Analysis of the *Escherichia coli* Genome: DNA Sequence of the Region from 84.5 to 86.5 Minutes

Donna L. Daniels, Guy Plunkett III, Valerie Burland, Frederick R. Blattner

The DNA sequence of 91.4 kilobases of the *Escherichia coli* K-12 genome, spanning the region between *rrnC* at 84.5 minutes and *rrnA* at 86.5 minutes on the genetic map (85 to 87 percent on the physical map), is described. Analysis of this sequence identified 82 potential coding regions (open reading frames) covering 84 percent of the sequenced interval. The arrangement of these open reading frames, together with the consensus promoter sequences and terminator-like sequences found by computer searches, made it possible to assign them to proposed transcriptional units. More than half the open reading frames correlated with known genes or functions suggested by similarity to other sequences. Those remaining encode still unidentified proteins. The sequenced region also contains several RNA genes and two types of repeated sequence elements were found. Intergenic regions include three "gray holes," 0.6 to 0.8 kilobases, with no recognizable functions.

Complete genomic sequences, including those of viruses, plasmids, organelles, and recently one of the smaller yeast chromosomes (*1*), have provided valuable insights even though none of these encode all the functions required for life. The complete sequence of the *Escherichia coli* genome, however, would in principle contain information sufficient to define an independent life form. The segment of the *E. coli* K-12 genome described in this article is an initial step toward such an analysis (*2*).

The circular *E. coli* genome of 4.7×10^3 kb corresponds to a genetic map of 100 minutes (*3*). The physical map has been defined by the overlapping set of bacteriophage lambda (λ) clones of Kohara *et al.* (*4*), and maps are now available for ten restriction endonuclease sites (*4–6*). The genetic map includes more than 1400 genes, many of which have been placed on the physical map (*2*, *7*). Approximately one-third of the genome has been sequenced in patches by individual investigators in studies addressing specific genes of interest, many regions having been sequenced several times in different strains and mutants. Available sequences have been collated by Rudd (*7*) and by Kröger (*8*) and assembled into sequenced regions, whose average size is 3.8 kb and the largest is 32.3 kb. Although more than 1200 *E. coli* genes have been sequenced, perhaps another 2000 remain to be discovered.

We have chosen the *E. coli* K-12 strain MG1655 to represent the wild type for sequencing (*9*). It was derived from the original K-12 (isolated in 1934) by curing it of lambda prophage and F factor without treatment by mutagens. Other common laboratory strains of *E. coli* have all been obtained by mutagenic methods, making them unsuitable for sequencing. The strain used by Kohara *et al.* (*4*), W3110, has several known point mutations, a large inversion, several transpositions, and many deviations from the wild-type restriction map (*5–7*). From MG1655 we constructed an overlapping lambda clone bank (*6*, *10*) similar to the Kohara set. Nine lambda clones covering about 100 kb (2 percent of the genome) were sequenced while technical approaches for mass sequencing and data analysis were being developed.

For sequencing of *E. coli* within 3 to 5 years, rates of sequencing and analysis in excess of 10^3 kb per year are required. A low rate of insertion and deletion errors is essential, since one of the main objectives is the identification of ORF's (open reading frames) that code for proteins. We have developed a process with three overlapping stages—production, finishing, and analysis. The production stage includes automation of most steps in the Sanger dideoxy method (*11*). Libraries were constructed from lambda clones in an M13 vector. The M13 library clones were grown in microtiter dishes, single-stranded DNA template was isolated in a parallel process, sequence reactions incorporating internal ^{35}S were performed robotically by a pipetting machine, electrophoresis through large format gels was used to resolve sequence, and autoradiograph films were scanned photoelectrically into computers where individual sequences were merged into overlapping contiguous segments (the assembly process). The production stage was effected by a relatively small team of technicians aided by student workers. At this point examination of the sequence data was limited to quality control checks. Ambiguities, where several determinations of an individual nucleotide (nt) differed (*12*), averaged about 1 per 100 nt.

A second team, working with computer assistance, conducted the finishing stage (*13*). Human editing of the computer-generated alignments reduced ambiguities to about 1 in 200 nt and the autoradiograph lanes where data required proofreading were identified. Deferral of proofreading until after initial assembly saved time and reduced costs. In regions where data remained ambiguous, the finishing team requested additional data, which could involve special treatments, from the data production team. Next, a computer-aided examination for ORF's, codon usage frequencies, and similarities to database entries was used to further refine the sequence. A translated frame could often be distinguished by its codon distribution or by similarity of its predicted amino acid sequence to a known protein. The sequence was scrutinized for potential insertion or deletion mistakes where the preferred translation frame shifts inexplicably or where the match to a protein in a database required a frameshift in our sequence. Caution was required since many database entries were themselves in error, and in some cases the authentic sequence contained a mutation. At this stage an attempt was made to resolve differences between our sequence and those reported by others; sometimes those individuals who had reported conflicting data were consulted. Altogether, finishing reduced the rate of ambiguities to about 1 per 600 nt.

During the analysis, the annotations for submission to the EMBL and GenBank sequence databases were prepared. These included the ORF's, identified where possible with known genes or proposed functions, and proposed transcriptional units. Features of interest such as repetitive elements were also noted. Genes or functions were identified by comparisons with Bachmann's genetic map (*3*) and the databases, and occasionally by correlating known restriction maps with sequence-based predictions. Proposing the start points of genes was particularly difficult; we normally chose the ATG or GTG codon farthest upstream. A collection of amino-terminal sequences of *E. coli* proteins compiled by Church and Link (*14*) proved useful in this regard.

Transcription units were suggested by the arrangement of genes. To locate promoters of transcriptional units, a matrix search derived from in vitro measurements of Moyle *et al.* (*15*) was used to obtain consensus sequence matches which were

The authors are in the Laboratory of Genetics, University of Wisconsin, 445 Henry Mall, Madison, WI 53706.

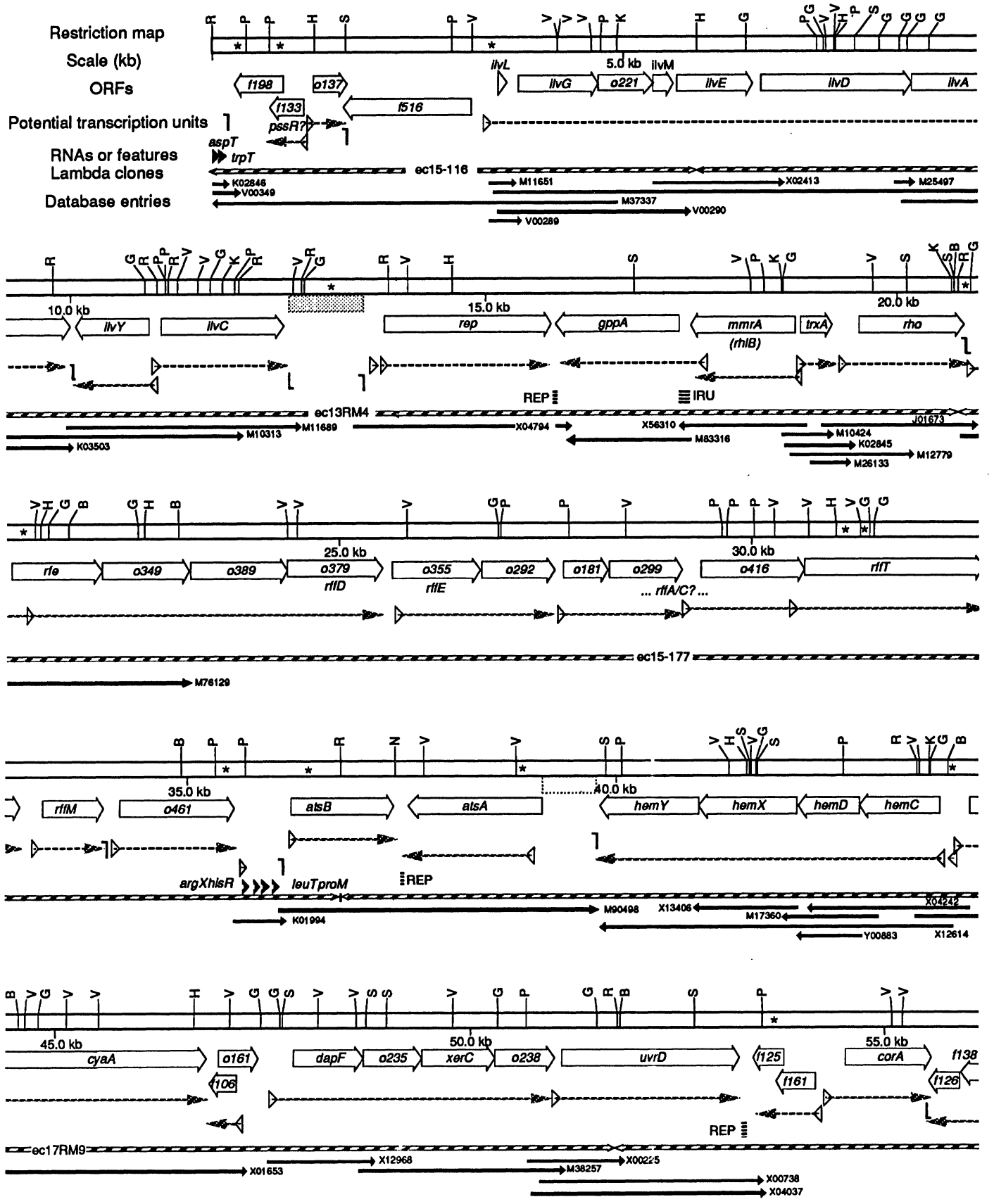
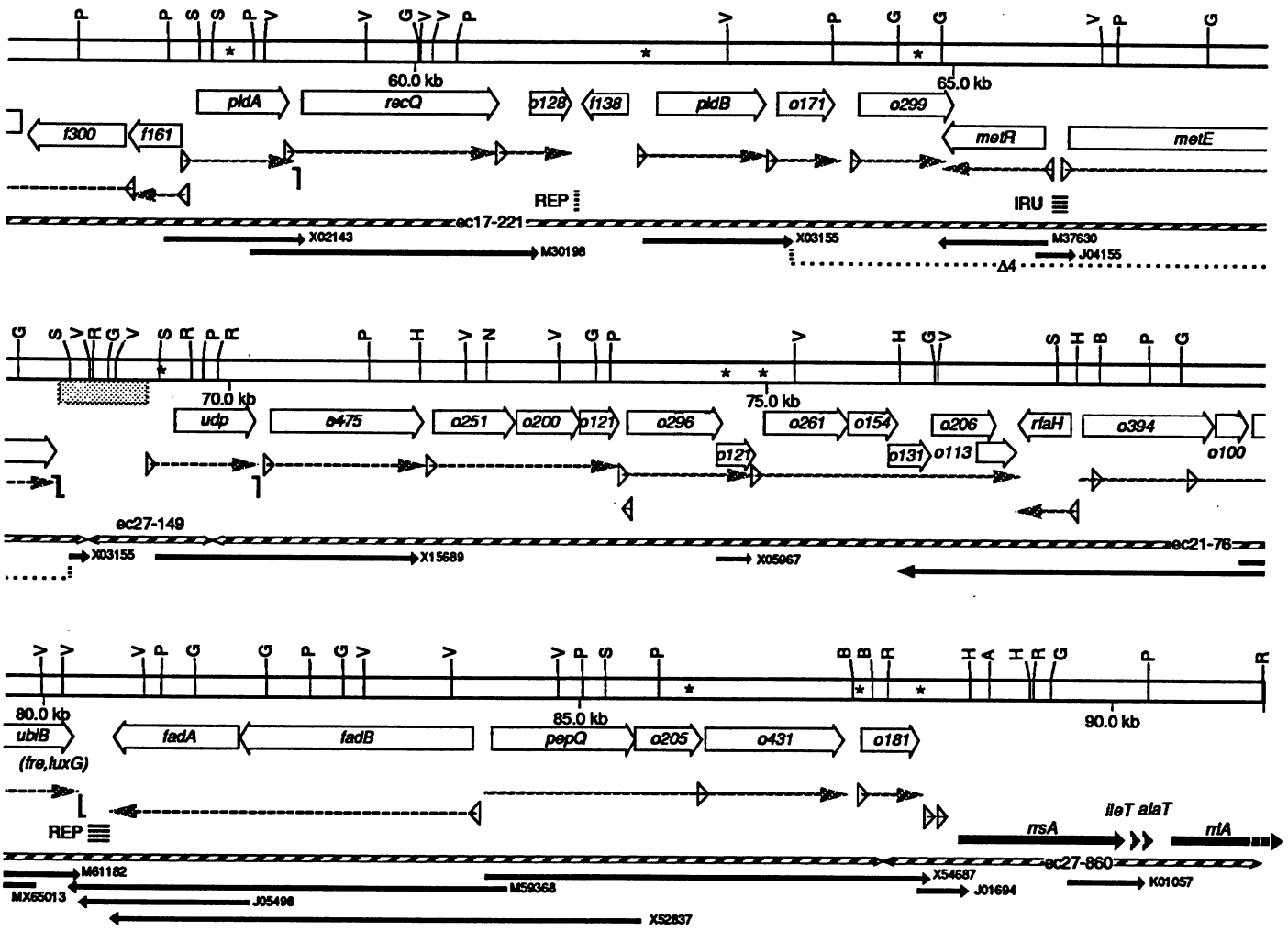

Restriction map
Scale (kb)
ORFs
Potential transcription units
RNAs or features
Lambda clones
Database entries
5.0 kb
ilvL
ilvM
f198
o137
ilvG
o221
ilvE
ilvD
ilvA
f133
f516
pssR?
aspT
trpT
ec15-116
K02646
V00349
M11651
X02413
M25497
M37337
V00290
V00289
10.0 kb
15.0 kb
20.0 kb
ilvY
ilvC
rep
gppA
mmrA
(rhlB)
trxA
rho
REP
IRU
ec13RM4
M11689
M10313
K03503
X04794
X56310
M83316
J01673
M10424
K02845
M12779
M26133
25.0 kb
30.0 kb
rfe
o349
o389
o379
rfID
o355
rfIE
o292
o181
o299
... rfIA/C? ...
o416
rfIT
ec15-177
M76129
35.0 kb
40.0 kb
rfIM
o461
atsB
atsA
hemY
hemX
hemD
hemC
argXhisR
leuTproM
REP
K01994
M90498
X13406
M17360
X04242
Y00883
X12614
45.0 kb
50.0 kb
55.0 kb
cyaA
o161
dapF
o235
xerC
o238
uvrD
f125
corA
f138
f106
f161
f126
REP
ec17RM9
X01653
X12968
M38257
X00225
X00738
X04037

Fig. 1. (*continued*).



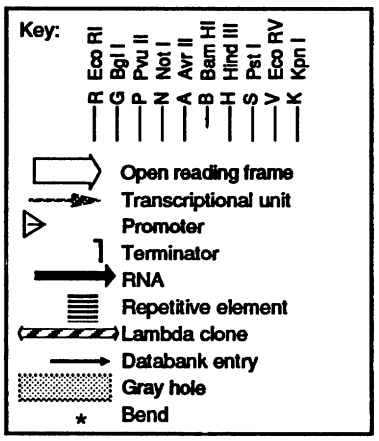


Fig 1. Map of the finished sequence and its features, proceeding left to right in eight tiers. Top line shows restriction sites for ten enzymes and calculated bend sites (*). Large open arrows, ORF's identified with the gene name if known; unidentified ORF's are designated *o* or *f* for the two transcriptional orientations with the numbers indicating the predicted peptide length. Tentative identifications are indicated beneath by a question mark. Dashed arrows represent proposed transcriptional units. Filled thick arrows, RNA genes; horizontal stripes, repeated sequence elements; shaded boxes, gene-free regions (gray holes); hatched lines, lambda clones; filled thin arrows, database (GenBank-EMBL) entries with accession number. Promoters are denoted by open triangles in the appropriate orientation, terminators by L (inverted for the reverse orientation). The dotted line below the features of 65 kb indicates a deletion in GenBank entry X03155. The sequence derived from the map is in GenBank as M87049, in which all of these features and the conflicts with other sequences are annotated.

scored from 0 to 100. More than 6100 promoters were found that might function in vitro. This list was pruned by a four-step algorithm that considered both promoter score and position relative to ORF's; it removed weak promoters near strong ones on the same strand, identified the strongest one or two promoters at the beginning of each proposed gene regardless of actual strength, eliminated all but the strongest promoters within proposed genes on either strand, and removed weak promoters in gene-free regions. The pruning algorithm produced about 200 promoter candidates that were examined individually by eye, resulting in 54 proposed promoters ranging in score from 39 to 88. A search for Rho-independent terminators (*16*) located appropriately spaced inverted repeats followed by a succession of T residues.

Static bend sites were located by calculating the trajectory of the 91-kb DNA segment (*17*) and then scoring the angular deviation of the path over 100-nt segments. We noted 23 sites where the predicted deviation from straight exceeded 72 degrees.

The restriction map and features of the sequence are shown in Fig. 1. Features include RNA genes, repeats, bend sites, and gray holes as well as the ORF's, pro-

moters and transcription terminators that make up the proposed transcription units. A list of the genes and their assigned names (*18*) and the peptides coded and their physical characteristics was compiled (Table 1). The 82 ORF's cover 84.2 percent of the sequence. Of these, 34 are genes whose functions have yet to be determined. Some of these are probably the ORF's for the 12 genes listed in Table 2, which map to this region by genetic experiments, but are not yet sufficiently characterized to correlate with ORF's. Eight transfer RNA (tRNA) genes accounted for 0.7 percent and ribosomal RNA (rRNA) genes accounted for 2.6 percent of the sequence. Intergenic regions represented 12.5 percent.

Sequence features. The following five items can be considered as the map is examined.

1) The sequence begins with tRNA genes *aspT* and *trpT* at the 3′ end of *rrnC*. The tRNA genes *argX*, *hisR*, *leuT*, and *proM* are found between sequence coordinates 35486 and 35922, and at the end of the sequence are genes for the 16S rRNA, two spacer tRNA's, and part of the 23S rRNA of the *rrnA* locus.

2) The repeated sequences include repetitive extragenic palindromic (REP) elements (not to be confused with the *rep* gene); these are scattered throughout the *E. coli* genome and are made up of consensus inverted repeats called REP sequences. Several functions have been attributed to them, such as a role as structural units involved in chromosome architecture, and as DNA gyrase binding sites (*19*). There are REP elements in five locations in this sequence (Fig. 1), composed of one to six REP sequences each. In all cases, adjacent genes are transcribed in opposite directions converging on an REP element. The *ubiB* and *fadBA* transcription units converge on an REP element consisting of six copies of the REP sequence; this is the most complex REP element in this sequence. In these cases, the REP elements may function as transcription terminators.

The genomes of various enterobacteria contain a family of repeated DNA sequence elements known as either IRU's (intergenic repeat units) or ERIC's (enterobacterial repetitive intergenic consensuses) (*20*). There are two IRU's in the sequence that we describe, and both have been noted previously; they are in the promoter region between *metR* and *metE* (*20*) and between *mmrA* and *gppA* (*21*). The function of this sequence element is unknown.

3) There are 23 bend sites, with predicted deviations of more than 72 degrees per 100 nt. Of these, 15 are found in coding regions, often near one end. Although this analysis suggests that the sites are frequent in *E. coli*, their significance is unknown. One bend was previously documented: near the ORF *o121* (Fig. 1) is a GenBank-EMBL entry X05967 containing the sequence "bent19," isolated in an experiment designed to identify fragments with static bends. It is actually two separate DNA fragments accidentally joined during cloning. The first fragment is from about 95 minutes on the *E. coli* map, while the second fragment is located within the region described here. Our computer analysis shows that this fragment and not the one at 95 minutes contains a static bend of 72 degrees.

4) There are three areas in this sequence, ranging in size from 0.6 to 0.8 kb and covering 2.3 percent of the sequence, which contain no potential coding regions of more than 100 amino acids (aa). Smaller ORF's in these regions might actually encode small proteins (Ilv M is only 87 aa). However, these sequences may have no function, or they may be pseudogenes or gene remnants (*22*). Other identifiable

Table 1. The 92 genes in the described sequence, and predicted characteristics of encoded proteins. Gene names are those used in the most recent *E. coli* genetic map (*3*); alternative names and tentative assignments are listed in parentheses. In addition, mapped genes have been assigned "identifiers" (CGSC site numbers) in a database maintained by the *E. coli* Genetic Stock Center (*18*). The sequences of 37 of the genes have the following GenBank-EMBL accession numbers (when a gene has been sequenced more than once, only a single database entry is

Gene name	CGSC No.	Endpoints	First...last codon	Molecular size (kD)	Molecular size (aa)	pI	Ref.
aspT	989	10 > 86		tRNA			
trpT	66	95 > 170		tRNA			
f198		862 < 269	GTG...TAA	22.4	198	6.1	
f133 (pssR?)	18010	1104 < 706	GTG...TAA	15.3	133	10.0	(25)
o137		1223 > 1633	ATG...TAG	16.3	137	10.0	
f516		3135 < 1588	ATG...TAA	56.2	516	7.3	(27)
ilvL		3458 > 3553	ATG...TAG	3.2	32	11.0	
ilvG	603	3696 > 4676	ATG...TGA	34.5	327	5.1	
o221		4675 > 5337	...TGA	24.8	221	5.5	
ilvM	18214	5337 > 5597	ATG...TGA	9.7	87	8.6	
ilvE	605	5620 > 6546	ATG...TAA	34.2	309	5.8	
ilvD	606	6645 > 8459	ATG...TAA	64.4	605	5.3	
ilvA	609	8465 > 10006	ATG...TAG	56.2	514	5.7	
ilvY	598	10954 < 10064	GTG...TGA	33.2	297	6.7	
ilvC	607	11104 > 12576	ATG...TAA	54.1	491	5.2	
rep	303	13803 > 15821	ATG...TAA	76.9	673	6.9	(30)
gppA	664	17355 < 15871	ATG...TAA	54.9	494	6.3	
mmrA (rhlB)	18151	18756 < 17494	ATG...TAA	47.1	421	7.3	(33)
trxA	65	18833 > 19213	ATG...TAA	14.0	127	5.2	
rho	288	19543 > 20800	ATG...TAA	47.0	419	7.1	
rfe	294	21042 > 22142	GTG...TAA	40.9	367	10.0	(38)
o349		22154 > 23200	GTG...TAG	39.5	349	5.7	
o389		23217 > 24383	GTG...TGA	43.5	389	6.4	
o379 (rffD?)		24383 > 25519	ATG...TGA	41.5	379	5.5	(39)
o355 (rffE?)		25643 > 26707	ATG...TAA	39.7	355	5.8	(40)
o292		26729 > 27604	ATG...TGA	32.7	292	5.3	(41)
o181 (rffA or rffC?)		27714 > 28256	GTG...TGA	19.6	181	8.8	
o299 (rffA or rffC?)		28264 > 29160	ATG...TGA	33.2	299	7.8	(42)
o416 (rffA or rffC?)		29392 > 30639	ATG...TGA	44.9	416	9.9	
o716 (rffT)		30639 > 32786	ATG...TGA	82.0	716	9.6	
o246 (rffM)		33067 > 33804	ATG...TGA	27.8	246	9.5	
o461		34098 > 35380	ATG...TAA	50.6	461	9.6	(43)
argX	17749	35486 > 35562		tRNA			
hisR	625	35620 > 35696		tRNA			
leuT	565	35717 > 35803		tRNA			
proM	17626	35846 > 35922		tRNA			
atsB (aslB)		36069 > 37304	ATG...TAG	46.6	411	6.8	(45)
atsA (aslA; gppB)		39118 < 37463	ATG...TAA	60.7	551	6.2	(45)
hemY		40991 < 39798	ATG...TAG	45.2	398	8.5	
hemX		42175 < 40997	ATG...TAA	43.0	393	4.5	
hemD	645	42937 < 42200	ATG...TAA	27.8	246	6.2	
hemC	646	43896 < 42937	ATG...TGA	34.6	320	5.3	
cyaA	902	44262 > 46805	TTG...TGA	97.6	848	6.0	
f106		47168 < 46851	ATG...TAA	12.2	106	4.1	
o161 (cyaX)		46959 > 47441	ATG...TAA	17.3	161	7.7	(48)
dapF	17713	47868 > 48692	ATG...TGA	30.4	275	6.1	

functions have not been detected, at least by the available methods with our current criteria. The term "gray hole" is coined for such areas, not previously described in bacteria, although a similar feature is found in phage lambda (*23*).

5) Identified genes and transcription units are described in order from left to right along the map in Fig. 1. Only those coding regions with unexpected or specific points of interest are discussed. Their locations are listed in Table 1.

The first potential protein coding region is a leftward transcription unit that contains two ORF's. Of these, the predicted *f133* product shows similarity to the LysR family of bacterial regulatory proteins (*24*); with *E. coli* IlvY, OxyR, CysB, and LysR, scores range from 26 to 41 percent (*25*). At 133 aa, this candidate peptide is smaller than other members of the LysR family. *f198* just downstream and overlapping has no promoter and may be part of *f133*. This could be explained by an undetected error in our sequence, or by a mutation, as in *ilv*G (discussed below), or by a frame-shifting codon. The LysR family proteins activate transcription of other genes, and in most cases one activated gene is transcribed divergently from the same promoter region (for example, *ilvY* and *ilv*C; *metR* and *metE*).

The next ORF, *o137*, is thus a candidate for such a regulated gene. We suggest that *f133* may be *pssR*, a regulatory gene of phosphatidylserine synthetase, which has a role in biosynthesis of a minor membrane phospholipid and maps to this area (*26*). In that case *o137* may serve some function in phospholipid synthesis.

The ORF *f516* shows weak similarity to a subunit of magnesium chelatase from *Rhodobacter capsulatus* (*27*). However, Coppola *et al.* (*28*) did not detect the product of *f516* (their ORF III) in maxicells and suggest a smaller in-frame peptide (their ORF I).

The well-studied *ilv* locus (3,458 to 12,576) codes for seven proteins, IlvG, D, M, E, A, Y, and C plus a leader peptide, IlvL. The K-12 strains carry a frameshift mutation in *ilv*G, resulting in premature termination of the protein (*29*). *o221* is the "missing piece" eliminated from the carboxy terminus of *ilv*G. *ilv*Y, encoding the regulatory protein, and *ilv*C are transcribed divergently.

The genes *rep*, *gppA*, *mmrA*, *trx*A, and *rho* are interspersed with two repetitive DNA elements, REP and IRU. The first gene codes for the *rep* helicase. Our data differ from the reported sequence by a single nucleotide insertion, lengthening the protein at the -COOH end and extending its previously noted similarity with *uvrD* (*30*). We identified *gppA* from the restriction map in Kalman *et al.* and from complementation data (*21*). This gene codes for guanosine pentaphosphatase, part of the stringent response system of regulation. The next ORF is identified as *mmrA* (*31*), which maps near *rep* and functions in postreplication repair.

The peptide sequence shares similarities with members of the DEAD (Asp-Glu-Ala-Asp) family of proteins that exhibit RNA-dependent adenosine triphosphatase (ATPase) and, in some cases, helicase activity (*32*, *33*). The appropriateness of these characteristics for a role in postreplication repair, along with the absence of other candidates in the region, is the basis for this gene assignment. The same ORF (termed *rhlB*) was also sequenced by Kalman *et al.* (*21*) in a specific search for DEAD-family genes. The gene *trx*A (*34*) codes for thioredoxin and is transcribed divergently from the same promoter region as *mmrA*; the *rho* gene codes for the transcription termination factor Rho (*35*).

Table 1 (*continued*).
noted): *aspT, trpT* [K02846]; *ilvL, ilvG, ilvM, ilvE, ilvD, ilvA* [M10313]; *ilvY, ilvC* [M11689]; *rep* [X04794]; *gppA* [M83316]; *rhlB* [X56310]; *trxA* [M12779]; *rho* [J01673]; *rfe* [M76129]; *argX, hisR, leuT, proM* [K01994]; *aslB, aslA* [M90498]; *hemD, hemC* [X12614]; *cyaA* [X01653]; *dapF* [X12968]; *xerC [M38257]; uvrD* [X00738]; *pldA* [X02143]; *recQ* [M30198]; *pldB* [X03155]; *metR* [M37630]; *udp* [X15689]; *fre* [M61182]; *fadA, fadB* [M59368]; *pepQ* [X54687].

Gene name	CGSC No.	Endpoints	First...last codon	Molecular size (kD)	(aa)	pI	Ref.
o235		48692 > 49396	ATG...TGA	26.7	235	6.3	
xerC		49396 > 50289	ATG...TAA	33.8	298	9.4	
o238		50292 > 51005	ATG...TAA	27.1	238	6.0	
uvrD	18	51092 > 53251	ATG...TAA	82.0	720	6.0	
f125		53778 < 53404	GTG...TAA	13.9	125	9.6	
f161		54164 < 53682	ATG...TGA	17.8	161	8.5	
corA	911	54533 > 55564	ATG...TAG	39.9	344	4.9	(50)
f126		55905 < 55528	ATG...TGA	14.6	126	9.6	
f138		56335 < 55922	ATG...TAG	15.8	138	8.6	
f300		57296 < 56397	ATG...TAA	33.7	300	9.6	
f161		57820 < 57338	ATG...TGA	17.9	161	6.8	
pldA	384	57967 > 58833	ATG...TGA	33.2	289	5.1	
recQ	17959	58963 > 60792	GTG...TAG	68.3	610	6.8	
o128		61091 > 61474	GTG...TGA	13.9	128	8.7	
f138		61955 < 61542	GTG...TGA	15.3	138	10.9	
pldB	5001	62268 > 63287	ATG...TAA	38.7	340	7.1	
o171		63410 > 63922	GTG...TGA	19.4	171	4.9	
o299		64158 > 65054	GTG...TAA	33.7	299	9.7	
metR	18163	65898 < 64948	ATG...TAA	35.7	317	8.1	
metE	512	66135 > 68393	ATG...TAA	85.1	753	6.6	(53)
udp	41	69498 > 70256	ATG...TAA	27.2	253	6.0	
o475		70400 > 71824	GTG...TAG	54.7	475	5.2	
o251		71922 > 72674	ATG...TGA	28.1	251	8.0	
o200		72691 > 73290	ATG...TGA	22.2	200	8.0	
o121		73290 > 73652	ATG...TGA	14.0	121	10.2	
o296		73737 > 74624	GTG...TGA	34.4	296	7.9	
o121		74561 > 74923	ATG...TGA	13.9	121	9.7	
o261		75009 > 75791	GTG...TAA	27.7	261	5.2	
o154		75797 > 76258	ATG...TAA	17.5	154	9.7	
o131		76177 > 76569	GTG...TAA	14.5	131	4.4	
o206		76602 > 77219	ATG...TGA	22.7	206	4.7	
o113		77056 > 77394	GTG...TAG	12.9	113	10.5	
rfaH (sfrB)	164	77882 < 77397	ATG...TAA	18.3	162	8.4	
o394		78049 > 79230	ATG...TAA	43.9	394	6.1	
o100		79239 > 79538	GTG...TGA	11.2	100	4.2	
ubiB (fre; luxG)	49	79587 > 80285	ATG...TGA	26.3	233	5.4	(56)
fadA	794	81830 < 80670	ATG...TAA	40.9	387	6.5	
fadB	793	84029 < 81843	ATG...TAA	79.6	729	6.0	
pepQ		84219 > 85547	ATG...TGA	50.2	443	5.8	(58)
o205		85547 > 86161	GTG...TAA	21.9	205	5.4	(60)
o431		86203 > 87495	ATG...TGA	47.7	431	9.8	(61)
o181		87662 > 88204	GTG...TAA	21.2	181	9.7	
rrsA	189	88585 > 90126		rRNA			
ileT	612	90195 > 90271		tRNA			
aldT	1038	90314 > 90389		tRNA			
rrlA	203	90564 >>91408		rRNA (partial)			

Table 2. These genes have been genetically mapped to this general region (*3*) but have not been correlated with particular ORF's. For example, *kdsB* (CGSC No. 18202; CTP:CMP-3-deoxy-D-mannooctulosonate cytidylyltransferase) was mapped at 85 minutes and has been sequenced (*62*); however, the sequence does not match any of our data, and examination of the restriction map derived from the data indicates that *kdsB* actually maps at about 20.5 minutes [(*7*); our unpublished observations].

Gene	CGSC No.	Phenotypic trait affected
prlB	18034	Protein export
ridB	17935	Transcription and translation; rifampicin dependence
bfm	966	Phage BF23 multiplication
bioP (birB)	953	Biotin transport
fcsA	787	Cell division; septation
fexB	778	*fexA* phenotype affected
fipB	18334	Morphogenesis of phage F1
hemG	642	Protoporphyrinogen oxidase activity
tabC	125	Development of phage T4
chlB	921	Nitrate reductase; biosynthesis of molybdopterin
ubiD	47	3-Octaprenyl-4-hydroxybenzonate → 2-octaprenylphenol
ubiE	46	2-Octaprenyl-6-methoxy-1,4-benzoquinone → 2-octaprenyl-3-methyl-6-methoxy-1,4-benzoquinone

Region 21,040 to 33,870 contains 11 ORF's, all transcribed rightward, and several transcriptional units are proposed within which the genes are arranged head to tail with very little intergenic space. This area contains the two loci, *rfe* and *rff*, which participate in the synthesis of the enterobacterial common antigen (ECA) (*36*). ECA consists of multiple repeats of a three-sugar polysaccharide (*N*-acetyl-D-glucosamine, *N*-acetyl-D-mannosaminuronic acid, and 4-acetamido-4,6-dideoxy-D-galactose) linked to a glycerophosphatidyl residue anchoring it to the membrane. The first ORF in this region has been sequenced (*37*) and identified as *rfe*; it shows a short similarity to the peptidoglycan biosynthetic enzyme phospho-*N*-acetylmuramyl-pentapeptide transferase, encoded by the *E. coli mraY* (or *murX*) gene (*38*). Some of the succeeding ORF's can be tentatively assigned to *E. coli rff* genes on the basis of restriction mapping and complementation data (*37*) and similarities of the predicted proteins with known sugar processing enzymes.

The product of *o379* displays similarity to the GDP-mannose dehydrogenase (AlgD) of *Pseudomonas aeruginosa* (*39*); it is probably *rffD*, which encodes UDP-*N*-acetyl-D-mannosaminuronic acid dehydrogenase. The *o355* sequence encodes a protein similar to UDP glucose 4-epimerase (GalE) of *Streptomyces lividans*, galactose transferase (GAL10) of *Kluyveromyces lactis*, and CDP-2–tyvelose epimerase (RfbE) of *Salmonella typhimurium* (*40*); it is probably *rffE*, which encodes UDP-*N*-acetyl-D-glucosamine-2-epimerase. The product of *o292* is similar to a putative sugar-activating enzyme (StrD) of *Streptomyces griseus* (*41*). Mapping data (*37*) indicate that two of the following three ORF's, *o181*, *o299*, and *o416*, correspond to *rrfA* and *rrfC*. The *o299* product is similar to the pleiotropic regulatory protein DegT of *Bacillus stearothermophilus* and the erythromycin biosynthesis enzyme EryC1 of *Saccharopolyspora erythaea* (*42*). Finally, *o716* and *o246* are assigned to genes *rffT* and *rffM*.

The peptide of *o461* is similar to amino acid transport proteins of *E. coli* (AroP), *S. cerevisiae* (PUT4), and *Emericella nidulans* (proline transport protein) and is probably a previously unidentified amino acid transport gene (*43*). The biotin transport gene *bioP* genetically maps to this area (*3*).

Arylsulfatase activity is reportedly absent in *E. coli*, but it produces a protein that cross-reacts with antiserum to *Klebsiella aerogenes* AtsA and that is regulated similarly (*44*). Two of our ORF sequences match the structural and regulator genes (*atsA* and *atsB*) of *K. aerogenes* very well (*45*). Thus *atsA*, for the defective arylsulfatase of *E. coli*, and the regulator, *atsB*, have been assigned. Although *atsA* probably codes for defective arylsulfatase, genetic evidence suggests another function (*45*). In *E. coli*, the *ats* genes are on convergent transcripts, and an REP element is located between them; the putative *atsA* promoter matches the consensus for operons that are under the overall control of *flhD* and *flhC*, involved in flagellar synthesis or chemotaxis (*46*). In contrast, the *K. aerogenes* genes are encoded by a single *atsB-atsA* transcript, which is followed by a Rho-independent terminator.

The *hem* operon contains two uncharacterized ORF's, *hemY* and *hemX*, and two genes, *hemD* and *hemC*, which code for uroporphyrinogen III synthase and porphobilinogen deaminase. Urogen III methylase may be coded by *hemX* (*47*). Following *cyaA*, two potential ORF's overlap in opposite directions. Although this is common because of the structure of the genetic code, it would be unusual if both strands were actually used to code functional proteins. Previously, *o161* has been reported as "*cyaX*" (*48*). On the complementary strand, however, *f106* has an extremely high scoring promoter consensus and so is more likely to be a gene.

Two well-studied genes, *dapF* for diaminopimelate epimerase and *xer*C, required for site-specific recombination of ColE1, share a proposed transcription unit with two unassigned ORF's, *o235* and *o238*. The gene for DNA helicase II, *uvrD*, follows with its own transcription unit followed by an REP element consisting of two copies of the repeat sequence. The *corA* sequence encodes a membrane-associated protein with a role in Mg^{2+} and Co^{2+} ion transport (*49*) and maps near *uvrD*. It was identified by comparison with the *corA* sequence of *Salmonella typhimurium* (*50*).

The reported sequence of *pldB*, encoding lysophospholipase L2, is followed by a spontaneous deletion of 4.94 kb spanning the *metR-metE* region, termed *pldBΔ4* (*51*). The divergently transcribed genes coding for a regulator, *metR*, and the gene regulated, *metE* [encoding tetrahydro-pteroyltriglutamate methyltransferase (*3*)] were identified from the published sequence of *metR* and the beginning of *metE* (*52*). An otherwise uncharacterized yeast sequence may be the yeast homolog of *metE* (*53*). The 7.7-kb segment starting at *udp*, which codes for uridine phosphorylase (*3*), is transcribed rightward and contains 11 ORF's in six transcriptional units. There is one high-scoring promoter consensus in the opposite direction, but it is not associated with an ORF. The gene *rfaH* (*sfrB*), a regulator of the *tra* operon of F plasmids and the *rfa* genes for lipopolysaccharide synthesis (*3*), was assigned by comparison of our data on the restriction map and the predicted peptide size with that of others (*7*, *54*).

The *fre* gene, coding for flavin reductase, has been sequenced by Spyrou *et al.* (*55*), who cloned the gene with a probe designed from the amino acid sequence of the enzyme. We believe that this gene is identical with *ubiB*, defined by a mutation that blocks ubiquinone synthesis. Similarities were found to the xylene monooxygenase subunit of *Pseudomonas putida* and a subunit of a methane monooxygenase from *Methylococcus capsulatus* (*56*). Most striking, however, are the similarities to LuxG, encoded by an uncharacterized ORF in the *lux* operons of all marine strains of luminescent bacteria (*57*). Similarity indices range from 38 to 41 percent when the *ubiB* protein is compared to LuxG proteins from *Vibrio harveyi*, *Vibrio fischeri*, and *Photobacterium phosphoreum* (*56*). In addition, the sequences from these species are as similar to each other as they are to that from *E. coli*. These similarities indicate that *lux*G

encodes the flavin reductase of the bioluminescence pathway; this flavin reductase had not previously been correlated with any gene.

Our sequence for *pepQ*, encoding proline peptidase, differs from that reported (*58*) by an insertion of 5 nt, resulting in a shorter predicted protein that nonetheless includes all the motifs for the proline peptidase family, and is closer in size to other proline peptidases such as human proline dipeptidase (*59*). The "extra" ORF *o205* created by this difference is similar to a *B. subtilis* ORF of unknown function (*60*). The promoter search failed to detect a promoter for *pepQ*. The only other ORF for which a promoter was not found was *recQ*, where a poor candidate far upstream was separated from the gene by a terminator.

Genome sequencing strategy. In the course of this work we designed a data production system that can be used with many sequencing strategies, varying from completely random (sequencing randomly chosen clones from a library) to completely directed (sequencing specific clones chosen to cover known positions in the genome). Purely random strategies require collection of sequence data from many clones, while purely directed strategies instead require mapping or screening many clones prior to sequencing only a small number. The most efficient combination of random and directed strategies depends on the technologies available. We developed a variation of the combined strategy made possible by construction of the M13 cloning vector Janus. This vector contains elements at which an inducible site-specific recombination system acts to invert the orientation of the cloned insert with respect to the sequencing primer site ("flipping"). Thus, either strand of an insert can be sequenced efficiently as single-stranded DNA. Janus was used as the vector for DNA libraries constructed from lambda clones, and random clones were sequenced. The sequences were assembled (aligned), and candidates for flipping were identified near areas requiring improvement such as (i) less than fourfold coverage, (ii) data from one strand only, or (iii) data at the ends of contiguous segments. Sequence obtained from the opposite ends of the inserts by flipping was then added to the assembly. Finally, persistent poor areas or small gaps were sequenced by primers designed from adjacent determined sequences ("walking").

Optimum efficiency should be achieved by choosing the correct point to switch from the random phase to flipping so that a minimum number of walking steps is needed. Analysis of our data suggests that sixfold redundancy in the random phase is appropriate, followed by flipping at onefold, after which this sequence segment needed only one walking step for closure (more recent data confirm that on average, fewer than five walking steps per 200 kb are required). The critical advantage of this strategy compared to directed methods is that, in random sequencing, successful finishing does not depend on the success of any individual sequence reaction but on the total quantity of data obtained.

To assure accuracy in the finished sequence, our goal is that each nucleotide be included in at least four determinations and at least once on each strand. The average depth of coverage in this segment was 9.2. More than 95 percent of the sequence was determined at least four times, and 90 percent was sequenced at least once on both strands. A weighting feature in the assembly software reduced the frequency of errors in that it emphasized the portion of the gel where the data were clearest. We began sequencing in a well-studied region in order to develop technical approaches and to assess sequence accuracy. The whole of the region was sequenced in this work; previously reported data were used only for comparison. Where more than one database entry existed, the sequences differed from each other and from our determination by about 1 residue per 600, a value agreeing well with our internal estimate of 1 error per 500 residues.

The search programs used in the analysis phase were effective but required extensive editing by the scientists to reduce the computer "finds" to a reasonable list. The activity of promoters, terminators, and other features depends on undetermined biological factors as well as on the DNA sequence, such as binding proteins or transcription-translation interactions, which have not yet been defined by algorithms. The pruning regime used in the promoter search approached the human editing function in this regard, but much more experimental data are needed before sequence analysis can be automated or, indeed, before the analyzed genome can be fully interpreted.

REFERENCES AND NOTES

1. Examples include bacteriophage lambda, F. R. Sanger, A. R. Coulson, G. F. Hong, D. F. Hill, G. B. Petersen, *J. Mol. Biol.* **162**, 729 (1982); bacteriophage ϕX174, F. Sanger *et al.*, *Nature* **265**, 687 (1977); bacteriophage T7, J. J. Dunn and F. W. Studier, *J. Mol. Biol.* **148**, 303 (1981); Epstein-Barr virus, R. Baer *et al.*, *Nature* **310**, 207 (1984); pBR322, J. G. Sutcliffe, *Cold Spring Harbor Symp. Quant. Biol.* **43**, 77 (1978); rice chloroplast, J. Hiratsuka *et al.*, *Mol. Gen. Genet.* **217**, 185 (1989); and yeast chromosome III, S. G. Oliver *et al.*, *Nature* **357**, 38 (1992).
2. The goal of the *E. coli* Genome Project at the Laboratory of Genetics, University of Wisconsin–Madison, is to determine the complete DNA sequence of *E. coli* K-12 strain MG1655, while developing technology for sequencing and analysis.
3. B. J. Bachmann, in *Escherichia coli and Salmonella typhimurium: Cellular and Molecular Biology*, F. C. Neidhardt *et al.*, Eds. (American Society for Microbiology, Washington, DC, 1987), pp. 807–876; *Microbiol. Rev.* **54**, 130 (1990).
4. Y. Kohara, K. Akiyama, K. Isono, *Cell* **50**, 495 (1987).
5. C. L. Smith, J. G. Econome, A. Schutt, S. Klco, C. R. Cantor, *Science* **236**, 1448 (1987).
6. D. L. Daniels, *Nucleic Acids Res.* **18**, 2649 (1990); in *The Bacterial Chromosome*, K. Drlica and M. Riley, Eds. (American Society for Microbiology, Washington, DC, 1990), pp. 43–52.
7. K. E. Rudd *et al.*, *Nucleic Acids Res.* **19**, 637 (1991); K. E. Rudd, in *A Short Course in Bacterial Genetics: A Laboratory Manual and Handbook for Escherichia coli and Related Bacteria*, J. H. Miller, Ed. (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, 1992), pp. 2.3–2.43.
8. M. Kröger, *Nucleic Acids Res.* **17** (suppl.), 283 (1989); ———, R. Wahl, P. Rice, *ibid.* **19** (suppl.), 2023 (1991).
9. B. J. Bachmann, in *Escherichia coli and Salmonella typhimurium: Cellular and Molecular Biology*, F. C. Neidhart *et al.*, Eds. (American Society for Microbiology, Washington, DC, 1987), pp. 1190–1219.
10. D. L. Daniels, *Nature* **325**, 831 (1987).
11. M13 DNA preparation, C. H. Olson, F. R. Blattner, D. L. Daniels, *Methods* **3**, 27 (1991); electrophoresis, D. L. Daniels, L. Marr, R. L. Brumley, F. R. Blattner, in *Structure & Methods, Volume I: Human Genome Initiative & DNA Recombination*, R. H. Sarma and M. H. Sarma, Eds. (Adenine, Guilderland, NY, 1990), pp. 29–35. The film scanner was a prototype developed in collaboration with DNASTAR Inc. The pipetting robot was a modified Gilson 212B liquid handler.
12. Ambiguities are nucleotides of uncertain identity symbolized according to the International Union of Biochemistry nucleotide code, recognized by the databases.
13. For sequence assembly we used the *SeqMan* programs (DNASTAR). Codon usage statistics [M. Gribskov, J. Devereux, R. R. Burgess, *Nucleic Acids Res.* **12**, 539 (1984); R. Staden and A. D. McLachlan, *ibid.* **10**, 141 (1982)] were graphically displayed by DNASTAR Geneplot program to aid gene identification and to locate frameshift errors. The FIND-IT program [J. Shavlik, *Technical Report 988*, University of Wisconsin Computer Science Department (1990); S. Henikoff, J. C. Wallace, J. P. Brown, *Methods Enzymol.* **183**, 111 (1990)] was used for BLAST searches against the protein databases; it detected matches even when there were phaseshifting errors in the DNA sequence. Promoter searching was done by the Patterns program (DNASTAR), with pruning by a routine written for the purpose in this laboratory. Translation, molecular size, and isoelectric point calculations and FASTA database searches [D. J. Lipman and W. R. Pearson, *Science* **227**, 1435 (1985)] were done with DNASTAR packages. Align for the Macintosh (DNASTAR) was used for protein sequence alignments; "similarity index" is a calculation of the ratio of percent identity to the length of the alignment and was corrected for gaps introduced in the alignment.
14. G. Church and A. Link, personal communication.
15. H. Moyle, C. Waldburger, M. M. Susskind, *J. Bacteriol.* **173**, 1944 (1991).
16. R. Hagstrom, D. Joerg, R. Overbeek, M. Price, personal communication.
17. S. Levene and D. M. Crothers, *J. Biomol. Struct. Dynam.* **1**, 429 (1983).
18. The *E. coli* Genetic Stock Center site numbers assigned to the genes: M. Berlyn, personal communication.
19. Y. Yang and G. F.-L. Ames, in *The Bacterial Chromosome*, K. Drlica and M. Riley, Eds. (American Society for Microbiology, Washington, DC, 1990), pp. 211–225.
20. G. J. Sharples and R. G. Lloyd, *Nucleic Acids Res.* **18**, 6503 (1990); C. S. J. Hulton, C. F. Higgins, P. M. Sharp, *Mol. Microbiol.* **5**, 825 (1991).
21. M. Kalman, H. Murphy, M. Cashel, *New Biol.* **3**, 886 (1991); M. Cashel, personal communication.
22. E. V. Koonin has suggested that, for example, the sequence between 12660 and 12920 is a pseudogene related to the 33-kD lipoprotein of *B. subtilis* (personal communication).
23. D. L. Daniels, F. Sanger, A. R. Coulson, *Cold Spring Harbor Symp. Quant. Biol.* **47**, 1009 (1983).

24. S. Henikoff, G. W. Haughn, J. M. Calvo, J. C. Wallace, *Proc. Natl. Acad. Sci. U.S.A.* **85**, 6602 (1988).
25. The predicted product of *f133* (*pssR*) displays the following similarities to the LysR family of regulatory proteins: 42.0 percent over 69 aa to an *E. coli* hypothetical regulatory protein [T. J. Goss and P. Datta, *Mol. Gen. Genet.* **201**, 308 (1985)]; 40.5 percent over 74 aa to *E. coli* IlvY [R. C. Wek and G. W. Hatfield, *J. Biol. Chem.* **261**, 2441 (1986)]; 37.3 percent over 67 aa to *E. coli* OxyR [M. F. Christman, G. Storz, B. N. Ames, *Proc. Natl. Acad. Sci. U.S.A.* **86**, 3484 (1989); M. Boelker and R. Kahmann, *EMBO J.* **8**, 2403 (1989)]; 33.3 percent over 89 aa to *Pseudomonas putida* CatR [R. K. Rothmel *et al.*, *J. Bacteriol.* **172**, 922 (1990)]; 29.2 percent over 70 aa to both *S. typhimurium* and *E. coli* CysB [J. Ostrowski, G. Jagura-Burdzy, N. M. Kredich, *J. Biol. Chem.* **262**, 5999 (1987)]; and 26.3 percent over 75 aa to *E. coli* LysR [P. Stragier and J. C. Patte, *J. Mol. Biol.* **168**, 333 (1983)].
26. C. P. Sparrow and C. R. H. Raetz, *J. Biol. Chem.* **258**, 9963 (1983).
27. The predicted product of *f516* displays 19.7 percent similarity over 329 aa with the *Rhodofacter capsulatus* 38-kD subunit of Mg chelatase (*bchI*) [GenBank Z11165].
28. G. Coppola *et al.*, *Gene* **97**, 21 (1991).
29. R. P. Lawther *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **78**, 922 (1981).
30. The reported sequence of the Rep helicase [C. A. Gilchrist and D. T. Denhardt, *Nucleic Acids Res.* **15**, 465 (1987)] displays 37.1 percent similarity over 621 aa with *E. coli* helicase II (UvrD) [Y. Yamamoto *et al.*, *J. Biochem.* **99**, 1579 (1986)]; our predicted sequence displays 37.3 percent similarity over 637 aa.
31. R. C. Sharma, N. J. Sargentini, K. C. Smith, *J. Bacteriol.* **154**, 743 (1983); R. C. Sharma and K. C. Smith, *Mutation Res.* **184**, 23 (1987).
32. D. A. Wassarman and J. A. Steitz, *Nature* **349**, 463 (1991).
33. The predicted product of *mmrA* (*rhlB*) displays the following similarities to the DEAD family of helicases: 36.6 percent over 328 aa to *E. coli* DbpA [R. Iggo, S. Picksley, J. Southgate, J. McPheat, D. P. Lane, *Nucleic Acids Res.* **18**, 5413 (1990)]; 36.4 percent over 384 aa to *E. coli* DeaD [W. M. Toone, K. E. Rudd, J. D. Friesen, *J. Bacteriol.* **173**, 3291 (1991)]; 35.8 percent over 417 aa to *E. coli* SrmB [K. Nishi, F. Morel-Deville, J. W. B. Hershey, T. Leighton, J. Schnier, *Nature* **336**, 496 (1988); _____, Erratum, *ibid.* **340**, 246 (1989)]; 34.6 percent over 365 aa to *Mus musculus* PL10 protein [P. Leroy, P. Alzari, D. Sassoon, D. Wolgemuth, M. Fellous, *Cell* **57**, 549 (1989)]; 34.6 percent over 357 aa to *Drosophila melanogaster* RNA helicase [D. R. Dorer, A. C. Christensen, D. H. Johnson, *Nucleic Acids Res.* **18**, 5489 (1990)]; 33.7 percent over 351 aa to *D. melanogaster* vasa protein [P. F. Lasko and M. Ashburner, *Nature* **335**, 611 (1988); B. Hay, L. Y. Jan, Y. N. Jan, *Cell* **55**, 577 (1988)]; 33.2 percent over 414 aa to *Homo sapiens* p68 protein [P. Hloch, G. Schiedner, H. Stahl, *Nucleic Acids Res.* **18**, 3045 (1990)]; 32.1 percent over 383 to *S. cerevisiae* translation initiation factor [P. F. Linder and P. P. Slonimski, *ibid.* **16**, 10359 (1988)]; 32.1 percent over 384 aa to *Mus musculus* initiation factor EIF-4A-I [P. J. Nielsen, G. K. McMaster, H. Trachsel, *ibid.* **13**, 6867 (1985)]; 28.2 percent over 399 aa to *S. cerevisiae* RNA helicase MSS116 [B. Seraphin, M. Simon, A. Boulet, G. Faye, *Nature* **337**, 84 (1989)].
34. Y. Matsumoto, K. Shigesada, M. Hirano, M. Imai, *J. Bacteriol.* **166**, 945 (1986).
35. J. L. Pinkham and T. Platt, *Nucleic Acids Res.* 11, 3531 (1983).
36. H. Nikaido and M. Vaara, in *Escherichia coli and Salmonella typhimurium: Cellular and Molecular Biology*, F. C. Neidhardt *et al.*, Eds. (American Society for Microbiology, Washington, DC, 1987), pp. 7–22.
37. M. Ohta *et al.*, *Mol. Microbiol.* **5**, 1853 (1991); U. Meier-Dieter, K. Barr, R. Starman, L. Hatch, P. D. Rick, *J. Biol. Chem.* **267**, 746 (1992).
38. The predicted Rfe protein displays a short internal match (29.6 percent similarity over 54 aa) with *E. coli* phospho-*N*-acetylmuramoylpentapeptide transferase [M. Ikeda, M. Wachi, F. Ishino, M. Matsuhashi, *Nucleic Acids Res.* **18**, 1058 (1990)].
39. The predicted product of *o379* displays 25.2 percent similarity over 355 aa with the *Pseudomonas aeruginosa* AlgD protein [V. Deretic, J. F. Gill, A. M. Chakrabarty, *ibid.* **15**, 4567 (1987)].
40. The predicted product of *o355* displays similarities to several proteins: 30.0 percent over 192 aa to *Corynebacterium diphtheriae* orf3, downstream of toxin repressor (*dtxR*) gene [J. Boyd, M. N. Oza, J. R. Murphy, unpublished (1990), GenBank-EMBL entry M34239]; 29.4 percent over 171 aa to *Kluyveromyces lactis* galactose transferase (GAL10) [T. D. Webster and R. C. Dickson, *Nucleic Acids Res.* **16**, 8192 (1988)]; 28.7 percent over 232 aa to *Streptomyces lividans* UDP glucose 4-epimerase (*galE*) [C. W. Adams, J. A. Fornwald, F. J. Schmidt, M. Rosenberg, M. E. Brawner, *J. Bacteriol.* **170**, 203 (1988)]; and 25.3 percent over 155 aa to *S. typhimurium* CDP-2-tyvelose epimerase (*rfbE*) [N. Verma and P. Reeves, *J. Bacteriol.* 171, 5694 (1989)]. In addition, a shorter NH_2-terminal similarity is shared with several proteins: 39.5 percent over 42 aa to *S. typhimurium* CDP-2-tyvelose epimerase (non-contiguous with the longer similarity mentioned above); 33.3 percent over 48 aa to *E. coli* ADP-L-glycero-D-mannoheptose-6-epimerase (*rfaD*) [J. C. Pegues, L. Chen, A. W. Gordon, L. Ding, W. G. Coleman, Jr., *J. Bacteriol.* **172**, 4652 (1990)]; and 57.9 percent over 19 aa to *S. typhimurium* paratose synthase (*rfbS*) [N. Verma and P. Reeves, *ibid.* **171**, 5694 (1989)].
41. The predicted product of *o292* displays 34.9 percent similarity over 244 aa with a putative sugar activating enzyme, encoded by the *Steptomyces griseus strD* gene [J. Distler *et al.*, *Nucleic Acids Res.* **15**, 8041 (1987)].
42. The predicted product of *o299* displays 38.0 percent similarity over 202 aa with *B. stearothermophilus* pleiotropic regulatory protein gene (*degT*) [T. Takagi, H. Takada, T. Imanaka, *J. Bacteriol.* **172**, 411 (1990)] and 34.7 percent similarity over 221 aa with the *Saccharopolyspora erythraea (Streptomyces erythraeus) eryCI* protein [N. Dhillon, R. S. Hale, J. Cortes, P. F. Leadlay, *Mol. Microbiol.* **3**, 1405 (1989)].
43. The predicted product of *o461* displays the following similarities to amino acid transport proteins: 38.0 percent over 443 aa to *E. coli* aromatic amino acid transport protein (*aroP*) [N. Honore and S. T. Cole, *Nucleic Acids Res.* **18**, 653 (1990)]; 30.5 percent over 467 aa to *S. cerevisiae* arginine permease (CAN1) [W. Hoffmann, *J. Biol. Chem.* **260**, 11831 (1985)]; 28.3 percent over 474 aa to *S. cerevisiae* proline-specific permease (PUT4) [M. Vandenbol, J.-C. Jauniaux, M. Grenson, *Gene* **83**, 153 (1989)]; and 24.9 percent over 475 aa to *A. nidulans* proline transport protein [V. Sophianopoulou and C. Scazzocchio, *Mol. Microbiol.* **3**, 705 (1989)]. There is at least one instance of a transport protein being initially identified on the basis of sequence similarity and predicted properties [W. Seol and A. J. Shatkin, *Proc. Natl. Acad. Sci. U.S.A.* **88, 3802 (1991)].**
44. T. Yamada, Y. Murooka, T. Harada, *J. Bacteriol.* **133**, 536 (1978).
45. The predicted product of *atsB* displays 37.8 percent similarity over 384 aa with the *Klebsiella pneumoniae* putative regulatory protein AtsB [Y. Murooka *et al.*, *J. Bacteriol.* **172**, 2131 (1990)]. The predicted product of *atsA* displays 30.8 percent similarity over 131 aa with the *K. pneumoniae* arylsulfatase (*atsA*) [Y. Murooka, as above] and 27.7 percent similarity over 345 aa with the *Homo sapiens* arylsulfatase A precursor [C. Stein *et al.*, *J. Biol. Chem.* **264**, 1252 (1989)]. Genetic data from M. Cashel, personal communication.
46. D. H. Bartlett, B. B. Frantz, P. Matsumura, *J. Bacteriol.* **170**, 1575 (1988).
47. A. Sasarman, Y. Echelard, J. Letowski, D. Tardif, M. Drolet, *ibid.*, p. 1575.
48. H. Aiba *et al.*, *Nucleic Acids Res.* **12**, 9427 (1984).
49. M. H. Park, B. B. Wong, J. E. Lusk, *J. Bacteriol.* **126**, 1096 (1976).
50. M. Maguire, personal communication.
51. T. Kobayashi *et al.*, *J. Biochem.* **98**, 1017 (1985).
52. M. E. Maxon *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **86**, 85 (1989); M. E. Maxon, J. Wigboldus, N. Brot, H. Weissbach, *ibid.* **87**, 7076 (1990).
53. The predicted sequence of MetE displays 41.9 percent similarity over 490 aa to an uncharacterized sequence from yeast [*Saccharomyces cerevisiae* promoter fragment delta P8; Y. Ohtake *et al.*, *Agric. Biol. Chem.* **52**, 2753 (1988)].
54. A. Rehemtulla, S. K. Kadam, K. E. Sanderson, *J. Bacteriol.* **166**, 651 (1986).
55. G. Spyrou *et al.*, *ibid.* **173**, 3673 (1991).
56. The predicted sequence of UbiB displays the following similarities: 23.9 percent over 202 aa with *Methylococcus capsulatus* methane monooxygenase component C [A. C. Stainthorpe, V. Lees, G. P. C. Salmond, H. Dalton, J. C. Murrell, *Gene* **91**, 27 (1990)]; 23.2 percent over 227 aa with the monooxygenase XylA subunit encoded by the TOL plasmid of *P. putida* [M. Suzuki, T. Hayakawa, J. P. Shaw, M. Rekik, S. Harayama, *J. Bacteriol.* **173**, 1690 (1991)]; 40.2 percent over 226 aa with LuxG of *Vibrio fischeri* [E. Swartzman, S. Kapoor, A. Graham, E. Meighen, *ibid.* **172**, 6797 (1990)]; 37.7 percent over 234 aa with LuxG of *V. harveyi*, and 40.9 percent over 147 aa with a partial sequence of *Photobacterium phosphoreum* LuxG [E. Swartzman, C. Miyamoto, A. Graham, E. Meighen, *J. Biol. Chem.* **265**, 3513 (1990)].
57. E. A. Meighen, *Microbiol. Rev.* **55**, 123 (1991).
58. K. Nakahigashi and H. Inokuchi, *Nucleic Acids Res.* **18**, 6439 (1990).
59. F. Endo *et al.*, *J. Biol. Chem.* **264**, 4476 (1989).
60. The predicted product of *o205* displays 34.2 percent similarity over 160 aa with *B. subtilis* predicted protein L [D. J. Henner, M. Yang, E. Ferrari, *J. Bacteriol.* **170**, 5102 (1988)].
61. In the Swiss-Prot database entry P21166 this ORF is identified as *trkH*, part of the Trk potassium uptake system; no evidence or reference is given. Sequence differences extend our reading frame by 10 aa at the COOH-terminal end.
62. R. C. Goldman, T. J. Bolling, W. E. Kohlbrenner, Y. Kim, J. L. Fox, *J. Biol. Chem.* **261**, 15831 (1986).

63. This is paper 3222 from the Laboratory of Genetics. Supported by award HG00301 from the NIH Human Genome Project. A preliminary report was presented at the Cold Spring Harbor Meeting "The Genome of *E. coli*," 16 to 19 October 1991. We thank D. Rose, B. Fritz, L. Marr, C. Moynihan, C. Olson, M. Schwid, E. Sommers, S. Subramanian, R. Talley, and S. Xiong for technical help; N. Peterson for administration; D. Joseph, M. Livny, J. Shavlik, R. Hagstrom, and R. Overbeek for help with computing; G. Rabin, A. Kondrashov, and G. Bouriakov for programming; S. Baldwin, J. Schroeder, J.-X. Wang, and S.-C. Lin from DNASTAR; G. Church, M. Maguire, W. Reznikof, R. Roberts, K. Rudd, G. Stormo, M. Borodovsky, M. Cashel, E. Koonin, and M. Berlyn for discussions and unpublished results. Finally, we thank our team of University of Wisconsin undergraduates, summer interns, and cooperating graduate students from other departments for their contributions: D. Baxter, P. Bekx, K. Blouke, C. R. Boardman, J. Brandt, K. Cherkauer, L. Cheung, D. Ciske, M. Craven, H. Dahi, W. Davis, T. Delaney, N. Dibben, A. Doubles, H. Eisenberg, J. Eisner, C. Elliott, A. Feldman, S. Foley, J. Freund, M. Gigot, P. Gorski, J. Grant, L. Grota, A. Grumann, T. Gu, L. Guyer, L. Hammes, E. Harsay, A. Hefty, T. Heim, T. Hinton, B. Hoeft, M. Jensen, C. Johnson, G. Jurgella, K. Kadner, J. Kenan, H. Kirkpatrick, H. Kirkpatrick, J. Klamerus, S. Kleiner, J. Klemmer, D. Klinzing, R. Kotarski, A. Kryder, J. Lang, B. H. Lee, D. Luchini, J. Macek, A. Madsen, M. Maguire, K. Merwin, M. Myers, A. Mohamed, S. Mohamed, L. Morrison, A. Nakano, T. Nelson, D. Pochan, M. Polka, T Richmond, G. Rifkin, B. Robella, N. Roskos, M. Rothman, R. Russell, T. Ryan, C. Sanok, M. Sauter, M. Schmelzer, A. Schmidt, L. Schricker, T. Schricker, L. Schroeder, W. Schurer, B. Shiekholeslami, C. Snyder, H. Steltzer, P. Stevens, J. Stroebel, A. Tsang, T. Uhrman, C. Vang, R. Verma, E. Whitford, C. Wipperman, A. Yamane, and Y. Yussman.

27 January 1992; accepted 29 June 1992