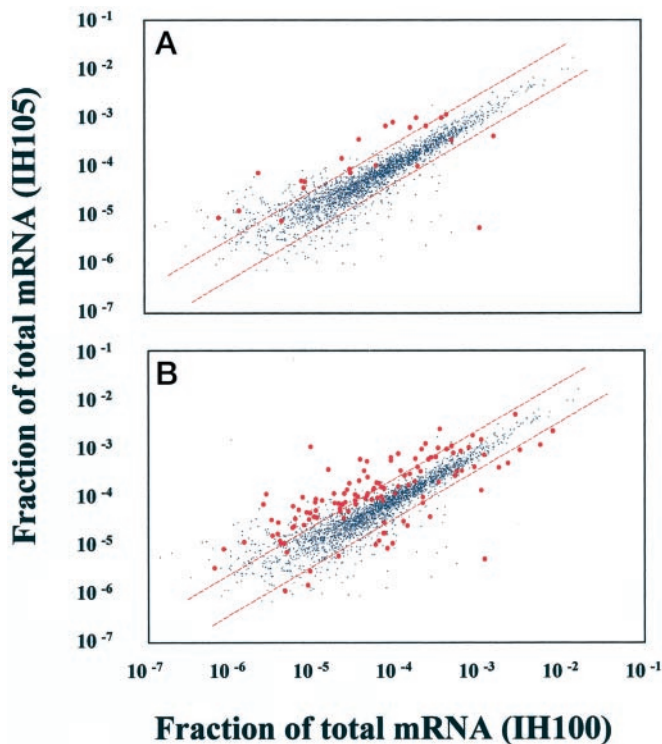


FIG. 3. Scatter plots showing the mean of the fractional mRNA levels obtained from eight filters hybridized with <sup>33</sup>P-labeled cDNA probes prepared from total RNA preparations extracted from *E. coli* K12 strains IH100 (IHF<sup>+</sup>) and IH105 (IHF<sup>-</sup>). A, the larger red dots identify 23 genes differentially expressed between strains IH100 and IH105 with *p* values less than 0.0001 (Table I). B, the larger red dots identify 124 genes differentially expressed between strains IH100 and IH105 with *p* values less than 0.005 (Table II). The dashed red lines demarcate the limits of 2-fold differences in expression levels.

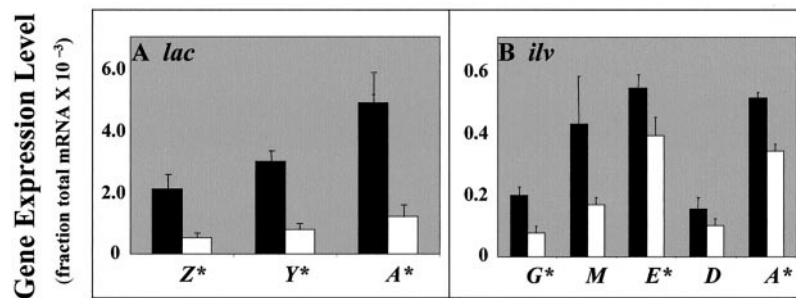
23 genes are expressed at different levels in strains IH100 and IH105. However, given the experimental error in these experiments, it is obvious that other genes differentially expressed between these strains will fail to pass this rigorous statistical test. Nevertheless, two genes of the *ilvP<sub>C</sub>::lacZYA*, and *ilvG*-

*MEDA* operons (*lacY* and *ilvA*) previously shown to be regulated by IHF do appear in this limited list (18–21, 32, 33). Therefore, if our measurements are meaningful, it is expected that the remaining genes of these two operons should exhibit similar expression levels and be similarly regulated. The data in Fig. 4, panels A and B, show that this is the case. The *p* values for these differences range from 0.00253 for the *lacA* gene to 0.165 for the *ilvM* gene. This suggests that reliable data can be obtained for genes that differ with *p* values considerably higher than 0.0001. Therefore, to facilitate the interpretation of our results, the statistical criterion for significant differences in gene expression levels was lowered to *p* < 0.005. This results in the identification of 124 genes that are differentially expressed between strains IH100 and IH105 (Fig. 3B) and includes genes for which we expect an IHF effect (Table II; see Fig. 6). However, raising the *p* value to < 0.005 identifies an average of four false positives in the control data sets (IH100 versus IH100 or IH105 versus IH105). Thus, for this experiment we expect four false positives in our set of 124 differentially expressed genes. This means that at this level of statistical accuracy we can be at least 97% confident of the differences we observe.

It should be emphasized that this level of global confidence (97%) is less than the local confidence of each measurement (99.5%) based on the *p* value (0.005) because of the false positives expected by chance alone given the large number of genes tested for expression differences from high density arrays. It should also be emphasized that increasing the *p* value threshold to higher levels rapidly increases the number of false positives in the control data sets (IH100 versus IH100 or IH105 versus IH105) relative to the number of genes differentially expressed at the same *p* value in the experimental (IH105 versus IH100) set and, therefore, decreases the confidence with which differentially expressed genes can be identified.

*There Is Little Correlation between the Fold Difference and the Accuracy of Differential Gene Expression Levels Obtained from DNA Microarray Measurements*—There is a popular tendency to equate the magnitude of the fold difference between the expression levels of a gene obtained under two experimental conditions and the accuracy of those measurements. This is exemplified by the fact that the manufacturer of the DNA

FIG. 4. Effects of IHF on the differential expression of the genes of the *ilvP<sub>G</sub>::lacZYA* and *ilvGMEDA* operons in *E. coli* K12 strains IH100 and IH105. The mean  $\pm$  S.D. expression levels in *E. coli* K12 strain IH100 (black bars, IHF<sup>+</sup>) and IH105 (white bars, IHF<sup>-</sup>) expressed as a fraction of total mRNA. Asterisks identify genes differentially expressed with *p* values less than 0.005.



microarrays used in these experiments instructs the users of these arrays, “When comparing differences in expression levels between two arrays, a 2-fold difference in expression levels is considered significant.” The data reported in Table II demonstrate that this need not be the case. These data illustrate that *p* values for small differential expression ratios can be much lower than *p* values for large differential expression ratios. This lack of a positive correlation between fold difference and significance is even more dramatically illustrated in the scatter plot shown in Fig. 3B. Here the expression level for each of the 2,592 expressed genes in IH100 cells is plotted against the expression level of these genes in IH105 cells (small blue dots). The 124 genes differentially expressed with a probability value based on a *t* test distribution less than *p* = 0.005 (Table II) are identified as large red dots. The parallel lines demarcate the 2-fold boundary on either side of the mean for all measurements. 23 of the 124 differentially expressed genes (16%) with a *p* value < 0.005 show less than a 2-fold difference. Conversely, only 29 of the 197 genes that are more than 2-fold decreased and only 95 of the 366 genes that are increased more than 2-fold in strain IH105 can be ascertained to be differentially expressed with a 97% or greater level of confidence. Finally, a further demonstration that there is little correlation between accuracy and fold difference is provided by the observation that there is only a 25% correspondence between a list of the 100 genes with the greatest fold difference in expression levels between strains IH100 and IH105 and a list of the 100 genes with the lowest *p* values. Thus, the significance of differential gene expression measurements cannot be assessed simply by the magnitude of the fold difference between two experimental conditions.

*A Positive Correlation between mRNA Level and Protein Abundance Is Observed in E. coli.*—VanBogelen *et al.* (2) used metabolic labeling and high resolution two-dimensional gel electrophoresis to measure the cellular levels of 80 proteins growing in the same medium used for the experiments reported here. A comparison of these protein levels with our mRNA measurements shows a good correlation ( $r_p = 0.67$ ; Fig. 5). Thus, at least with this comparison with a limited number of highly expressed proteins, we observe a reasonable correspondence between cellular mRNA levels and protein abundance in *E. coli*. A similar correspondence between cellular mRNA levels and protein abundance has been reported for *Saccharomyces cerevisiae* (34). These results support our suggestion that the enzymatic clearance of mRNA degradation intermediates to free nucleotides proceeds at a relatively constant rate and that the steady-state levels of these degradation intermediates are proportional to their rates of synthesis and, therefore, their relative abundance in the cell (see above).

*Reliability of Gene Expression Profile Results Observed between (IHF<sup>+</sup>) IH100 and (IHF<sup>-</sup>) IH105 Strains*—To assess the reliability of the mRNA expression levels inferred from the DNA microarray experiments reported here, we examined the effects of IHF on the expression of the genes of the *ilvGMEDA*

operon of *E. coli*. We have previously demonstrated that the promoter regulatory region of the *ilvGMEDA* operon contains two IHF-binding sites. IHF binding to a site located 92 base pairs upstream of the transcriptional start site activates *in vitro* and *in vivo* transcription initiation from the downstream promoter 3–5-fold by a DNA supercoiling-dependent mechanism (19, 20). We have also shown that IHF binding to another site in the leader region reduces *in vitro* and *in vivo* transcription through the leader-attenuator region into the structural genes of this operon about 2-fold by enhancing transcription termination at the attenuator site (32). Thus, in an IHF<sup>-</sup> strain, transcription into the attenuator is decreased about 4-fold, but transcription through the attenuator into the structural genes is increased about 2-fold, resulting in an overall increase in the expression of the downstream genes of about 2-fold. To monitor these effects of IHF at both of its binding sites in the *ilvGMEDA* operon, we replaced the *lac* promoter regulatory region of the *lac* operon with a portion of the *ilvP<sub>G</sub>* promoter regulatory region of the *ilvGMEDA* operon that contains the upstream IHF site but lacks the attenuator and the downstream IHF-binding site. Therefore, in an IHF<sup>-</sup> strain, transcription from the *ilvP<sub>G</sub>* promoter directly into the structural genes of the *lac* operon is expected to be decreased about 4-fold, but transcription through the attenuator into the structural genes of the wild-type *ilvGMEDA* operon is expected to be decreased only 2-fold. The data in Fig. 4 show that these expected results are observed. Transcription from the *ilvP<sub>G</sub>* promoter directly into the structural genes of the *lac* operon is decreased 4.14  $\pm$  0.16-fold (Fig. 4A), but transcription through the attenuator into the first two structural genes of the *ilvGMEDA* operon is decreased only 2.52  $\pm$  0.06-fold (Fig. 4B). Also, as expected, the transcriptional level of the promoter-attenuator distal genes (*ilvE*, *-D*, and *-A*) of this operon are decreased only 1.47  $\pm$  0.13-fold due to the activity of a previously characterized internal, IHF-independent promoter, *ilvP<sub>E</sub>*, preceding the *ilvE* structural gene (35). These results agree very well with independent transcript level measurements (32) and reporter enzyme assays (21). Furthermore, our ability to detect the effect of the internal promoter on the expression levels of the promoter distal genes of this operon demonstrates that small differences can be accurately measured employing the methods described in this work.

To confirm our ability to accurately measure small differences in gene expression levels and to further verify the correspondence between our measured mRNA levels and protein levels, we assayed the activities of several enzymes in strains IH100 and IH105 and compared the ratios of these activities to the ratios of their cognate mRNA levels in these two strains. These data are presented in Table III.

Since the observations reported here accurately describe the experimentally determined effects of IHF on the expression of genes of the *ilvGMEDA* and *ilvP<sub>G</sub>::lacZYA* operons and since the differential expression levels of three of the five genes of the *ilvGMEDA* operon and two of the genes of the *ilvP<sub>G</sub>::lacZYA*

TABLE II

Genes differentially expressed between *E. coli* K12 strains IH100 (IHF<sup>+</sup>) and IH105 (IHF<sup>-</sup>) with a *p* value less than 0.005

The data are presented as the average (Avg) and S.D. of four independent gene expression measurements expressed as a fraction of the total hybridization signal (total mRNA) on each DNA microarray filter. The *p* values are calculated on the basis of the t test distribution. Positive fold differences indicate increased gene expression in strain IH105. Negative fold differences indicate decreased gene expression in strain IH105.

Gene	Avg		S.D.		<i>p</i> value	Fold	Gene product	Function
	IH 100	IH 105	IH 100	IH 105				
<i>glnA</i>	2.91E - 03	9.39E - 04	6.80E - 04	1.33E - 04	1.30E - 03	-3.1	Glutamine synthetase	Amino acid biosynthesis: glutamine
<i>hisB</i>	5.67E - 04	9.00E - 04	7.38E - 05	1.13E - 04	2.60E - 03	1.59	Histidinol-phosphate phosphatase	Amino acid biosynthesis: histidine
<i>ilvA</i>	5.06E - 04	3.42E - 04	1.86E - 05	2.26E - 05	3.00E - 05	-1.48	Threonine deaminase (dehydratase)	Amino acid biosynthesis: isoleucine
<i>ilvE</i>	5.81E - 04	3.58E - 04	4.70E - 05	5.77E - 05	9.80E - 04	-1.62	Branched-chain amino-acid aminotransferase	Amino acid biosynthesis: isoleucine, valine
<i>ilvG</i>	1.97E - 04	7.67E - 05	2.65E - 05	2.23E - 05	4.40E - 04	-2.57	Acetolactate synthase II large subunit cryptic	Amino acid biosynthesis: isoleucine, valine
<i>leuA</i>	6.99E - 04	1.07E - 03	9.21E - 05	9.23E - 05	1.30E - 03	1.53	2-Isopropylmalate synthase	Amino acid biosynthesis: leucine
<i>thrA</i>	9.94E - 04	1.52E - 03	9.14E - 05	2.10E - 04	3.60E - 03	1.53	Aspartokinase I homoserine dehydrogenase I	Amino acid biosynthesis: threonine
<i>menA</i>	5.34E - 05	1.07E - 05	3.29E - 05	9.67E - 06	3.10E - 03	-4.99	1,4-Dihydroxy-2-naphthoate → dimethylmenaquinone	Biosynthesis of cofactors
<i>nadB</i>	8.15E - 06	1.59E - 06	3.04E - 06	1.31E - 06	2.10E - 03	-5.11	Quinolate synthetase B protein	Biosynthesis of cofactors
<i>b2103</i>	7.15E - 06	1.03E - 04	2.46E - 06	2.51E - 05	2.70E - 04	14.4	Phosphomethylpyrimidine kinase	Biosynthesis of cofactors prosthetic groups and carriers
<i>cobT</i>	1.00E - 05	7.97E - 05	7.82E - 06	2.13E - 05	8.50E - 04	7.97	Dimethylbenzimidazole-P phosphoribosyl transferase	Biosynthesis of cofactors prosthetic groups and carriers
<i>cobU</i>	4.26E - 05	1.22E - 04	1.79E - 05	1.95E - 05	9.90E - 04	2.85	Cobinamide kinase/phosphate guanylyltransferase	Biosynthesis of cofactors prosthetic groups and carriers
<i>lacA</i>	5.14E - 03	1.21E - 03	1.54E - 03	3.52E - 04	2.50E - 03	-4.24	Thiogalactoside acetyltransferase	Carbon compound catabolism
<i>lacZ</i>	2.10E - 03	5.14E - 04	3.77E - 04	1.34E - 04	2.20E - 04	-4.08	β-D-galactosidase	Carbon compound catabolism
<i>ddl</i>	1.85E - 04	3.42E - 04	5.62E - 05	3.87E - 05	3.60E - 03	1.85	D-Lactate dehydrogenase FAD protein	Carbon compound catabolism
<i>gmhA</i>	6.20E - 05	1.64E - 04	2.11E - 05	3.75E - 05	3.30E - 03	2.64	Phosphoheptose isomerase	Cell structure
<i>ompF</i>	7.23E - 03	2.35E - 03	1.90E - 03	3.69E - 04	2.40E - 03	-3.07	Outer membrane protein Ia (Ia;b;F)	Cell structure
<i>rfaD</i>	1.97E - 04	1.01E - 04	1.35E - 05	1.28E - 05	4.90E - 05	-1.95	ADP-L-glycero-D-mannoheptose-6-epimerase	Cell structure
<i>b1651</i>	1.71E - 04	4.31E - 04	3.62E - 05	6.98E - 05	5.70E - 04	2.53	Lactoylglutathione lyase	Central intermediary metabolism
<i>cysQ</i>	1.21E - 05	8.80E - 05	5.71E - 06	1.71E - 05	3.00E - 03	7.27	Regulator of sulfite synthesis	Central intermediary metabolism
<i>gltD</i>	9.91E - 04	1.40E - 04	1.88E - 04	3.06E - 05	1.10E - 04	-7.1	Glutamate synthase small subunit	Central intermediary metabolism
<i>hdhA</i>	4.20E - 05	5.48E - 04	2.11E - 05	1.33E - 04	2.80E - 04	13.05	NAD-dependent 7 α-hydroxysteroid dehydrogenase	Central intermediary metabolism
<i>lpdA</i>	1.07E - 03	7.60E - 04	1.17E - 04	7.75E - 05	4.60E - 03	-1.41	Lipoamide dehydrogenase (NADH)	Central intermediary metabolism
<i>rffT</i>	5.81E - 06	3.65E - 05	4.66E - 06	2.86E - 05	9.40E - 04	6.28	TDP-Fuc4NAc:lipid II transferase	Central intermediary metabolism
<i>himD</i>	3.81E - 04	6.54E - 04	9.88E - 05	4.14E - 05	2.20E - 03	1.72	IHF β subunit	DNA replication, recombination, gene expression
<i>ndh</i>	5.03E - 05	1.46E - 04	1.94E - 05	3.29E - 05	2.50E - 03	2.9	Respiratory NADH dehydrogenase	Energy metabolism
<i>pta</i>	1.38E - 04	3.04E - 04	2.46E - 05	6.54E - 05	3.10E - 03	2.21	Phosphotransacetylase	Energy metabolism
<i>b0598</i>	2.01E - 05	5.08E - 05	1.06E - 05	8.03E - 06	4.70E - 03	2.52	Carbon starvation protein	Environmental, metabolic response
<i>cheR</i>	1.29E - 04	2.68E - 05	2.07E - 04	1.75E - 05	1.30E - 03	-4.82	Glutamate methyltransferase	Environmental, metabolic response
<i>cspC</i>	7.40E - 05	9.00E - 06	6.08E - 05	8.40E - 06	5.90E - 04	-8.22	Cold shock protein	Environmental, metabolic response
<i>htpX</i>	5.54E - 05	1.99E - 04	2.61E - 05	4.27E - 05	1.20E - 03	3.59	Integral membrane heat shock protein	Environmental, metabolic response
<i>mdaB</i>	8.55E - 05	1.16E - 05	8.22E - 05	5.78E - 06	1.30E - 04	-7.37	Modulator of drug activity	Environmental, metabolic response
<i>sodA</i>	3.80E - 04	9.74E - 04	1.06E - 04	6.26E - 05	7.00E - 05	2.57	Superoxide dismutase manganese	Environmental, metabolic response
<i>sodB</i>	7.80E - 04	1.91E - 03	2.45E - 04	4.11E - 04	3.30E - 03	2.44	Superoxide dismutase iron	Environmental, metabolic response
<i>cpdB</i>	1.92E - 05	7.56E - 05	1.24E - 05	1.40E - 05	9.50E - 04	3.94	2',3'-Cyclic-nucleotide 2'-phosphodiesterase	Nucleotide biosynthesis and metabolism

TABLE II—continued

Gene	Avg		S.D.		<i>p</i> value	Fold	Gene product	Function
	IH 100	IH 105	IH 100	IH 105				
<i>guaA</i>	8.25E-04	4.31E-04	5.43E-05	1.34E-04	1.60E-03	-1.91	GMP synthetase (glutamine-hydrolyzing)	Nucleotide biosynthesis and metabolism
<i>b0255</i>	5.43E-06	6.73E-05	7.35E-06	8.45E-06	6.90E-04	12.38	IS911 hypothetical protein	Phage
<i>intB</i>	2.41E-06	7.33E-05	2.28E-06	8.27E-06	3.10E-06	30.38	Prophage P4 integrase	Phage
<i>b0281</i>	1.01E-04	7.97E-04	4.16E-05	1.37E-04	6.90E-05	7.91	Putative phage integrase	Phage transposon
<i>ogrK</i>	8.84E-06	1.10E-03	5.89E-06	1.62E-04	8.90E-04	124.65	Prophage P2 <i>ogr</i> protein	Phage transposon
<i>tra8_2</i>	2.48E-04	6.58E-04	3.59E-05	5.57E-05	1.70E-05	2.65	IS30 transposase	Phage transposon
<i>tra8_3</i>	4.39E-04	1.13E-03	5.97E-05	1.26E-04	5.80E-05	2.58	IS30 transposase	Phage transposon
<i>tra8_1</i>	2.81E-04	6.92E-04	2.86E-05	1.06E-04	2.90E-04	2.47	IS30 transposase	Phage, transposon, or plasmid
<i>b1835</i>	1.91E-05	6.35E-06	1.47E-05	1.73E-06	6.70E-04	-3.02	Putative nucleolar protein	Putative cell structure
<i>b0877</i>	4.75E-05	8.92E-05	1.73E-05	8.08E-06	4.70E-03	1.88	Putative enzyme	Putative enzymes
<i>b1035</i>	3.30E-05	1.03E-04	1.09E-05	2.64E-05	2.60E-03	3.14	Putative oxidoreductase component	Putative enzymes
<i>b1511</i>	7.98E-07	8.68E-06	5.53E-07	7.95E-06	9.50E-05	10.87	Putative kinase	Putative enzymes
<i>b2255</i>	4.80E-04	2.81E-04	4.68E-05	1.97E-05	2.30E-04	-1.71	Putative transformylase	Putative enzymes
<i>yhjS</i>	1.33E-04	1.78E-04	6.34E-06	1.44E-05	1.40E-03	1.33	Putative protease	Putative enzymes
<i>yjhO</i>	3.81E-06	1.29E-05	2.02E-06	1.66E-05	2.90E-03	3.38	Putative lyase/synthase	Putative enzymes
<i>b1360</i>	6.77E-05	3.10E-05	1.54E-05	6.42E-06	4.60E-03	-2.18	Putative DNA replication factor	Putative factors
<i>b1377</i>	6.34E-06	5.55E-05	7.02E-06	2.05E-05	4.00E-03	8.75	Putative outer membrane protein	Putative membrane protein
<i>b1284</i>	5.84E-05	1.96E-04	3.72E-05	5.07E-05	4.70E-03	3.36	Putative deoR-type transcriptional regulator	Putative regulatory proteins
<i>b2381</i>	4.33E-06	1.16E-05	1.88E-06	7.68E-07	3.30E-04	2.67	Putative 2-component transcriptional regulator	Putative regulatory proteins
<i>b2556</i>	3.92E-05	3.57E-04	7.99E-06	6.07E-05	4.70E-05	9.13	Putative 2-component sensor protein	Putative regulatory proteins
<i>b3515</i>	2.30E-04	6.15E-04	4.86E-05	1.62E-04	3.80E-03	2.68	Putative <i>araC</i> -type regulatory protein	Putative regulatory proteins
<i>creA</i>	2.39E-06	7.26E-05	2.49E-06	3.16E-05	4.50E-03	30.38	ORF hypothetical protein	Putative regulatory proteins
<i>yagP</i>	2.43E-05	1.46E-04	7.12E-06	2.44E-05	7.70E-05	5.98	Putative transcriptional regulator <i>lysR</i> -type	Putative regulatory proteins
<i>yghB</i>	2.58E-05	1.97E-04	1.87E-05	3.65E-05	1.60E-04	7.63	Putative 2-component regulator	Putative regulatory proteins
<i>yiaJ</i>	3.47E-05	6.15E-04	1.74E-05	1.64E-04	4.10E-04	17.74	Putative regulator	Putative regulatory proteins
<i>yiaU</i>	3.38E-05	2.24E-04	1.12E-05	4.49E-05	1.80E-04	6.61	Putative transcriptional regulator <i>lysR</i> -type	Putative regulatory proteins
<i>b4115</i>	3.01E-06	3.50E-05	2.91E-06	1.93E-05	2.10E-04	11.59	Putative amino acid/amine transport protein cryptic	Putative transport proteins
<i>dsdX</i>	1.05E-05	3.88E-05	5.23E-06	2.44E-05	1.70E-03	3.7	Transport system permease (serine?)	Putative transport proteins
<i>oppD</i>	2.32E-05	8.02E-05	1.81E-05	1.66E-05	3.50E-03	3.46	Putative ATP-binding protein, ABC peptide transport	Putative transport proteins
<i>sapA</i>	8.17E-05	5.76E-04	2.78E-05	1.42E-04	4.90E-04	7.05	Peptide transport periplasmic protein	Putative transport proteins
<i>sapD</i>	4.28E-05	2.68E-04	3.70E-05	5.93E-05	6.70E-04	6.25	Putative ATP-binding protein for peptide transport	Putative transport proteins
<i>sapF</i>	1.02E-05	9.33E-05	7.68E-06	1.16E-05	7.30E-04	9.17	Putative ATP-binding protein for peptide transport	Putative transport proteins
<i>yedE</i>	6.55E-05	8.73E-05	5.21E-06	3.94E-06	5.50E-04	1.33	Putative transport system permease protein	Putative transport proteins
<i>yeeE</i>	8.20E-05	6.64E-04	1.83E-05	6.37E-05	2.20E-06	8.11	Putative transport system permease protein	Putative transport proteins
<i>yeeF</i>	3.04E-04	2.06E-04	2.00E-05	3.62E-05	3.30E-03	-1.47	Putative amino acid/amine transport protein	Putative transport proteins
<i>yjiJ</i>	3.58E-06	1.83E-05	2.51E-06	3.88E-06	7.20E-04	5.1	Putative transport protein	Putative transport proteins
<i>air</i>	3.61E-06	3.10E-05	4.99E-06	4.81E-05	1.60E-03	8.59	Aerotaxis sensor receptor flavoprotein	Regulatory function
<i>dniR</i>	1.91E-04	9.75E-04	4.20E-05	1.09E-04	1.10E-05	5.1	Transcriptional regulator for nitrite reductase	Regulatory function
<i>ebgR</i>	5.06E-05	1.55E-04	1.92E-05	1.27E-05	1.00E-04	3.06	Regulator of <i>ebg</i> operon	Regulatory function
<i>emrR</i>	5.82E-05	1.31E-05	7.00E-05	6.46E-06	8.00E-04	-4.45	Regulator of plasmid <i>mcrB</i> operon	Regulatory function
<i>glnL</i>	2.41E-04	3.99E-05	4.81E-05	2.81E-05	3.60E-04	-6.04	Sensor for <i>GlnG</i> regulator (nitrogen regulator II, NRII)	Regulatory function
<i>rscF</i>	1.20E-04	2.77E-04	4.41E-05	3.69E-05	1.50E-03	2.32	Regulator in colanic acid synthesis; interacts with <i>RcsB</i>	Regulatory function
<i>arcA</i>	3.01E-05	1.16E-04	1.98E-05	3.18E-05	3.80E-03	3.86	Regulator of genes in aerobic pathways	Regulatory function
<i>lacY</i>	1.62E-03	4.08E-04	2.53E-04	7.95E-05	9.80E-05	-3.96	Galactoside permease (M protein)	Transport and binding proteins

TABLE II—continued

Gene	Avg		S.D.		<i>p</i> value	Fold	Gene product	Function
	IH 100	IH 105	IH 100	IH 105				
<i>oppA</i>	2.54E - 03	5.06E - 03	1.72E - 04	5.68E - 04	1.40E - 04	2	Oligopeptide transport, periplasmic-binding protein	Transport and binding proteins
<i>oppB</i>	1.06E - 04	3.57E - 04	3.06E - 05	6.22E - 05	3.60E - 04	3.35	Oligopeptide transport permease protein	Transport and binding proteins
<i>potC</i>	9.07E - 05	1.23E - 04	1.13E - 05	7.27E - 06	3.10E - 03	1.35	Spermidine/putrescine transport system permease	Transport and binding proteins
<i>proV</i>	2.50E - 05	5.30E - 05	7.34E - 06	9.57E - 06	3.60E - 03	2.12	Component for glycine betaine and proline transport	Transport and binding proteins
<i>rbsC</i>	4.20E - 05	1.12E - 04	1.47E - 05	2.70E - 05	3.90E - 03	2.67	D-Ribose high affinity transport system	Transport and binding proteins
<i>b0280</i>	2.99E - 04	1.04E - 03	7.23E - 05	3.23E - 04	4.30E - 03	3.47	ORF hypothetical protein	Unknown
<i>b0926</i>	1.16E - 04	3.27E - 05	2.74E - 05	1.71E - 05	2.10E - 03	-3.55	ORF hypothetical protein	Unknown
<i>b0927</i>	6.61E - 05	1.83E - 05	1.60E - 05	1.25E - 05	3.30E - 03	-3.61	ORF hypothetical protein	Unknown
<i>b1034</i>	1.58E - 04	2.13E - 04	1.38E - 05	2.12E - 05	4.50E - 03	1.35	ORF hypothetical protein	Unknown
<i>b1111</i>	1.44E - 05	3.72E - 04	1.67E - 05	1.26E - 04	1.30E - 03	25.92	ORF hypothetical protein	Unknown
<i>b1285</i>	3.04E - 05	8.74E - 05	7.35E - 06	5.64E - 06	1.80E - 05	2.87	ORF hypothetical protein	Unknown
<i>b1375</i>	3.10E - 05	7.29E - 05	5.44E - 06	6.97E - 06	7.90E - 05	2.35	ORF hypothetical protein	Unknown
<i>b1381</i>	9.70E - 06	4.05E - 05	4.65E - 06	9.07E - 06	9.30E - 04	4.17	ORF hypothetical protein	Unknown
<i>b1383</i>	1.72E - 05	7.66E - 05	8.53E - 06	1.96E - 05	1.50E - 03	4.45	ORF hypothetical protein	Unknown
<i>b1434</i>	4.55E - 06	7.66E - 06	7.22E - 07	1.88E - 07	6.10E - 05	1.68	ORF hypothetical protein	Unknown
<i>b1585</i>	8.69E - 06	4.86E - 05	4.44E - 06	5.62E - 06	3.10E - 05	5.59	ORF hypothetical protein	Unknown
<i>b1667</i>	6.21E - 05	1.04E - 04	4.01E - 06	4.71E - 06	1.00E - 05	1.67	ORF hypothetical protein	Unknown
<i>b1725</i>	2.20E - 05	3.78E - 05	3.64E - 06	2.13E - 06	2.90E - 04	1.72	ORF hypothetical protein	Unknown
<i>b1762</i>	2.23E - 05	1.22E - 04	1.19E - 05	3.79E - 05	2.40E - 03	5.46	ORF hypothetical protein	Unknown
<i>b1783</i>	6.88E - 05	1.83E - 05	1.22E - 06	1.74E - 05	1.20E - 03	-3.75	ORF hypothetical protein	Unknown
<i>b2006</i>	8.49E - 06	3.58E - 05	3.25E - 06	4.05E - 06	4.30E - 05	4.22	ORF hypothetical protein	Unknown
<i>b2226</i>	1.41E - 06	1.23E - 05	1.82E - 07	1.67E - 05	1.30E - 06	8.69	ORF hypothetical protein	Unknown
<i>b2295</i>	5.60E - 05	9.84E - 05	7.08E - 06	7.89E - 06	2.00E - 04	1.76	ORF hypothetical protein	Unknown
<i>b2433</i>	3.09E - 06	1.66E - 05	1.44E - 06	2.47E - 05	3.40E - 04	5.36	ORF hypothetical protein	Unknown
<i>b2873</i>	2.61E - 06	1.17E - 04	2.01E - 06	2.93E - 05	2.30E - 04	44.99	ORF hypothetical protein	Unknown
<i>b2876</i>	7.94E - 06	5.01E - 05	6.05E - 06	3.45E - 06	1.90E - 05	6.3	ORF hypothetical protein	Unknown
<i>b2883</i>	1.02E - 05	5.50E - 05	1.04E - 05	1.03E - 05	8.70E - 04	5.39	ORF hypothetical protein	Unknown
<i>b2973</i>	2.13E - 05	7.40E - 05	5.65E - 06	1.25E - 05	2.50E - 04	3.48	ORF hypothetical protein	Unknown
<i>b2983</i>	1.99E - 05	5.87E - 05	9.46E - 06	6.07E - 06	4.60E - 04	2.95	ORF hypothetical protein	Unknown
<i>hdeB</i>	1.09E - 03	5.51E - 06	1.80E - 04	3.47E - 06	2.00E - 05	-198.5	ORF hypothetical protein	Unknown
<i>ybaA</i>	2.06E - 05	7.19E - 05	1.39E - 05	1.25E - 05	3.70E - 03	3.5	ORF hypothetical protein	Unknown
<i>yciO</i>	4.03E - 05	9.54E - 05	5.26E - 06	1.43E - 05	3.60E - 04	2.37	ORF hypothetical protein	Unknown
<i>yebE</i>	3.90E - 06	1.14E - 05	1.35E - 06	1.61E - 05	1.60E - 03	2.93	ORF hypothetical protein	Unknown
<i>yecH</i>	6.19E - 07	3.54E - 06	3.26E - 07	4.00E - 06	3.80E - 04	5.72	ORF hypothetical protein	Unknown
<i>yeeD</i>	1.61E - 04	6.26E - 04	2.38E - 05	5.05E - 05	3.00E - 06	3.9	ORF hypothetical protein	Unknown
<i>yefM</i>	4.63E - 04	8.12E - 04	5.02E - 05	6.07E - 05	1.10E - 04	1.75	ORF hypothetical protein	Unknown
<i>yhiE</i>	8.72E - 06	3.12E - 06	1.40E - 06	1.96E - 06	3.50E - 03	-2.8	ORF hypothetical protein	Unknown
<i>yieI</i>	5.62E - 06	2.56E - 05	3.03E - 06	4.87E - 06	4.40E - 04	4.55	ORF hypothetical protein	Unknown
<i>yifR</i>	1.68E - 05	1.20E - 04	9.59E - 06	4.61E - 05	4.60E - 03	7.14	ORF hypothetical protein	Unknown
<i>yifY</i>	4.39E - 06	1.23E - 06	1.22E - 06	7.13E - 07	3.00E - 03	-3.58	ORF hypothetical protein	Unknown
<i>yjiI</i>	7.34E - 06	2.70E - 05	4.56E - 06	4.60E - 06	9.00E - 04	3.68	ORF hypothetical protein	Unknown
<i>yjiU</i>	6.74E - 05	1.53E - 04	1.41E - 05	2.40E - 05	8.70E - 04	2.26	ORF hypothetical protein	Unknown
<i>b1743</i>	5.33E - 06	2.40E - 05	2.61E - 06	6.81E - 06	2.20E - 03	4.51	Periplasmic protein related to spheroplast formation	Unknown
<i>bcp</i>	1.00E - 04	1.68E - 04	2.33E - 05	1.24E - 05	2.20E - 03	1.67	Bacterioferritin comigratory protein	Unknown

operon in strains IH100 and IH105 are measured with *p* values less than 0.005, it is reasonable to conclude that equally accurate measures of differential expression can be obtained for other genes at this level of statistical rigor. The mean, S.D., *p* values, and fold change for the 124 genes differentially expressed between strains IH100 and IH105 with a *p* value less than 0.005 are listed according to their metabolic functions in Table II. The functions of 64 of these genes have been determined, and putative functions have been assigned to 11 of the remaining 60 genes. IHF negatively regulates 95 and positively regulates 29 of these genes.

**Identification of Putative IHF-binding Sites in the Promoter Regulatory Regions of Genes Differentially Expressed in Strains IH100 and IH105**—To identify those genes in Table II differentially expressed in strains IH100 and IH105 most likely as a direct consequence of IHF-mediated effects on transcription initiation, 500 base pairs upstream of the ORF for each of these genes were examined for high affinity IHF-binding sites. Because the core consensus recognition sequence for IHF-binding

sites is degenerate (5'-(A/T)ATCAANNNTTR-3'; N = any nucleotide and R = purine) and because sequence variable structural properties of the DNA duplex are important for high affinity IHF binding, these sites are difficult to identify (8). For example, a search of the *E. coli* chromosome for an 11 out of 13 match for the degenerate IHF core consensus sequence identified nearly 40,000 potential binding sites. It was, therefore, necessary to define more restrictive criteria for the identification of putative IHF-binding sites. We, therefore, demanded a 12 out of 13 match with the core consensus sequence and required that at least 10 out of 15 of the base pairs immediately upstream of the core sequence were AT base pairs. These are very stringent criteria that certainly identify high affinity IHF-binding sites; in fact, they exceed the criteria for many documented IHF-binding sites such as the IHF site in the leader-attenuator region of the *ilvGMEDA* operon. Also, IHF-binding sites that perform a DNA looper function located farther than 500 base pairs upstream of an ORF would not be identified in our search. Nevertheless, these stringent criteria identify

IHF-binding sites upstream of 46 genes (operons) differentially regulated in strains IH100 and IH105 with a  $p$  value less than 0.005 (Table II). The locations of these putative IHF-binding sites as well as previously documented sites are shown in (Fig. 6).

**Examples of Genes Only Expressed in Either Strain IH100 or Strain IH105**—The single genotypic difference between the strains used for the studies reported here is a deletion in strain IH105 of the *himA* gene, the structural gene for the  $\alpha$ -subunit of IHF. As expected, the data in Table IV show that no *himA* mRNA is detected in strain IH105, but it is detected in the *himA* wild-type strain IH100. Indeed, *himA* mRNA was detected in all four IH100 mRNA pools, but no *himA* mRNA was detected in any of the IH105 mRNA pools. Although no  $p$  value can be calculated when no mRNA for a gene is detected, these data suggest that the expression of a gene in all of the mRNA samples of one strain and no expression in any of the mRNA samples of the other strain likely represents a gene that is either strongly repressed by IHF or strongly dependent upon IHF for its expression.

In addition to *himA*, the data in Table IV show that six other genes are consistently expressed only in strain IH100 (putative IHF-dependent genes), and eight are consistently expressed only in strain IH105 (putative IHF-repressed genes). Two genes in Table IV that appear to require IHF for expression are the *fimI* and *fimC* genes. These genes are members of a large operon (*fimBEAICDFGH*). Although none of the genes of this operon are differentially expressed with a  $p$  value less than 0.005, they are all expressed in strain IH100 and are either nondetectable or expressed at very much lower levels in strain IH105. This operon is expressed from two documented transcription start sites located approximately 150 and 250 base pairs upstream of the first ORF of this operon, and a putative IHF site is observed near the upstream promoter (Fig. 6). This arrangement of tandem promoters and an IHF-binding site is similar to that observed in the  $\lambda$  P<sub>L</sub> and *ilvGMEDA* tandem promoter regions. In these cases, transcription from the up-

stream promoter is repressed, and transcription from the downstream promoter is activated (36). These observations demonstrate that potentially important biological information might be missed if data analysis is limited to strict statistical measures.

Dps is an abundant, nonspecific *E. coli* DNA-binding protein important for protection from hydrogen peroxide-induced DNA damage. In log phase cells growing in rich medium, expression of the *dps* gene from a  $\sigma^{70}$  promoter requires the OxyR protein (37). However, it has been shown that as cells enter the stationary phase, the *dps* gene is not induced by oxidative stress (even though OxyR is present), and its expression requires IHF binding to a documented site upstream of a  $\sigma^S$  promoter (Fig. 6). Our results show that during log phase growth in glucose-supplemented minimal MOPS medium in a wild-type strain (IH100) no oxyR gene expression can be detected, and as previously reported (38), the *rpoS* ( $\sigma^S$ ) gene is expressed at an intermediate level. Under this same growth condition, *dps* is expressed in an IHF<sup>+</sup> strain but is not expressed in an IHF<sup>-</sup> strain (Table IV). Thus, our data suggest that during log phase, as in stationary phase growth in minimal medium, IHF is required for low level transcription of the *dps* gene from a  $\sigma^S$  promoter. The criteria used in this report did not identify any putative high affinity IHF-binding sites upstream of the remaining genes in Table IV that require IHF for expression (*rfaJ*, *nac*, *yfiG*).

Putative IHF-binding sites were identified upstream of three of the eight genes completely repressed by IHF in strain IH100 (*b1112*, *b3592*, and *b2984*; Table IV and Fig. 6). One of these, *b1112*, is a member of what appears to be a divergently transcribed operon (Fig. 6). The location of the transcription start sites and the putative IHF-binding site in the approximately 200-base pair promoter regulatory between the oppositely oriented *b1111* and *b1112* ORFs is nearly the same as the promoter and IHF site arrangement between the similarly spaced divergently transcribed *cpd* and *cysQ* genes (Fig. 6). In both cases, both genes are repressed by IHF binding in the divergently transcribed region. This pattern of gene regulation suggests that the *cpd-cysQ* and *b1111-b1112* genes are divergently transcribed operons. Typical of other genes whose expression is affected by IHF, no functional correlation is apparent for these two sets of genes.

**Examples of Direct Effects of IHF on Gene Expression Profiles in Strains IH100 and IH105**—Although a remarkably few IHF-binding sites that affect gene expression during exponential growth in minimal medium have been documented, several of these genes appear in Tables II and IV. For example, it is known that the manganese superoxide dismutase *sodA* gene is repressed by IHF binding to four sites (Fig. 6) in the promoter region (39, 40). However, no evidence for the regulation of the iron superoxide dismutase *sodB* gene has been reported. Our data shows that the expression of both *sodA* and *sodB* is increased in strain IH105 (Table II).

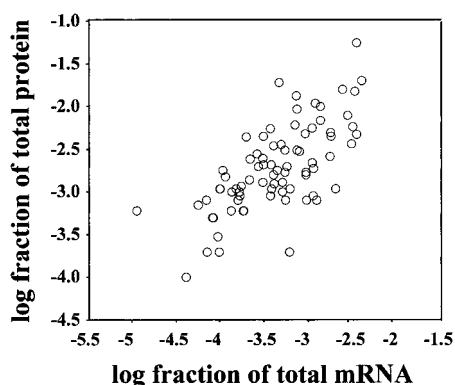


FIG. 5. Scatter plot of the relative abundance of mRNAs and their cognate proteins in *E. coli*.

TABLE III  
mRNA and enzyme activity levels in *E. coli* K12 strains IH100 (IHF<sup>+</sup>) and IH105 (IHF<sup>-</sup>)

See Methods and Materials section for enzyme assay conditions.

Enzyme	Gene	Message level		Fold	Specific activity		Fold
		IH100	IH105		IH100	IH105	
		<i>fraction total mRNA</i>			<i>nmol/min/mg protein</i>		
Imidazolylacetolphosphate:L-glutamate aminotransferase	<i>hisC</i>	$2.6 \times 10^{-4}$	$4.2 \times 10^{-4}$	1.6	$183 \pm 2$	$330 \pm 2$	1.8
Glutamine synthetase	<i>gluA</i>	$2.7 \times 10^{-3}$	$1.4 \times 10^{-3}$	-1.9	$209 \pm 4$	$125 \pm 7$	-1.7
$\alpha,\beta$ -Dihydroxyacid dehydratase	<i>ilvD</i>	$1.3 \times 10^{-4}$	$1.2 \times 10^{-4}$	-1.1	$35 \pm 2$	$38 \pm 2$	1.1
$\beta$ -Galactosidase	<i>lacZ</i>	$2.1 \times 10^{-3}$	$5.1 \times 10^{-4}$	-4.1	$4698 \pm 930$	$1407 \pm 150$	-3.3
Cystathionase	<i>metC</i>	$1.1 \times 10^{-4}$	$1.1 \times 10^{-4}$	1.0	$117 \pm 11$	$139 \pm 6$	1.2

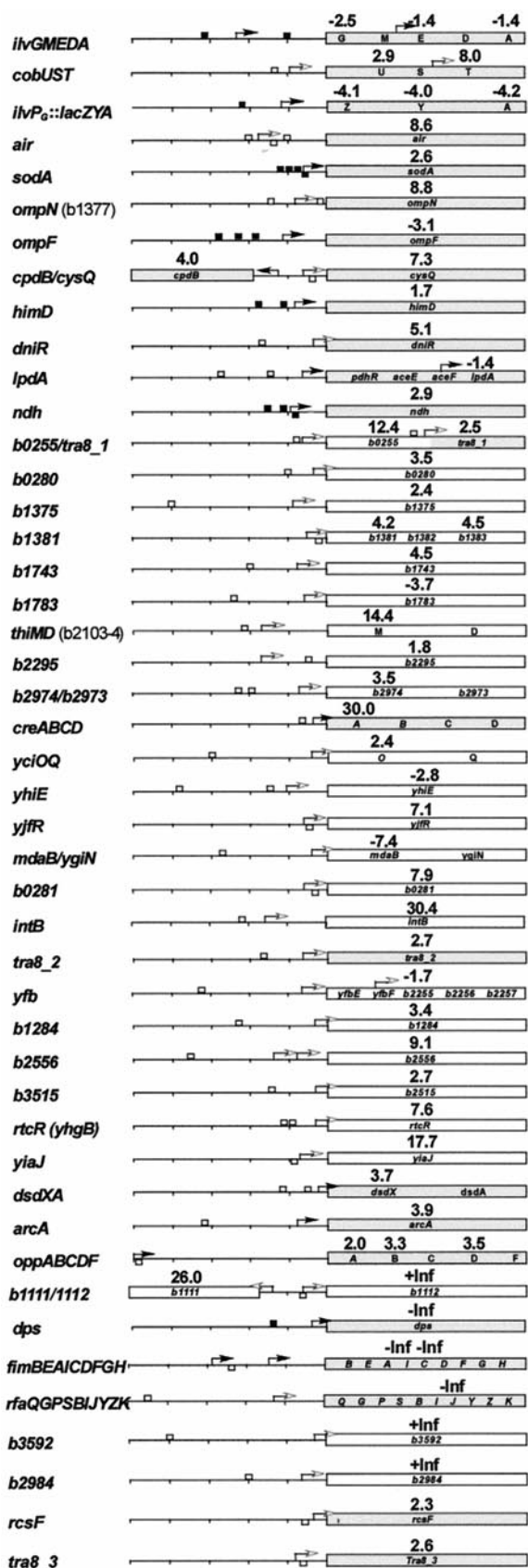


FIG. 6. Genes differentially regulated in *E. coli* K12 strains IH100 and IH105 with a *p* value less than 0.005 that contain documented or putative high affinity IHF-binding sites. ORFs of genes (operons) are represented as bars. The 500 base pairs upstream of each ORF is represented by a straight line; tick marks are spaced at 100-base pair intervals. Open bars denote predicted genes. Gray bars denote documented genes. Open arrows identify the position of pre-

TABLE IV  
Gene expression detected only in *E. coli* K12 strain IH100 (IHF<sup>+</sup>) or IH105 (IHF<sup>-</sup>)

The data are presented as the average (Avg) and standard deviation (S.D.) of four independent gene expression measurements expressed as a fraction of the total DNA hybridization signal (total mRNA) on each filter.

Gene	Avg IH100	Avg IH105	S.D. IH100	S.D. IH105
<i>himA</i>	5.08E - 04	0.00E + 00	2.02E - 04	0.00E + 00
<i>fimC</i>	2.45E - 04	0.00E + 00	1.53E - 04	0.00E + 00
<i>fimI</i>	2.23E - 04	0.00E + 00	1.37E - 04	0.00E + 00
<i>dps</i>	2.26E - 05	0.00E + 00	3.23E - 05	0.00E + 00
<i>rfaJ</i>	1.95E - 05	0.00E + 00	9.42E - 06	0.00E + 00
<i>nac</i>	1.64E - 05	0.00E + 00	6.99E - 06	0.00E + 00
<i>yfiG</i>	4.01E - 06	0.00E + 00	2.47E - 06	0.00E + 00
<i>b1112</i>	0.00E + 00	3.25E - 05	0.00E + 00	1.83E - 05
<i>b1314</i>	0.00E + 00	2.03E - 05	0.00E + 00	1.00E - 05
<i>b3592</i>	0.00E + 00	1.30E - 05	0.00E + 00	1.13E - 05
<i>fucU</i>	0.00E + 00	1.27E - 05	0.00E + 00	1.35E - 05
<i>yfiE</i>	0.00E + 00	7.26E - 06	0.00E + 00	7.87E - 06
<i>b3011</i>	0.00E + 00	6.68E - 06	0.00E + 00	1.00E - 05
<i>b2984</i>	0.00E + 00	4.90E - 06	0.00E + 00	2.64E - 06
<i>yjcH</i>	0.00E + 00	1.67E - 06	0.00E + 00	1.13E - 06

The *ndh* gene, which encodes NADH dehydrogenase II, is an example of a gene that is known to be regulated by both ArcA and IHF. A direct role for *ndh* repression by IHF binding at three sites in the promoter region (Fig. 6), consistent with its increased expression in strain IH105 (Table II), has been reported by Green *et al.* (41).

Freundlich *et al.* (13) present evidence that IHF mutants exhibit increased expression and altered osmoregulation of OmpF, a major *E. coli* outer membrane protein. They have identified upstream IHF-binding sites centered at base pair positions -179 and -61 of the *ompF* promoter regulatory region. They have also reported that the addition of IHF to a purified *in vitro* transcription system inhibits *ompF* transcription by altering how OmpR, a positive activator required for *ompF* expression, interacts with the *ompF* promoter. In contrast, our results suggest that IHF acts as an activator of *ompF* expression (Table II).

Two IHF-binding sites have been documented upstream of the promoter for the *himD* gene. Aviv *et al.* (42) show that expression of the monocistronic *himD* operon, which encodes the structural gene for the  $\beta$  subunit of IHF, is negatively autoregulated by an intact, heterodimeric IHF, and our results show that the expression of the *himD* gene is increased in the IHF<sup>-</sup> strain IH105 (Table II). However, since this derepression is only 1.7-fold, it is possible that in the absence of the *himA*-encoded  $\alpha$ -subunit for the formation of *himA-himD* encoded  $\alpha\beta$ -heterodimers, *himD*-encoded  $\beta\beta$ -homodimers can function to autoregulate *himD* expression. Indeed, several examples of the *in vivo* formation of  $\beta\beta$ -homodimers functionally competent to recognize and bind to natural IHF-binding sites have been reported (43–46).

In addition to the above examples with documented IHF-binding sites, reporter gene and other assays have been used to identify other IHF-regulated genes. Some of these genes appear in Table II, and we have identified putative IHF-binding sites for several of them (Fig. 6). For example, a well known

dicted transcriptional start sites. Black arrows identify the position of documented transcriptional start sites. Open boxes identify the position of predicted IHF-binding sites. Black boxes identify the position of documented IHF-binding sites. The level of expression of each gene in strain IH105 relative to its level of expression in strain IH100 is shown above each gene. The operon organizations, the positions of the transcriptional start sites, and the documented IHF-binding sites were obtained either from GenBank<sup>®</sup> or RegulonDB.

regulatory function for IHF involves its role in conjugal transfer of F plasmid DNA by affecting transcription of the transfer (*tra*) operon (47). Indeed, our results show that all three of the chromosomally encoded *tra8\_1*, *tra8\_2*, and *tra8\_3* genes contained in the chromosome of our strains are repressed by IHF, and putative IHF-binding sites are observed in the promoter region of each gene.

The data in Table II show that IHF affects the expression of one known global regulatory gene (*arcA*), several operon-specific regulatory proteins, and 11 genes encoding putative regulatory proteins. This suggests that IHF might indirectly affect the expression of many genes via its effect on the expression of regulatory genes. At this time, we are aware of a previously demonstrated direct effect of IHF on the regulation of only one of these regulatory genes, *arcA*. ArcA is a global regulator protein for genes involved in anaerobic carbon metabolism (48). Reporter enzyme assays and nested deletion experiments have suggested the presence of an IHF site in the *arcA* promoter regulatory region and shown that *arcA* gene expression is elevated more than 3-fold in an IHF<sup>-</sup> *E. coli* strain growing in glucose minimal medium under aerobic conditions.<sup>5</sup> We have identified an IHF-binding site in the *arcA* promoter regulatory region, and our data show a 3.9-fold increase in *arcA* mRNA expression in strain IH105 growing under similar conditions. Gunsalus and co-workers (49–52) also use reporter enzyme assays to examine the expression of several tricarboxylic acid cycle (*fumA*, *gltA*, *icdA*, *mdh*, *mur*, and *sucA*) and respiratory genes (*atp*, *cydA*, and *cyoA*) in IHF<sup>+</sup> and IHF<sup>-</sup> strains. In each case the small mRNA expression level differences (often less than 2-fold) between our IHF<sup>+</sup> and IHF<sup>-</sup> strains agree well with their reporter enzyme assay data. However, because the expression of each of these genes is also affected by ArcA protein, it is unclear whether or not these small IHF effects are direct or indirect. On the other hand, *ndh* is an example of a gene that is known to be regulated by both ArcA and IHF. In this case, a direct role for *ndh* repression by IHF, consistent with its increased expression in strain IH105 (Table II), has been reported by Green *et al.* (41).

Other studies in *Salmonella typhimurium* establish that the *arcA* gene product activates the expression of the *cob* operon required for cobalamin synthesis (53). Therefore, it is again unclear whether or not the increases in the *cobU* and *cobT* genes reported in Table II are the result of direct or indirect effects of IHF. However, the fact that we find a putative IHF site near the promoter of the *cob* operon suggests a direct role for IHF in the regulation of this operon.

The remaining genes in Fig. 6 are genes identified by this work that are differentially expressed between strains IH100 and IH105 with a *p* value less than 0.005 and contain putative high affinity IHF-binding sites. The remaining genes in Table II that do not contain either documented or putative IHF-binding sites may represent genes that are indirectly affected by IHF.

**Acknowledgments**—We gratefully acknowledge the many helpful discussions, advice, and computational assistance received from Dr. Suzanne B. Sandmeyer and the University of California at Irvine Experimental and Computational Genomics Group. We also acknowledge the helpful comments and suggestions of an anonymous reviewer.

**Note Added in Proof**—Following the submission of this work for publication, confirmation that, as reported here, the expression of the *sodB* gene of *E. coli* is repressed approximately 2-fold by IHF has been published by Dubrac and Touati (54).

## REFERENCES

- Jacob, F. & Monod, J. (1961) *Cold Spring Harbor Symp. Quant. Biol.* **25**, 193–209
- VanBogelen, R. A., Abshire, K. Z., Pertsemilidis, A., Clark, R. L. & Neidhardt, F. C. (1996) in *Escherichia coli and Salmonella Cellular and Molecular Biology* (Neidhardt, F. C., Curtis, R. I., Ingraham, J. L., Lin, E. C. C., Low, K. B., Magasanik, B., Reznikoff, W. S., Riley, M., Schaechter, M., and Umbarger, H. E., eds) pp. 2067–2117, American Society for Microbiology, Washington, D. C.
- Neidhardt, F. C. & Savageau, M. A. (1996) in *Escherichia coli and Salmonella Cellular and Molecular Biology* (Neidhardt, F. C., Curtis, R. I., Ingraham, J. L., Lin, E. C. C., Low, K. B., Magasanik, B., Reznikoff, W. S., Riley, M., Schaechter, M., and Umbarger, H. E., eds) pp. 1310–1324, American Society for Microbiology, Washington, D. C.
- Kolb, A., Busby, S., Buc, H., Garges, S. & Adhya, S. (1993) *Annu. Rev. Biochem.* **62**, 749–795
- Ditto, M. D., Roberts, D. & Weisberg, R. A. (1994) *J. Bacteriol.* **176**, 3738–3748
- Oberto, J., Drlaca, K. & Rouviere-Yaniv, J. (1994) *Biochimie (Paris)* **76**, 901–908
- Ellenberger, T. & Landy, A. (1997) *Structure (Lond.)* **5**, 153–157
- Goodman, S. D., Velten, N. J., Gao, Q., Robinson, S. & Segall, A. M. (1999) *J. Bacteriol.* **181**, 3246–3255
- Craig, N. L. & Nash, H. A. (1984) *Cell* **39**, 707–716
- Goodrich, J. A., Schwartz, M. L. & McClure, W. R. (1990) *Nucleic Acids Res.* **18**, 4993–5000
- Miller, H. I. & Friedman, D. I. (1980) *Cell* **20**, 711–719
- Friedman, D. I. (1988) *Cell* **55**, 545–554
- Freundlich, M., Ramani, N., Mathew, E., Sirko, A. & Tsui, P. (1992) *Mol. Microbiol.* **6**, 2557–2563
- Moitosa, D. V., Kim, S. & Landy, A. (1989) *Science* **244**, 1457–1461
- Claverie-Martin, F. & Magasanik, B. (1992) *J. Mol. Biol.* **227**, 996–1008
- Hoover, T. R., Santero, E., Porter, S. & Kustu, S. (1990) *Cell* **63**, 11–22
- Yang, S. W. & Nash, H. A. (1995) *EMBO J.* **14**, 6292–6300
- Parekh, B. S., Sheridan, S. D. & Hatfield, G. W. (1996) *J. Biol. Chem.* **271**, 20258–20264
- Sheridan, S. D., Benham, C. J. & Hatfield, G. W. (1998) *J. Biol. Chem.* **273**, 21298–21308
- Sheridan, S. D., Benham, C. J. & Hatfield, G. W. (1999) *J. Biol. Chem.* **274**, 8169–8174
- Parekh, B. S. & Hatfield, G. W. (1996) *Proc. Natl. Acad. Sci. U. S. A.* **93**, 1173–1177
- Benham, C. J. (1996) *Comput. Appl. Biosci.* **12**, 375–381
- Neidhardt, F. C., Bloch, P. L. & Smith, D. F. (1974) *J. Bacteriol.* **119**, 736–747
- Dwivedi, C. M., Ragin, R. C. & Uren, J. R. (1982) *Biochemistry* **21**, 3064–3069
- Arfin, S. M. (1969) *J. Biol. Chem.* **244**, 2250–2251
- Bender, R. A., Janssen, K. A., Resnick, A. D., Blumenberg, M., Foor, F. & Magasanik, B. (1977) *J. Bacteriol.* **129**, 1001–1009
- Martin, R. G. & Goldberger, R. F. (1967) *J. Biol. Chem.* **242**, 1168–1174
- Bradford, M. M. (1976) *Anal. Biochem.* **72**, 248–254
- Carpousis, A. J., Vanzo, N. F. & Raynal, L. C. (1999) *Trends Genet.* **15**, 24–28
- Higgins, C. F., Peltz, S. W. & Jacobson, A. (1992) *Curr. Opin. Genet. Dev.* **2**, 739–747
- Hauser, C. A. & Hatfield, G. W. (1984) *Proc. Natl. Acad. Sci. U. S. A.* **81**, 76–79
- Pagel, J. M. & Hatfield, G. W. (1991) *J. Biol. Chem.* **266**, 1985–1996
- Pagel, J. M., Winkelman, J. W., Adams, C. W. & Hatfield, G. W. (1992) *J. Mol. Biol.* **224**, 919–935
- Futcher, B., Latter, G. I., Monardo, P., McLaughlin, C. S. & Garrels, J. I. (1999) *Mol. Cell. Biol.* **19**, 7357–7368
- Wek, R. C. & Hatfield, G. W. (1986) *Nucleic Acids Res.* **14**, 2763–2777
- Giladi, H., Koby, S., Gottesman, M. E. & Oppenheim, A. B. (1992) *J. Mol. Biol.* **224**, 937–948
- Altuvia, S., Almiron, M., Huisman, G., Kolter, R. & Storz, G. (1994) *Mol. Microbiol.* **13**, 265–272
- Hengge-Aronis, R. (1996) in *Escherichia coli and Salmonella Cellular and Molecular Biology* (Neidhardt, F. C., Curtis, R. I., Ingraham, J. L., Lin, E. C. C., Low, K. B., Magasanik, B., Reznikoff, W. S., Riley, M., Schaechter, M., and Umbarger, H. E., eds) pp. 1497–1512, American Society for Microbiology, Washington, D. C.
- Presutti, D. G. & Hassan, H. M. (1995) *Mol. Gen. Genet.* **246**, 228–235
- Hassan, H. M. & Sun, H. C. (1992) *Proc. Natl. Acad. Sci. U. S. A.* **89**, 3217–3221
- Green, J., Anjum, M. F. & Guest, J. R. (1997) *Microbiology* **143**, 2865–2875
- Aviv, M., Giladi, H., Schreiber, G., Oppenheim, A. B. & Glaser, G. (1994) *Mol. Microbiol.* **14**, 1021–1031
- Hiszczynska-Sawicka, E. & Kur, J. (1997) *Plasmid* **38**, 174–179
- Zablewska, B. & Kur, J. (1995) *Gene* **160**, 131–132
- Werner, M. H., Clore, G. M., Gronenborn, A. M. & Nash, H. A. (1994) *Curr. Biol.* **4**, 477–487
- Zulianello, L., de la Gorgue de Rosny, E., van Ulsen, P., van de Putte, P. & Goosen, N. (1994) *EMBO J.* **13**, 1534–1540
- Gamas, P., Caro, L., Galas, D. & Chandler, M. (1987) *Mol. Gen. Genet.* **207**, 302–305
- Gunsalus, R. P. & Park, S.-J. (1994) *Res. Microbiol.* **145**, 437–450
- Park, S.-J., Cotter, P. A. & Gunsalus, R. P. (1995) *J. Bacteriol.* **177**, 6652–6656
- Park, S.-J. & Gunsalus, R. P. (1995) *J. Bacteriol.* **177**, 6255–6262
- Chao, G., Shen, J., Tseng, C. P., Park, S.-J. & Gunsalus, R. P. (1997) *J. Bacteriol.* **179**, 4299–4304
- Park, S.-J., Chao, G. & Gunsalus, R. P. (1997) *J. Bacteriol.* **179**, 4138–4142
- Lawrence, J. G. & Roth, J. R. (1995) *J. Bacteriol.* **177**, 6371–6380
- Dubrac, S. & Touati, D. (2000) *J. Bacteriol.* **182**, 3802–3808

<sup>5</sup> S.-J. Park and R. P. Gunsalus, personal communication.