DOI: 10.1534/genetics.104.027532

Genetics Software

Genome Image Programs: Visualization and Interpretation of *Escherichia coli* Microarray Experiments

Daniel P. Zimmer,*[,1] Oleg Paliy,*[,2] Brian Thomas,[†] Prasad Gyaneshwar* and Sydney Kustu*[,3]

**Department of Plant and Microbial Biology, University of California, Berkeley, California 94720 and*
[†]*College of Natural Resources, University of California, Berkeley, California 94720*

Manuscript received February 11, 2004
Accepted for publication April 21, 2004

ABSTRACT

We have developed programs to facilitate analysis of microarray data in *Escherichia coli*. They fall into two categories: manipulation of microarray images and identification of known biological relationships among lists of genes. A program in the first category arranges spots from glass-slide DNA microarrays according to their position in the *E. coli* genome and displays them compactly in genome order. The resulting genome image is presented in a web browser with an image map that allows the user to identify genes in the reordered image. Another program in the first category aligns genome images from two or more experiments. These images assist in visualizing regions of the genome with common transcriptional control. Such regions include multigene operons and clusters of operons, which are easily identified as strings of adjacent, similarly colored spots. The images are also useful for assessing the overall quality of experiments. The second category of programs includes a database and a number of tools for displaying biological information about many *E. coli* genes simultaneously rather than one gene at a time, which facilitates identifying relationships among them. These programs have accelerated and enhanced our interpretation of results from *E. coli* DNA microarray experiments. Examples are given.

DURING the past decade research in the field of molecular biology has gradually shifted from the analysis of single genes to the analysis of whole genomes, transcriptomes, and proteomes. The availability of full genome sequences for many organisms together with the development of microarray technology has allowed researchers to compare simultaneously mRNA levels for each gene in an organism under different conditions or in different cell types or strains. Given that even a single experiment generates thousands of spots and numerical values (*e.g.*, ~4400 for the genes of *Escherichia coli*), analysis of the data has necessitated the development of a variety of tools. More than 50 different commercial, shareware, and free software products are currently available (for a brief summary, see Goodman 2002), most of which focus on statistical normalization and analysis of numerical data. Powerful statistical clustering algorithms have been developed for interpreting data from many experiments (reviewed in Sherlock 2000). To complement the tools available, we have developed simple programs for visualizing gene expression patterns in *E. coli* in their genomic context and for identifying known biological relationships among lists of genes (Zimmer *et al.* 2000; Wendisch *et al.* 2001; Soupene *et al.* 2003). These are particularly helpful to biologists who wish to interpret relatively small numbers of experiments.

MATERIALS AND METHODS

Experimental methods, data acquisition, and storage of data in AMAD: Growth of *E. coli* cultures, isolation of total RNA, cDNA synthesis and labeling with Cy3 (green fluorescence) or Cy5 (red fluorescence), hybridization to glass-slide DNA microarrays, and scanning of the data were carried out as described (Zimmer *et al.* 2000). TIFF images (~7 MB each, 10-μm resolution) representing fluorescence intensities for the Cy3- and Cy5-labeled cDNAs hybridized to slides were generated using a GenePix scanner (Axon Instruments, Union City, CA). These images were overlaid and analyzed in Scanalyze 2.x (http://rana.lbl.gov/EisenSoftware.htm) or GenePix 3.0. Global intensity normalization was used to calculate a normalization factor for each pair of images (Schena *et al.* 1995) and the image intensities were normalized accordingly in Scanalyze. The normalized overlaid image was then saved as a bitmap image, which was converted to an 8-bit color GIF image and then to a portable network graphic file (PNG)

[1] *Present address:* Microbia, Cambridge, MA 02139.

[2] *Present address:* Department of Biochemistry and Molecular Biology, Wright State University, Dayton, OH 45435.

[3] *Corresponding author:* Plant and Microbial Biology, 111 Koshland Hall, University of California, Berkeley, CA 94720-3102.
E-mail: kustu@nature.berkeley.edu

using standard image manipulation software. At this point, some of the quantitative information is lost.

Glass-slide DNA microarray data were stored in AMAD (Another MicroArray Database; http://www.microarrays.org/AMADFaq.html). AMAD, developed by Joe DeRisi, is a flat file database written in Perl and JavaScript that allows storage and retrieval of raw scanned slide images and extraction of numerical data.

Generation of genome images: Genome images were built on AMAD as a core component. All image-manipulating scripts were written in Perl programming language (Wall *et al.* 2000), with the CGI.pm and the GD.pm modules of Lincoln Stein (http://stein.cshl.org/WWW/software/) installed.

From the PNG files described above, the program that generates genome images extracts rectangles containing the spots and arranges them according to their *E. coli* b number (Blattner *et al.* 1997). For *E. coli* microarrays, the resulting genome images contain 45 rows of 100 spots/row, with each spot in a 10-pixel square (the original size of the scanned spot). The output of the program is a PNG "genome image" file and an HTML document containing an image map of the b numbers, gene names, gene descriptions, and links to the raw data. The image is stored in a local database along with raw data files, and both can be easily accessed through a web-based interface that is provided. By clicking on a spot in a genome image, the user is transferred to the *E. coli* Entry Point (see below), which allows quick access to biological information on the gene corresponding to this spot.

On a separate page the user can display a list of genes corresponding to spots that fulfill certain criteria (see below). The spots can then be outlined in blue boxes on the genome image and can also be transferred directly to the *E. coli* Entry Point (see below), from which other biological information can be accessed.

A generalized version of the program for generating genome (and other sorts of) images uses as input: (1) a tab-delimited "ORDER" text file containing the headings (ORDER, TOP, LEFT, RIGHT, BOTTOM, NAME, DESC, LINK), where the TOP, LEFT, RIGHT, and BOTTOM parameters refer to the corresponding pixel positions of each spot in the original microarray image, and (2) a PNG image file. The program aligns the spots according to the order specified in the ORDER file, yielding an HTML page and an image. An image map identifies each spot and includes a user-specified hyperlink. The generalized genome image program is written in Perl and requires that all of the appropriate Perl modules (GD, CGI) be installed. It can be accessed at http://coli.berkeley.edu/genomeimages/ and the stand-alone version can be downloaded from the same site.

A GenePix-specific version of the program for generating genome images was written to accommodate the large number of users of the Axon GenePix software. This program uses as input: (1) a GenePix results file (GPR), (2) a PNG or JPEG image file, and (3) an ORDER file. For the tab-delimited ORDER file, the ORDER and ID fields are required, and the NAME, DESC, and LINK fields are optional but recommended. The TOP, LEFT, RIGHT, and BOTTOM fields are not used by the GenePix version of this program because they are calculated from fields that are present in the GPR file. The program joins the ORDER table to the GPR table by the ID field present in both files. For *E. coli*, we have used the b number as the ID in both the GPR file and the ORDER file. The GenePix-specific program can also be accessed at http://coli.berkeley.edu/genomeimages.

Alignment of genome images: The program that aligns genome images takes as input a list of genome images in PNG format (each assumed to be 100 cells wide with each 10-pixel cell containing a spot). The program vertically concatenates corresponding rows from each of the genome images to generate a new larger image of the data with rows of spot images aligned. A generalized version of the genome image alignment program works as follows: (1) The user is first prompted for the number of genome images he would like to align, and then (2) on a second page the user must upload all of the images to be aligned, preferably in PNG format, and must upload the ORDER file described above. Currently a maximum of 12 images can be aligned, but this can be reconfigured at local installations. The generalized version of the genome image alignment program can be run or downloaded from http://coli.berkeley.edu/genomeimages/.

The AMAD core database of DeRisi (accessible through a web-based interface) allows the user to extract from multiple experiments lists of genes corresponding to spots that fulfill specified criteria, *e.g.*, have a normalized median red-to-green (*R/G*) ratio higher than a specified cutoff value. Outputs can be saved directly to the local computer.

The *E. coli* Entry Point database: The *E. coli* Entry Point programs are written in Perl using the CGI.pm module. Data are stored in a MySQL database (http://www.mysql.com/) and accessed using the DBI.pm and DBD::MySQL Perl modules (Descartes and Bunce 2000; http://www.cpan.org/modules/by-module/DBI/; http://www.cpan.org/modules/by-module/DBD/). The *E. coli* Entry Point is composed of a main script and several subsidiary scripts. Their functions, features, and data resources, which can be accessed at http://coli.berkeley.edu/genomeimages/, are outlined below, along with those of the additional databases to which the Entry Point has links.

The main page allows the user to display annotation information for lists of *E. coli* genes. The primary source of data that was used is the ecoli.ptt file (NC_000913.ptt), which was compiled as part of the *E. coli* sequencing effort and downloaded from the National Center for Biotechnology Information (NCBI) (ftp://ftp.ncbi.nih.gov/genomes/Bacteria/Escherichia_coli_K12/). The user begins by entering or selecting a list of genes using any one of several nomenclatures. The heading "Display standard fields" allows the user to display basic annotation information on these genes, including b number, gene name, gene position on chromosome (left and right), strand orientation, protein length, GenBank ID, functional description (called gene description), and operon ID. We have updated some of the gene names on the basis of evidence in the primary literature. The heading "Sorting" allows the user to sort and group the genes being displayed by genome position or functional category (see below). When genes are sorted by position, the background color of the row (alternating between yellow and white) is used to indicate different operons. Similarly, when genes are sorted by category, genes belonging to the same category are indicated with the same color. Additional fields that can be displayed are:

1. "Show functional category" (Riley-Labedan). Gives the superheading, heading, and category as defined by Riley and Labedan (1996).
2. "Show Blattner groups." Gives the category as defined by Blattner *et al.* (1997).
3. Operon [A Systematic Annotation Package for Community Analysis of Genomes (ASAP)]. Gives the name of the operon and whether it is documented or predicted according to Glasner *et al.* (2003); http://asap.ahabs.wisc.edu/annotation/php/ASAP1.htm).
4. "Show COG (NCBI)." Gives the COG (clusters of orthologous groups) description of Tatusov *et al.* (2000), with a link to the NCBI web site (http://www.ncbi.nlm.nih.gov/COG/).
5. EcoGene and external links. Gives the number or name used by various databases along with a direct link to infor-

mation on the particular gene in each external database. The databases are: EcoGene (http://bmb.med.miami.edu/EcoGene/EcoWeb/; Rudd 2000); SwissProt (http://www.expasy.ch/; Bairoch and Apweiler 2000); EcoCyc (http://biocyc.org/ecocyc/; Karp *et al.* 2000); the NCBI *E. coli* genome page (http://www.ncbi.nlm.nih.gov/cgi-bin/Entrez/framik?db=Genome&gi=115); GenProtEC (http://genprotec.mbl.edu/; Riley and Space 1996); Colibri (http://genolist.pasteur.fr/Colibri/; Medigue *et al.* 1993); RegulonDB (http://www.cifn.unam.mx/Computational_Genomics/regulondb/; Salgado *et al.* 2001); and the *E. coli* Genetic Stock Center (http://cgsc.biology.yale.edu/; Berlyn and Letovsky 1992).

6. Ecogene bibliography. Provides a hyperlink to the gene-specific bibliographies in EcoGene (Rudd 2000).
7. Protein binding sites (ASAP). Gives the names of transcriptional regulatory proteins that have documented or predicted binding sites within 2000 nt of each gene according to Glasner *et al.* (2003; http://asap.ahabs.wisc.edu/annotation/php/ASAP1.htm).
8. Promoters (ASAP). Gives the names of σ-factors that have documented or predicted binding sites within 2000 nt of each gene according to Glasner *et al.* (2003; http://asap.ahabs.wisc.edu/annotation/php/ASAP1.htm).
9. *E. coli* BLAST neighbors. Gives the number of genes in the *E. coli* genome with BLAST homology to each gene in the list, with a hyperlink to the b numbers (seqid), names, expect scores (*E*-values), and percent identities. The only BLAST hits stored in the database and reported are those with an *E*-value <0.001 in a blastp search against ORFs of the *E. coli* annotated genome (NC-000913.ptt).

After a list of genes has been generated, a set of clickable buttons at the bottom of the main *E. coli* Entry Point page allows access to information from the subsidiary programs and to supplementary data taken from external sources. The fields available are (in order of presentation): chromosome position, operons, functional category, COG description, protein binding sites, selfBLAST, features, gene sequences, and AMAD (takes the user to Genome Images/AMAD database). A brief description of each follows.

1. Chromosome position. For a list of genes selected on the *E. coli* Entry Point page, the program generates a PNG image of the circular chromosome of *E. coli* with gene names marked at the appropriate positions on the circle.
2. Operons. For a list of genes selected on the *E. coli* Entry Point page, this tool displays diagrammatically all of the genes that are members of the corresponding operons (predicted and documented). Each operon is on a separate line. Genes that were part of the original list are shown on a pink background. The user then has the option to return to the *E. coli* Entry Point with a new list that includes all genes in the operons. The operons are annotated largely as defined at ASAP (Glasner *et al.* 2003; http://asap.ahabs.wisc.edu/annotation/php/ASAP1.htm).
3. Functional category. This option overlaps with the additional field "Show Functional Category" described above but also includes the category number.
4. COG description. This option overlaps with the additional field "Show COG (NCBI)" but also gives the COG number.
5. Protein binding sites. For each gene in the query list this program identifies documented (dark blue background) and putative (light blue background) regulatory proteins that bind within a user-specified number of base pairs (default is 2000) upstream of the start site for the gene or corresponding multigenic operon. Note that every gene is considered a member of an operon, whether or not it is multigenic, and that the start site is the *translational* start for the first gene in the documented or predicted operon. The output is a table in which regulatory proteins are in rows and gene names in columns. The source of the protein binding data is the ASAP database (Glasner *et al.* 2003; http://asap.ahabs.wisc.edu/annotation/php/ASAP1.htm).
6. SelfBLAST. At the top of the page, this tool displays the results of sequence comparisons between each gene in the selected list (pink background, separate row) and all other *E. coli* genes. The names of proteins with BLAST homology to the gene of interest (*E*-value <0.001) are listed in the same row. At the bottom of the page the tool displays BLAST homology scores for comparisons between all members (n) of the selected list in an $n \times n$ matrix.
7. Features. This tool allows the user to search the nucleotide sequence upstream of each gene in a list for binding sites for selected σ-factors and/or selected protein transcriptional regulators. The left and right ends of the binding site for each σ-factor or regulatory protein are specified, along with the status of the site (documented or predicted). The distance to the starting ATG of the gene and the position of the transcriptional start are also indicated (Glasner *et al.* 2003).
8. Gene sequences. This tool displays the primary nucleotide sequences for all genes in a list.

RESULTS

Genome images—visualization of microarray data: *E. coli* genome images are generated by arranging the spots in the original image of a glass-slide DNA microarray in genome order (see materials and methods). The spots are ordered in a grid that is 100 columns wide by 45 rows tall and are read from left to right and then from top to bottom, as one would read words on a page (Figure 1). An *E. coli* genome image, which carries primary expression data for all the genes of the organism, can be viewed on a single page or computer screen. When the user holds a cursor over a particular spot, the corresponding gene name and description are displayed in a web browser. Blank areas represent genes/PCR products that were not printed on the slides, which in our case are stable RNAs.

Figure 1 shows an example of a genome image for a wild-type *E. coli* K12 strain grown with taurine (2-aminoethanesulfonate) as the sole sulfur source (Cy5; red fluorescent label) or with sulfate, an optimal sulfur source (Cy3; green). Spots with an *R/G* median ratio of ≥ 3 are boxed in blue (see below). The strain grows slightly less rapidly on taurine than on sulfate and appears to perceive some degree of sulfur limitation. As expected from previous work (van der Ploeg *et al.* 2001), two operons under control of the regulators CysB and Cbl (CysB-like) were more highly expressed on taurine. These are *tauABCD* (b0365–b0368), a catabolic operon for taurine, and *ssuEADCB* (b0937–b0933), a catabolic operon for utilization of alkanesulfonates. They are easily identified on the image as striking strings of red spots. A number of single red spots are also clearly visible. Two that are easily understood are a spot corresponding to the *cbl* regulatory gene (b1987) and one corresponding to *sbp* (b3917), the gene for a peri-

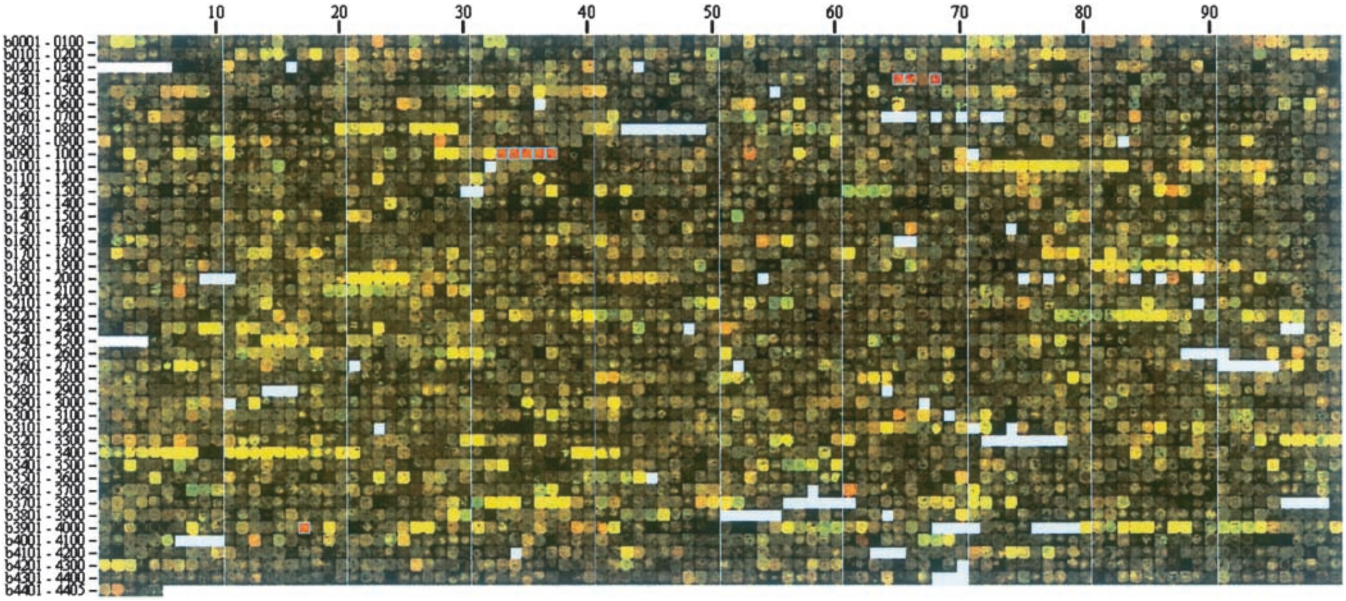


FIGURE 1.—Genome image of an *E. coli* cDNA microarray. *E. coli* wild-type strain NCM3722 was grown in $N^-C^-S^-$ minimal medium (GUTNICK *et al.* 1969) with either 0.25 mM taurine (Cy5; red fluorescence) or 0.25 mM sulfate (Cy3; green fluorescence) as the sulfur source. The carbon and nitrogen sources were glycerol and ammonium, respectively. As described in the text, the image was generated by rearranging spots in the scanned image of the glass slide in genome order. The b number centuries are indicated to the left and the decades on the top. Blanks represent b numbers that do not correspond to ORFs. Genes with *R/G* median ratio ≥3 are highlighted in blue boxes (a feature of the Genome Image/AMAD programs; see MATERIALS AND METHODS). They are (in order): *tauACD* (b0365, b0367, and b0368), *ssuBCDAE* (b0933–b0937), and *sbp* (b3917). *R/G* median refers to the ratio of *R* median/*G* median for all of the pixels constituting a spot. These values are determined after global normalization.

plasmic sulfate transport component known to be highly expressed under sulfur-limiting conditions (QUADRONI *et al.* 1996). Note that *tauB* and *cbl* are not boxed because their *R/G* ratios were <3. The reproducibility and significance of other red spots is currently being assessed (P. GYANESHWAR, unpublished results).

Also available from genome images is visual information on spot intensities, information that may be lost in some higher-level analyses of the data (*e.g.*, clustering based on *R/G* ratios). By displaying the spots rather than pseudocolors representing *R/G* ratios, we can discern, for example, bright yellow spots, genes for which there is probably a large amount of mRNA in both cultures. Several long strings of bright yellow spots in Figure 1 correspond to operons of ribosomal protein genes or clusters of such operons (*e.g.*, b3294–3298, b3299–3310, b3311–3321, b3339–3342, b3983–3984, b3985–3986), which are always highly expressed in *E. coli* (NEIDHARDT *et al.* 1990). Other strings of bright yellow spots correspond to genes of the flagellar and chemotaxis regulon (b1070–1083, b1881–1892, b1920–1926) and to operons encoding the F_1F_0 ATPase (b3731–3739) and the tricarboxylic acid cycle enzymes succinate dehydrogenase (b0721–0724) and 2-oxoglutarate dehydrogenase and succinyl-CoA synthetase (b0726–b0729). Although high intensity is probably an indication of high mRNA levels, long length of a gene and/or a large amount of DNA (PCR product) attached to the slide may contribute. Low intensity of a spot may have many causes but low intensity of a group of adjacent spots corresponding to an operon(s) probably indicates that the operon is not highly expressed under either condition chosen for the comparison and hence *R/G* ratios should be evaluated accordingly.

The quality of spots can be assessed directly on genome images without the need for complex statistical procedures because the images are composed of the actual scanned pixels. Dark spots within operons can be seen easily when they are surrounded by spots that are otherwise red, green, or bright yellow. Such spots often indicate failed PCR products or damaged print tips. In Figure 1 there are two black spots (b3309 and b3310) in the middle of the string of ribosomal protein genes between b3294 and b3321. They were reproducibly black in several prints and hence probably are failed PCR products.

Finally, in conjunction with analyses at the *E. coli* Entry Point (see below), genome images can be helpful in detecting misannotated operons and artifactual differential expression. For example, we determined that the *gltIJK* and *L* genes probably constitute a single operon, as do *yhdWXY* and *Z*, although *gltI* was not originally included with the other *glt* genes and the *yhd* operon was split in half (ZIMMER *et al.* 2000). Likewise we showed that apparent overexpression of the *cynX* gene upon IPTG induction of the lactose operon was an artifact of

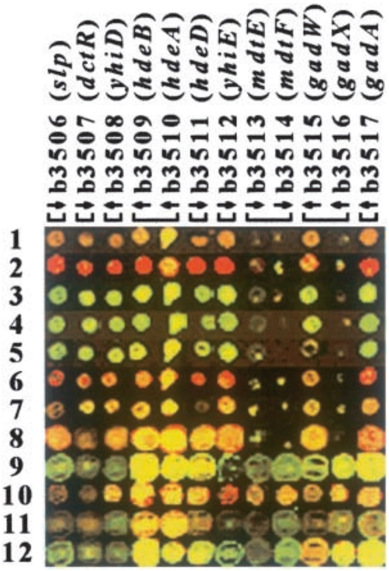


Figure 2.—The *slp-gadA* region (b3506–b3517) of 12 aligned genome images. Rows correspond to different experiments and columns indicate the genes in the portion of the images shown (b numbers and gene names). Brackets denote operon boundaries and arrows indicate the direction of transcription of each gene. The data were taken from the *E. coli* Entry Point. In experiments 1–8, b3513, b3514, and b3517 appear to be bad PCR products. A table with the descriptions of the experiments and a table with functional descriptions of the genes are available at http://nature.berkeley.edu/~opaliy/papers/GenomeImages.html.

readthrough transcription (Wendisch *et al.* 2001, Figure 3C). The *cynX* gene is adjacent to *lacA* and is the last gene in the *cynTSX* operon, which is transcribed toward *lac*. The high signal seen in IPTG-induced cells is probably due to the presence of antisense RNA because many transcripts for the *lac* operon terminate at least one-third of the way into the *cynX* gene (Hediger *et al.* 1985; McCormick *et al.* 1991).

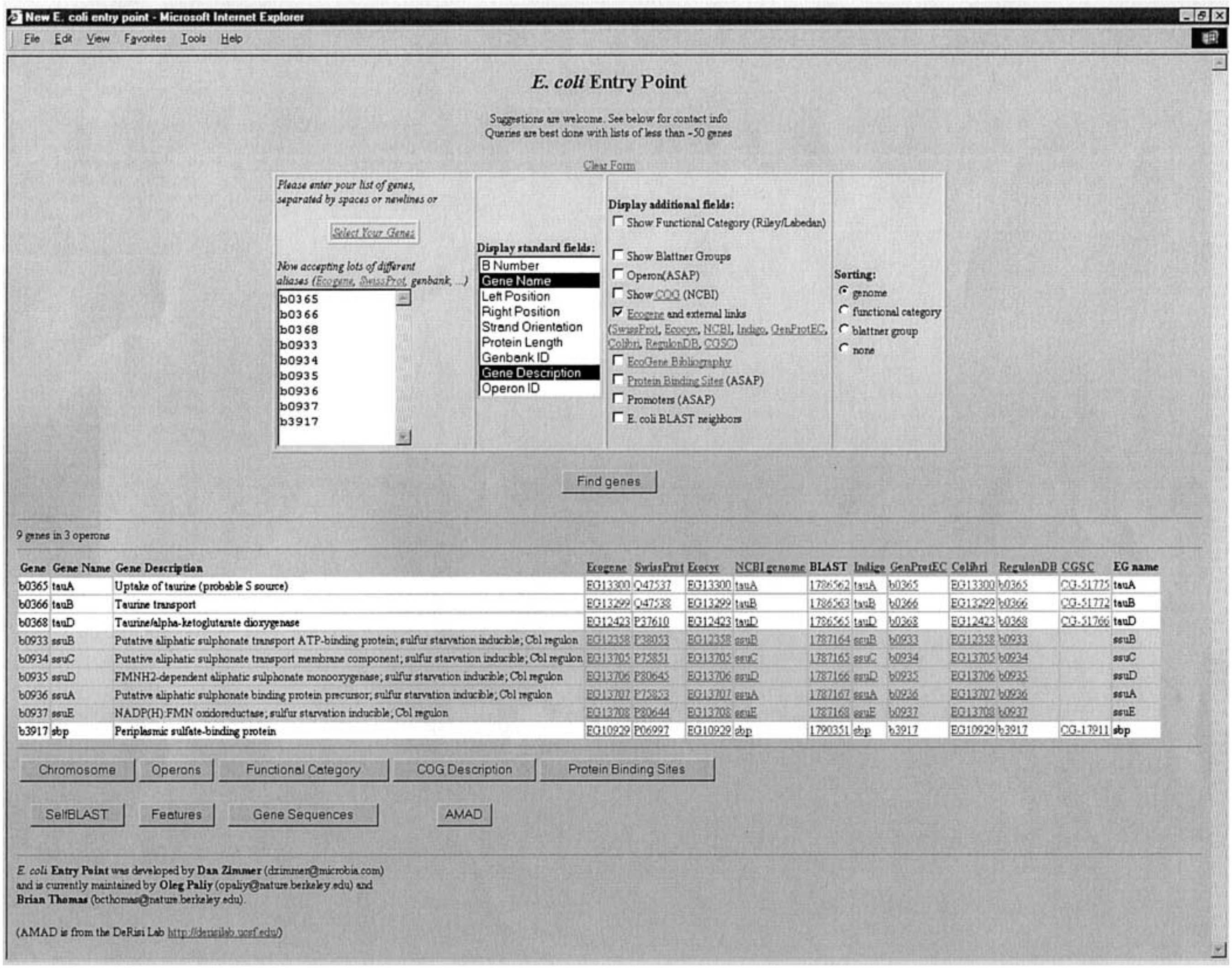
As indicated in materials and methods, the AMAD database, in which our genome images are stored, allows extraction of the corresponding numerical data. AMAD was developed by Joe DeRisi. Extraction of numerical data can, of course, also be accomplished with other microarray data analysis programs.

Alignment of genome images: Aligned genome images are used to identify similarities and differences in gene expression (mRNA levels) in several experiments and to assess reproducibility of these differences (see, for example, an alignment of four images for an *E. coli* K12 strain grown on taurine *vs.* sulfate at http://nature.berkeley.edu/~opaliy/papers/GenomeImages.html). An alignment of two *E. coli* genome images can be viewed on a single screen at 1280 × 1024 resolution. Alignments of several images (we have used up to a dozen; *e.g.*, see Figure 2) must be scrolled. As for single images, image maps allow ready identification of gene names and functions corresponding to particular spots.

We are currently using aligned genome images to identify the intersections and unions of genes induced upon limitation of sulfur or nitrogen (P. Gyaneshwar, unpublished results). Previously, we have used them to analyze a regulatory cascade that controls the homeostatic response to nitrogen limitation and to note simultaneous changes in expression of the 12 genes in the *slp-gadA* region (b3506–b3517), whose expression appears to be elevated under conditions of slow growth (Zimmer *et al.* 2000). Behavior of these 12 genes in a dozen independent experiments is shown in Figure 2, which represents only the *slp-gadA* region of the dozen aligned genome images. Although the genes are annotated as members of nine different operons (Figure 2 and information from the *E. coli* Entry Point at http://nature.berkeley.edu/~opaliy/papers/GenomeImages.html), their expression appears to change in parallel in all of the experiments. The visual analysis was confirmed by calculating a pairwise correlation matrix of log-transformed *R/G* ratios, which showed a good positive relationship among expression of all these genes (average pairwise correlation of 0.91). Similar effects have been seen for clusters of operons with common regulatory control, *e.g.*, clusters of ribosomal protein genes and the flagellar and chemotaxis regulon (Soupene *et al.* 2003). Regulation of genes in the *slp-gadA* region has been intensively studied recently (Masuda and Church 2003 and references cited therein). This region is probably also subjected to common transcriptional control. In addition, one or more structural proteins, *e.g.*, H-NS, may control access of RNA polymerase to this region of the genome, which would be analogous to regional effects observed in eukaryotic organisms (Lercher *et al.* 2002; Roy *et al.* 2002; Spellman and Rubin 2002). To our knowledge such effects have not been documented in bacteria.

***E. coli* Entry Point—tools for identifying relationships among *E. coli* genes:** The *E. coli* Entry Point is a set of tools for identifying known biological relationships among groups of genes. A user can enter any list and then display various sorts of biological information for each gene, including information on chromosome position and inclusion in an operon, promoter and σ-factor controlling expression, regulatory proteins that bind upstream and their binding sites, sequence, function of the gene product, and homology relationships to other gene products (see materials and methods). The full list of genes together with the information requested is shown on one web page, allowing fast comparisons and interpretations. In addition, the information can easily be copied into a spreadsheet program such as Microsoft Excel for further analysis locally.

As an illustration (Figure 3) we show a screen shot of the *E. coli* Entry Point page displaying information on genes that were highly expressed on taurine *vs.* sulfate in the experiment of Figure 1. Data for spots with *R/G* ratio ≥3 were first extracted from AMAD and the corre-



E. coli Entry Point

Suggestions are welcome. See below for contact info
Queries are best done with lists of less than ~50 genes

Clear Form

Please enter your list of genes, separated by spaces or newlines or

Now accepting lots of different aliases (Ecogene, SwissProt, genbank, ...)

b0365
b0366
b0368
b0933
b0934
b0935
b0936
b0937
b3917

Display standard fields:

B Number
Gene Name
Left Position
Right Position
Strand Orientation
Protein Length
Genbank ID
Gene Description
Operon ID

Display additional fields:

□ Show Functional Category (Riley/Labedan)
□ Show Blattner Groups
□ Operon(ASAP)
□ Show COG (NCBI)
☑ Ecogene and external links (SwissProt, Ecocyc, NCBI, Indigo, GenProtEC, Colibri, RegulonDB, CGSC)
□ EcoGene Bibliography
□ Protein Binding Sites (ASAP)
□ Promoters (ASAP)
□ E. coli BLAST neighbors

Sorting:

● genome
○ functional category
○ blattner group
○ none

9 genes in 3 operons

Gene	Gene Name	Gene Description	Ecogene	SwissProt	Ecocyc	NCBI genome	BLAST	Indigo	GenProtEC	Colibri	RegulonDB	CGSC	EG name
b0365	tauA	Uptake of taurine (probable S source)	EG13300	Q47537	EG13300	tauA	1786562	tauA	b0365	EG13300	b0365	CG-51775	tauA
b0366	tauB	Taurine transport	EG13299	Q47538	EG13299	tauB	1786563	tauB	b0366	EG13299	b0366	CG-51772	tauB
b0368	tauD	Taurine/alpha-ketoglutarate dioxygenase	EG12423	P37610	EG12423	tauD	1786565	tauD	b0368	EG12423	b0368	CG-51766	tauD
b0933	ssuB	Putative aliphatic sulphonate transport ATP-binding protein; sulfur starvation inducible; Cbl regulon	EG12358	P38053	EG12358	ssuB	1787164	ssuB	b0933	EG12358	b0933		ssuB
b0934	ssuC	Putative aliphatic sulphonate transport membrane component; sulfur starvation inducible; Cbl regulon	EG13705	P75851	EG13705	ssuC	1787165	ssuC	b0934	EG13705	b0934		ssuC
b0935	ssuD	FMNH2-dependent aliphatic sulphonate monooxygenase; sulfur starvation inducible; Cbl regulon	EG13706	P80645	EG13706	ssuD	1787166	ssuD	b0935	EG13706	b0935		ssuD
b0936	ssuA	Putative aliphatic sulphonate binding protein precursor; sulfur starvation inducible; Cbl regulon	EG13707	P75853	EG13707	ssuA	1787167	ssuA	b0936	EG13707	b0936		ssuA
b0937	ssuE	NADP(H):FMN oxidoreductase; sulfur starvation inducible; Cbl regulon	EG13708	P80644	EG13708	ssuE	1787168	ssuE	b0937	EG13708	b0937		ssuE
b3917	sbp	Periplasmic sulfate-binding protein	EG10929	P06997	EG10929	sbp	1790351	sbp	b3917	EG10929	b3917	CG-17911	sbp

E. coli **Entry Point** was developed by **Dan Zimmer** (dzimmer@microbia.com) and is currently maintained by **Oleg Paliy** (opaliy@nature.berkeley.edu) and **Brian Thomas** (bcthomas@nature.berkeley.edu).

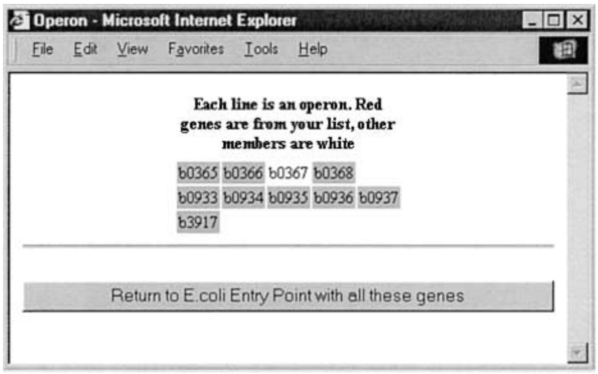
(AMAD is from the DeRisi Lab http://derisilab.ucsf.edu/)

Figure 3.—A screen shot of the main page of the *E. coli* Entry Point. Genes corresponding to spots with an *R/G* median ratio ≥3 in the experiment of Figure 1 were entered in the selection box and information fields for gene name, gene description, and a number of web links to external databases were displayed. Background shading of the rows alternates between operons. Other information available from the *E. coli* Entry Point is discussed in the text.

sponding list of genes was transferred directly to the *E. coli* Entry Point. The screen shot shows some of the basic annotation information available from the options "Display Standard Fields" and "Display Additional Fields." Note that the background shading of the rows alternates between operons. Note, too, that there are direct links to the other major *E. coli* databases listed. Thus, if the user wishes additional information on a gene(s) of interest, he or she can go to a gene-specific page of any of these databases with one click of the mouse button. A screen shot of all the additional information available from the option "Display Additional Fields" is provided at http://nature.berkeley.edu/∼opaliy/papers/GenomeImages.html, along with comments.

Figure 4 shows the information available from the clickable button "Operon" at the bottom of the *E. coli* Entry Point page for the gene list of Figure 3. Use of the "Operon" option shows that one gene of the *tau* operon (b0367 is *tauB*, white background) was not included in the original list of those with *R/G* ratio ≥3. By clicking the button at the bottom of the operon page the user can now return to the *E. coli* Entry Point with all genes of the operons being considered and obtain additional information for all of them. Entering b0367 in AMAD allows the user to determine that the *R/G* median for this gene was 2.0, whereas ratios for the 9 genes originally in the list were between 3.2 and 14.8. Screen shots obtained by using all of the clickable buttons at the bottom of the *E. coli* Entry Point page for the expanded list of 10 genes are given at http://nature.berkeley.edu/∼opaliy/papers/GenomeImages.html, along with comments on the timeliness and accuracy of the information currently available.

An important feature of the *E. coli* Entry Point is that other parts of the genome images/AMAD database are interactively linked to it. For example, when a genome


Operon - Microsoft Internet Explorer
File
Edit
View
Favorites
Tools
Help
Return to E.coli Entry Point with all these genes

Each line is an operon. Red genes are from your list, other members are white

b0365 b0366 b0367 b0368
b0933 b0934 b0935 b0936 b0937
b3917

FIGURE 4.—Screen shot of the operon tool for the gene list of Figure 3. The data shows that the genes (shaded background) were in three operons (indicated in separate rows) and that only one gene (b0367, white background) in one of the operons was missing in the original list. The button at the bottom of the page allows the user to return to the *E. coli* Entry Point with all of the genes listed.

image is displayed in AMAD, clicking on a spot of interest transfers the user directly to the *E. coli* Entry Point with the corresponding gene already entered in the gene selection field. From the gene one can proceed to its operon and all of the other information described above. Similarly, when a user decides to highlight a number of spots on a genome image, *e.g.*, those whose *R/G* ratio is above a certain cutoff value (see MATERIALS AND METHODS), he or she can, in a separate operation, also transfer the corresponding list of genes to the *E. coli* Entry Point.

Finally, if the user wishes first to determine the gene set meeting a certain criterion, *e.g.*, all the genes containing "*tau*" in their name, he or she can begin with the option "Select Genes" at the Entry Point and then return to the Entry Point with the resulting list. Criteria for selecting genes include gene name, description, b number, position on the genome, and length. The user can also generate lists of genes from other programs or *E. coli* resources on the internet and import them into the *E. coli* Entry Point.

DISCUSSION

Genome images and aligned genome images: We developed genome images to visualize microarray data in a way that would facilitate comprehensive qualitative analysis of one or a few experiments. In the RESULTS we present two new examples of their use, along with the use of the *E. coli* Entry Point database. Previously, we have used genome images and aligned images to aid in determining the regulons controlled by nitrogen regulatory protein C (NtrC) and the nitrogen assimilation control protein (Nac; ZIMMER *et al.* 2000) to assess the responses of freshly isolated urinary tract and intestinal commensal strains of *E. coli* to nitrogen limitation in comparison to those of a laboratory strain and to compare gene expression between different laboratory strains of *E. coli* grown in the same medium (SOUPENE *et al.* 2003). In the latter case, comparison between a robust *E. coli* K12 wild-type strain and MG1655 (CGSC 6300) illustrated strikingly the low expression of flagellar and chemotaxis genes in MG1655 (LEHNEN *et al.* 2002) because these are arranged in several large clusters on the genome. After we initially employed genome images (ZIMMER *et al.* 2000; WENDISCH *et al.* 2001), several other programs that present microarray data in genome order, *e.g.*, GeneSpring (SiliconGenetics http://www.silicongenetics.com/cgi/SiG.cgi/Products/GeneSpring/index.smf), also became available. However, information about primary data—overall quality of the experiment and/or the occurrence of missing spots in operons whose expression differs under the two conditions chosen—must be assessed less directly because expression differences are presented in artificial color. In addition, these new programs are often costly.

A major goal of aligning genome images is similar to that of powerful statistical methods for data analysis such as hierarchical clustering (EISEN *et al.* 1998). The two approaches are complementary, with alignment having two distinct strengths for small numbers of experiments. First, the alignment of several genome images can be viewed on a single page, whereas the complete cluster analysis for *E. coli* requires many more pages. (Results of the latter are usually organized into a figure/table that is *L* experiments wide and *N* genes high, where *N* is ~4400 for *E. coli.*) The compactness of genome images facilitates rapid qualitative analysis of the data and reduces its complexity by allowing immediate consideration of operons without the need to sift through lists of hundreds of genes. Apart from the 1125 genes that are transcribed separately in *E. coli*, the remaining 3100 protein-coding genes are partitioned into only about one-quarter as many operons (~750; GLASNER *et al.* 2003; http://asap.ahabs.wisc.edu/annotation/php/ASAP1.htm). In the case of the NtrC and Nac regulons, the 75 genes involved were members of only 25 operons (ZIMMER *et al.* 2000). A second advantage of genome images is that members of operons are contiguous whereas this often is not the case in a hierarchical cluster. The nested NtrC and Nac regulons provide an interesting example. Many of the operons under control of these regulatory proteins encode ABC transport systems for nitrogen-containing compounds, and, biologically, members of each operon are their own closest neighbors because they must function together. Genome images showed clearly that expression of all genes in each operon changed in the same direction when various strains and growth conditions were compared (ZIMMER *et al.* 2000). Nevertheless, members of different operons were intermingled in hierarchical clusters (http://nature.berkeley.edu/~opaliy/papers/GenomeImages.html) at least partly because expression (mRNA levels) of the genes in each operon apparently did not change to the same extent. [As discussed previously, we think this has a biological explanation (Zimmer *et al.* 2000).] Apart from problems with operons, however, the results of hierarchical clustering and interpretation of genome images (Zimmer *et al.* 2000) were remarkably congruent, illustrating the complementarity of the two means of analysis. A cluster of only 39 genes contained 32 genes in operons directly under NtrC control and a second cluster of only 24 genes contained 17 genes in operons under Nac control. In all, two-thirds of the 75 genes we had identified previously were in these two clusters (http://nature.berkeley.edu/∼opaliy/papers/GenomeImages.html).

In their masterful study using glass-slide DNA microarrays and hierarchical clustering to analyze tryptophan metabolism in *E. coli*, Khodursky *et al.* (2000) mentioned that only five known multigene operons were fully represented in the set of 169 genes that they selected to analyze, whereas 37 operons were represented by only a single gene. For example, expression of only a few of the 50 genes in the flagellar and chemotaxis regulon appeared to respond to tryptophan availability. However, examination of the data in genome images showed that expression of genes in many operons, including those of the flagellar and chemotaxis regulon, differed in the same direction in particular comparisons between growth conditions or strains (strings of contiguous red or green spots; see example at http://nature.berkeley.edu/∼opaliy/papers/GenomeImages.html). One image revealed a striking artifact: an apparent difference in expression of the flagellar and chemotaxis regulon between a wild-type strain (W3110) and a strain lacking the tryptophan repressor CY15682. The W3110 wild-type strain is in the same lineage as MG1655 (Bachmann 1996) and expresses flagellar genes poorly (see above), whereas the particular *trpR2* strain used for this and one other experiment is apparently noncongenic with W3110 and expresses these genes well. The difference was not seen when congenic *trpR2* and wild-type strains were compared (Khodursky *et al.* 2000; P. Gyaneshwar, unpublished results). The use of genome images to examine the data of Khodursky *et al.* (2000), which we are analyzing in further detail elsewhere (P. Gyaneshwar, A. Jones, A. Khodursky and S. Kustu, unpublished results), illustrated the value of these images as an adjunct to hierarchical clustering.

The *E. coli* Entry Point: After examining genome images and using data sorting and filtering methods to determine a list of genes whose expression differs in a microarray experiment, an investigator can use the *E. coli* Entry Point to extract biological information about these genes. Examples of the uses of the Entry Point are given in the results. We previously used the Entry Point to determine relationships among the genes of the NtrC and Nac regulons and to update their annotations. In addition, we used it recently to export a list of all ∼4400 *E. coli* genes together with appropriate functional information into a spreadsheet file that was used to compare the protein and mRNA profiles of *E. coli* on a global scale (Corbin *et al.* 2003). The comparison was also visualized in artificial color in an analog of an aligned genome image (http://coli.berkeley.edu/protein_profile/).

The Entry Point consists of simple programs that extract and visualize data, which can be downloaded from a variety of publicly available sources (see materials and methods). The capacity to visualize this data in new ways rests on the flexibility given by being able to access it from a MySQL database that was implemented locally. As illustrated in the results and at http://nature.berkeley.edu/∼opaliy/papers/GenomeImages.html, the quality of the information obtained from the Entry Point depends on whether information in other databases is current and accurate. One very useful feature of the Entry Point is that it facilitates access to primary literature from PubMed (EcoGene Bibliography) and to information from other databases. Data from these sources can be cross-checked to obtain the best possible information on a list of genes at any given time.

Conclusions: Global expression technologies have led to a rapid increase in our knowledge and understanding of metabolic pathways and regulatory networks in a variety of microbes and other organisms. As the use of DNA microarrays becomes more widespread among biologists of all generations, it will be useful to have biologist-friendly software and visualization tools available to supplement more mathematical tools. Genome images and the *E. coli* Entry Point should be useful in this regard. Our current efforts are directed at improving these tools for *E. coli*, making them widely available, and generalizing them to other microorganisms.

D.P.Z. thanks Arkady Khodursky, Brian Peter, and Volker Wendisch for stimulating discussions of technical aspects of bacterial microarray experiments and the merits of various methods for interpreting the data; David Botstein, Patrick Brown, and Joseph DeRisi for guidance in all aspects of microarray experiments; and Daniel Rokhsar, Nik Putnam, and David Schweisguth for ideas and advice on computation. We thank Charles Yanofsky for access to his data in AMAD and Jon McAuliffe and Michael I. Jordan for advice on analysis of the *slp-gadA* region. This work was supported by National Institutes of Health (NIH) fellowship GM19862 to D.P.Z. and NIH grant GM38361 and a grant from the Torrey Mesa Research Institute, Syngenta Research and Technology, La Jolla, California, to S.K.

LITERATURE CITED

Bachmann, B. J., 1996 Derivations and genotypes of some mutant derivatives of *Escherichia coli* K-12, pp. 2460–2488 in *Escherichia coli and Salmonella: Cellular and Molecular Biology*, Ed. 2, edited by F. C. Neidhardt, R. Curtiss III, E. C. C. Lin, J. Ingraham, B. K. Low *et al.* ASM Press, Washington, DC.

Bairoch, A., and R. Apweiler, 2000 The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. Nucleic Acids Res. **28:** 45–48.

Berlyn, M. B., and S. Letovsky, 1992 Genome-related datasets within the E. coli Genetic Stock Center database. Nucleic Acids Res. **20:** 6143–6151.

Blattner, F. R., G. Plunkett, III, C. A. Bloch, N. T. Perna, V.

Burland *et al.*, 1997 The complete genome sequence of *Escherichia coli* K-12. Science **277:** 1453–1474.

Corbin, R. W., O. Paliy, F. Yang, J. Shabanowitz, M. Platt *et al.*, 2003 Toward a protein profile of *Escherichia coli*: comparison to its transcription profile. Proc. Natl. Acad. Sci. USA **100:** 9232–9237.

Descartes, A., and T. Bunce, 2000 *Programming the Perl DBI.* O'Reilly & Associates, Cambridge, MA.

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

Glasner, J. D., P. Liss, G. Plunkett, III, A. Darling, T. Prasad *et al.*, 2003 ASAP, a systematic annotation package for community analysis of genomes. Nucleic Acids Res. **31:** 147–151.

Goodman, N., 2002 A dim summary of microarray software. Genome Technol. **19:** 58–64.

Gutnick, D., J. M. Calvo, T. Klopotowski and B. N. Ames, 1969 Compounds which serve as the sole source of carbon or nitrogen for *Salmonella typhimurium* LT-2. J. Bacteriol. **100:** 215–219.

Hediger, M. A., D. F. Johnson, D. P. Nierlich and I. Zabin, 1985 DNA sequence of the lactose operon: the lacA gene and the transcriptional termination region. Proc. Natl. Acad. Sci. USA **82:** 6414–6418.

Karp, P. D., M. Riley, M. Saier, I. T. Paulsen, S. M. Paley *et al.*, 2000 The EcoCyc and MetaCyc databases. Nucleic Acids Res. **28:** 56–59.

Khodursky, A. B., B. J. Peter, N. R. Cozzarelli, D. Botstein, P. O. Brown *et al.*, 2000 DNA microarray analysis of gene expression in response to physiological and genetic changes that affect tryptophan metabolism in *Escherichia coli*. Proc. Natl. Acad. Sci. USA **97:** 12170–12175.

Lehnen, D., C. Blumer, T. Polen, B. Wackwitz, V. F. Wendisch *et al.*, 2002 LrhA as a new transcriptional key regulator of flagella, motility and chemotaxis genes in *Escherichia coli*. Mol. Microbiol. **45:** 521–532.

Lercher, M. J., A. O. Urrutia and L. D. Hurst, 2002 Clustering of housekeeping genes provide a unified model of gene order in the human genome. Nat. Genet. **31:** 180–183.

Masuda, N., and G. M. Church, 2003 Regulatory network of acid resistance genes in *Escherichia coli*. Mol. Microbiol. **48:** 699–712.

McCormick, J. R., J. M. Zengel and L. Lindahl, 1991 Intermediates in the degradation of mRNA from the lactose operon of *Escherichia coli*. Nucleic Acids Res. **19:** 2767–2776.

Medigue, C., A. Viari, A. Henaut and A. Danchin, 1993 Colibri: a functional data base for the *Escherichia coli* genome. Microbiol. Rev. **57:** 623–654.

Neidhardt, F. C., J. L. Ingraham and M. Schaechter, 1990 *Physiology of the Bacterial Cell: A Molecular Approach.* Sinauer Associates, Sunderland, MA.

Quadroni, M., W. Staudemann, M. Kertesz and P. James, 1996 Analysis of global responses by protein and peptide fingerprinting of proteins isolated by two dimensional gel electrophoresis: application to sulfate starvation responses of *Escherichia coli*. Eur. J. Biochem. **239:** 773–781.

Riley, M., and B. Labedan, 1996 *E. coli* gene products: physiological functions and common ancestries, pp. 2118–2202 in *Escherichia coli and Salmonella: Cellular and Molecular Biology*, Ed. 2, edited by F. C. Neidhardt, R. Curtiss III, E. C. C. Lin, J. Ingraham, B. K. Low *et al.* ASM Press, Washington, DC.

Riley, M., and D. B. Space, 1996 Genes and proteins of *Escherichia coli* (GenProtEC). Nucleic Acids Res. **24:** 40.

Roy, P. J., J. M. Stuart, J. Lund and S. K. Kim, 2002 Chromosomal clustering of muscle-expressed genes in *Caenorhabditis elegans*. Nature **418:** 975–979.

Rudd, K. E., 2000 EcoGene: a genome sequence database for *Escherichia coli* K-12. Nucleic Acids Res. **28:** 60–64.

Salgado, H., A. Santos-Zavaleta, S. Gama-Castro, D. Millan-Zarate, E. Diaz-Peredo *et al.*, 2001 RegulonDB (version 3.2): transcriptional regulation and operon organization in *Escherichia coli* K-12. Nucleic Acids Res. **29:** 72–74.

Schena, M., D. Shalon, R. W. Davis and P. O. Brown, 1995 Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science **270:** 467–470.

Sherlock, G., 2000 Analysis of large-scale gene expression data. Curr. Opin. Immunol. **12:** 201–205.

Soupene, E., W. C. van Heeswijk, J. Plumbridge, V. Stewart, D. Bertenthal *et al.*, 2003 Physiological studies of *Escherichia coli* strain MG1655: growth defects and apparent cross-regulation of gene expression. J. Bacteriol. **85:** 5611–5626.

Spellman, P. T., and G. M. Rubin, 2002 Evidence for large domains of similarly expressed genes in the *Drosophila* genome. J. Biol. **1:** 5.

Tatusov, R. L., M. Y. Galperin, D. A. Natale and E. V. Koonin, 2000 The COG database: a tool for genome-scale analysis of protein functions and evolution. Nucleic Acids Res. **28:** 33–36.

van der Ploeg, J. R., E. Eichhorn and T. Leisinger, 2001 Sulfonate-sulfur metabolism and its regulation in *Escherichia coli*. Arch. Microbiol. **176:** 1–8.

Wall, L., T. Christiansen and J. Orwant, 2000 *Programming Perl.* O'Reilly & Associates, Cambridge, MA.

Wendisch, V. F., D. P. Zimmer, A. Khodursky, B. Peter, N. Cozzarelli *et al.*, 2001 Isolation of *Escherichia coli* mRNA and comparison of expression using mRNA and total RNA on DNA microarrays. Anal. Biochem. **290:** 205–213.

Zimmer, D. P., E. Soupene, H. L. Lee, V. F. Wendisch, A. B. Khodursky *et al.*, 2000 Nitrogen regulatory protein C-controlled genes of *Escherichia coli*: scavenging as a defense against nitrogen limitation. Proc. Natl. Acad. Sci. USA **97:** 14674–14679.

Communicating editor: E. W. Jones