Research

A functional update of the *Escherichia coli* K-12 genome

Margrethe H Serres*, Shuba Gopal†, Laila A Nahum*, Ping Liang*, Terry Gaasterland† and Monica Riley*

Addresses: *The Josephine Bay Paul Center for Comparative Molecular Biology and Evolution, Marine Biological Laboratory, Woods Hole, MA 02543, USA. †Laboratory of Computational Genomics, Rockefeller University, New York, NY 10021, USA.

Correspondence: Monica Riley. E-mail: mriley@mbl.edu

Published: 20 August 2001

Genome ***Biology*** 2001, **2(9)**:research0035.1–0035.7

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2001/2/9/research/0035

(Print ISSN 1465-6906; Online ISSN 1465-6914)

Received: 2 April 2001
Revised: 8 June 2001
Accepted: 10 July 2001

Abstract

Background: Since the genome of *Escherichia coli* K-12 was initially annotated in 1997, additional functional information based on biological characterization and functions of sequence-similar proteins has become available. On the basis of this new information, an updated version of the annotated chromosome has been generated.

Results: The *E. coli* K-12 chromosome is currently represented by 4,401 genes encoding 116 RNAs and 4,285 proteins. The boundaries of the genes identified in the GenBank Accession U00096 were used. Some protein-coding sequences are compound and encode multimodular proteins. The coding sequences (CDSs) are represented by modules (protein elements of at least 100 amino acids with biological activity and independent evolutionary history). There are 4,616 identified modules in the 4,285 proteins. Of these, 48.9% have been characterized, 29.5% have an imputed function, 2.1% have a phenotype and 19.5% have no function assignment. Only 7% of the modules appear unique to *E. coli,* and this number is expected to be reduced as more genome data becomes available. The imputed functions were assigned on the basis of manual evaluation of functions predicted by BLAST and DARWIN analyses and by the MAGPIE genome annotation system.

Conclusions: Much knowledge has been gained about functions encoded by the *E. coli* K-12 genome since the 1997 annotation was published. The data presented here should be useful for analysis of *E. coli* gene products as well as gene products encoded by other genomes.

Background

The field of genomics has been expanding at a rapid pace since the annotated *Escherichia coli* K-12 genome was published in 1997 [1], with the current number of published genomes exceeding 66 and with another 364 on their way according to the Genomes OnLine Database (GOLD) [2]. Deciphering the functions encoded by all gene products of the genomes is the next big challenge in the field. Function attributions through experimental, biochemical and genetic analyses and through bioinformatic studies are continuing, and microarray technology is shedding additional light on the functions associated with the gene products of the organism in question. The wealth of biological information on *E. coli* is still increasing [3] and is contributing to a better understanding of this organism as well as of functions encoded in other organisms. It is therefore important that the most up-to-date information on *E. coli* gene products is available and used by researchers.

Several databases have been assembled for various areas of knowledge about the *E. coli* genome [4-9]. Each compilation has a different emphasis and collects different sets of information related to the function of the gene products. In the GenProtEC database, we have been curating information on physiological function and modular construction of gene products. Other databases most closely related to ours include EcoCyc, with emphasis on metabolic pathways [6], the CGSC database, with information on the genotypes and phenotypes of mutant strains [8], and EcoGene, which includes information on gene reconstructions, alternative gene boundaries and verified amino-terminal amino-acid sequences of the mature proteins [5]. The *E. coli* genome project at the University of Wisconsin-Madison presents genome data on *E. coli* K-12 and pathogenic enterobacteria [9].

We present a functional update for *E. coli* K-12 gene products that incorporates information from the literature and referenced databases obtained since the 1997 GenBank deposit. Our focus has been the biological function of the gene products. Coding sequences (CDSs) encoding proteins whose function previously was imputed or not known were re-evaluated, and putative functions were assigned by manually evaluating the results from BLAST and DARWIN (data analysis and retrieval with indexed nucleotide/peptide sequences) analyses. The MAGPIE (multipurpose automated genome project investigation environment) genome annotation system [10] was also applied. MAGPIE detected alternative boundaries for some of the open reading frames (ORFs).

Results

Number of genes in the *E. coli* K-12 genome

For the initial annotation of the *E. coli* K-12 genome [1], 4,404 genes were identified with Blattner numbers (Bnums). Among the genes, 4,288 were believed to encode proteins and 116 to encode RNAs. Since then six Bnums have been retired: b0322, b0395, b0663, b0667, b0669 and b0671 (G. Plunkett, personal communication). In addition, three new genes have been identified and assigned to Bnums. These include the protein-coding b4406 (*yaeP*, SWISS-PROT P52099) and b4407 (*thiS*, SWISS-PROT O32583) and the RNA encoding b4408. The current number of *E. coli* genes is 4,401, with 4,285 encoding proteins and 116 encoding RNAs.

MAGPIE identified 5,527 candidate CDSs that were assigned to MAGPIE identifiers (Magnums) (see MAGPIE [11] for details). The 4,285 CDSs identified by Bnums were also identified with Magnums. Variations were detected for either the start or stop positions for 1,077 of these CDSs resulting in differences in the encoded proteins ranging from 1 to 147 amino acids, the latter in PtsA (Bnum b3947, Magnum ec_6103). The other Magnum-identified candidate CDSs include retired Bnums (six Magnums), CDSs located between the boundaries of Bnums (506 Magnums), and CDSs overlapping existing Bnums (730 Magnums). Among the Magnums located between the boundaries of Bnums are 21 CDSs that encode proteins of 80 or more amino acids. One such CDS identified by MAGPIE (Magnum ec_2510) is located between b1624 and b1625 and encodes a protein of 66 amino acids. The carboxy-terminal 41 amino acids of this CDS are identical to the amino-acid sequence of the recently characterized beta-lactam resistance protein Blr (SWISS-PROT P56976) located at the same position [12]. Other Magnums located between Bnum boundaries may correspond to short *E. coli* proteins.

Functional annotation of *E. coli* K-12 gene products

The functional assignments of the *E. coli* gene products in the November 97 GenBank U00096 deposit represented an accumulation of information retrieved from the literature (collected in the GenProtEC and EcoCyc databases) as well as imputed functions based on similarity of a known protein to the translated sequences [1]. Since the deposit to GenBank was made, our database GenProtEC has continually been updated with knowledge on *E. coli* gene products appearing in the literature [3,13]. Information on transcriptional regulators has been incorporated from the work of J. Collado-Vides [14,15], and transport protein information has been adapted from the work of M.H. Saier and I.T. Paulsen [16,17]. GenProtEC also contains imputed function assignments based on sequence similarity to orthologous or paralogous proteins, on gene (operon) location and on phenotypes of mutants [18].

Gene products whose functions were known were not considered further for the functional update. The remaining 2,294 CDSs whose gene products had a putative or unknown function assignment were analyzed using BLAST and DARWIN. BLAST analyses were carried out for both the Bnum- and the Magnum-derived protein sequences. The results for the Bnum-derived protein sequences and the automatic functions predicted by MAGPIE or HERON (human-emulated reasoning for objective notations) were manually evaluated and imputed functions were assigned. Although the manual annotation step could not compete with the speed of the automatic annotation process of HERON, it provided us with more useful function descriptions. A comparison of the manually assigned putative functions with the HERON predicted functions showed that when leaving aside issues of specificity, a nearly equivalent function was predicted in 46% of the cases, whereas in 52% of the cases less information was obtained with HERON.

After the function update of the 2,294 CDSs, 1,306 gene products were assigned a putative function and 126 gene products were described by a phenotype. The remaining gene products were given one of the following three assignments: 'conserved protein', where sequence-similar matches were found but the function could not be determined in the

absence of consistent functions reported for the matching sequences; 'conserved hypothetical protein', where sequence-similar matches existed but these had no associated function; 'unknown CDS', where the translated sequence had no known sequence match outside *E. coli*. The current function description includes 256 conserved proteins, 282 conserved hypothetical proteins and 324 unknown CDSs. The 862 gene products with no function assignment represent 19.6% of the *E. coli* chromosomal genes, and the unknown CDSs at this time represent 7.4% of *E. coli* genes.

A sample of the annotated *E. coli* K-12 genes is shown in Table 1. Each gene is identified by a Bnum, Magnum, gene product type and gene product (Function April 2001). A complete table of the current 4,401 Bnums is available as an additional data file online and at MAGPIE [11]. In this table the genes are identified by their Bnum, Bnum_module, Bnum start and stop position, Magnum, Magnum start and stop position, and gene product type. The functions of the gene products are described by the currently annotated function (Function April 2001) and by the function in the GenBank deposit (Function November 1997). A continually updated table that contains the functions of *E. coli* gene products is available through GenProtEC [4].

Many changes are evident when comparing the updated annotation to that of 1997. The number of CDSs without function assignment has been reduced from 1,354 to 862. This reduction is due to functions being experimentally determined (77 CDSs), assignment of putative functions (367 CDSs), phenotype-associated functions (14 CDSs), and genes identified as belonging to phages (138 CDSs). In addition, inferred function assignments were withdrawn for 104 CDS-coded proteins whose functions remain unknown.

The number of gene products with putative function assignments has changed from 1,120 to 1,306. New functions were inferred for 473 CDSs. Putative function assignments were also removed as a result of new experimental data (175 CDSs), assignment of phenotype (8 CDSs) or reassessment of putative function assignments (104 CDSs).

Proteins as modular entities

Some of the proteins encoded in the *E. coli* genome have arisen through fusion of two or more genes. Examples of such gene fusions are the multifunctional enzymes Aas (2-acylglycerophospho-ethanolamine acyl transferase and acyl-acyl carrier protein synthetase) and GlmU (*N*-acetyl glucosamine-1-phosphate uridyltransferase and glucosamine-1-phosphate acetyl transferase) [19,20]. We have chosen to deal with

Table 1

A sample of annotated *E. coli* K-12 genes

Bnum	Magnum	Gene	Gene product type*	Gene product†
b0038	ec_0059	*caiB*	e	l-carnitine dehydratase, NAD(P)-binding
b0039	ec_0061	*caiA*	pe	Putative acyl-CoA dehydrogenase
b0019	ec_0026	*nhaA*	t	Na^+/H^+ antiporter, NhaA family
b0040	ec_0062	*caiT*	pt	Putative betaine/carnitine/choline transport protein, BCCT family
b0064	ec_0098	*araC*	r	Transcriptional regulator of arabinose catabolism, AraC/XylS family
b0076	ec_0116	*leuO*	pr	Putative transcriptional regulator of leucine biosynthesis, LysR family
b0814	ec_1234	*ompX*	m	Outer membrane protease, receptor for phage OX2
b0117	ec_0171	*yacH*	pm	Putative membrane protein
b0170	ec_0246	*tsf*	f	Protein chain elongation factor EF-Ts
b0236	ec_0334	*prfH*	pf	Putative peptide chain release factor
b0023	ec_0031	*rpsT*	s	30S ribosomal subunit protein S20
b0138	ec_0200	*yadM*	ps	Putative fimbrial-like protein
b0684	ec_1032	*fldA*	c	Flavodoxin 1
b1697	ec_2618	*ydiQ*	pc	Putative electron transfer flavoprotein
b0251	ec_0359	*yafY*	h	CP4-6 prophage
b0054	ec_0083	*imp*	ph	Organic solvent tolerance
b0201		*rrsH*	n	16S rRNA
b0001	ec_G0001	*thrL*	l	*thr* operon leader peptide
b0050	ec_0078	*apaG*	o	Conserved protein
b0081	ec_0123	*mraZ*	o	Conserved hypothetical protein
b0005	ec_G0005	*yaaX*	o	Unknown CDS

*Gene product type: c, carrier; e, enzyme; f, factor; h, extrachromosomal origin; l, leader peptide; m, membrane component; n, RNA; o, ORF of unknown function; pc, putative carrier; pe, putative enzyme; pf, putative factor; ph, phenotype; pm, putative membrane component; pr, putative regulator; ps, putative structure; pt, putative transporter; r, regulator; s, structure; t, transporter. †Gene products consisting of one identified module.

proteins as modular entities where a module is defined as a protein element that has at least 100 amino-acid residues, carries a biological function and is presumed to have an independent evolutionary history [21]. Most modules in *E. coli* are individual proteins. They can, however, also be part of a protein where multiple modules have been joined by gene fusion, as is the case for Aas and GlmU. Other protein types in *E. coli* such as transporters and regulators also involve gene fusion events. The current modular assignments are based on analysis of protein sequences within *E. coli* K-12 (P. Liang and M. Riley, unpublished data).

There are at present 287 compound genes identified in the *E. coli* genome, each containing two to four modules. Table 2 contains a list of multimodular proteins where each module encodes a distinct function. Enzymes, transporters and regulators are all present in the list. The majority of modular proteins, 217, contain modules belonging to different paralogous groups (data not shown). Other multimodular proteins appear to be a result of internal duplication (56 genes) or a combination of gene fusion and duplication (14 genes). The *E. coli* chromosome is currently represented by 4,401 genes encoding 116 RNAs and 4,616 protein modules. Additional modules are expected to be identified upon analysis of protein sequences from other genomes (P. Liang and M. Riley, unpublished data). Examples are the bifunctional proteins ThrA (aspartokinase I and homoserine dehydrogenase I) and MetL (aspartokinase II and homoserine dehydrogenase II) where only the amino-terminal modules representing the kinase activities have been identified on the basis of their sequence similarity to the *E. coli* unimodular aspartokinase III (LysC). Both the amino-terminal aspartokinase and the carboxy-terminal homoserine dehydrogenase activities of ThrA and MetL have been verified with biochemical and genetic tools [22,23]. The module representing the dehydrogenase activity has not been identified by matching internal paralogs as *E. coli* itself does not contain a unimodular sequence-similar dehydrogenase. S*accharomyces cerevisiae*, however, does contain a unimodular sequence-similar homoserine dehydrogenase (DhoM, SWISS-PROT P31116), which can be used in identifying the carboxy-terminal module. Thus, by detecting orthologous matches to parts of genes we will be able to identify additional multimodular proteins.

Current status

Table 3 presents a summary of the gene products encoded in the *E. coli* K-12 genome represented as modular entities. Half of the modules have been experimentally characterized. Enzymes are the largest gene product type, representing 43.9% of the characterized gene products and 34.2% of the total gene products. Other major gene product types are transporters and regulators. Among the remaining modules, 60% have function predictions. The gene products without a function assignment still constitute a significant portion of the *E. coli* genome (19% of modules). A summary of the development of information on *E. coli* gene products over the past eight years is shown in Table 4. It is evident that much knowledge has been gained since these analyses began in 1993 [3,24].

Discussion

An updated version of the function assignments for *E. coli* K-12 gene products has been presented using the genes identified in the GenBank U00096 deposit. Alternative gene boundaries were produced by MAGPIE. The MAGPIE genome annotation system also identified candidate CDSs that may represent gene products not identified in the GenBank U00096 deposit. Small ORFs with biological activity are likely to be abundant in the organism but await verification by biological data. Undoubtedly, the intergenic regions of *E. coli* K-12, as studied by Rudd [25] and Bachellier *et al.* [26], are also important for the function and regulation of gene products.

The percentage of identified chromosomal gene products without a function assignment is decreasing and is currently 19.6%. Only 7.4% of *E. coli* genes have no match in current sequence databases. This number will be further reduced with the release of the annotated genomes of *Salmonella*, *Shigella* and other closely related organisms. Preliminary data show that the number of unknown CDSs (ORFs encoding proteins without sequence-similar matches) will be less than 170 after data on the *Salmonella typhimurium* genome is included (M.H.S., unpublished data).

The function assignments presented here mainly represent the molecular functions of the gene products. With the generation of microarray data, gene products will also be characterized to a greater degree by the role they play in the cell under specific conditions. We have recently developed a classification system for cellular functions of *E. coli* K-12 gene products and have assigned more than one cellular role to some gene products where this is appropriate [27]. There is also a need for a more uniform way of describing both the molecular and cellular roles of gene products among diverse organisms, and this issue is currently being addressed by the Gene Ontology Consortium [28].

Conclusions

We have presented a functional update of the gene products encoded by the genes of *E. coli* K-12 identified in the GenBank Accession U00096 deposit. The *E. coli* proteins were treated as modular entities where a module is at least 100 amino acids, carries a biological function, and has an independent evolutionary history. The functional update was performed by manual evaluation of the data obtained from GenProtEC, BLAST and DARWIN analyses, and MAGPIE annotation. A table containing the updated function assignments of *E. coli* K-12 gene products is available as

Table 2

A sample of multimodular gene products of *E. coli* K-12

Module	Magnum	Gene	Gene product type*	Gene product
b0149_2	ec_0214	*mrcB*	e	Glycosyl transferase of penicillin-binding protein 1b (2nd module)
b0149_3	ec_0214	*mrcB*	e	Transpeptidase of penicillin-binding protein 1b (3rd module)
b0679_1	ec_1018	*nagE*	t	PTS family enzyme IIC, n-acetylglucosamine-specific (1st module)
b0679_2	ec_1018	*nagE*	t	PTS family enzyme IIB, n-acetylglucosamine-specific (2nd module)
b0679_3	ec_1018	*nagE*	t	PTS family, enzyme IIA, n-acetylglucosamine-specific (3rd module)
b0886_1	ec_1338	*cydC*	t	ABC superfamily (membrane) cytochrome-related transporter (1st module)
b0886_2	ec_1338	*cydC*	t	ABC superfamily (atp_bind) cytochrome-related transporter (2nd module)
b1241_1	ec_1883	*adhE*	e	Acetaldehyde-CoA dehydrogenase (1st module)
b1241_2	ec_1883	*adhE*	e	Iron-dependent alcohol dehydrogenase (2nd module)
b1439_1	ec_2214	*ydcR*	pr	Putative transcriptional regulator, GntR family (1st module)
b1439_2	ec_2214	*ydcR*	pt	Putative ATP-binding component of a transport system (2nd module)
b1621_1	ec_2503	*malX*	t	PTS family enzyme IIC, maltose and glucose-specific (1st module)
b1621_2	ec_2503	*malX*	t	PTS family enzyme IIB, maltose and glucose-specific (2nd module)
b2463_1	ec_3778	*maeB*	pe	Putative malic oxidoreductase (1st module)
b2463_3	ec_3778	*maeB*	pe	Putative phosphate acetyl transferase (3rd module)
b2537_1	ec_3905	*hcaR*	r	Transcriptional activator of hca cluster, LysR family (1st module)
b2537_2	ec_3905	*hcaR*	pe	Putative oxidoreductase (2nd module)
b2836_1	ec_4348	*aas*	e	2-acylglycerophospho-ethanolamine acyl transferase (1st module)
b2836_2	ec_4348	*aas*	e	Acyl-acyl carrier protein synthetase (2nd module)
b3464_1	ec_5327	*ftsY*	m	Membrane-binding component of cell division protein (1st module)
b3464_2	ec_5327	*ftsY*	e	GTPase component of cell division membrane protein (2nd module)
b3692_1	ec_5701	*dgoA*	e	2-dehydro-3-deoxygalactonate 6-phosphate aldolase (1st module)
b3692_2	ec_5701	*dgoA*	e	Galactonate dehydratase (2nd module)
b3730_1	ec_5762	*glmU*	e	*N*-acetyl glucosamine-1-phosphate uridyltransferase (1st module)
b3730_2	ec_5762	*glmU*	e	Glucosamine-1-phosphate acetyl transferase (2nd module)
b3846_1	ec_5942	*fadB*	e	3-hydroxybutyryl-coa epimerase; delta(3)-cis-delta(2)-trans-enoyl-coa-isomerase;enoyl-coa-hydratase (1st module)
b3846_2	ec_5942	*fadB*	e	3-hydroxyacyl-coa dehydrogenase (2nd module)
b4035_2	ec_6229	*malK*	r	Phenotypic repressor of mal operon (2nd module)
b4035_1	ec_6229	*malK*	t	ABC superfamily (atp_bind) maltose transport protein (1st module)

*Gene product type: e, enzyme; m, membrane; pe, putative enzyme; pr, putative regulator; pt, putative transporter; r, regulator; t, transporter.

an additional data file online, and at GenProtEC [4] and MAGPIE [11]. We believe these data will be valuable for analysis of *E. coli* K-12 itself as well as for the analysis of gene products encoded by other genomes.

Materials and methods

Automated annotation

MAGPIE ORF prediction

A three-step approach to ORF prediction was taken to prepare the MAGPIE project for *E. coli*. GLIMMER 2.0 with a minimum ORF length of 80 nucleotides was initially used to create the base set of predictions [29]. Glimmer 2.0 was run with all default parameters, as recommended in the documentation [29] and trained on the annotated set of ORFs from the Blattner *et al.* release of 1997 [1]. Because GLIMMER selectively identifies ORFs that match a statistical model of a gene for the organism [29], GLIMMER may miss genes that were laterally transferred or acquired more recently from other genomes. We therefore chose to combine the GLIMMER predictions with those of a syntactic tool encoded within MAGPIE. This tool identifies stop codons and then 'backtracks' to the farthest upstream acceptable in-frame start codon and defines this as the ORF [10]. A non-redundant set of all GLIMMER ORFs plus syntactic ORFs between GLIMMER ORFs was generated. Finally, ORFs annotated by Blattner *et al.* that were not present in the non-redundant set were added to the MAGPIE project.

BLAST analysis

The CDSs were compared to the NCBI nucleotide (nt) and non-redundant protein (nr) databases using gapped BLAST [30]. Protein-sequence motifs were identified by PROSITE

Table 3

Gene products encoded by the *E. coli* K-12 chromosome

Gene product type	Characterized	Putative assignment	Total (%)
Enzyme	1,042	578	1,620 (34.2)
Transport	382	364	746 (15.8)
Regulator	238	167	405 (8.5)
Membrane	53	158	211 (4.4)
Factor	117	33	150 (3.2)
Structure	92	35	127 (2.7)
Carrier	35	25	60 (1.3)
Extrachromosomal*	288		288 (6.1)
Phenotype			98 (2.1)
RNA	116		116 (2.4)
Leader	12		12 (0.3)
ORF			899 (19.0)
Total	2,375	1,360	4,732 (100.0)

Proteins are represented as modules. *Extrachromosomal origin.

[31]. A search against the MAGPIE-predicted proteins of over 40 completed genomes, including the previously annotated *E. coli* set, was also performed.

Functional annotation

Automated function annotation was provided using HERON. Description lines with low information content (for example, descriptions containing words such as ''hypothetical'' or ''putative'') were filtered out. HERON then calculated word frequencies in the remaining descriptions, identified the top three most common words, and selected the description of the highest-scoring sequence match (for homology comparisons) with one or more high-frequency words. The selected description became the automated annotation for the coding region.

Manual annotation

BLAST analysis

The protein sequences collected from GenBank Accession U00096 were compared to the nr database using gapped BLAST [30].

DARWIN analysis

DARWIN (version 2.0) was used to detect sequence-similar proteins within *E. coli* K-12 and in 20 additional microbial genomes [32] (P. Liang and M. Riley, unpublished data). In addition to orthologous matches, groups of paralogous proteins of *E. coli* K-12 were generated on the basis of the DARWIN results. In our hands, DARWIN is particularly successful in identifying distant sequence similarities, a consequence no doubt of the application of multiple substitution matrices optimized for the organism and to each sequence pair.

Table 4

History of distribution of gene product types for *E. coli* K-12

Gene product type	1993*	1998†	2001
Enzyme	748‡	906	990
Putative enzyme		452	550
Transporter	221	257	310
Putative transporter		281	298
Regulator	164	204	213
Putative regulator		168	151
Membrane		37	47
Putative membrane		55	132
Factor	36	68	109
Putative factor		52	33
Structure	113§	83	90
Putative structure		58	35
Carrier		17	35
Putative carrier		6	25
Extrachromosomal origin		56	282
Putative extrachromosomal		15	
Phenotype	314	148	98
RNA	104	112	116
Putative RNA		4	
Leader		12	12
ORF		1,413	886
Total	1,700	4,404	4,412¶

*Adapted from Riley [24]. †Data from July 1998 record of GenProtEC. ‡Includes enzymes, leader peptides and enzyme activity. §Includes membrane components. ¶This number includes overlap situations where modules of a gene belong to different gene product type categories. The total number of genes is 4,401.

Functional annotation

Functions were assigned to gene products on the basis of a manual evaluation of the results from the BLAST and DARWIN analyses. The automatic function prediction was also taken into account. In addition to incorporating recent experimental information, a substantial amount of human judgment was brought to bear.

Additional data files

A complete table of the current 4,401 Bnums is provided online.

Acknowledgements

This work was supported by NIH grant RO1 RR07861, the NASA Astrobiology Institute grant NCC2-1054, grants from the Edward Mallinckrodt, Jr Foundation and the Sinsheimer Foundation, and NSF grants NSF DBI - 9984882 and NSF IIS - 9996304. We thank Alastair Kerr for help on data retrieval and Edward A. Adelberg for help on monitoring the *E. coli* literature. We thank Guy Plunkett 3rd for information on Blattner number status, Mark Schroeder for assistance with the MAGPIE analysis, and Peter Karp and Stefan Bekiranov for suggestions regarding the design and implementation of HERON.

References

1. Blattner FR, Plunkett G 3rd, Bloch CA, Perna NT, Burland V, Riley M, Collado-Vides J, Glasner JD, Rode CK, Mayhew GF, *et al.*: **The complete genome sequence of *Escherichia coli* K-12.** *Science* 1997, **277:**1453-1474.
2. **GOLD: Genomes OnLine Database homepage** [http://igweb.integratedgenomics.com/GOLD/]
3. Riley M, Serres MH: **Interim report on genomics of *Escherichia coli*.** *Annu Rev Microbiol* 2000, **54:**341-411.
4. **GenProtEC database** [http://genprotec.mbl.edu/]
5. Rudd KE: **EcoGene: a genome sequence database for *Escherichia coli* K-12.** *Nucleic Acids Res* 2000, **28:**60-64.
6. Karp PD, Riley M, Saier M, Paulsen IT, Paley SM, Pellegrini-Toole A: **The EcoCyc and MetaCyc databases.** *Nucleic Acids Res* 2000, **28:**56-59.
7. Thomas GH: **Completing the *E. coli* proteome: a database of gene products characterised since the completion of the genome sequence.** *Bioinformatics* 1999, **15:**860-861.
8. **CGSC: *E. coli* Genetic Stock Center** [http://cgsc.biology.yale.edu/]
9. ***E. coli* genome project University of Wisconsin-Madison** [http://www.genome.wisc.edu/]
10. Gaasterland T, Sensen CW: **Fully automated genome analysis that reflects user needs and preferences. A detailed introduction to the MAGPIE system architecture.** *Biochimie* 1996, **78:**302-310.
11. **MAGPIE automated genome project investigation environment** [http://genomes.rockefeller.edu/magpie/ecoli/]
12. Wong RS, McMurry LM, Levy SB: **'Intergenic' blr gene in *Escherichia coli* encodes a 41-residue membrane protein affecting intrinsic susceptibility to certain inhibitors of peptidoglycan synthesis.** *Mol Microbiol* 2000, **37:**364-370.
13. Serres MH, Riley M: **Genomics and metabolism in *Escherichia coli*.** In *The Prokaryotes: An Evolving Electronic Database for the Microbiological Community*. Edited by Dworkin M, *et al.* New York: Springer-Verlag, 2000. Available at [http://www.prokaryotes.com]
14. Perez-Rueda E, Collado-Vides J: **The repertoire of DNA-binding transcriptional regulators in *Escherichia coli* K-12.** *Nucleic Acids Res* 2000, **28:**1838-1847.
15. **RegulonDB** [http://www.cifn.unam.mx/regulondb/]
16. Saier MH Jr: **A functional-phylogenetic classification system for transmembrane solute transporters.** *Microbiol Mol Biol Rev* 2000, **64:**354-411.
17. **Genomic Comparisons of Membrane Transport Systems** [http://www.biology.ucsd.edu/~ipaulsen/transport/]
18. Riley M: **Genes and proteins of *Escherichia coli* K-12.** *Nucleic Acids Res* 1998, **26:**54.
19. Jackowski S, Jackson PD, Rock CO: **Sequence and function of the *aas* gene in *Escherichia coli*.** *J Biol Chem* 1994, **269:**2921-2928.
20. Mengin-Lecreulx D, van Heijenoort J: **Copurification of glucosamine-1-phosphate acetyltransferase and N- acetylglucosamine-1-phosphate uridyltransferase activities of *Escherichia coli*: characterization of the *glmU* gene product as a bifunctional enzyme catalyzing two subsequent steps in the pathway for UDP-N-acetylglucosamine synthesis.** *J Bacteriol* 1994, **176:**5788-5795.
21. Riley M, Labedan B: **Protein evolution viewed through *Escherichia coli* protein sequences: introducing the notion of a structural segment of homology, the module.** *J Mol Biol* 1997, **268:**857-868.
22. Vickers LP, Ackers GK, Ogilvie JW: **Aspartokinase I-homoserine dehydrogenase I of *Escherichia coli* K12. Concentration-dependent dissociation to dimers in the presence of L-threonine.** *J Biol Chem* 1978, **253:**2155-2160.
23. Truffa-Bachi P, Van Rapenbusch R, Gros C, Cohen GN, Janin J: **The threonine-sensitive homoserine dehydrogenase and aspartokinase activities of *Escherichia coli* K-12. Subunit structure of the protein catalyzing the two activities.** *Eur J Biochem* 1969, **7:**401-407.
24. Riley M. **Functions of the gene products of *Escherichia coli*.** *Microbiol Rev* 1993, **57:**862-952.
25. Rudd KE: **Novel intergenic repeats of *Escherichia coli* K-12.** *Res Microbiol* 1999, **150:**653-664.
26. Bachellier S, Clement JM, Hofnung M: **Short palindromic repetitive DNA elements in enterobacteria: a survey.** *Res Microbiol* 2000, **150:**627-639.
27. Serres MH, Riley M: **MultiFun, a multifunctional classification scheme for *Escherichia coli* K-12 gene products.** *Microb Comp Genomics* 2001, **5:**205-222.
28. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT: **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.** *Nat Genet* 2000, **25:**25-29.
29. Delcher AL, Harmon D, Kasif S, White O, Salzberg SL: **Improved microbial gene identification with GLIMMER.** *Nucleic Acids Res* 1999, **27:**4636-4641.
30. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Res* 1997, **25:**3389-3402.
31. Bairoch A: **PROSITE: a dictionary of sites and patterns in proteins.** *Nucleic Acids Res* 1992, Suppl 20:2013-2018.
32. Gonnet GH, Cohen MA, Benner SA: **Exhaustive matching of the entire protein sequence database.** *Science* 1992, **256:**1443-1445.