Nucleic Acids Research, 2006, Vol. 34, No. 1 **1–9**
doi:10.1093/nar/gkj405

Escherichia coli K-12: a cooperatively developed annotation snapshot—2005

Monica Riley*, Takashi Abe[1], Martha B. Arnaud[2], Mary K.B. Berlyn[3], Frederick R. Blattner[4], Roy R. Chaudhuri[5], Jeremy D. Glasner[4], Takashi Horiuchi[6], Ingrid M. Keseler[7], Takehide Kosuge[1], Hirotada Mori[8,9], Nicole T. Perna[4], Guy Plunkett III[4], Kenneth E. Rudd[10], Margrethe H. Serres, Gavin H. Thomas[11], Nicholas R. Thomson[12], David Wishart[13] and Barry L. Wanner[14]

Josephine Bay Paul Center, Marine Biological Laboratory, Woods Hole, MA 02543, USA, [1]Center for Information Biology and DNA Data Bank of Japan, National Institute of Genetics, Research Organization of Information and Systems, Yata 1111, Mishima, Shizuoka 411-8540, Japan, [2]Department of Genetics, Candida Genome Database Stanford University School of Medicine, Stanford, CA 94305-5120, USA, [3]Department of Molecular, Cellular and Developmental Biology, Yale University, New Haven, CT 06520-8103, USA, [4]Genome Center of Wisconsin, 425 Henry Mall, University of Wisconsin, Madison, WI 53706, USA, [5]Division of Immunity and Infection, University of Birmingham, Edgbaston, Birmingham B15 2TT, UK, [6]National Institute for Basic Biology, Nishigonaka 38, Myodaiji, Okazaki 444-8585 Aichi, Japan, [7]SRI International, 333 Ravenswood Avenue, Menlo Park, CA 94025-3493, USA, [8]Graduate School of Biological Sciences, Nara Institute of Science and Technology, Ikoma, Nara 630-0101, Japan, [9]The Institute of Advanced Biosciences, Keio University, Tsuruoka, Yamagata 997-0017, Japan, [10]Department of Biochemistry and Molecular Biology, The University of Miami Miller School of Medicine, Miami, FL 33140, USA, [11]Department of Biology, University of York, PO Box 373, York YO10 5YW, UK, [12]The Wellcome Trust Sanger Institute, Genome Campus, Hinxton, Cambridge CB10 1SA, UK, [13]Department of Computing Science and Biological Sciences, 2-21 Athabasca Hall University of Alberta, Edmonton, Alberta, Canada T6G 2E8 and [14]Department of Biological Sciences, Purdue University, 915 W. State Street, West Lafayette, IN 47907-2054, USA

Received November 5, 2005; Revised and Accepted December 5, 2005

ABSTRACT

The goal of this group project has been to coordinate and bring up-to-date information on all genes of *Escherichia coli* K-12. Annotation of the genome of an organism entails identification of genes, the boundaries of genes in terms of precise start and end sites, and description of the gene products. Known and predicted functions were assigned to each gene product on the basis of experimental evidence or sequence analysis. Since both kinds of evidence are constantly expanding, no annotation is complete at any moment in time. This is a snapshot analysis based on the most recent genome sequences of two *E.coli* K-12 bacteria. An accurate and up-to-date description of *E.coli* K-12 genes is of particular importance to the scientific community because experimentally determined properties of its gene products provide fundamental information for annotation of innumerable genes of other organisms. Availability of the complete genome sequence of two K-12 strains allows comparison of their genotypes and mutant status of alleles.

INTRODUCTION

Escherichia coli strain K-12 is arguably the single organism about which the most is known. Originally isolated in 1922, it was catapulted to prominence by the discovery of strain K-12's ability to carry out genetic recombination by conjugation (1) and, soon after, by generalized transduction (2). The strain K-12 has been widely distributed to laboratories across the world. Over the ensuing years it became the primary model organism for basic biology, molecular genetics and physiology of bacteria, and was the founding workhorse of the biotechnology industry.

*To whom correspondence should be addressed. Tel: +1 508 269 7388; Fax: +1 508 457 4727; Email: mriley@mbl.edu
Correspondence may also be addressed to Barry L. Wanner. Tel: +1 765 494 8034; Fax: +1 765 494 0876; Email: blwanner@purdue.edu

Annotation of *E.coli* has not only served the *E.coli* community, but has formed a basis for extrapolation of gene functions to virtually every other prokaryotic, as well as eukaryotic, genome through analogy based on protein sequence similarities. As such, the accuracy and completeness of the *E.coli* information is of great importance to the community of biologists working in all disciplines and with all organisms. We report here the work of a group of scientists dedicated to full review and update of the annotation of *E.coli* K-12.

The entire genome sequence of K-12 strain MG1655 was first completed and annotated by a group assembled by F. R. Blattner (3). The genome of a second K-12 strain, W3110, was completed recently under the direction of Takashi Horiuchi at the National Institute for Basic Biology in Japan (4). At the same time the sequence of the genome of MG1655 was corrected and updated. MG1655 was chosen for its close relationship with the original *E.coli* strain K-12 (called EMG2), whereas W3110 was chosen because it has been widely used as a 'wild-type' strain by many investigations worldwide from the 1950s. Both had been cured of the λ prophage and lack the F^+ fertility factor of ancestral *E.coli* K-12 EMG2. MG1655 and W3110 are 1- and 2-step descendents of *E.coli* K-12 W1485 (F^+, λ^-), respectively, which is in turn a direct descendent of EMG2 (4,5).

By comparing and re-sequencing regions of discrepancies between MG1655 and W3110, highly accurate genomes have now been created for both strains (4). Corrections to the original MG1655 genome (3) are at 243 sites (totaling 358 nt), a correction rate 8 years later of $\sim$7 in 10^5. Work done by the participants of an *E.coli* annotation workshop held in November 2003 reconciled sequence differences that led to deposit of a corrected MG1655 genome sequence entry (GenBank™ U00096.2, released in June 2004). Subsequent work done in a March 2005 workshop introduced additional changes. The participants of these workshops have co-authored this manuscript.

Although both MG1655 and W3110 are isolates of the *E.coli* K-12 strain, their genomes are not identical. The different lengths of the MG1655 (4 639 675 nt) and W3110 (4 646 332 nt) genomes reflect a larger number of insertion sequence (IS) elements and absence of a defective phage in the W3110 genome. Other differences are found in the occurrence of mutations, reflecting changes that presumably occurred during maintenance of the cultures in separate laboratories.

Genome annotation, of necessity, is an ongoing process. In the interim from 1997, many scientists, not organized as a group, but united intellectually by their interest in developing a unified vision of the organism, have continued to upgrade, update and collate new information about *E.coli* as it has emerged. This has resulted in a number of public databases with information on genes, genomics and proteins of *E.coli* K-12, none identical, each with a different emphasis. Other more general databases contain information relevant to many organisms, helpful in interpretation of gene sequences.

The goal of the current project was to consolidate the work of scientists who have been working independently by developing our best consensus on the status and properties of each of the genes of *E.coli* K-12 at the present moment. The goal was decidedly not to create a new database, but instead, to present to the public a comprehensive, updated annotation of *E.coli* K-12 which would be presented both in spreadsheet and simple flat-file formats. The latter can easily be parsed by computers and readers alike and therefore can be incorporated into extant databases by their providers. These are available as Supplementary Table 1.xls, Supplementary Table 1.txt and, to aid in interpreting the data, Supplementary Table 1 Explanatory Notes. Less extensive information from the new MG1655 and W3110 annotations have been included in new GenBank™ and DNA Data Bank of Japan (DDBJ) entries, accession number U00096.3 and DDBJ AP009048, respectively.

We refer to this outcome as a 'snapshot' to emphasize that information about *E.coli* genes and their products are a moving target, and overtaken rapidly with more recent information. The authors have made no plans to develop this snapshot further. Highly desirable would be the establishment of an accessible community resource of data on *E.coli* K-12 with community participation, ongoing maintenance and continuous updating of all information. At this moment interested members of the *E.coli* community are applying to NIH for support to establish a 'K-12 information resource'.

RESULTS AND DISCUSSION

The workshops

The need to consolidate the efforts of scientists who had been working independently was a subject of discussion at an informal '*E.coli* consortium' meeting organized by Barry Wanner and coworkers, which was held in early March 2003 at the University of California San Diego Supercomputer Center. Following on from this, Monica Riley and Margrethe 'Gretta' Serres organized two Annotation Workshops held at the Marine Biological Laboratory, Woods Hole, MA on November 14–18, 2003 and March 19–24, 2005 followed by a wrap-up meeting hosted by Fred Blattner in Madison, WI on June 2, 2005.

Pooling information, annotation and reconciliation

Two distinct aspects of *E.coli* annotation had to be addressed. One related to issues arising from sequence corrections that necessitated annotation changes in the sense of establishing new boundaries for some genes and transcripts. The other quite separate operation, functional annotation, entailed developing up-to-date descriptions of all gene products.

In establishing gene boundaries, by convention start and end sites of eubacterial genes encoding proteins are the first nucleotide translated in the mRNA and the last nucleotide of the stop codon. The start and end sites of RNA genes are the first and last nucleotides of the processed species. Determining gene boundaries in the two *E.coli* genomes entailed close nucleotide by nucleotide inspection of entire genomic sequences by members of the sequencing teams. Many changes to gene boundaries were made on the basis of experimental evidence collated in databases, e.g. EcoGene (Table 1). Others were made on the basis of conservation of CDSs in distinct but related organisms, e.g. *Erwinia* or *Salmonella* species. Others were dictated by the W3110 sequence and the 243 corrected sites in MG1655 which changed the length of 84 open reading frames (ORFs), mostly due to frame shifting.

Table 1. Information gathered on genes of *E.coli* K-12 and their sources

Column heading	Column content	Sources of information
Feature	Type of genetic element (e.g. CDS, RNA, pseudogenes)	WP[1]
Locustag K-12	New K-12 specific gene identifier (ECK number)	WP
Gene Name K-12	Name in Demerec format	WP, CGSC[2], EcoGene[3], GenoBase[4], GenProtEC[5], Entrez[6], personal communications
Locus Name K-12	Name, including non-Demerec conforming format	WP, CGSC, EcoGene, Entrez, GenoBase, GenProtEC, personal communications
Synonyms of Locus Name	Other names of same locus	CGSC, EcoGene, Entrez, GenProtEC
Locus Tag MG1655	Identifier in MG1655 (b number)	WP, GenBank™ U00096.2, ASAP[7], EcoGene
Left nucleotide MG1655	Left boundary of gene	WP, GenBank™ U00096.2
Right nucleotide MG1655	Right boundary of gene	WP, GenBank™ U00096.2
Direction of transcription MG1655	Direction described as clockwise (+) or counterclockwise (−)	WP, GenBank™ U00096.2
Comment on gene boundary MG1655		WP
Locus Tag W3110	Identifier in W3110 (JW number)	WP, GenoBase
Left nucleotide W3110	Left boundary of gene	WP
Right nucleotide W3110	Right boundary of gene	WP
Direction of transcription W3110	Direction described as clockwise (+) or counterclockwise (−)	WP
Comment on gene boundary W3110		WP
Type of gene product	Code for class of molecule in Table 3	GenProtEC
Gene product description	Name of encoded protein, RNA or site	WP, ASAP, BLAST[8], Brenda[9], CCDB[10], coliBASE[11], EchoBASE[12], EcoCyc[13], EcoGene, Entrez, GeneMark[14], GenProtEC, Highwire[15], IUBMB[16], PORES[17], RegulonDB[18], SwissProt[19]
Comment gene product description	More detail on description and function of gene product	WP
Evidence	Basis for assignment of function, E (experimental) or C (computational prediction)	WP
Literature	Literature citations, PMID or abbreviated format if unavailable	GenProtEC, CCDB, EcoGene, PubMed[20]
Cell location	Location of gene product based on evaluation of literature and computational predictions	WP, EchoBASE, HMMTOP[21], LipoP[22], SignalP[23], TMHMM[24]
Context (genetic element)	Location of gene within a genetic element such as prophage, IS	WP, Entrez
Enzyme nomenclature	EC number	IUBMB
Cofactor		EcoCyc
Protein complex	Name of complex with component units listed	EcoCyc
Transporter classification	Superfamily assignment from Transport Classification Database	TCDB[25]
Transcription regulator family	Self explanatory	EcoCyc, RegulonDB
Proteases	Known and predicted in MEROPS database	MEROPS[26]
Signal peptide predictions		SignalP
Signal peptide cleavage sites		EcoGene
No. of transmembrane segments 1	Predicted with HMMTOP	HMMTOP
No. of transmembrane segments 2	Predicted with TMHMM	TMHMM
TM protein C-term location	Experimentally based determination of location of the C-terminal end of transmembrane proteins as in or out of the cytoplasm	Publication[27]
Transcriptional unit(s) regulated	Gene(s) transcriptionally regulated, known and predicted	EcoCyc, RegulonDB
Operons with attenuation regulation	Genes predicted to be regulated by transcriptional attenuation	Attenuator website[28]
Fused genes	Genes identified as encoding more than one function as a result of gene fusion	GenProtEC
Structure (PDB) id	Structure identifier from the Protein Data Bank	PDB[29]
COG assignment	Sequence similarity to cluster of orthologous groups	COG[30]
SCOP assignment	Sequence similarity to SCOP superfamily structural domains	Superfamily[31]
PFAM assignment	Sequence similarity to PFAM families and domains	Pfam[32]

Table 1. Continued

Column heading	Column content	Sources of information
TIGRFAM assignment	Sequence similarity to TIGRFAM protein families	TIGRFAM[33]
GO cellular component	Mapping of location prediction to GO terms (this study)	WP, GO[34]
GO cellular process	Mapping of function to GO terms	WP, MultiFun2GO[35]
GO molecular function	Mapping of function to GO terms	WP, MultiFun2GO

[1]Workshop participants.
[2]http://cgsc.biology.yale.edu (6).
[3]http://ecogene.org (7).
[4]http://ecoli.aist-nara.ac.jp/GB5/search.jsp (8).
[5]http://genprotec.mbl.edu/ (9,10).
[6]http://www.ncbi.nlm.nih.gov/Entrez (11).
[7]https://asap.ahabs.wisc.edu/annotation/php/ASAP1.htm (12).
[8]http://highwire.stanford.edu (13).
[9]http://www.ncbi.nlm.nih.gov/BLAST/ (11).
[10]http://www.brenda.uni-koeln.de (14).
[11]http://redpoll.pharmacy.ualberta.ca/CCDB/ (15).
[12]http://colibase.bham.ac.uk (16).
[13]http://www.ecoli-york.org (17).
[14]http://www.ecocyc.org (18).
[15]http://www.ebi.ac.uk/genemark/ (19).
[16]http://www.chem.qmul.ac.uk/iubmb/enzyme/ (20).
[17]http://garlic.mefos.hr/pores/ (21).
[18]http://www.cifn.unam.mx/Computational_Genomics/regulondb/ (22).
[19]http://us.expasy.org/sprot/ (23).
[20]http://www.pubmedcentral.nih.gov/ (11).
[21]http://www.enzim.hu/hmmtop (24).
[22]http://www.cbs.dtu.dk/services/LipoP/ (25).
[23]http://www.cbs.dtu.dk/services/SignalP (26).
[24]http://www.cbs.dtu.dk/services/TMHMM/ (27).
[25]http://www.tcdb.org/ (28).
[26]http://merops.sanger.ac.uk/ (29).
[27]http://www.sciencemag.org/cgi/content/full/308/5726/1321/DC1 (30).
[28]http://cmgm.stanford.edu/%7Emerino (31).
[29]http://www.rcsb.org/pdb/ (32).
[30]http://ncbi.nlm.nih.gov/COG (11).
[31]http://supfam.mrc-lmb.cam.ac.uk/SUPERFAMILY/ (33).
[32]http://www.sanger.ac.uk/Software/Pfam/ (34).
[33]http://www.tigr.org/TIGRFAMs/ (35).
[34]http://www.geneontology.org/ (36).
[35]http://geneontology.org/external2go/multifun2go (37).

Functional annotation was carried out by small groups of the Workshop participants incorporating extensive new experimental data from the literature, melding and reconciling collections of data from several sources (Table 1). When no experimental data beyond the sequence were available, these groups reached consensus after surveying predictions previously made by others with new predictions based on sequence similarity, domain content and other predictive techniques and information. Supplementary Table 1 presents 44 discrete types of information about each gene where applicable. One column indicates whether the function of a gene product is experimentally known or predicted. Workshop participants made an effort to use consistent vocabularies to indicate degrees of uncertainty about those gene products not known experimentally. Literature citations underpinning experimental information are provided so the history can be traced to its source. Most citations are given as PMID keys to the PubMed electronic abstracting service (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=PubMed). Some, especially older biochemical and genetic information, are in literature that predates abstracting services. For these, abbreviated spelled-out references are given.

RESULTS

Complete genome sequences for the two strains of *E.coli* K-12 allow comparison of the current sequence for the MG1655 genome with the original 1997 version. It also allows comparison of the gene content of two K-12 strains which have had different histories since their isolation in the early 1950s in the laboratory of Joshua and Esther Lederberg at the University of Wisconsin. Their common ancestor was an isolate of the original K-12 cured of lambda and F. MG1655 was stored most of the time before it was sequenced in 1997. Cultures were maintained variously lyophilized, frozen and on stab. In contrast W3110 would have undergone many more generations over this period of time as it was used actively for research over these years, passing from laboratory to laboratory. For more detail on these histories see Ref. (4).

Inspection of the two genomic sequences and consultation resulted in changing many designated start codons which led to elimination of some old genes and formation of some new ones. Compared with the content of GenBank™ entry U00096.2 there were 682 changes in start codon assignments of previously identified genes, 31 old genes have been

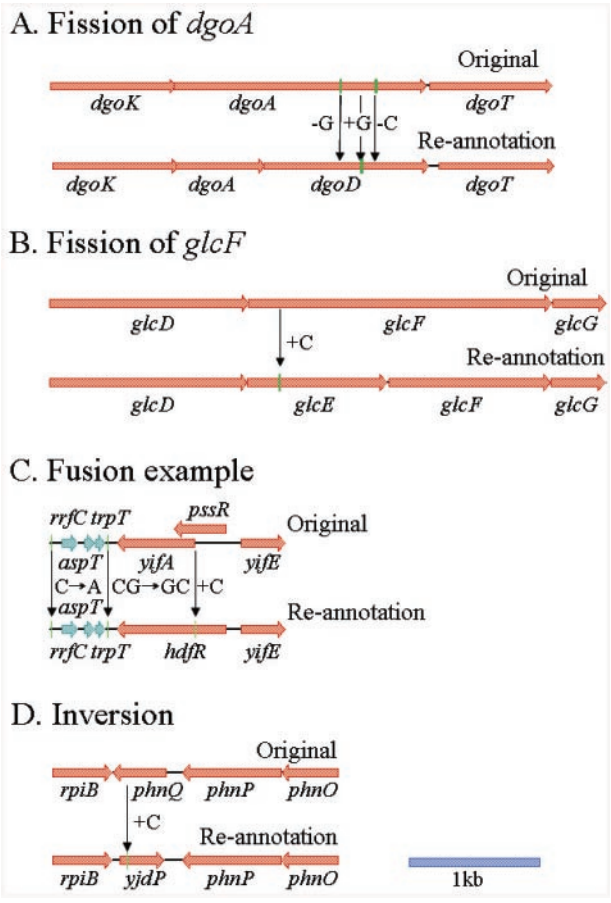

Figure 1. Gene fissions, fusions and an inversion resulting from 1 nt indel corrections. Of 78 frameshift corrections, two 1 nt indels led to fissions (splitting) of genes (**A** and **B**), 23 resulted in gene fusions, similar to the example in (**C**), and 1 led to an inversion (D) (4). (**D**) The original annotation of the *rpiB* region showed a gene called *phnQ*, whose sequence is not conserved. A 1 nt insertion created a CDS for a conserved protein (*yjdP*) in the opposite orientation. While *phnQ* was originally thought to be a downstream gene in the large phosphonate (*phn*) operon (39), mutational studies later revealed no role for it in phosphonate metabolism (40).

eliminated (Supplementary Table 6) and 66 new ones mostly of unknown function have been recognized (48 CDSs, 17 pseudogenes and 1 RNA). Even small differences in start sites affect important matters such as the design of probes for microarray experiments, quantifying distance relationships to upstream regulatory elements, and the design of primers for gene amplification and gene deletion, e.g. as used for the construction of a complete set of *E.coli* K-12 in-frame, single-gene knockout mutants (38).

Several corrections had dramatic consequences for a gene(s). In addition to changing the reading frame, 2 frameshifts led to fissions (Figure 1A and B), 23 led to fusions of adjacent or overlapping ORFs into single proteins, as shown for *hdfR* in Figure 1C, and 1 led to an inversion, i.e. recognition of a conserved protein encoded on the opposite strand (Figure 1D). Other corrections led to missense changes (62), were silent (17) or in intergenic regions (73) or RNA genes (2).

Inspecting and comparing the sequence data for both MG1655 and W3110, we can ask how the genomes of these isolates, both K-12 strains, differ. Owing to extra copies

Table 2. Genes not common to strains MG1655 and W3110

Type of gene product	W3110	MG1655
Pseudogene	6	
IS	17	2
Prophage genes		9
Total	23	11

of IS elements, there are 17 more genes for IS element proteins (more IS1, IS2 and IS5 genes) in W3110 than in MG1655. In return there are 11 genes, 9 encoding the CPZ-55 prophage and 2 encoding IS1 proteins, in MG1655 that are absent in W3110 (Table 2). As both IS elements and temperate phages are horizontally transmitted genetic elements, differences of this kind in the two *E.coli* genomes are not unexpected. The presence of more IS elements in W3110 could reflect its role as a prime experimental strain that has experienced more exposure and more generations than has MG1655.

Predicting pseudogenes requires an extremely high level of nucleotide accuracy. Pseudogenes are caused by frameshifts, in-frame stops, or insertions or deletions which divide a gene into fragments. Most of the pseudogenes are broken into two fragments (18 ancestral genes that are now 36 pseudogene fragments), a few are broken into three fragments (3 ancestral genes that are now 9 pseudogene fragments) and several exist as single fragments (41) in both strains. In addition, W3110 contains six genes with an IS insertion resulting in either split genes (four ancestral genes that are now eight pseudogenes) or truncated genes (two pseudogenes). Hence the number of pseudogenes differs between the two strains.

We do not know the full phenotypic consequences of the genetic differences between the two K-12 isolates. Functions of the four genes that are pseudogenes in W3110, split by insertion, are known. These are genes for the galactitol PTS enzyme II (GatA), aerobic and anaerobic C4-dicarboxylate transporters (DcuC), a hybrid sensory kinase (RcsC), and a low-affinity tryptophan permease in the tryptophanase operon (TnaB). Each of these may affect metabolism, for instance growth on galactitol or succinate would be affected unless redundant systems are present. The use of tryptophan as a carbon and nitrogen source may also be affected. These testable characteristics illustrate the breadth of phenotypic difference possible between isolates of one strain of a single bacterial species maintained separately for several decades.

With the updated annotation in hand, in terms of the biology of the organism we can ask how much we have learned about the *E.coli* cell in terms of the functions of its gene products. How many genes encode enzymes, how many genes encode a transporter function, regulator function or have cellular roles? Surveying the content of the two genomes that is in common (4453 genes), the numbers of gene products of different types in our snapshot are listed in Table 3.

Comparing the number of genes in Table 3 with earlier counts, we find that in 1993, before the genome sequence was known, only 1700 genes were listed (41). Upon completing the genome sequence in 1997, the number of MG1655 genes was 4289 (3), a number that is close to today's total of 4464 (for the 4453 genes in common see Table 3 and for MG1655-specific genes see Table 2). The increase is due in

Table 3. Numbers and types of known and predicted gene products of *E.coli* K-12[1]

Code	Gene product type	Number	Percentage[2]
e	Enzyme	1094	33.3
pe	Enzyme, predicted	390	
t	Transporter	337	13.3
pt	Transporter, predicted	254	
r	Regulator	241	9.1
pr	Regulator, predicted	164	
m	Membrane	43	5.7
pm	Membrane, predicted	210	
f	Factor	150	4.7
pf	Factor, predicted	60	
s	Structural component	89	2.8
ps	Structural component, predicted	37	
c	Carrier	77	2.7
pc	Carrier, predicted	42	
n	RNA	156	3.5
lp	Lipoprotein	46	1.0
cp	Cell process	56	1.3
l	Leader peptide	11	0.3
su	Pseudogenes in common	74	1.6
i	Site (*oriC*)	1	<0.1
h	Phage/IS in common (including 15 pseudogenes)	304	6.8
d	Partial information	146	3.3
o	Unknown function	471	10.6
Total		4453[1]	100.0

[1]Genes in common to strains MG1655 and W3110.
[2]The percentage is calculated from the sum of known and predicted gene types.

large part to identifying small proteins and small RNAs (42,43).

We looked at the proportions of types of molecular functions of the genes and compared these values with assessments of the same kind collected at earlier stages of knowledge of the genome. One needs to be aware that gene products can serve more than one cell role, thus choosing to identify a gene with a single category is sometimes arbitrary and can shift between assessments. In spite of this potential variability, we see that over a period of 12 years the proportion of enzymes, transporters, regulators and undefined membrane proteins has remained remarkably stable at ~33, 13, 9 and 6%, respectively. The proportion of the genome occupied by phage and IS genes also has remained steady at ~7%. Changes in other categories reflect new discoveries and/or redefinitions of a role category. The category called 'factors', although a small category, has increased in size over 10-fold from the earliest assessment because of discoveries of new factors such as transcription and translation factors and chaperones. An increase in size of another small category, 'carriers', results in large part from redefining the category 'carriers' to include specialized electron-carrying proteins and specialized electron-carrying subunits of enzymes. We drew an arbitrary line defining cytochrome and iron–sulfur proteins and subunits as 'carriers', but retaining definition of NAD(P)H-binding proteins and flavoproteins as 'enzymes' as the latter often have the catalytic site in the same polypeptide chain. Finally, numbers of known RNA genes have risen from 104 reported in 1993 through 116 reported in 2004, to 156 today. The increase in the numbers results from the identification of new 'small RNAs' many of which have regulatory function. Future experimental characterizations of the cellular functions of presently unknown genes will complete the picture of the contents and proportions of all types of macromolecules in an *E.coli* cell.

Unique identifiers

Beyond the annotation activities, a third aim was to produce a gene identification system for *E.coli* K-12 genes that is consistent between the two strains over the vast regions where they are essentially identical while also making accessible those genes that are strain specific or have different map locations. Owing to use of slightly different coordinate systems, more copies of IS elements in W3110, a defective phage only in MG1655, and the large W3110 inversion (44), there is no simple formula relating the positions of corresponding nucleotides in the two K-12 genomes. The problem being that the genomes do not have the same length, and there is a gene order reversal due an inversion. Consequently, consistent sequential numbering of sequence and features is impossible.

Our solution was to provide a tripartite system of identifiers for each annotated feature: 'b' numbers for MG1655, 'JW' numbers for W3110, and 'ECK' (*E.coli* K-12) numbers for reference to *E.coli* K-12 as a composite strain. The b and JW numbers are indexed to the nucleotide sequences of the respective genomes and ECK numbers point to the corresponding b and/or JW numbers depending on whether the gene exists in one or both genomes. In updating the MG1655 genome, we retained the original b numbers if the gene was not substantially changed. Otherwise, the original b number was permanently retired and a new number was taken from the end of the series. The JW numbers were similarly styled. We chose this approach over one that would introduce decimal extensions to existing numbers as a process more easily applied in cases of future changes. Single ECK numbers were assigned for each unique CDS of an IS element, resulting in a one to many mapping for these CDSs. We limited the 'one to many' nomenclature to mobile elements so, for example, ribosomal RNA genes are each assigned separate ECK numbers. Genes interrupted by an IS element or frameshift were given unique b and JW numbers for each gene segment and the same ECK number for all gene segments. The ECK unique identifiers are numbered sequentially in the order of the MG1655 map beginning with *thrL*.

Gene names. The *E.coli* community uses Demerec format (45) for gene names consisting of a unique three-letter abbreviation intended to suggest a function, followed by a capital letter to distinguish different genes related to the same function. 'Official' gene names are managed by the Coli Genetic Stock Center (CGSC) [(11) and Table 1)]. The 'y gene' system (46) follows a unique Demerec format with names beginning with the letter 'y' as a way to name genes of unknown function. Although intended for only temporary use until a function was unraveled, y gene names have been retained in the literature for many genes whose function is now well understood. We updated the nomenclature in two ways: (i) Mary Berlyn of the CGSC at Yale University provided new Demerec names, from the literature and personal communications, to replace y gene names for which functions have now been discovered, and has resolved conflicts and redundancies resulting from multiple name assignments made to a single gene or class of genes or the same name assigned to multiple genes, (ii) Kenn Rudd assigned y gene names for some newly delineated

genes of unknown function. In all cases, both the canonical name and synonyms are in Supplementary Table 1. For some genes, informal names that do not comply with the Demerec rules are also given as locus names. These include names for fragmented pseudogenes (each fragment named by adding on '_1', '_2' and so on, numbering from the N-terminal end of the full length protein) and multiple copies of IS proteins (each copy assigned an extension of '-1', '-2', '-3' and so on, depending on its chromosomal location).

Some genes are clearly inactivated by deletion, frameshift or IS element insertion. In an attempt to connect terminology with genetic nomenclature of eukaryotes, we refer to these as pseudogenes and pseudogene fragments. Individual fragments of divided pseudogenes are given the same ECK identifier but locus names are modified as described above. In addition to specification of the fragments, an entry under the same ECK identifier for individual fragments, provides the range of nucleotides of the entire (ancestral) pseudogene. Unique locus identifiers have only been assigned to the predicted ancestral pseudogenes in MG1655.

The output data

The main table, Supplementary Table 1, has a row for each gene or gene fragment and 44 data columns. Because this table has empty spaces where a property does not apply to a particular gene type, separate more compact tables are provided for enzymes (Supplementary Table 2), transport proteins (Supplementary Table 3), regulatory proteins (Supplementary Table 4) and the remainder (Supplementary Table 5). All five tables are provided in both spreadsheet (Microsoft Excel) and text formats. The text format offers a seldom-seen advantage in the presentation of genomic data in that the information is not presented one gene at a time, but the information can be addressed as a whole. This format lends itself to importation into relational or other database management systems and to exploration using query languages.

The information in the data columns is given in Table 1 (vide supra), which has a description of the type of information in each column and the major sources used in the annotation process. Text notes with definitions and explanations of the types of data in the table and descriptions of how they were generated are in Supplementary Document 1 Explanatory Notes. Table contents are not exhaustive. Most entries could be expanded. For instance only the coarsest granularity of terms that are available in the Gene Ontology (GO) system were applied to each gene product. Time did not permit taking proper advantage of the rich detail of the ontology. Application of fine detail awaits future work by member(s) of the *E.coli* community.

We can ask where we stand in having definite facts about every gene in the organism. Figure 2 summarizes how many gene products have functions that have been demonstrated experimentally, how many have functions that can be predicted by similarity to known genes and how many are still of unknown function. Unknown gene products were divided into those that are conserved in the sense of having similarity to the sequence of at least one other protein in current databases, and those that are not. Of the least known, there is useful information for some, such as presence of a predicted domain within the sequence. Only 5.3% of *E.coli* K-12 genes remain totally unknown without even a predicted domain, no clue to their identity or function at this time. The larger category of unknowns having some information about them constitutes an additional 8.6%. These CDSs require proper characterization to learn the identity and function of their gene products. It seems likely the number of genes of unknown function of various kinds will continue to fall as experimental findings continue to accumulate in the future.

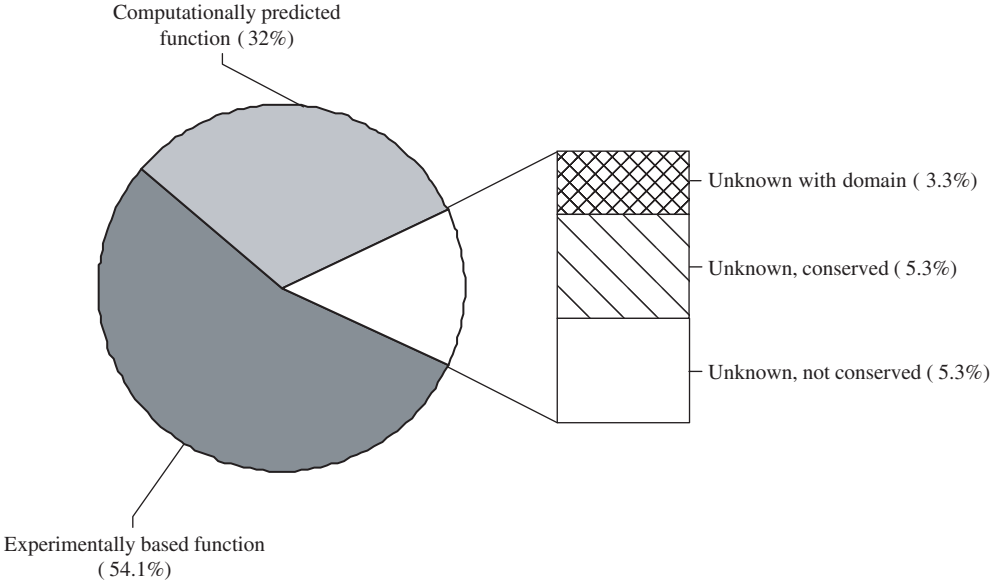

Figure 2. Status of annotation of *E.coli* gene products. The total number of gene products present in both MG1655 and W3110, 4452 excluding oriC, are categorized according to their function assignment. Evidence code and gene type assignments available in the Supplementary Table 1 were used to group the gene products. The annotation groups include gene products whose function is experimentally determined (2403, 54.1%), predicted by computational analysis (1425, 32%), or unknown (616, 13.9%). The gene products of unknown function are further separated into those containing a conserved domain (145, 3.3%), those with (233, 5.3%) or without (238, 5.3%) a detectable homolog in the sequence databases.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

Invaluable technical assistance was provided by Daniella Wilmot (Marine Biological Laboratory). A special thank you to Drs Heladia Salgado Osorio and Julio Collado-Vides (UNAM) for providing information on transcriptional regulators. Also, thank you to Dr Richard Horler (University of York) for data analysis and Jie Shao (Purdue University) for making Figure 1. Thank you to Dr Edward A. Adelberg for providing up-to-date *E.coli* references. We gratefully acknowledge conference grant support from the US National Institute of General Medical Sciences, 1 R13 GM74562-01. Workshop participants are supported by a variety of government and private funding agencies in their home countries, including the British Biotechnology and Biological Sciences Research Council (R.R.C. and G.H.T.), CREST, Japan Science and Technology and New Energy and Industrial Technology Development Organization (T.H, H.M.), Genome Canada (D.W.), Ministry of Education, Culture, Sports, Science and Technology of Japan (T.A., T.K.), The Wellcome Trust (N.R.T.), US Department of Energy (M.R., M.H.S), US National Institutes of Health (M.B.A., F.R.B., J.D.G., I.M.K., N.T.P., G.P.,III, K.E.R., B.L.W.) and the US National Science Foundation (M.K.B.). The Open Access publication charges for this article were waived by Oxford University Press.

Conflict of interest statement. None declared.

REFERENCES

1. Lederberg,J. and Tatum,E.L. (1946) Gene recombination in *Escherichia coli*. *Nature*, **158**, 558.
2. Lennox,E.S. (1955) Transduction of linked genetic characters of the host by bacteriophage P1. *Virology*, **1**, 190–206.
3. Blattner,F.R., Plunkett,G.III, Bloch,C.A., Perna,N.T., Burland,V., Riley,M., Collado-Vides,J., Glasner,J.D., Rode,C.K., Mayhew,G.F. *et al*. (1997) The complete genome sequence of *Escherichia coli* K-12. *Science*, **277**, 1453–1462.
4. Hayashi,K., Morooka,N., Yamamoto,Y., Choi,S., Ohtsubo,E., Baba,T., Wanner,B.L., Mori,H. and Horiuchi,T. (2006) Highly accurate genome sequences of the *Escherichia coli* K-12 strains MG1655 and W3110. *Mol. Syst. Biol.* (in press).
5. Bachmann,B.J. (1996) Derivations and genotypes of some mutant derivatives of *Escherichia coli* K-12.. In Neidhardt,F.C., Curtiss,R.III, Ingraham,J.L., Lin,E.C.C., Low,K.B.Jr, Magasanik,B., Reznikoff,W.S., Riley,M., Schaechter,M. and Umbarger,H.E. (eds), *Escherichia coli and Salmonella typhimurium Cellular and Molecular Biology*. ASM Press, Washington, D.C., pp. 2460–2488.
6. Berlyn,M.K.B. (1996) Accessing *E.coli* genetic stock center database. In Neidhardt,F.C., Curtiss,R.III, Ingraham,J.L., Lin,E.C.C., Low,K.B., Magasanik,B., Reznikoff,W.S., Riley,M., Schaechter,M. and Umbarger,H.E. (eds), *Escherichia coli and Salmonella Cellular and Molecular Biology*. ASM Press, Washington, DC, pp. 2489–2495.
7. Rudd,K.E. (2000) EcoGene: a genome sequence database for *Escherichia coli* K-12. *Nucleic Acids Res.*, **28**, 60–64.
8. Mori,H., Isono,K., Horiuchi,T. and Miki,T. (2000) Functional genomics of *Escherichia coli* in Japan. *Res. Microbiol.*, **151**, 121–128.
9. Serres,M.H., Goswami,S. and Riley,M. (2004) GenProtEC: an updated and improved analysis of functions of *Escherichia coli* K-12 proteins. *Nucleic Acids Res.*, **32**, D300–D302.
10. Serres,M.H. and Riley,M. (2005) Gene fusions and gene duplications: relevance to genomic annotation and functional analysis. *BMC Genomics*, **6**, 33.
11. Wheeler,D.L., Barrett,T., Benson,D.A., Bryant,S.H., Canese,K., Church,D.M., DiCuccio,M., Edgar,R., Federhen,S., Helmberg,W. *et al*. (2005) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **33**, D39–D45.
12. Glasner,J.D., Liss,P., Plunkett,G.III, Darling,A., Prasad,T., Rusch,M., Byrnes,A., Gilson,M., Biehl,B., Blattner,F.R. *et al*. (2003) ASAP, a systematic annotation package for community analysis of genomes. *Nucleic Acids Res.*, **31**, 147–151.
13. HighWire press is 5 years old. *J. Biol. Chem.*, **275**, 13165.
14. Schomburg,I., Chang,A., Ebeling,C., Gremse,M., Heldt,C., Huhn,G. and Schomburg,D. (2004) BRENDA, the enzyme database: updates and major new developments. *Nucleic Acids Res.*, **32**, D431–D433.
15. Sundararaj,S., Guo,A., Habibi-Nazhad,B., Rouani,M., Stothard,P., Ellison,M. and Wishart,D.S. (2004) The CyberCell Database (CCDB): a comprehensive, self-updating, relational database to coordinate and facilitate *in silico* modeling of *Escherichia coli*. *Nucleic Acids Res.*, **32**, D293–D295.
16. Chaudhuri,R.R., Khan,A.M. and Pallen,M.J. (2004) coliBASE: an online database for *Escherichia coli*, *Shigella* and *Salmonella* comparative genomics. *Nucleic Acids Res.*, **32**, D296–D299.
17. Misra,R.V., Horler,R.S., Reindl,W., Goryanin,I.I. and Thomas,G.H. (2005) EchoBASE: an integrated post-genomic database for *Escherichia coli*. *Nucleic Acids Res.*, **33**, D329–D333.
18. Keseler,I.M., Collado-Vides,J., Gama-Castro,S., Ingraham,J., Paley,S., Paulsen,I.T., Peralta-Gil,M. and Karp,P.D. (2005) EcoCyc: a comprehensive database resource for *Escherichia coli*. *Nucleic Acids Res.*, **33**, D334–D337.
19. Besemer,J. and Borodovsky,M. (2005) GeneMark: web software for gene finding in prokaryotes, eukaryotes and viruses. *Nucleic Acids Res.*, **33**, W451–W454.
20. Osborn,M. (2005) IUBMB: the first half-century. *Trends Biochem. Sci.*, **30**, 273–275.
21. Juretic,D., Zoranic,L. and Zucic,D. (2002) Basic charge clusters and predictions of membrane protein topology. *J. Chem. Inf. Comput. Sci.*, **42**, 620–632.
22. Salgado,H., Gama-Castro,S., Martinez-Antonio,A., Diaz-Peredo,E., Sanchez-Solano,F., Peralta-Gil,M., Garcia-Alonso,D., Jimenez-Jacinto,V., Santos-Zavaleta,A., Bonavides-Martinez,C. *et al*. (2004) RegulonDB (version 4.0): transcriptional regulation, operon organization and growth conditions in *Escherichia coli* K-12. *Nucleic Acids Res.*, **32**, D303–D306.
23. Boeckmann,B., Bairoch,A., Apweiler,R., Blatter,M.C., Estreicher,A., Gasteiger,E., Martin,M.J., Michoud,K., O'Donovan,C., Phan,I. *et al*. (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.*, **31**, 365–370.
24. Tusnady,G.E. and Simon,I. (2001) The HMMTOP transmembrane topology prediction server. *Bioinformatics*, **17**, 849–850.
25. Juncker,A.S., Willenbrock,H., Von,H.G., Brunak,S., Nielsen,H. and Krogh,A. (2003) Prediction of lipoprotein signal peptides in Gram-negative bacteria. *Protein Sci.*, **12**, 1652–1662.
26. Bendtsen,J.D., Nielsen,H., Von Heijne,G. and Brunak,S. (2004) Improved prediction of signal peptides: SignalP 3.0. *J. Mol. Biol.*, **340**, 783–795.
27. Krogh,A., Larsson,B., Von,H.G. and Sonnhammer,E.L. (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.*, **305**, 567–580.
28. Saier,M.H., Hvorup,R.N. and Barabote,R.D. (2005) Evolution of the bacterial phosphotransferase system: from carriers and enzymes to group translocators. *Biochem. Soc. Trans.*, **33**, 220–224.
29. Rawlings,N.D., Tolle,D.P. and Barrett,A.J. (2004) MEROPS: the peptidase database. *Nucleic Acids Res.*, **32**, D160–D164.
30. Daley,D.O., Rapp,M., Granseth,E., Melen,K., Drew,D. and Von,H.G. (2005) Global topology analysis of the *Escherichia coli* inner membrane proteome. *Science*, **308**, 1321–1323.
31. Merino,E. and Yanofsky,C. (2005) Transcription attenuation: a highly conserved regulatory strategy used by bacteria. *Trends Genet.*, **21**, 260–264.
32. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.

33. Madera,M., Vogel,C., Kummerfeld,S.K., Chothia,C. and Gough,J. (2004) The SUPERFAMILY database in 2004: additions and improvements. *Nucleic Acids Res.*, **32**, D235–D239.
34. Bateman,A., Coin,L., Durbin,R., Finn,R.D., Hollich,V., Griffiths-Jones,S., Khanna,A., Marshall,M., Moxon,S., Sonnhammer,E.L. *et al.* (2004) The Pfam protein families database. *Nucleic Acids Res.*, **32**, D138–D141.
35. Haft,D.H., Selengut,J.D. and White,O. (2003) The TIGRFAMs database of protein families. *Nucleic Acids Res.*, **31**, 371–373.
36. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genet.*, **25**, 25–29.
37. Serres,M.H. and Riley,M. (2000) MultiFun, a multifunctional classification scheme for *Escherichia coli* K-12 gene products. *Microb. Comp. Genomics*, **5**, 205–222.
38. Baba,T., Ara,T., Okumura,Y., Hasegawa,M., Takai,Y., Okumura,Y., Baba,M., Datsenko,K.A., Tomita,M., Wanner,B.L. *et al.* (2006) Construction of *Escherichia coli* K-12 in-frame, single-gene knock-out mutants—the Keio collection. *Mol. Syst. Biol.* (in press).
39. Chen,C.-M., Ye,Q., Zhu,Z., Wanner,B.L. and Walsh,C.T. (1990) Molecular biology of carbon-phosphorus bond cleavage: cloning and sequencing of the *phn* (*psiD*) genes involved in alkylphosphonate uptake and C-P lyase activity in *Escherichia coli* B. *J. Biol. Chem.*, **265**, 4461–4471.
40. Metcalf,W.W. and Wanner,B.L. (1993) Mutational analysis of an *Escherichia coli* fourteen-gene operon for phosphonate degradation using Tn*phoA′* elements. *J. Bacteriol.*, **175**, 3430–3442.
41. Riley,M. (1993) Functions of the gene products of *Escherichia coli*. *Microbiol. Rev.*, **57**, 862–952.
42. Wagner,E.G.H. and Vogel,J. (2003) Noncoding RNAs encoded by bacterial chromosomes. In Barciszewski,J. and Erdmann,V. (eds), *Noncoding RNAs*. Landes Bioscience, Georgetown, pp. 243–259.
43. Gottesman,S. (2004) The small RNA regulators of *Escherichia coli*: roles and mechanisms. *Annu. Rev. Microbiol.*, **58**, 303–328.
44. Hill,C.W. and Harnish,B.W. (1981) Inversions between ribosomal RNA genes of *Escherichia coli*. *Proc. Natl Acad. Sci. USA*, **78**, 7069–7072.
45. Demerec,M., Adelberg,E.A., Clark,A.J. and Hartman,P.E. (1966) A proposal for a uniform nomenclature in bacterial genetics. *Genetics*, **54**, 61–76.
46. Rudd,K.E. (1998) Linkage map of *Escherichia coli* K-12, edition 10: The physical map. *Microbiol. Mol. Biol. Rev.*, **62**, 985–1019.