Nucleic Acids Research, 1993, Vol. 21, No. 15 3391–3398

Analysis of the *Escherichia coli* genome. III. DNA sequence of the region from 87.2 to 89.2 minutes

Guy Plunkett, III, Valerie Burland, Donna L.Daniels and Frederick R.Blattner
Laboratory of Genetics, University of Wisconsin, 445 Henry Mall, Madison, WI 53706, USA

Received May 8, 1993; Revised and Accepted June 15, 1993 GenBank accession no. L19201

ABSTRACT

The DNA sequence of 96.5 kilobases of the *Escherichia coli* K-12 genome has been determined, spanning the region between *rrnA* at 87.2 minutes and *katG* at 89.2 minutes on the genetic map. The sequence includes 84 open reading frames, of which 46 code for unidentified proteins. Six previously mapped but unsequenced genes have been identified in this span: *mob, fdhD, rhaD, rhaA, rhaB,* and *kdgT*. In addition, five new genes have been assigned: the heat shock genes *hslU* and *hslV*, and the genes *fdoG, fdoH, and fdoI*, which encode the three subunits of formate dehydrogenase-O. The arrangement of the genes relative to possible promoters and terminators suggests 57 potential transcription units. Other features include the precise location of the bacteriophage P2 attachment site *attP2II* , and eleven REP elements, including one containing 9 REP sequences—one of the largest such elements known. This segment brings the total length of contiguous finished sequence to 325 kilobases.

INTRODUCTION

Whole genome analysis is a major current endeavor of molecular biology. *E.coli*, despite having one of the smaller and simpler genomes among the organisms being studied in this way, is a rewarding genome for sequencing.This paper is the third in a series from the *E.coli* Genome Project, whose goal is to determine the complete DNA sequence of *Escherichia coli* (1, 2). We describe 96.5 kilobases (kb), systematically determined as one continuous sequence. Together with the first two segments, the analysis of 325 kb of contiguous DNA have now been completed, locating known genes, new open reading frames, promoters, terminators, repeated regions, Chi sites and other physical features. The density of information gathered continues to be high, with less than 2 percent of the sequence containing no identifiable features. In spite of the existence of a great body of information regarding this organism, we have been able to identify less than half of the potential genes discovered in the sequence so far. Thus *E.coli* provides an appropriate model for the development of genome analysis techniques.

Continual improvements in technical and organizational efficiency have enabled our raw-data gathering teams to reach a rate in excess of a megabase per year. Analysis and interpretation of data produced at this rate present a challenge, involving the constant review of a massive and increasing volume of *E.coli* literature. Even with the help of databases, search and alignment programs and other computing tools, this effort still requires much human judgement.

MATERIALS AND METHODS

All of the sequence reported here has been newly determined. Previously reported data was compared to our own after the sequence was assembled, and conflicts were resolved wherever possible. The starting material for *E.coli* sequencing was a mapped set of strain MG1655 clones in bacteriophage lambda-derived vectors (3). Random libraries for sequencing were prepared from the lambda clones in M13mp19 (4) or in Janus, an engineered M13 vector. The Janus strategy for data collection and procedures for preparation of random libraries in Janus are described in an accompanying paper (5).

DNA template preparation, sequencing, data collection, assembly and finishing, as well as identification and assessment of features in the sequence, were performed as described previously (1, 2). M. Borodovsky's GENMARK program (6) was used to aid in identifying potential reading frames. A database of experimentally determined N-terminal amino acid sequences, obtained from A.Link and G.Church (personal communication), was used to confirm the assignment of the correct start for some genes. Comparisons to the PROSITE database of protein patterns (7) were performed using the program MacPattern (8). A specially written program was used to detect signal sequences based on the weight-matrix method of von Heijne (9).

RESULTS AND DISCUSSION

The 96,484 base sequence presented here has been deposited in the sequence databases and assigned the accession number L19201. This sequence lies to the right (*i.e.*, clockwise) of the 91,408 bases that comprised the first segment of our project (GenBank accession M87049; 1) and overlaps it by three bases of an *Eco*RI site. As in the case of the segment 1—segment 2 overlap (2), the region spanning this junction was examined to ensure that no missing sequences lay between the segments. Figure 1 is a map of the sequence, showing the features identified. Genes and putative genes (ORFs), along with the predicted molecular weights and isoelectric points of the protein products, are listed in Table 1. Gene names are from published sequences or the last edition of the *E.coli* genetic map (10), unless otherwise noted.

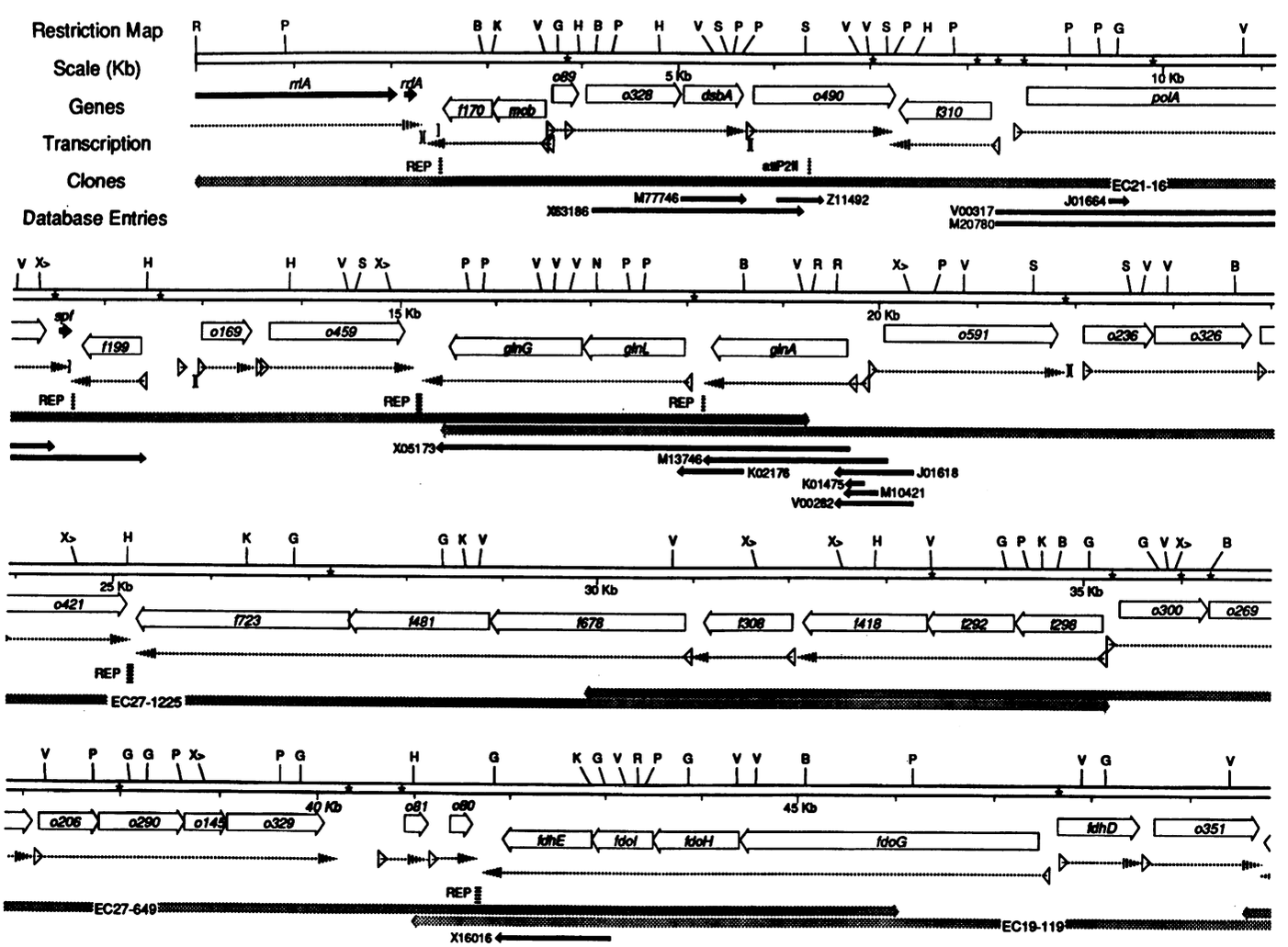
Eighty-eight percent of the sequence codes for either structural RNA or protein. We have identified 84 open reading frames (ORFs), including the partial gene *katG* at the right end. Forty-six of the ORFs are potential new genes, still unidentified despite searches of the databases for similarities suggestive of function. Three of these ORFs have been sequenced previously but remain unidentified. Similarities of predicted ORFs with database protein sequences, and matches to PROSITE patterns, are summarized in Table 2. Six genes previously mapped to this region have been newly sequenced in this span, and five more new genes have been assigned by physical and genetic data. The 27 remaining genes had been previously characterized.

Newly identified genes

Formate dehydrogenase-O. In addition to the sequenced FDH-H and FDH-N enzymes, *E.coli* is known to possess a third unmapped formate dehydrogenase activity (11). Recently it has been further characterized by biochemical and immunological data (12). Moreover, a selenocysteine tRNA-dependent FDH activity is synthesized both aerobically and anaerobically in the presence of nitrate (13) and has been called FDH-O (14). Comparison with the DNA sequence of *fdnGHI* (15) as well as the sequences of the predicted proteins, allowed us to identify the new genes *fdoGHI*. Three correctly sized ORFs between coordinates 42846 and 47461 correspond to the α, β and γ subunits of the enzyme, *fdoG, H* and *I*, respectively. Like *fdnG*, coding for the α subunit of FDH-N, *fdoG* contains an in-frame TGA (opal) codon that specifies selenocysteine, and the sequences contributing to the mRNA context required for decoding UGA as selenocysteine are identical with those demonstrated for *fdnG* (16). Our identification of these genes is also consistent with an analysis of formate oxidase activity of plasmid subclones of the *fdo* locus (H. Abaibou and M.-A. Mandrand, personal communication).

Two FDH-associated genes are also located here. We identified *fdhD* by comparison of the restriction map with published data (17, 18). *fdhE*, adjacent to and cotranscribed with *fdoGHI*, was previously known (19). The products of these two genes are required for active FDH-N but do not regulate transcription or translation of the structural genes (20). Their precise function is unknown, but may be in assembly or localization of the subunits, or associated with cofactor(s). An FDH-associated gene of *Wolinella succinogenes* (*fdhD*, function unknown) shows similarity to the *fdhD* in this sequence (Table 2).

Heat shock genes. Two previously unknown members of the heat shock regulon have been identified between 82049 and 83920 by DNA sequence and protein characterization (21). They are designated *hslU* and *V*. The gene products have been identified as the heat shock induced polypeptides HtpI and O respectively



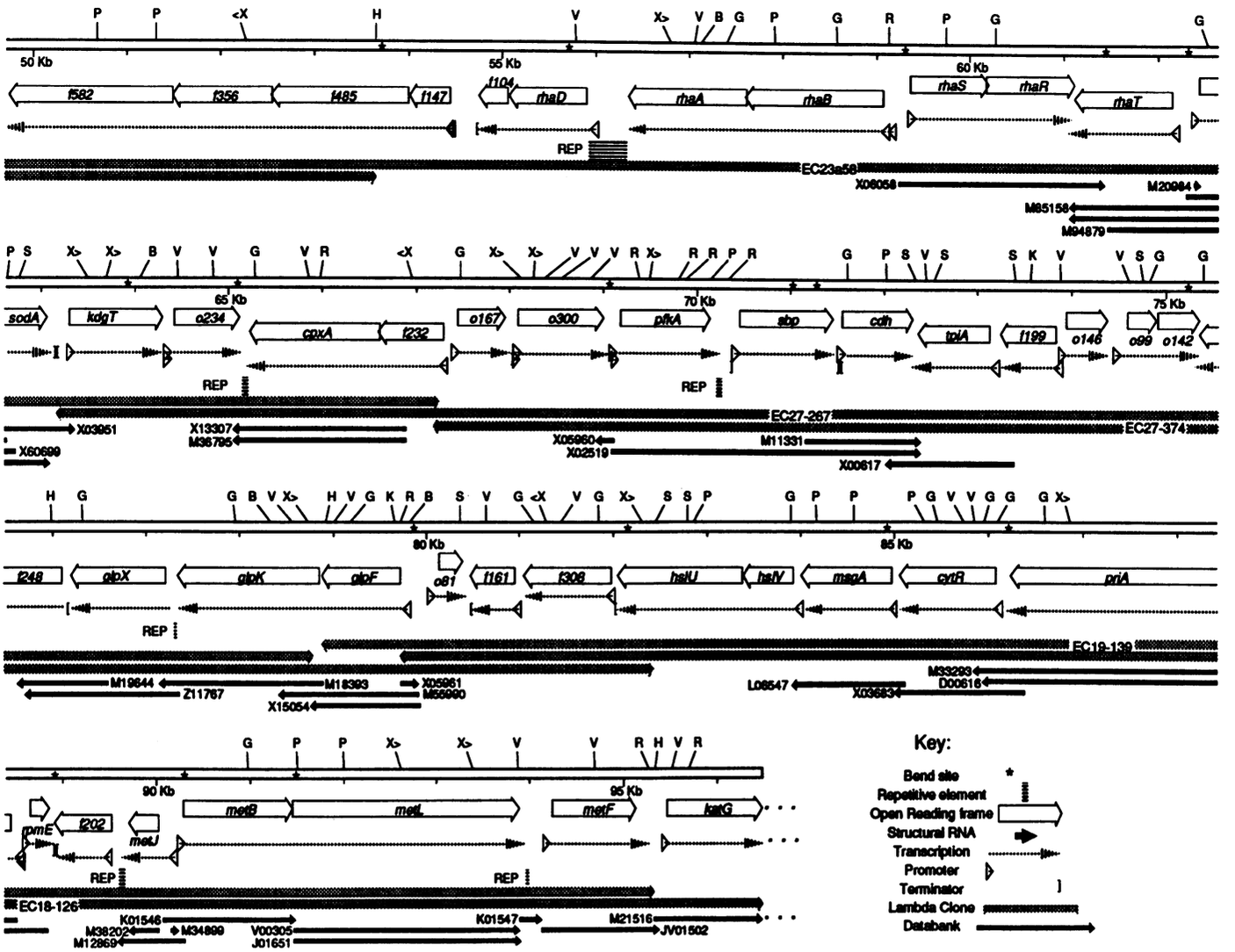


Figure 1. Map of the finished sequence and its features, proceeding left to right in eight tiers. The top line shows restriction sites for ten enzymes: *Avr*II (A), *Bam*HI (B), *Bgl*I (G), *Eco*RI (R), *Eco*RV (V), *Hin*dIII (H), *Kpn*I (K), *Not*I (N), *Pst*I (S) and *Pvu*II (P). In addition, the location and orientation of Chi sites are indicated by X> and X<. Gene names are indicated for identified RNA-coding regions and open reading frames. Unidentified open reading frames are designated *o#* or *f#* for the two transcriptional orientations, where the numbers indicate the predicted peptide length. Other sequence determinations spanned by this segment are indicated, with their database accession numbers. This sequence is in GenBank (accession L19201), annotated with these features as well as conflicts with other sequence determinations.

(22), by correlation with induced protein spots analysed by 2-D PAGE.

Identification of mapped genes

mob. The gene product of the *mob* (*chlB*, *narB*) locus is essential for the formation of molybdopterin guanine dinucleotide, a late step in the biosynthesis of molybdenum cofactor (23). Of several candidate ORFs in the region, the correct one was located by comparing restriction map data (24). The predicted product of the ORF fitting the restriction map was similar to the protein size observed experimentally (25).

rhaBAD. The genes of the L-rhamnose regulon mapped to this region, linked to *metB* (26), but only the transcriptional regulators *rhaSR* and the permease gene *rhaT* had been identified previously (27, 28). The protein products of three more genes, *rhaABD* were characterized in maxicells: rhamnose isomerase, rhamnulose kinase and rhamnulose 1-phosphate aldolase at 47, 52−54, and 32 kDa respectively (29). This data together with the restriction map of the three genes and some partial DNA sequence data for the analogous genes in *Salmonella typhimurium* (30), enabled us to identify *rhaBAD* adjacent to and divergently transcribed from *rhaSR*.

kdgT. This gene encodes the permease for 3-deoxy-2-oxo-D-gluconate. It was located by genetic data, protein size, and comparison with the restriction map for the *E.coli* gene (31). The predicted gene product also showed similarity to the amino acid sequence of the analogous gene in *Erwinia chrysanthemi* (32) (Table 2).

dsbA. This gene (also called *dsf* or *ppfA*) encodes a protein responsible for disulphide bond formation *in vivo*. It was recently identified and sequenced using data from this project (33).

Table 1. Genes and predicted characteristics of deduced protein products

Gene[a]	Site No.[b]	Endpoints in sequence	first...last codon	molecular (kD)	size (aa)	pI
rrlA[c]	203	1 > 2063	(23S rRNA)		2905 bp	
rrfA	210	2157 > 2276	(5S rRNA)		120 bp	
f170		3058 < 2546	ATG...TAA	18.9	170	5.2
mob (chlB)	921	3639 < 3055	GTG...TGA	21.6	194	6.1
o89		3709 > 3978	ATG...TAA	10.3	89	5.0
o328		4056 > 5042	ATG...TAA	38.1	328	4.9
dsbA (ppfA)		5059 > 5685	ATG...TAA	23.1	208	6.2
o490		5797 > 7269	GTG...TAA	54.2	490	5.2
f310		8242 < 7310	ATG...TAA	36.3	310	9.6
polA	375	8606 > 11392	ATG...TAA	103.1	928	5.4
spf	157	11539 > 11647	(spot 42 RNA)		109 bp	
f199 (yihA)		12372 < 11773	GTG...TAA	22.2	199	7.3
o169		12987 > 13496	ATG...TAA	19.1	169	6.7
o459		13679 > 15058	GTG...TAA	53.0	459	5.9
glnG (ntrC)	702	16918 < 15509	ATG...TGA	52.3	469	6.3
glnL (ntrB)	701	17979 < 16930	ATG...TAA	38.6	349	5.5
glnA	705	19674 < 18265	ATG...TAA	51.9	469	5.3
o591		20047 > 21822	GTG...TGA	65.5	591	5.0
o236		22087 > 22797	ATG...TAG	26.9	236	6.8
o326		22805 > 23785	ATG...TAA	36.9	326	5.1
o421		23887 > 25152	ATG...TAA	46.3	421	9.1
f723		27414 < 25243	GTG...TGA	81.8	723	8.6
f481		28856 < 27411	ATG...TGA	53.1	481	9.1
f678		30915 < 28879	ATG...TAA	77.3	678	5.0
f308		32040 < 31114	ATG...TAA	34.0	308	6.4
f418		33410 < 32154	ATG...TAA	48.0	418	6.1
f292		34290 < 33412	ATG...TAA	32.0	292	5.8
f298		35210 < 34314	ATG...TAA	31.2	298	6.1
o300		35372 > 36274	ATG...TAA	32.0	300	4.8
o269		36284 > 37093	GTG...TAG	29.5	269	5.9
o206		37171 > 37791	GTG...TAA	23.5	206	5.5
o290		37785 > 38657	ATG...TGA	32.8	290	9.0
o145		38654 > 39091	ATG...TGA	16.0	145	4.7
o329		39088 > 40077	ATG...TAG	37.1	329	6.2
o81		40903 > 41148	ATG...TGA	9.4	81	9.8
o80		41366 > 41608	ATG...TAA	9.2	80	4.3
fdhE		42867 < 41938	ATG...TAA	34.8	309	4.9
fdoI		43499 < 42864	ATG...TGA	24.6	211	10.2
fdoH		44398 < 43496	ATG...TGA	33.1	300	5.1
fdoG		47461 < 44411	ATG...TAA	112.7	1016	7.5
fdhD		47655 > 48488	GTG...TAA	30.6	277	6.3
o351		48641 > 49696	ATG...TAA	39.3	351	4.9
f582		51495 < 49747	ATG...TGA	66.0	582	6.2
f356		52565 < 51495	ATG...TAA	38.7	356	5.7
f485		54012 < 52555	GTG...TGA	51.3	485	8.7
f147		54460 < 54017	ATG...TAA	16.1	147	7.8
f104		55075 < 54761	ATG...TAA	12.3	104	5.3
rhaD	289	55909 < 55085	ATG...TAA	30.1	274	5.7
rhaA	292	57615 < 56356	ATG...TAA	47.2	419	5.6
rhaB	291	59081 < 57612	ATG...TGA	54.1	489	5.0
rhaS (rhaC2)	17950	59369 > 60205	ATG...TAA	32.3	278	6.5
rhaR (rhaC1)	290	60189 > 61127	ATG...TAA	35.7	312	6.7
rhaT		62158 < 61124	ATG...TAA	37.3	344	9.5
sodA	17593	62443 > 63063	ATG...TAA	23.1	206	6.8
kdgT	589	63314 > 64306	GTG...TAA	34.1	330	8.8
o234		64425 > 65129	ATG...TAA	26.6	234	5.9
cpxA	908	66608 < 65235	ATG...TAA	51.6	457	5.7
f232 (yiiA)		67303 < 66605	ATG...TGA	26.3	232	5.4
o167		67450 > 67953	ATG...TAG	19.1	167	6.7
o300		68102 > 69004	ATG...TAA	32.9	300	6.4
pfkA	413	69185 > 70147	ATG...TAA	34.8	320	5.5
sbp	17911	70467 > 71456	ATG...TGA	36.7	329	7.0
cdh	931	71563 > 72318	ATG...TAA	28.5	251	8.5
tpiA	88	73140 < 72373	ATG...TAA	27.0	255	5.8
f199		73847 < 73248	ATG...TGA	21.8	199	8.9
o146		73948 > 74388	ATG...TAA	16.5	146	9.9
o99		74600 > 74899	ATG...TGA	10.8	99	4.4
o142		74926 > 75354	ATG...TAA	16.3	142	6.7
f248 (mvrA)		76105 < 75359	ATG...TAA	27.8	248	6.5
glpX		77212 < 76202	ATG...TGA	35.9	336	5.3

glpK	691	78855 < 77347	ATG...TAA	56.2	502	5.3
glpF	692	79723 < 78878	ATG...TAA	29.8	281	6.5
o81		80148 > 80393	ATG...TGA	9.6	81	4.6
f161		80963 < 80478	ATG...TGA	17.4	161	3.9
f308		81982 < 81056	ATG...TAA	33.6	308	9.0
hslU (htpI)		83380 < 82049	ATG...TAA	49.6	443	5.2
hslV (htpO)		83920 < 83390	GTG...TAA	19.1	176	6.2
msgA (ftsN)		84972 < 84013	GTG...TGA	35.8	319	10.2
cytR	887	86089 < 85064	GTG...TAA	37.8	341	6.3
priA		88443 < 86245	ATG...TAA	81.7	732	8.9
rpmE	237	88646 > 88858	ATG...TAA	7.9	70	9.3
f202		89527 < 88919	ATG...TAA	23.1	202	9.6
metJ	508	90028 < 89711	ATG...TAA	12.1	105	5.4
metB	515	90305 > 91465	ATG...TAA	41.6	386	6.3
metL	506	91468 > 93900	ATG...TAA	88.9	810	5.4
metF	511	94249 > 95139	ATG...TAA	33.1	296	6.2
katG'[d]	14983	95468 > 96484	ATG...	80.0	726	5.1

[a]Gene names are those used in the most recent *E.coli* genetic map (10), unless the genes are not present on that map; alternate gene names are indicated in parentheses.
[b]Mapped genes have been assigned Site numbers in a database maintained by the *E.coli* Genetic Stock Center (61; M.Berlyn, personal communication).
[c]The sequence of *rrlA* extends beyond the sequence presented here; our previous determination of the balance of the sequence (1) permitted calculation of the length of the intact rRNA.
[d]The sequence of *katG* extends beyond the sequence presented here; data from the overlapping GenBank entry M21516 was used to calculate the protein size and isoelectric point.

msgA. This gene, at 84013−84972, was identified by comparison with the DNA sequence ECOGRPESUP (34) which has several differences from our determination that affect the reading frame. More recently the sequence of *ftsN* (35) was found to match the same open reading frame. Both *msgA* and *ftsN* were isolated as multi-copy suppressors of ts mutations, in *grpE* and *ftsA* respectively; the actual function of this gene remains unclear.

Unidentified genes

Two additional ORFs show striking similarities to other proteins, but the data is insufficient for a firm identification. *o591* is similar to elongation factor G (EF-G) from a variety of organisms, as well as TetM/TetO tetracycline-resistance proteins (Table 2). The similarity is especially striking in the first 150 aa, with similarities ranging from 38.4% to 45.3%. This amino-terminal region of the sequences contains the PROSITE patterns PS00017 (ATP/GTP-binding site motif A) and PS00301 (GTP-binding elongation factors signature). EF-G is involved in the GTP-dependent translocation of the nascent protein chain from the A-site to the P-site of the ribosome (36). The TetM/TetO proteins abolish the inhibitory effect of tetracycline on protein synthesis, apparently by a non-covalent modification of the ribosomes (37). We conclude that the product of *o591* probably interacts with the ribosomes in a GTP dependent manner.

The 39.1% similarity between *o300* at 68102−69004 and protein P34 of *Rickettsia rickettsi* (38) might suggest some functional identity but neither ORF has been associated with an activity or a phenotype.

Four other genes mapping to this region have not yet been identified(*hemG, fcsA, manC, menA*), and three more loci (*ecfB*, *ssd* and *eup*) were reported to be identical with *cpxA* (39). As well as *fcsA* (40), four other genes concerned with regulation of cell division or chromosome partitioning appear to map at 88−89 minutes, although not yet assigned to specific genetic map locations. These are *mukC26* (41), *divA* (42), *parD* (43) and *mbrB* (44). There is no evidence to allow identification of these at present.

The phenotypes of two genes mapping to this region, *rimD* and *rit*, suggest they are involved in 50S ribosomal subunit structure (45, 46). It is possible that these are alleles of *rpmE*.

Signal peptides

A computer search for signal peptide-like sequences identified six candidates. Three were in ORFs of unknown function: *o167* (between *cpxA* and *pfkA*), *f199* (to the right of *tpiA*), and *f202* (between *rpmE* and *metJ*). The other three were found in the *cdh, sbp* and *dsbA* genes, whose products are all periplasmic proteins (47, 48, and 33 respectively).

Differences from published sequence

Comparison of our sequence and that published for the *tpiA* gene shows a 50 bp fragment missing from X00617 as well as numerous smaller differences. *tpiA* is not affected but 24 bases are deleted from the adjacent gene *f199*. The missing sequence is between two Hinf I sites. Since this enzyme was used in the sequencing process, an error in the strategy is suggested.

Several differences from previously reported data are present in *glnA* and *G*. Some of these cause changes in the predicted amino acid sequence of the MG1655 gene although at each of these locations a perfect match with the sequence of *glnG* from *Klebsiella pneumoniae* or *glnA* from *S. typhimurium* is maintained. These differences may reflect strain differences between MG1655 and the *E.coli* strains from which the published sequences were obtained, or they may be errors.

The ribosomal gene *rpmE* encoding the L31 protein was identified by comparison with the L31 amino acid sequence from *E.coli* B (49). Translation of our sequence results in an RpmE which is longer at the carboxy terminus than the previously reported protein (our sequence data is clear at this point). Although one study found no differences in 50S proteins from different strains (50), the difference we note between the L31 predicted for MG1655 and the L31 from *E.coli* B does suggest either a strain difference or post-translational processing.

Other features

RNA genes. The sequence of *rrlA* and *rrfA* at the beginning of this segment completes the ribosomal operon *rrnA*, part of which

Table 2. Similarity of predicted ORF products to other proteins. (newly identified genes and unidentified ORFs, presented in map order)

ORF	length	PROSITE match and/or sequence similarity to	score
f170	170 aa	PS00017: ATP_GTP_A [ATP/GTP-binding site motif A]	100% match
f199	199 aa	PS00017: ATP_GTP_A [ATP/GTP-binding site motif A]	100% match
o591	591 aa	PS00017: ATP_GTP_A [ATP/GTP-binding site motif A]	100% match
		PS00301: EFACTOR_GTP [GTP-binding elongation factors signature]	100% match
		similar to elongation factor G, and TetM/TetO tetracycline-resistance proteins; examples:	
		EFG_MICLU *Micrococcus luteus* elongation factor G (EF-G)	35.6% 175 aa
		EFG_THETH *Thermus aquaticus* elongation factor G (EF-G)	27.3% 323 aa
		EFG_ECOLI *Escherichia coli* elongation factor G (FusA)	25.8% 353 aa
		EFG_ANANI *Anacystis nidulans* elongation factor G (EF-G)	23.3% 471 aa
		EFG_SYNY3 *Synechocystis* sp. elongation factor G (EF-G)	22.0% 454 aa
		TET9_ENTFA *Enterococcus faecalis* tetracycline resistance protein TetM	28.9% 220 aa
		TETM_UREUR *Ureaplasma urealyticum* tetracycline resistance protein TetM	28.0% 220 aa
		the similarities are greater over the amino-terminal 150 amino acids:	
		EFG_ANANI *Anacystis nidulans* elongation factor G (EF-G)	45.3% 136 aa
		EFG_SYNY3 *Synechocystis* sp. elongation factor G (EF-G)	38.4% 145 aa
		EFG_MICLU *Micrococcus luteus* elongation factor G (EF-G)	39.0% 145 aa
		EFG_THETH *Thermus aquaticus* elongation factor G (EF-G)	43.1% 142 aa
		EFG_ECOLI *Escherichia coli* elongation factor G (FusA)	40.4% 143 aa
		TET9_ENTFA *Enterococcus faecalis* tetracycline resistance protein TetM	40.3% 133 aa
		TETM_UREUR *Ureaplasma urealyticum* tetracycline resistance protein TetM	40.3% 133 aa
o236	236 aa	PS00043: HTH_GNTR_FAMILY [GntR family of transcriptional regulators]	close match
		matches all but position 11 of the pattern; has K instead of N,S, or T	
		P30_ECOLI *Escherichia coli* hypoth. 30 kDa protein, adjacent to *suc* operon	29.7% 74 aa
f723	723 aa	MELB_ECOLI *Escherichia coli* melibiose carrier protein (melibiose permease)	24.6% 452 aa
f481	481 aa	MELB_ECOLI *Escherichia coli* melibiose carrier protein (melibiose permease)	26.2% 439 aa
		F723 and F481 are also very similar to each other:	62.1% 468 aa
f298	298 aa	MMSB_PSEAE *Pseudomonas aeruginosa* 3-hydroxyisobutyrate dehydrogenase	36.0% 284 aa
		YHAE_ECOLI *Escherichia coli* hypoth. 31 kDa protein in *rnpB* 3′ region	35.7% 280 aa
		D3HI_RAT *Rattus norvegicus* 3-hydroxyisobutyrate dehydrogenase	28.0% 292 aa
o300	300 aa	PS00584: PFKB_KINASES_2 [PfkB family of carbohydrate kinases, signature 2]	100% match
		but contains no match to PS00583: PFKB_KINASES_1	
		similar to members of the PfkB family of carbohydrate kinases; examples:	
		SCRK_KLEPN *Klebsiella pneumoniae* fructokinase	30.4% 251 aa
		RBSK_ECOLI *Escherichia coli* ribokinase	28.6% 297 aa
		SCRK_SALTY *Salmonella typhimurium* fructokinase	28.0% 238 aa
		SCRK_VIBAL *Vibrio alginolyticus* fructokinase	27.6% 280 aa
o269	269 aa	GLPR_ECOLI *Escherichia coli* glycerol-3-phosphate regulon repressor	31.4% 140 aa
o80	80 aa	PS00659: GLYCOSYL_HYDROL_F5 [Glycosyl hydrolases family 5 signature]	100% match
fdoI	211 aa	FDNI_ECOLI *Escherichia coli* formate dehydrogenase-N gamma subunit	44.6% 156 aa
fdoH	300 aa	PS00198: 4FE4S_FERREDOXIN	100% match
		[4Fe-4S ferredoxins, iron-sulfur binding region signature]	
		FDNH_ECOLI *Escherichia coli* formate dehydrogenase-N beta subunit	76.2% 294 aa
fdoG	1016 aa	FDNG_ECOLI *Escherichia coli* formate dehydrogenase-N alpha subunit	75.0% 1016 aa
fdhD	277 aa	PIR: S18216 *Wolinella succinogenes* formate dehydrogenase D	27.1% 242 aa
f485	485 aa	similar to fructose-specific phosphotransferase enzyme II; examples:	
		PT2F_XANCP *Xanthomonas campestris* FruA	34.0% 468 aa
		PT2F_RHOCA *Rhodobacter capsulatus* FruA	33.6% 470 aa
		PT2F_ECOLI *Escherichia coli* FruA (PtsF)	31.2% 470 aa
f147	147 aa	limited similarity to hypoth. protein downstream of *rpoN* (sigma-54);	
		(similarity includes a 12/21 bp identity among all 3 sequences)	
		PIR: D38179 *Bradyrhizobium japonicum* hypoth. protein *rpoN2* 3′ region (fragment)	38.2% 67 aa
		YRP2_KLEPN *Klebsiella pneumoniae* hypoth. 17.7 kDa protein *rpoN* 3′ region	22.1% 137 aa
rhaA	419 aa	PS00102: PHOSPHORYLASE [Phosphorylase pyridoxal-phosphate attachment site]	100% match
		YRHB_SALTY *Salmonella typhimurium* hypoth. protein *rhaB* 3′ region (fragment)	92.2% 64 aa
rhaB	489 aa	RHAB_SALTY *Salmonella typhimurium* rhamnulokinase (RhaB)	81.0% 489 aa
kdgT	330 aa	KDGT_ERWCH *Erwinia chrysanthemi* 2-keto-3-deoxygluconate permease (KdgT)	88.8% 320 aa
f232	232 aa	OMPR_SALTY *Salmonella typhimurium* transcriptional regulatory protein OmpR	38.2% 233 aa
o300	300 aa	P34_RICRI *Rickettsia rickettsii* protein P34	39.1% 283 aa
hslU	443 aa	PS00017: ATP_GTP_A [ATP/GTP-binding site motif A]	100% match
		GB: PASLEUTREP_1 *Pasteurella haemolytica* hypoth. protein ORF1	74.9% 389 aa
hslV	176 aa	PRCU_YEAST *Saccharomyces cerevisiae* potential proteasome component	36.8% 67 aa
f202	202 aa	YEBB_ECOLI *Escherichia coli* hypoth. 26.8 kDa protein, ruvA-ruvC region	26.0% 178 aa

Sequences from the SwissProt (release 24), NBRF-PIR (release 35) and translated GenBank databases were aligned with the predicted proteins, using DNASTAR's Align for the Macintosh with a gap penalty of 4 and a gap length penalty of 12; the score is the% amino acid identity and the length of the alignment. Sequences were compared to the PROSITE database (release 10) using MacPattern (8). The products of previously sequenced and characterized genes are not listed in this table.

we reported in the first segment (1). This sequence also contains the Spot 42 RNA gene *spf* adjacent to *polA*.

Transcription signals. Promoters and terminators were identified by computer searches, with final assessment of each feature by eye (1, 2). Their arrangement indicated 57 transcription units (Figure 1) including five groups of at least three unidentified genes. Thirteen 'alternate' promoters were found (regulated by sigmas other than sigma 70) four of which were the only promoter candidate for the adjacent gene. The promoter candidate for *o160* is an *flhDC* type and those for *fdhD* and *rhaD* are *rpoS* (starvation) types. A heat shock promoter was defined experimentally for *hslVU* (21).

Non-coding features. An attachment site for bacteriophage P2, *attP2II,* was located within o490 at 6243–6269 by sequence comparison. The previously sequenced *att* was from *E.coli* C (51). There are 4 differences between the MG1655 and the *E.coli* C sequences. Two of the differences are in the *att P2* core, and correspond to the previously characterized *saf* variant (51) and a variant detected in K12 strain C600, presumably strain differences.

Twelve REP (Repeated Extragenic Palindromic) elements (52) were found in the 96.5 kb sequence. Seven of these have been identified previously, either experimentally or by sequence analysis (53, 54) and five are new. REP elements are variable in structure and do not always match the consensus well. Thus there are differences between described REPs depending on the method of search; some 'by-eye' assessment is frequently involved. They are all in intergenic spaces and often between convergently transcribed genes (Figure 1), as though the REP secondary structure might act as a transcription block in both orientations. The REP element between *rhaA* and *rhaD* is one of the largest known, containing nine REP sequences separated by two different inter-REP sequences. This distinctive structure must surely have some effect on expression of the *rha* genes, perhaps acting as an attenuator or in the processing of a transcript. The promoter candidate for *rhaD* is a weak match with the *rpoS* type and is completely contained within the REP element but there is no evidence that this might be functional. An experimental examination of expression is clearly required. A search of the sequence for ERIC or IRU (enterobacterial repetitive intergenic consensuses or intergenic repeat units) (55) found none.

Static bends in the DNA were predicted by calculating the trajectory by the method of Levene and Crothers (56) and calculating the degree of deviation from straight for all overlapping 100 nucleotide spans. The locations of the bend sites greater than 72 degrees are shown in Figure 1, and include the two previously reported sequences ECOBENT5 and 6 (X05960 and X05961).

A search for Chi sites found 21, oriented consistently with the directions of replication and translation as discussed at length previously (2). The translational orientation of genes in this segment deviates slightly from the asymmetric distribution found previously (40 with replication and 44 against) but when this data is added to all the accumulated data we have sequenced so far, two thirds of genes are oriented with replication (2).

Updates

The purpose of this section is to advise of changes and corrections to the published *E.coli* Genome Project sequences and their annotations. A merged entry will be maintained by the project at Wisconsin with all corrections, and deposited with the databases as a separate entry.

Segment 1 (1) covers 84.5 to 87.3 minutes on the current genetic map (10). The minute coordinates used in the title of the first segment (1) were taken from the 1987 version of the map (58). There is no gap between the segments 1 and 2. A reassessment of the data alters our sequence of *metE* to remove a short internal frame shift compared with ECOCDMS, maintaining the original endpoints (57). In the *corA* gene, our data now agrees with that of ECOCRA (L11042), giving a polypeptide of 316 amino acids (59). The gene o716, tentatively identified as *rrfT*, has been extended by 92 amino acids by insertion of a single base near the 3′ end. The extended reading frame continues the match with the amino acid sequence of the *S. typhimurium rrfT* gene STYCARABA_1. A previous report (60) contained a restriction map and data on the peptides encoded

Table 3. The sequence changes are presented in a context of 14 residues, with the altered residues in lower case. This will allow use of a simple text editor both to change the sequence portion of an entry and to verify which version is at hand

Alteration	Sequence change			Description (ECOUW85U coordinates)
extend *rffT*	ATCCCGCTCGGGCG	to	ATCCCGCTCgGGGCG	insert G at position 32769
shorten *corA*	TTTATGATCTCGCG	to	TTTATGATCcTCGCG	insert C at position 55432
switch frame in *metE*	TGGCGTGCGTGATG	to	TGGCGcTGCGTGATG	insert C at position 67875
resume frame in *metE*	CGAACCGGCGcCTG	to	CGAACCGGCGCTG	delete C at position 67940
intergenic compression	AATACCAcCCCGGT	to	AATACCACCCGGT	delete C at position 68424
intergenic compression	GCATGCCGGCGTCC	to	GCATGCCcGGCGTCC	insert C at position 68549
intergenic compression	TAATCTCTcgTTTC	to	TAATCTCTgcTTTC	change CG to GC at position 68575–68576
intergenic compression	GCCCGCACGCCTGG	to	GCCCGCAgCGCCTGG	insert G at position 68597
intergenic compression	AACGCTCTCTGCGG	to	AACGCTCTcCTGCGG	insert C at position 68623

Merger specification:
MG1655v1 = ECOUW82(1,REND) + ECOUW85U(7,32769) + 'G' + ECOUW85U(32770,55432) + 'C' + ECOUW85U(55433,67874) + 'C' + ECOUW85U(67875,67939) + ECOUW85U(67941,68423) + ECOUW85U(68425,68548) + 'C' + ECOUW85U(68549,68574) + 'GC' + ECOUW85U(68577,68596) + 'G' + ECOUW85U(68597,68622) + 'C' + ECOUW85U(68523,REND) + ECOUW87(4,REND).

The merger specification is in the DNASTAR splicing language (62). ECOUW82 is accession L10328 (2), ECOUW85U is accession M87049 (1), and ECOUW87 is accession L19201 (this paper).

in the *ilv–udp* region. The sequence changes are all detailed in Table 3.

In the 81.5–84.4 minute region the starts of the *pyrE* and *tnaA* genes were misannotated. The correct start codons are AUG at 5160 for *pyrE* (as in V01578) and AUG at 78127 for *tnaA* (as in K00032).

ACKNOWLEDGEMENTS

This is paper 3366 from the Laboratory of Genetics, supported by award HG000301 from the NIH Human Genome Project. Some of these results were reported at the 'Small Genomes' meeting, March 28–30 1993, Paris, France. We thank A.Link and G.Church for providing us with their amino-terminal sequence database,and S.En-Chuang, M.-A.Mandrand, C.Georgopolous and D.Ang for sharing data prior to publication. We are happy to acknowledge continuing cooperation from M.Kröger and K.Rudd in keeping abreast of the *E.coli* sequence data from other workers via their compendiums of such data. Thanks to H.Wirt for help with the sequence analysis, and to V.Stewart, K.Robison, R.Matthews, M.Maguire and other members of the *E.coli* community for helpful discussions. For software development and assistance we thank M.Borodovsky, G.Bouriakov, J.Shavlik, and the programming staff of DNASTAR. Excellent technical support was provided by B.Fritz, K.Kadner, M.Maguire, R.Mikkelson, C.Moynihan, D.Rose, E.Sommers, S.Subramanian, and R.Talley. Thanks also to N.Peterson for administration. Finally we thank our University of Wisconsin–Madison undergraduates and recent graduates for their excellent teamwork: H.Abernathy, S.Ahmad, A.Azman, L.Brennan, A.Broah, J.Champ, H.Cheng, W.Davis, K.Day, H.Demrow, T.Delaney, D.Duescher, J.Ehley, K.Griswold, E.Grotbeck, A.Grumann, A.Hamdan, L.Hammes, J.Hobbes, D.Hornung, D.Johanowicz, S.Johnson, J.Kamaruddin, J.Katcha, Y.Kim, Heather Kirkpatrick, Heidi Kirkpatrick, K.Klein, P.Lee, K.Li, S.Mohamad, A.Mohamed, Z.Othman, F.Pangil, N.Parwan, M.Polka, R.Ramamurthy, N.Rimmer, S.Rhode, T.Ryan, P.Scheskenbach, A.Schmidt, J.Schmidt, L.Schroeder, A.Shaya, K.Streif, C.Wan, W.Wilson, C.Whipperman, B.Yusoff, Z.Zahari, and M.Zhang.

REFERENCES

1. Daniels,D.L., Plunkett,G.,III, Burland,V. and Blattner,F.R. (1992) *Science* **257**, 771–778.
2. Burland,V., Plunkett,G.,III, Daniels,D.L. and Blattner,F.R. (1993) *Genomics* **16**, 551–561.
3. Daniels,D. (1990) In Drlica,K. and Riley,M. (eds.), *The Bacterial Chromosome.* American Society for Microbiology, Washington, D.C., pp. 43–52.
4. Yanisch-Perron,C., Vieira,J. and Messing,J. (1985) *Gene* **33**, 103–119.
5. Burland,V, Daniels,D.L., Plunkett,G.,III and Blattner,F.R. (1993) *Nucleic Acids Res.* **21**, 3385–3390.
7. Bairoch,A. (1992) *Nucleic Acids Res.* **20**, 2013–2018.
8. Fuchs,R. (1991) *Comput. Applic. Biosci.* **7**, 105–106.
9. von Heijne,G. (1986) *Nucleic Acids Res.* **14**, 4683–4690.
10. Bachmann,B.J. (1990) *Microbiol. Rev.* **54**, 130–197.
11. Pinset,J. (1954) *Biochem. J.* **57**, 10–16.
12. Pommier,J., Mandrand,M.-A., Holt,S.E., Boxer,D.H. and Giordano,G. (1992) *Biochim. Biophys. Acta* **1107**, 305–313.
13. Sawers,G., Heider,J., Zehelein,E. and Böck,A. (1991) *J. Bacteriol.* **173**, 4983–4993.
14. Böck,A., Forchhammer,K., Heider,J. and Baron,C. (1991) *Trends Biochem. Sci.* **16**, 463–467.
15. Berg,B.L., Li,J., Heider,J. and Stewart,V. (1991) *J. Biol. Chem.* **266**, 22380–22385.
16. Berg,B.L., Baron,C. and Stewart,V. (1991) *J. Biol. Chem.* **266**, 22386–22391.
17. Mandrand-Berthelot,M.-A., Couchoux-Luthaud,G., Santini,C.-L. and Giordano,G. (1988) *J. Gen. Microbiol.* **134**, 3129–3139.
18. Schlindwein,C., Giordano,G., Santini,C.-L. and Mandrand,M.-A. (1990) *J. Bacteriol.* **172**, 6112–6121.
19. Schlindwein,C. and Mandrand,M.-A. (1991) *Gene* **97**, 147–148.
20. Stewart,V., Lin,J.T. and Berg,B.L. (1991) *J. Bacteriol.* **173**, 4417–4423.
21. Chuang,S.-E., Burland,V., Plunkett,G.,III, Daniels,D.L. and Blattner,F.R. (1993) *Gene* (in press).
22. Neidhardt,F. C. and VanBogelen,R. A. (1987) In Neidhardt,F.C. (ed.) *Escherichia coli and Salmonella typhimurium. Cellular and Molecular Biology.* American Society for Microbiology, Washington, DC. pp 1334–1345.
23. Johnson,J. L., Indermaur,L. W. and Rajagopalan,K. V. (1991) *J. Biol. Chem.* **266**, 12140–12145.
24. Reiss,J., Kleinhofs,A. and Klingmüller,W. (1987) *Mol. Gen. Genet.* **206**, 352–355.
25. Santini,C.-L., Karibian,D., Vasishta,A., Boxer,D. and Giordano,G. (1989) *J. Gen. Micro.* **135**, 3467–3475.
26. Power,J. (1967) *Genetics* **55**, 557–568
27. Tobin,J.F. and Schleif,R. F. (1987) *J. Mol. Biol.* **196**, 789–799.
28. Baldomà,L., Badía,J., Sweet,G. and Aguilar,J. (1990) *FEMS Microbiol. Lett.* **72**, 103–108.
29. Badía,J., Baldomà,L., Aguilar,J. and Boronat,A. (1989) *FEMS Microbiol. Lett.* **65**, 253–258.
30. Nishitani,J. andWilcox,G. (1991) *Gene* **105**, 37–42.
31. Mandrand-Berthelot,M.-A., Ritzenthaler,P. and Mata-Gilsinger,M. (1984) *J. Bacteriol.* **160**, 600–606.
32. Allen,C., Reverchon,S. and Robert-Baudouy,J. (1989) *Gene* **83**, 233–241.
33. Bardwell,J.C.A., McGovern,K. and Beckwith,J. (1991) *Cell* **67**, 581–589.
34. Wu,B. and Ang,D. (unpublished) GenBank accession L06547.
35. Dai,K., Xu,Y. and Lutkenhaus,J. (unpublished) GenBank accession L14281.
36. Moldave,K. (1985) *Ann. Rev. Biochem.* **54**, 1109–1149.
37. Salyers,A.A., Speer,B.S.and Shoemaker,N.B. (1990) *Mol. Microbiol.* **4**, 151–156.
38. Anderson,B.E., Baumstark,B.R. and Bellini,W.J. (1990) *Nucleic Acids Res.* **18**, 7168–7168.
39. Rainwater,S. and Silverman,P.M. (1990) *J. Bacteriol.* **172**, 2456–2461
40. Kudo,T., Nagai,K. and Tamura,G. (1977) *Agric. Biol. Chem.* **41**, 97–107.
41. Hiraga,S., Niki,H., Imamura,R., Ogura,T., Yamanaka,K., Feng,J., Ezaki,B. and Jaffé,A. (1991) *Res. Microbiol.* **142**, 189–194.
42. Ciesla,Z., Bagdasarian,M., Szczurkiewicz,W., Przygonska,M. and Klopotowski,T. (1972) *Mol. Gen. Genet.* **116**, 107–125.
43. Hussain,K., Begg,K.J., Salmond,G.P.C. and Donachie,W.D. (1987) *Mol. Microbiol.* **1**, 73–81.
44. Trun,N.J. and Gottesman,S. (1990) *Genes Dev.* **4**, 2036–2047.
45. Bryant, R. B. and Sypherd,P. S. (1974) *J. Bacteriol.* **117**, 1082–1092.
46. Ono,M. and Kuwano,M. (1978) *J. Bacteriol.* **134**, 677–679.
47. Raetz,C.R.H., Hirschberg,C.B., Dowhan,W., Wickner,W.T.and Kennedy,E.P. (1972) *J. Biol. Chem.* **247**, 2245–2247.
48. Isihara,H. and Hogg,R.W. (1980) *J. Biol. Chem.* **255**, 4614–4618.
49. Brosius,J. (1978) *Biochemistry* **17**, 501–508.
50. Osawa,S., Takata,R. and Dekio,S. (1970) *Mol. Gen. Genet.* **107**, 32–38.
51. Yu,A., Bertani,L.E. and Haggîrd-Ljungquist,E. (1989) *Gene* **80**, 1–12.
52. Yang,Y. and Ames,G.F.-L. (1990) In Drlica,K. and Riley,M. (eds) *The Bacterial Chromosome.* American Society for Microbiology, Washington, D.C. pp. 211–225.
53. Goberdhan,P.D., Rudd,K.E., Morgan,M. K., Bayat,H. and Ames,G F.-L. (1992) *J. Bacteriol.* **174**, 4583–4593.
54. Gilson,E., Saurin,W., Perrin,D., Bachellier,S. and Hofnung,M. (1991) *Nucleic Acids Res.* **19**, 1375–1383.
55. Sharples,G.S. and Lloyd,R.G. (1990) *Nucleic Acids Res.* **18**, 6505–6508.
56. Levene,S. and Crothers,D.M. (1983) *J. Biomol. Struct. Dynam.* **1**, 429–435.
57. Gonzalez,J.C., Banerjee,R.V, Huang,S. and Matthews,R. (1992) *Biochemistry* **31**, 6045–6056.
58. Bachmann,B.J. (1987) In Neidhardt,F.C. (ed.) *Escherichia coli and Salmonella typhimurium. Cellular and Molecular Biology.* American Society for Microbiology, Washington, DC. pp. 1190–1219.
59. Smith,R.L., Banks,J., Snavely,M.D. and Maguire,M.E. (unpublished) GenBank accession L11042.
60. Aldea,M., Maples,V. F. and Kushner,S.R. (1988) *J. Mol. Biol.* **200**, 427–438.
61. Berlyn,M. and Letovsky,S. (1992) *Nucleic Acids Res.* **20**, 6143–6151.
62. Schroeder,J.L. and Blattner,F.R. (1982) *Nucleic Acids Res.* **10**, 69–73.