Analysis of the *Escherichia coli* genome. IV. DNA sequence of the region from 89.2 to 92.8 minutes

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

Received August 18, 1993; Revised and Accepted October 12, 1993

ABSTRACT

We present the sequence of 176 kilobases of the *Escherichia coli* K-12 genome, from *katG* at 89.2 to an open reading frame (ORF) of unknown function at 92.8 minutes on the genetic map. This brings the total of contiguous sequence from the *E. coli* genome project to 500 kb (81.5 to 92.8 minutes). This segment contains 134 putative coding genes (ORFs) of which 66 genes were previously identified. Eight new genes—*acs*, *pepE*, and *nrfB-G*—were identified as well as the previously mapped *gldA* and *alr* genes. Still, 58 ORFs remain unidentified despite literature and similarity searches. The arrangement of proposed genes relative to possible promoters and terminators suggests 55 potential transcription units. Other features include 13 REP elements, one IRU (ERIC) repeat, 59 computer-predicted bends, 42 Chi sites and one new grey hole. Sixteen signal peptides were found, including those of *lamB*, *btuB*, and *malE*. Two ribosomal RNA loci, *rrnB* and *rrnE*, are located in this segment, so we have now sequenced four of the seven *E. coli* rRNA loci. Comparison of the rRNA loci reveals some differences in the ribosomal structural RNAs which are generally compatible with the proposed secondary structures.

INTRODUCTION

This is the fourth DNA segment sequenced by the *E. coli* Genome Project (1,2,3). The 176,195 basepairs (bp) described here brings the project total to 500,243 contiguous bp, slightly over 10% of the genome. The density of information gathered for *E. coli* continues to be high, with 98% of the sequence coding for recognizable features. In the entire 500 kb we have identified 420 predicted genes (ORFs) of which 221 so far can be correlated with a known gene and 199 remain anonymous. Given the extensive history of research on *E. coli*, and the demonstrated success of similarity-based gene identification, the high percentage of unidentifiable genes may be surprising; however, similar results are emerging from global sequencing of yeast (4). Systematic experimental approaches to determining function may be needed in addition to systematic sequencing to complete a catalog of the proteins of a living cell. This will be an invaluable resource for the identification of genes in anonymous human sequences using computer-based similarity searches.

In this paper we report the sequence of two ribosomal RNA loci of *E. coli*, *rrnB* and *rrnE*, and in previous segments *rrnC*

and *rrnA* were completed. The ribosomal RNAs of *E. coli* are coded in seven loci dispersed around the chromosome (5). Each locus includes one or more tRNA genes and is oriented in the same direction as replication. Ribosomal transcription accounts for over 50% of *E. coli* RNA synthesis at high growth rates (6). Each locus has two tandemly arranged promoters P1 and P2. Prior to this project, only the *rrnB* locus has been sequenced completely (7), providing the 16S and 23S structural RNA sequences used for evolutionary comparisons and in the extensive experimental and computer work that has led to proposals of detailed secondary structure folding for these RNAs (8, 9, 10, 11). The seven copies are generally assumed to be very similar, although restriction mapping with *Avr* II has shown differences (12) and microheterogeneity among the ribosomal RNAs has been demonstrated by ribosomal RNA fingerprinting data (13).

The question of whether any of the *rrn* loci play a physiological role or whether the seven simply exist to provide a sufficient quantity of ribosomes has been considered over many years (5, 14) both by studying the regulation of each and by considering the relative growth advantage of bacteria with knockout mutations in competition with wild type. The effects observed so far are quite subtle. The *rrnE* gene exhibits differences in regulation and its promoter region sequence differs significantly from the others, yet the elimination of this locus from the genome has virtually no effect on growth relative to controls. Slight growth advantages are found with some of the knockouts in other loci, but no large effects are seen.

We have taken the opportunity in this paper to identify variations between the structural rRNAs, coded by the four loci that we have sequenced, to ascertain whether and how ribosomal RNAs encoded by the various loci might differ.

MATERIALS AND METHODS

The methods used in this project have been described in previous publications (1, 2, 3) although details of their application continue to evolve. In this segment we have increased the accuracy of the primary data determination, since extra effort in that area is repaid in simplification of the work of the analysis and annotation teams. To resolve compressions earlier in the process the resolution team resequences these regions using deoxyinosine triphosphate, at the same stage as sequencing the clones on the opposite strand by the Janus strategy (15). We find dITP resolves at least 90% of compressions we have encountered. The

resolution team also identifies and carries out PCR and cycle sequencing reactions for primer-walk gap filling although the level of redundancy we employ keeps the need for primer walking to a minimum.

The starting material for *E. coli* (strain MG1655) sequencing is a mapped set of clones in bacteriophage lambda-derived vectors (16). For this segment we sequenced 15 overlapping lambda clones. The sequence reported here has all been newly determined.

For prediction of protein coding genes in this segment we used DNASTAR's GENELOT, the graphical annotation tool for ORFs into which M. Borodovsky's GENMARK algorithm (17) has now been incorporated. Borodovsky's method, using the 5th order Markov chain calculation, gives clearly better discrimination for shorter ORFs than the calculations used previously. This software improvement has allowed us to propose ORFs as short as fifty codons with greater confidence when the codon usage statistics and other features were consistent. The validity of this approach was confirmed by the finding that the predicted amino-terminal sequence of o69 (69 residues) matched the experimentally determined amino-terminal sequence of a protein with an appropriate mobility in a 2D gel system, as measured by G. Church and A. Link (personal communication). Other software improvements include better tools for splicing together sequence segments while maintaining the numbering of annotations. A new program was written to search for REP elements which identifies REP sequences by searching for sequence similarity to half-REP sites, then screening these for strength of match and adjacency to the other half site, finally grouping closely spaced sequences for visual inspection. The presence of multiple REP elements in a lambda clone may cause deletions during propagation; regions containing REP elements are always carefully inspected for this problem.

RESULTS AND DISCUSSION

The 176,195 bp sequence presented here has been deposited in the sequence databases and assigned the GenBank accession number U00006. This sequence lies to the right (*i.e.* clockwise) of the first three segments of our project and overlaps the third by 93 bases to ensure that no sequences were missed 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 (18), unless otherwise noted.

Eighty-eight percent of this new sequence codes for either structural RNA or protein. We have identified 134 putative genes (ORFs), including the partial gene *katG* at the left end. Fifty-eight of the ORFs are potential new genes, still unidentified despite searches of the databases for similarities suggestive of function. Six 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.

Newly identified genes

Acetyl-CoA synthetase. A secondary synthetic pathway of acetyl-CoA from acetate has been identified in *E. coli* but better characterized in yeast (19). We find strong similarity between the *Phycomyces blakesleeanus* *facA* gene product and an ORF

of 652 amino acids expressed leftward beginning at residue 152612 of this segment. We term this gene *acs*.

Acs is identical over 330 amino acids to the hypothetical product of *orfX* reported in the *ubiG* 3' region (20). The authors noted the similarity of this predicted protein to fungal acetyl-CoA synthetases. They also noted that their sequence diverged from another sequence of the *ubiG* region at the location of PU sequences (*i.e.*, a REP element). Comparison of our completed sequence with both ECOUBIG (GenBank accession number M87509) and ECOGYRAAM (GenBank accession number Y00544) indicates that a recombination event could have occurred between 2 REP elements sharing a 36 bp identity, one located 3' to *ubiG* and the other located 3' to *acs*, to generate the clone sequenced by Wu *et al.*; it is unclear whether this was a chromosomal rearrangement or an event occurring during growth of the initial cosmid clones. We note that REP sequences seem to have been involved in a deletion in the *malF-malM* region in an isolate of one of our clones (Ec18-33).

Peptidase E. The *Salmonella* peptidase E is a dipeptidase which specifically hydrolyses dipeptides containing N-terminal aspartate residues (21). This is 89% similar at the protein level to the predicted product of *f229* at 95384–94695, and we therefore designate *f229* as *pepE*.

Formate-dependent nitrite reductase. Several ORFs in the group between 153004 and 159380 showed similarities of their predicted products to gene products of other organisms suggestive of electron transfer activity (Table 2). For example, *o552* and *o127* encode proteins which are similar to the *ccl1* and *ccl2* gene products of *Rhodobacter capsulatus*, involved in the ligation of cytochrome c to the apoprotein in the periplasm (22). Recently, the final component in the formate-dependent nitrite reductase pathway of *E. coli* was independently sequenced and identified as *nrfA*, encoding a tetraheme c-type cytochrome (23) which corresponds to the first of our ORFs in this group, *o478*. On further consultation with the authors (J. Cole, personal communication), our sequence data for the whole operon was found to match theirs. This together with their biochemical and genetic data, allowed the assignment of the other genes in this electron transfer operon, *nrfB-G*, to the six other ORFs in this group, *o190* through *o198* using nomenclature suggested by Cole.

Identification of mapped genes

Biosynthetic alanine racemase. *E. coli* is known to possess two alanine racemases under separate regulation- a catabolic enzyme employed in the utilization of alanine, and a biosynthetic gene which interconverts L- to D-alanine needed in cell wall biosynthesis (24, 25). In *Salmonella typhimurium* these proteins are encoded by the *dadB* gene near minute 36 and *alr* near minute 91, both of which have been sequenced (26, 27). Recently, partial sequence data located the *alr* gene of *E. coli* between *dnaB* and *tyrB* (28). The predicted product of our complete ORF in this position (358 amino acids beginning at 131026) shows a strong similarity score of with that of the *S. typhimurium* *alr* product. Together these data confirmed the identity of the *E. coli* *alr*.

Glycerol dehydrogenase. An ORF of 380 amino acids starting at base 3174 shows a high similarity score of 50% over most of its length with the glycerol dehydrogenase protein coded by *gldA* of *Bacillus stearothermophilus* (29). A glycerol dehydrogenase activity has been mapped to the 89 minute region

of *E. coli* (30). This activity has been purified by several groups and partly characterized, but its true substrate(s) and its physiological role remain uncertain (31). More recently the gene was cloned and partial sequence data obtained by one of these groups; their sequence is in agreement with ours (W. Boos, personal communication).

Unidentified genes

Some ORFs show strong similarities to other proteins but the data is insufficient for a firm identification. These are listed in Table 2. A few cases which deserve additional comment are described in this section.

Four of the proposed gene products in the region from 4962 to 12720 show similarity to fructose-specific PTS (phosphoenolpyruvate-dependent sugar phosphotransferase system) proteins of

various species, including *E. coli* (32, 33). The proposed gene *f711* is directed leftward and is related to Enzyme I. *o359*, *o106* and *o113* are directed rightward, are possibly in a single transcription unit, and are each related to Enzyme II which carries the sugar specificity. The predicted products of *o113* and *o106* are similar to one another and also are similar to the middle part of previously sequenced examples of enzyme II, while *o359* is similar to the carboxy terminal region of Enzyme II. This pattern may suggest a domain structure of Enzyme II compatible with coding these sections as separate polypeptides. In the same potential transcription unit as the enzyme II analogue are *o765* and *o315* whose predicted products are similar to pyruvate formate lyase and its activating enzyme, respectively (34).

The predicted product of *o225* (64025–64702) is similar to a mitochondrial protein involved in cytochrome c biogenesis in

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: *AvrII* (A), *BamHI* (B), *BglII* (G), *EcoRI* (R), *EcoRV* (V), *HindIII* (H), *KpnI* (K), *NotI* (N), *PstI* (S) and *PvuII* (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. Promoters marked S or N are regulated by starvation and nitrogen, respectively; unlabelled promoters are regulated by sigma 70. Other sequence determinations spanned by this segment are indicated, with their database accession numbers. This sequence is in GenBank (accession U00006), annotated with these features as well as conflicts with other sequence determinations.

Table 1. Genes and predicted characteristics of deduced protein products

Gene ^a	Site	No. ^b	Endpoints in sequence	first...last codon	molecular size (kD)	aa	pI
<i>katG</i> ^c	14983		1 > 1257	...TAA	80.0	726	5.1
<i>o312</i>			1317 > 2255	GTG...TGA	34.1	312	10.2
<i>f205</i>			2899 < 2282	ATG...TAA	23.0	205	9.0
<i>glcA</i>			4316 < 3174	ATG...TAA	40.1	380	4.8
<i>f220</i>			4950 < 4288	ATG...TAA	23.6	220	5.0
<i>f711</i>			7097 < 4962	GTG...TAA	78.3	711	5.3
<i>o109</i>			7099 > 7428	GTG...TAG	11.8	109	9.8
<i>o359</i>			7771 > 8850	ATG...TAA	37.1	359	7.2
<i>o106</i>			8865 > 9185	ATG...TAA	11.2	106	5.3
<i>o765</i>			9236 > 11533	ATG...TAA	86.0	765	5.7
<i>o315</i>			11430 > 12377	ATG...TAA	35.1	315	8.2
<i>o113</i>			12379 > 12720	ATG...TAG	12.6	113	5.7
<i>f283</i>			13558 > 12707	ATG...TAA	32.2	283	8.2
<i>f577</i>			15506 < 13773	ATG...TAA	66.6	577	7.2
<i>ppc</i>	368		18339 < 15688	ATG...TAA	99.1	883	5.5
<i>argE</i>	1017		20088 < 18937	ATG...TAA	42.4	383	5.7
<i>argC</i>	1019		20242 > 21246	ATG...TAA	36.0	334	5.8
<i>argB</i>	1020		21254 > 22030	ATG...TAA	27.2	258	5.4
<i>argH</i>	1014		22091 > 23464	ATG...TAA	50.3	457	5.0
<i>oxyR</i>			23731 > 24648	ATG...TAA	34.3	305	6.2
<i>udhA</i>			25965 < 24631	ATG...TAA	49.4	444	6.6
<i>o234</i> (yijC)			26309 > 27013	ATG...TAA	26.6	234	9.0
<i>o119</i> (yijD)			27013 > 27372	ATG...TAA	13.0	119	9.4
<i>trmA</i>	79		28512 < 27412	ATG...TAA	42.0	366	5.9
<i>btuB</i>	946		28881 > 30725	ATG...TGA	68.4	614	5.2
<i>murI</i> (dga)			30658 > 31527	ATG...TGA	31.5	289	5.4
<i>rrsB</i>	188		31901 > 33442	(16S rRNA)	1542 bp		
<i>glfT</i>	680		33614 > 33689	(rRNA)	76 bp		
<i>rrfB</i>	202		33883 > 36786	(23S rRNA)	2904 bp		
<i>rrfB</i>	209		36879 > 36998	(5S rRNA)	120 bp		
<i>murB</i>			37299 > 38327	ATG...TGA	37.9	342	6.0
<i>birA</i>	952		38324 > 39289	ATG...TAA	35.3	321	8.0
<i>coaA</i> (rfs)	17731		40268 < 39318	ATG...TAA	36.4	316	6.6
<i>f51</i>			40610 < 40455	ATG...TAA	5.7	51	7.7
<i>tuFB</i>	60		41186 > 42370	ATG...TAA	43.3	394	5.3
<i>secE</i>			42600 > 42983	ATG...TGA	13.6	127	10.6
<i>nusG</i>			42985 > 43530	ATG...TAA	20.5	181	6.6
<i>rplK</i>	255		43688 > 44116	ATG...TAA	14.9	142	9.6
<i>rplA</i>	263		44120 > 44824	ATG...TAA	24.7	234	9.7
<i>rplA</i>	256		45237 > 45734	ATG...TAA	17.7	165	9.0
<i>rplL</i>	254		45801 > 46166	ATG...TAA	12.3	121	4.5
<i>rpoB</i>	233		46486 > 50514	ATG...TAA	150.7	1342	5.1
<i>rpoC</i>	232		50591 > 54814	GTG...TAA	155.2	1407	6.8
<i>hcrC</i>			55027 > 55566	ATG...TAA	21.1	179	9.3
<i>thiH</i>			57109 < 55976	ATG...TGA	43.3	377	6.9
<i>thiG</i>			57951 < 57106	GTG...TGA	29.7	281	6.1
<i>thiF</i>			58799 < 58062	ATG...TAA	26.2	245	4.6
<i>thiE</i>			59445 < 58810	ATG...TGA	23.0	211	5.6
<i>thiC</i>	115		61340 < 59445	ATG...TGA	70.9	631	5.9
<i>f158</i>			62049 < 61573	ATG...TGA	18.2	158	5.8
<i>o257</i>			62144 > 62917	ATG...TGA	29.8	257	5.8
<i>hemE</i>	644		62957 > 64021	ATG...TAA	39.3	354	6.1
<i>o225</i>			64025 > 64702	GTG...TAA	24.9	225	8.3
<i>o196</i>			64745 > 65335	ATG...TAA	22.6	196	4.3
<i>hupA</i>			65522 > 65794	ATG...TAA	9.5	90	9.6
<i>o231</i>			65807 > 66502	GTG...TGA	26.3	231	6.8
<i>f188</i>			67070 < 66504	ATG...TAA	20.4	188	9.7
<i>hydH</i>			67167 > 68564	ATG...TGA	51.0	465	8.8
<i>hydG</i>			68561 > 69886	ATG...TAG	48.4	441	5.5
<i>purD</i>	342		71172 < 69883	ATG...TAA	45.9	429	4.9
<i>purH</i>	338		72773 < 71184	ATG...TAA	57.3	529	5.6
<i>rrsE</i>	185		73388 > 74929	(16S rRNA)	1542 bp		
<i>glvV</i>	678		75015 > 75090	(rRNA)	76 bp		
<i>rrfE</i>	199		75285 > 78188	(23S rRNA)	2904 bp		
<i>rrfE</i>	206		78282 > 78401	(5S rRNA)	120 bp		
<i>o127</i> (yjaA)			78476 > 78859	ATG...TAA	14.4	127	9.2
<i>f147</i> (yjaB)			79365 < 78922	ATG...TAG	16.4	147	4.8
<i>metA</i>	516		79522 > 80451	ATG...TAA	35.7	309	5.0
<i>aceB</i>	1051		80720 > 82321	ATG...TAA	60.3	533	5.4
<i>aceA</i>	1052		82351 > 83655	ATG...TAA	47.5	434	5.1
<i>aceK</i>	17770		83838 > 85574	ATG...TGA	67.7	578	6.3
<i>f728</i> (yjaC)			87729 < 85543	ATG...TAA	82.6	728	8.3
<i>iclR</i>	614		88909 < 88046	ATG...TGA	31.3	287	8.6
<i>metH</i>	509		89070 > 92753	GTG...TGA	136.0	1227	4.9
<i>o543</i>			92973 > 94604	GTG...TAA	59.5	543	7.8
<i>pepE</i>			95384 < 94695	ATG...TAA	24.6	229	5.8
<i>o290</i>			95596 > 96468	ATG...TGA	32.5	290	9.7
<i>f90</i>			96873 < 96601	ATG...TAA	10.5	90	10.2
<i>lysC</i>	539		98475 < 97126	ATG...TAA	48.5	449	4.9
<i>pgi</i>	409		99000 > 100649	ATG...TAA	61.5	549	6.1
<i>o80</i>			101148 > 101390	ATG...TAA	7.4	80	6.3
<i>o222</i>			101474 > 102142	GTG...TGA	25.0	222	5.1
<i>o245</i>			102139 > 102876	ATG...TAA	26.3	245	7.3
<i>o698</i>			102876 > 104972	ATG...TAA	78.5	698	6.3
<i>o136</i> (yjbA)			105567 > 105977	ATG...TAA	15.6	136	8.5
<i>xylE</i>	17776		107496 < 106021	ATG...TAA	53.6	491	8.1
<i>malG</i>	530		108758 < 107868	ATG...TAA	32.2	296	9.0
<i>malF</i>	531		110317 < 108773	ATG...TAA	57.0	514	8.4
<i>malE</i>	532		111661 < 110471	ATG...TAA	43.4	396	5.5
<i>malK</i>	529		112026 > 113141	ATG...TAA	41.0	371	6.5
<i>lamB</i>	575		113213 > 114553	ATG...TAA	49.9	446	4.7
<i>malM</i>	18178		114796 > 115716	ATG...TAA	32.0	306	8.5
<i>o442</i>			116197 > 117525	ATG...TAA	51.7	442	5.9
<i>ubiC</i>	48		117637 > 118245	ATG...TAA	23.1	202	8.0
<i>ubiA</i>	50		118258 > 119130	ATG...TGA	32.5	290	9.0
<i>plsB</i>	382		121768 < 119285	ATG...TAA	93.7	827	8.5
<i>dgaA</i>	862		121879 > 122247	ATG...TAA	13.2	122	6.3
<i>lexA</i>	558		122357 > 122965	ATG...TAA	22.4	202	6.5
<i>dinF</i>	854		122984 > 124363	ATG...TGA	49.7	459	9.8
<i>o69</i>			124479 > 124688	ATG...TAA	8.3	69	5.4
<i>f191</i>			125305 < 124730	GTG...TAA	21.7	191	6.8
<i>o84</i>			125565 > 125819	ATG...TAA	9.8	84	9.8
<i>o235</i>			125843 > 126550	ATG...TAA	26.7	235	4.3
<i>o345</i>			126913 > 127950	ATG...TAA	38.5	345	6.4
<i>o150</i>			127874 > 128326	GTG...TAA	17.4	150	10.6
<i>qor</i>			129475 < 128492	ATG...TAA	35.2	327	8.3
<i>dnaB</i>	850		129558 > 130973	ATG...TAA	52.4	471	4.8
<i>dir</i>	1034		131026 > 132102	ATG...TAA	39.1	358	6.9
<i>tyrB</i>	57		132355 > 133548	GTG...TAA	43.5	397	5.3
<i>o174</i>			134843 > 135367	ATG...TGA	19.6	174	5.8
<i>o138</i>			135478 > 135894	ATG...TAA	15.7	138	6.8
<i>o118</i>			135898 > 136254	ATG...TGA	13.4	118	6.0
<i>uvrA</i>	21		139111 < 136289	ATG...TAA	103.9	940	6.4
<i>ssb</i>	150		139365 > 139901	ATG...TGA	19.0	178	5.4
<i>f116</i>			140350 < 140000	GTG...TGA	13.0	116	8.6
<i>o528</i>			140711 > 142297	ATG...TAA	60.7	528	8.2
<i>soxS</i>			142623 < 142300	ATG...TAA	12.9	107	9.9
<i>soxR</i>			142709 > 143173	ATG...TAA	17.2	154	9.0
<i>o449</i>			143719 > 145068	ATG...TAA	45.7	449	6.7
<i>o549</i>			145220 > 146869	ATG...TAA	60.5	549	5.8
<i>f430</i>			148315 < 147023	ATG...TGA	49.4	430	7.0
<i>f549</i>			150142 < 148493	ATG...TAA	59.2	549	9.5
<i>f104</i>			150453 < 150139	ATG...TGA	11.7	104	7.3
<i>acs</i>			152611 < 150653	ATG...TAA	72.1	652	5.6
<i>rrfA</i>			153004 > 154440	ATG...TAA	53.7	478	7.1
<i>rrfB</i>			154479 > 155051	GTG...TGA	21.0	190	7.1
<i>rrfC</i>			155048 > 155719	ATG...TGA	24.6	223	8.6
<i>rrfD</i>			155716 > 156672	ATG...TAA	35.0	318	9.9
<i>rrfE</i>			156752 > 158410	GTG...TAA	61.3	552	9.1
<i>rrfF</i>			158403 > 158786	ATG...TGA	14.5	127	8.7
<i>rrfG</i>			158783 > 159379	ATG...TAA	22.8	198	8.6
<i>gluP</i>			159721 > 161034	ATG...TAA	47.2	437	8.5
<i>f229</i>			162365 < 161676	ATG...TAA	25.1	229	4.7
<i>fdhF</i>	18400		164606 < 162459	ATG...TAA	79.2	715	6.1
<i>f488</i>			166270 < 164804	ATG...TAA	53.5	488	7.2
<i>f636</i>			168177 < 166267	ATG...TGA	70.9	636	5.2
<i>f343</i>			169348 < 168317	ATG...TAA	36.9	343	7.3
<i>f665</i>			171848 < 169851	ATG...TAA	73.7	665	6.5
<i>f309</i>			173038 < 172109	ATG...TGA	33.8	309	6.8
<i>f231</i>			173717 < 173022	ATG...TAA	26.1	231	5.4
<i>f326</i>			174708 < 173728	ATG...TAA	34.3	326	9.0
<i>f502p</i> ^d			176195 < 174687	...TAA	>55.9	>502	?

^aGene names are those used in the most recent *E. coli* genetic map (18), unless the genes are not present on that map; alternate gene names are indicated in parentheses.

^bMapped genes have been assigned Site numbers in a database maintained by the *E. coli* Genetic Stock Center (55; M. Berlyn, personal communication).

^cThe sequence of *katG* extends beyond the sequence presented here; our previous determination of the balance of the sequence (3) permitted calculation of the protein size and isoelectric point.

^dThe sequence of *f502p* extends beyond the sequence presented here, and the balance of its sequence remains to be determined. The protein size calculated for the part sequenced is indicated.

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
<i>o312</i>	312 aa	P29086 <i>Desulfurolobus ambivalens</i> hypoth. 28.3 kD protein, <i>sor</i> 3' region	25.8% 264 aa
<i>gldA</i>	380 aa	JQ1474 <i>Bacillus stearothermophilus</i> glycerol dehydrogenase	50.3% 362 aa
<i>f220</i>	220 aa	P19669 <i>Bacillus subtilis</i> hypoth. 20 kD protein, <i>tsr</i> 3' region	38.8% 183 aa
<i>f711</i>	711 aa	similar to several phosphotransferase system enzyme I species; example: P30299 <i>Streptococcus salivarius</i> phosphoenolpyruvate-protein phosphotransferase	35.6% 571 aa
<i>o359</i>	359 aa	similar to several phosphotransferase system enzyme II species; example: B40944 <i>Xanthomonas campestris</i> phosphotransferase enzyme II, fructose-specific	46.7% 334 aa
<i>o106</i>	106 aa	similar to several phosphotransferase system enzyme II species; example: P23387 <i>Rhodobacter capsulatus</i> phosphotransferase enzyme II, fructose-specific	43.1% 102 aa
<i>o765</i>	765 aa	GB: X08035 <i>Escherichia coli</i> pyruvate formate-lyase	23.0% 808 aa
<i>o315</i>	315 aa	P09374 <i>Escherichia coli</i> pyruvate formate-lyase activating enzyme	28.1% 278 aa
<i>o113</i>	113 aa	similar to several phosphotransferase system enzyme II species; example: P23387 <i>Rhodobacter capsulatus</i> phosphotransferase enzyme II, fructose-specific	33.7% 95 aa
<i>f283</i>	283 aa	PS00041 Bacterial regulatory proteins, <i>araC</i> family signature similar to several members of the AraC family of positive transcriptional regulators; example: A42400 <i>Streptococcus mutans</i> MsmR regulatory protein	100% match 22.4% 294 aa
<i>f158</i>	158 aa	P15275 <i>Pseudomonas aeruginosa</i> alginate regulatory protein AlgR2	35.7% 157 aa
<i>o225</i>	225 aa	GB: X69555 <i>O. berteriana</i> mitochond. protein involved in cytochrome c biogenesis	23.3% 236 aa
<i>pepE</i>	229 aa	<i>Salmonella typhimurium</i> peptidase E	89.1% 229 aa
<i>o698</i>	698 aa	PS00107 Protein kinases ATP-binding region signature	100% match
<i>f191</i>	191 aa	P06975 <i>Escherichia coli</i> ferric uptake regulation protein	30.6% 144 aa
<i>abr</i>	358 aa	PS00395 Alanine racemase pyridoxal-phosphate attachment site similar to bacterial alanine racemases; example: P06655 <i>Salmonella typhimurium</i> biosynthetic alanine racemase, Alr	100% match 91.1% 359 aa
<i>o174</i>	174 aa	PS00215 Mitochondrial energy transfer proteins signature	100% match
<i>o528</i>	528 aa	S18623 <i>Azorhizobium caulinodans</i> hypoth. protein, <i>ntrC</i> 3' region	26.2% 530 aa
<i>o449</i>	449 aa	PS00017 ATP/GTP-binding site motif A similar to two unidentified ORFs from the <i>E. coli</i> 82 min. genomic region [GB L10328]: ORF f470 ORF f445	100% match 39.5% 431 aa 34.8% 438 aa
<i>o549</i>	549 aa	similar to several eukaryotic Na ⁺ /H ⁺ exchangers; example: P19634 Homo sapiens NA(+)/H(+) exchanger 1	23.6% 585 aa
<i>f549</i>	549 aa	GB: L01139 <i>Salmonella</i> proline permease (<i>putP</i>) gene, 5' end	27.4% 496 aa
<i>acs</i>	652 aa	PS00455 Putative AMP-binding domain signature PS00018 EF-hand calcium-binding domain similar to several acetyl-coenzyme A synthetases; example: GB: M94729 <i>Phycomyces blakesleeana</i> acetyl-coenzyme A synthetase (<i>facA</i>) gene	100% match 100% match 52.4% 654 aa
<i>nrfA</i>	478 aa	PS00190 Cytochrome c family heme-binding site signature	4x 100% matches
<i>nrfB</i>	190 aa	PS00190 Cytochrome c family heme-binding site signature	5x 100% matches
<i>nrfC</i>	223 aa	PS00198 4Fe-4S ferredoxins, iron-sulfur binding region signature PS00178 Aminoacyl-transfer RNA synthetases class-I signature	100% match 100% match
<i>nrfD</i>	318 aa	S23458 <i>Wolinella succinogenes</i> polysulfide reductase chain ParB	54.0% 189 aa
<i>nrfE</i>	552 aa	S23459 <i>Wolinella succinogenes</i> polysulfide reductase chain ParC PS00327 Bacterial rhodopsins retinal binding site	40.0% 315 aa 100% match
<i>nrfF</i>	127 aa	S23667 <i>Rhodobacter capsulatus</i> gene ccl1 protein	37.0% 570 aa
<i>nrfG</i>	198 aa	S23668 <i>Rhodobacter capsulatus</i> gene ccl2 protein	34.9% 129 aa
<i>f488</i>	488 aa	P19737 <i>Synechococcus</i> sp. hypoth. 28.7 kD protein, <i>recA</i> 3' region PS00443 Glutamine amidotransferases class-II active site	25.4% 185 aa 100% match
<i>f636</i>	636 aa	P24126 <i>Pseudomonas cepacia</i> fusaric acid resistance protein FusA	27.3% 395 aa
<i>f343</i>	343 aa	P15710 <i>Neurospora crassa</i> phosphate-repressible phosphate permease	22.1% 357 aa
<i>f665</i>	665 aa	P27303 <i>Escherichia coli</i> multidrug resistance protein A PS00743 Beta-lactamases class B signature 1 PS00225 Crystallins beta and gamma "Greek key" motif signature JC1118 <i>Pseudomonas</i> sp. alkyl sulfatase	27.6% 351 aa 100% match 100% match 32.1% 527 aa
<i>f309</i>	309 aa	P16557 <i>Bacillus subtilis</i> xylose repressor	23.6% 314 aa
<i>f231</i>	231 aa	GB: M64172 <i>Alcaligenes eutrophus</i> pHG1 D-ribulose-5-phosphate 3 epimerase	35.2% 219 aa
<i>f326</i>	326 aa	P04984 <i>Escherichia coli</i> high affinity ribose transport system component RbsC	39.0% 323 aa
<i>f502+</i>	>502 aa	PS00017 ATP/GTP-binding site motif A P23199 <i>Escherichia coli</i> galactoside-binding protein MglA	100% match 40.8% 502 aa

Sequences from the Swissprot (Release 25), NBRF-PIR (Release 36) and translated GenBank (Release 77) 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 (56). The products of previously sequenced and characterized genes are not listed in this table.

Figure 2. Predicted structures of mature RNA regions where differences between the loci affect base pairing. Bases marked * could be Gu instead of uG (data ambiguous). *rrmB* is the reference, numbered as in the RNA database (see text).

Oenothera berteriana. Interestingly, *o225* is adjacent to *hemE*, whose gene product is involved in porphyrin biosynthesis (35). Finally, this segment also contains *f191* at 124730–125305 whose predicted product is similar to the *E.coli* ferric uptake regulation protein (36).

Signal peptides

A computer search for signal peptide-like sequences identified sixteen candidates. Thirteen were in ORFs of unknown function: *o80*, *f549*, *o127*, *f229*, *o245*, *f665*, *f577*, *o190*, *o478*, *o134*, *o222*, *o106*, and *o119*. The predicted products of *o134* and *o106* may be membrane proteins, suggested by their similarity to the Enzyme II proteins in the PTS system (31). The predicted *f665* product is similar to alkyl sulfatase (37) which could reasonably be a secreted protein. Three genes whose products contain the consensus sequence were already characterized as periplasmic or outer membrane proteins: the *lamB*, *malE*, and *btuB* genes (38, 39, 40).

Ribosomal RNA analysis

To date, study of the 16S and 23S ribosomal RNAs has relied on the DNA sequence of the *rrmB* locus determined by Brosius (7). This sequence was used to predict the folding of these structural RNAs. Our determination of *rrmB* agrees exactly with this previous work. It is therefore of interest to identify variations between *rrmB* and the other loci that we have sequenced (*rrmA*, *rrmC* and *rrmE*) and to consider how these alterations might fit into the secondary structure folding proposed for the *rrmB* ribosomal RNA sequences. The beginning and ending residues of each of the structural RNAs are noted in Table 1.

The 16S and 5S structural RNAs coded by *rrmE* are identical to *rrmB*. The 16S RNA of *rrsA* is also identical to these. A number of differences were found in the 23S subunit RNAs coded by *rrlA*, *rrlC*, and *rrlE*, the 16S subunit of *rrsC* and the 5S RNAs of *rrfA* and *rrnC*. For simplicity in the discussion will use *rrmB*

as the reference, with the numbering system for each mature RNA as adopted by the RNA database (41). There are a few insertions and deletions that alter the lengths of the *rrlA* and *rrsC* sequences, shifting the numbering slightly in local regions. Residues that are different from *rrmB* are printed with capital letters. Examples of regions exhibiting differences in base pairing are shown in Figure 2.

23S RNA coded by *rrlA*. *rrlA* codes for a 23S RNA that is one base longer than the standard *rrlB* due to a missing u at position 1173, and insertion of a C following residue 1868 and a G following 1873. In addition there are 14 single base changes which do not disrupt the proposed secondary structure. In the 301–316 stem-loop a c > U at 316 changes a g:c to a g:U basepair in the stem. In the 1164–1185 stem-loop compensating changes g > A at 1171 and c > U at 1178 preserve the stem and the missing u at 1173 shortens the loop from 4 to 3 bases. At 1211 a c > U changes a residue in an unpaired bulge with no apparent effect on the secondary structure. In the 1722–1738 stem-loop there are eight changes: agguccucgcggaugg > aAgCGAcuUgcUCGugg. Six of these compensate for one another so the stem is unaffected (Figure 2a). At the tip of the 5 base loop there is a c > U change. In the 1865–1877 stem-loop the sequence uagcgcaagcgaa is altered by insertion of an additional g:c basepair into the stem as follows uagc/C/gcaa/G/gcgaa lengthening it from 2 to 3 residues (Figure 2b). Finally, in the stem-loop from 2791 to 2805 gaccuuuaaggguc is altered to gacUcCuuGagAguc (Figure 2c). These four single base changes shorten the stem from six to five basepairs and lengthen the loop from three to five bases.

23S RNA coded by *rrlC*. There are 18 single base changes in *rrlC* relative to *rrlB*. Two of the most substantial variations found in *rrlA* are also present in *rrlC*. These are the 7 base changes in the 1722–1738 stem-loop and the 4 changes in the 2791–2805

Table 3.

Alteration	Sequence change	Description (ECOUW85U coordinates)
resolve ambiguity in <i>rrsA</i>	GAGGGGkACCTTCG to GAGGGGgACCTTCG	change K to G at position 88788
resolve ambiguity in <i>rrsA</i>	TGCATGssTGTCGT to TGCATGgcTGTCGT	change SS to GC at position 89642–89643
resolve ambiguity in <i>rrlA</i>	AAAGGCGyGCGATA to AAAGGCGtGCGATA	change Y to T at position 90879
		Description (ECOUW82 coordinates)
lengthen <i>o519</i> (<i>kup</i>)	GAGCGGGTcGATTC to GAGCGGGTGATTC	delete C at position 122219
lengthen <i>o519</i> (<i>kup</i>)	GGTTGCGAGAAACG to GGTTGgCGAGAAACG	insert G at position 122330/122331
lengthen <i>o519</i> (<i>kup</i>)	GATCCTCGCAAACG to GATCCTCgGCAAACG	insert G at position 122438/122439
correction in <i>rrsC</i>	CCGCAAGTTAAAA to CCGCAAgGTTAAAA	insert G at position 132093/132094
correction in <i>rrlC</i>	TATGGGgGGGACCA to TATGGGGGGACCA	delete A at position 133496
correction in <i>rrlC</i>	GCGAAAGCtAGGTC to GCGAAAGCAGGTC	delete T at position 135469
correction in <i>rrlC</i>	GTACTCCGGGATA to GTACTCCgGGGATA	insert G at position 135532/135533
correction in <i>rrlC</i>	ACTCCTTTrAGrGT to ACTCCTTgAGaGT	delete T at position 135885
		change R to G at position 135888
		change R to A at position 135891

Merger specification:

MG1655v4 = ECOUW82 (1,122218) + ECOUW82 (122220,122330) + "G" + ECOUW82 (122331,122438) + "G" + ECOUW82 (122439,132093) + "G" + ECOUW82 (132094,133495) + ECOUW82 (133497,135468) + ECOUW82 (135470,135532) + "G" + ECOUW82 (135533,135884) + ECOUW82 (135886,135887) + "G" + ECOUW82 (135889,135890) + "A" + ECOUW82 (135892,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 (68623,88787) + "G" + ECOUW85U (88789,89641) + "GC" + ECOUW85U (89644,90878) + "T" + ECOUW85U (90880,REND) + ECOUW87 (4,96430) + ECOUW89 (40,REND) .

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. The merger specification is in the DNASTAR splicing language (57), and include both the changes noted above and previously reported changes (3) not yet updated in the database entries. ECOUW82 is accession L10328 (2), ECOUW85U is accession M87049 (1), ECOUW87 is accession L19201 (3), and ECOUW89 is accession U00006.

stem-loop noted above. In addition *rrlC* has differences where stretches of residues 281–cagugugu–288...352–acaugcug–359 fold back to pair with one another. The *rrlC* sequence is 281–caguAugu–288...352–acauAUug–359, which pairs slightly better. At position 1865 there is a u>C difference which changes a u:A basepair to u:C which would destabilize the stem of the 1855–1887 stem-loop. This area of the RNA varies in a different way in *rrlA* (Figure 2b). A u>G change at 2203 and a>G at 2211 would appear to have no impact on the ability of the 2200–2223 stem-loop to form. The 2520–2545 stem-loop is disrupted by a g>A at 2525.

23S RNA coded by *rrlE*. This sequence is identical to *rrlB* with one exception. *rrlE* also carries the 4 base changes of the 2791–2809 stem-loop present in *rrlA* and *C*.

16S rRNA coded by *rrsC*. The *rrsC* 16S ribosomal RNA has 5 single base alterations. Four of these change the stem-loop 76–gaagaagcuugcuuuu–93 to either 76–gaaACagcuugc–uGuuuC–93 or 76–gaaACagcuugcGuuuuC–93; because of a sequencing ambiguity we cannot be sure which. The first variant strengthens the proposed H bonding of the stem whereas the second variant significantly weakens it (Figure 2d). At 226 a g>A difference changes a g:u pair to a slightly better A:u pair in a proposed hydrogen bonded region.

5S RNAs. In *rrfA* there is a c>T change at position 92. In *rrfC* there is an a>C change at position 12. Otherwise all four 5S molecules are the same.

It is not known whether any of the structural differences seen could influence ribosomal function. All of the variations we found occur at positions where variation is seen phylogenetically. It is

interesting that the structural RNAs of *rrnE* are so close in sequence to those of *rrnB*, yet *rrnE* can be genetically inactivated without effect, while *rrnB* inactivation led to a strain that could be out-competed by wild type (5). These subtle effects are most likely related to several previously known differences in the promoter regions, spacer regions, or terminators (7).

***rrnH*.** After this manuscript had been submitted K. Miyamoto's GenBank entry ECORRNHK12 appeared (accession no. D15061) containing the 5' half of *rrsH*. This completed the *rrnH* locus, that had been partially sequenced previously (T. Nakayashiki, T. Adachi, T. Miki, and H. Inokuchi; ECORRNH, accession no. D12649). Comparison with these sequences shows the following: the 5S gene *rrfH* is the same as *rrfA*. 23S RNA coded by *rrlH* has loops at 1720 and 2795 like *rrlA* and *rrlC* and the single base change at 1211 is T, like *rrlA* rather than C as in *rrlB*, *rrlC* and *rrlE*. In addition there are 4 single base differences between *rrlH* and the other four *rrl* loci, located at positions 542 (c>T), 846 (t>C), 1220 (g>A) and 1229 (c>T). The latter two are complementary changes in a stem. The 16S RNA gene *rrsH* apparently codes for a 16S rRNA that is three residues shorter than the others as a result of single base deletions at positions 44, 192, and 418. In addition the complex stem-loop 1002–ggaagttttcagagatgagaatgtgccttcggaacc–1038 is changed to AgaaCtttCagagatgGATTGgtgccttcggaacT altering its secondary structure considerably.

Other features

Transcription signals. Promoters and terminators were identified by computer searches, assessed individually by eye and assigned to the ORFs taking into account transcription data already in the literature wherever possible. Their arrangement indicated 55 transcription units (Figure 1). Seventeen 'alternate' promoters

were assigned, twelve regulated by *rpoS*, in conditions of starvation and stationary phase, and nine by sigma 54, regulated by nitrogen availability. A sigma 32 (heat shock) regulated promoter was identified for *o549*, consistent with observed heat shock induction of transcription in this region (42).

Non-coding features. (All shown in Figure 1) Thirteen REP (Repeated Extragenic Palindromic) elements (43) were identified by a new search program and some human assessment. Eight of these were previously known, either experimentally or by sequence analysis (44, 45) and 5 are new. The largest we have observed, at 161076–161675, contains eleven REP sequences; this had been only partially sequenced before. One ERIC or IRU element (enterobacterial repetitive intergenic consensus or intergenic repeat unit) was identified at 44890–45014. The total numbers of these repeat elements found in the 500kb segment are 37 REPs and 3 IRUs.

The computer predicted 59 static bends in the DNA greater than 72 degrees, calculated as described previously. The orientation of genes and Chi sites in this segment is consistent with our previous findings (2). 42 Chi sites were found, of which 32 are oriented with the direction of replication, 33 with translation, and 29 with both replication and translation. Of the 134 genes in this segment, 83 (62%) are oriented with replication. In the 500 kb segment 259 out of 420 genes (also 62%) are oriented with replication. Finally, despite all efforts we were unable to find any features in the 1.29 kb segment 133550–134843, and therefore designate this area as a grey hole (1).

Differences from published sequence

There are several differences between the published sequence of the *thi* operon (46) and its corresponding GenBank entry M88701. Our determination differs from both of these and our annotations of *thiF* and *thiG* are different from the published gene assignments. The similarity between *thiF* and *chlN* (47) was helpful in assigning the start of the *thiF* protein and its reading frame. The start of *thiG* was chosen on the basis of codon usage statistics and uses the third of the possible start codons. These assignments are consistent with the experimental SDS-PAGE sizings of these proteins (46).

Our sequence of *hemE* predicts a protein that is one amino acid longer than that of the previous determination (34) and has three short segments which read in a different frame. One of these segments shows greater similarity than that previously published to the predicted *hemE* products of *B.subtilis* (BACHEMEHY; M97208) and *Synechococcus* (SYNHEME; Z11705). The other two segments are in regions which are not apparently conserved.

The *hydG* gene also has numerous differences from the previously published sequence (48). Where these sequences differ, our amino acid sequence matches that of *S.typhimurium* HydG (49).

The sequence between *aceK* and *iclR* containing *f728* (85575–88047) was previously determined (50). However, in that sequence, a 53 bp *Hind* III fragment near the end of the ORF was missing. In addition, the region contained many sequence differences including two ~50 bp regions of apparently unrelated sequence and an extra copy of a 25 bp sequence.

Our *metH* sequence differs from both previous determinations (ECOMETH, GenBank J04975, and ECMTHM, GenBank X16584), with a change in reading frame extending the length (and sequence) of the resulting protein at the COOH-terminus.

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.

Our examination of the ribosomal RNA loci led to the resolution of several sequence ambiguities in ECOUW85U (GenBank M87049; 1) and ECOUW82 (GenBank L10328; 2), as well as sequence corrections to *rrsC* and *rrlC* in ECOUW82. Also in ECOUW82, the ORF *o519* has now been identified as *kup* (or *trkD*), encoding a minor potassium uptake system (51; GenBank entry ECKUPRBS, accession no. X68551). We have reassessed the three regions in our sequence (all compressions) where we differed from their data, and have corrected our sequence; as a result *o519* becomes *o622* (*kup*, or *trkD*). The sequence changes are all detailed in Table 3.

The gene *prlB*, noted in (1) as having been mapped to this general region but not correlated with a particular ORF, is actually an allele of *rbsB* (52). We previously described two frameshifts in *atpA* relative to all five database entries covering the gene (2); our sequence agrees with a revision (53) which apparently was never entered in the databases. A recent report (54) shows that both *E.coli* MG1655 and W3110 have a frameshift mutation in *rph*, encoding RNase PH. This corresponds to an annotated difference between our sequence (2) and other database entries.

ACKNOWLEDGEMENTS

This is paper 3377 from the Laboratory of Genetics, supported by award HG000301 from the NIH Human Genome Project. We thank M.Kröger, K.Rudd, A.Link and G.Church for their continuing help with databases; N.Trun, S.Emr and T.Silhavy for helping track down the identification of *prlB*; B.Baum for bringing the *atpA* correction to our attention; E.P.Bakker, M.Schleyer, C.Miller, W.Boos and J.Cole for sharing data prior to publication and H.Noller, V.Stewart, C.Merriman and other members of the *E.coli* community for helpful discussions. For software development and assistance we thank M.Borodovsky, G.Bouriaikov, J.Shavlik, and the programming staff of DNASTAR. Excellent technical support was provided by B.Fritz, K.Kadner, M.Maguire, R.Mikkelsen, C.Moynihan, D.Rose, A.Schmidt, E.Sommers, S.Subramanian, 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, A.Broah, J.Champ, H.Cheng, W.Davis, K.Day, H.Demrow, T.Delaney, D.Duescher, J.Ehley, K.Griswold, E.Grotbeck, L.Hammes, J.Hobbes, D.Hornung, J.Kamaruddin, J.Katcha, Heather Kirkpatrick, Heidi Kirkpatrick, K.Li, A.Mohamed, Z.Othman, N.Parwan, M.Polka, R.Ramamurthy, S.Rhode, T.Ryan, P.Scheskenbach, J.Schmidt, L.Schroeder, K.Streif, W.Wilson, C.Whipperman, 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. Plunkett,G.P.III, Burland,V., Daniels,D.D. and Blattner,F.R. (1993) *Nucleic Acids Res.* 21, 3391–3398.

4. Bork, P., Ouzounis, C., Sander, C., Scharf, M., Schneider, R., Sonnhammer, E. (1992) *Nature* **358**, 287.
5. Condon, C., Philips, J., Fu, Z.-Y., Squires, C. and Squires, C.L. (1992) *EMBO* **11**, 4175–4185.
6. Bremer, H. and Dennis, P.P. (1987) In Neidhardt, F.C. (ed.) *Escherichia coli and Salmonella typhimurium. Cellular and Molecular Biology*. American Society for Microbiology, Washington, DC. pp 1527–1542.
7. Brosius, J., Dull, T.J., Sleeter, D.D. and Noller, H.F. (1981) *J. Mol. Biol.* **148**, 107–127.
8. Noller, H.F., (1984) *Annu. Rev. Biochem.* **53**, 119–162.
9. Woese, C.R., Gutell, R., Gupta, R., and Noller, H.F. (1983) *Microbiol. Rev.* **47**, 621–669.
10. De Rijk, P., Neefs, J.-M., Van de Peer, Y., and De Wachter, R. (1992) *Nucleic Acids Res.* **20**, 2075–2089.
11. Gutell, R.R., Schnare, M.N., and Gray, M.W. (1992) *Nucleic Acids Res.* **20**, 2095–2109.
12. Daniels, D.L. (1990) *Nucleic Acids Res.* **18**, 2649–2651.
13. Carbon, P., Ehresmann, C., Ehresmann, B., Ebel, J.P. (1979) *Eur. J. Biochem.* **100**, 399–410.
14. Ellwood, M. and Nomura, M. (1980) *J. Bacteriol.* **143**, 1077–1080.
15. Burland, V., Daniels, D.L., Plunkett, G., III and Blattner, F.R. (1993) *Nucleic Acids Res.* **21**, 3385–3390.
16. Daniels, D. (1990) In Drlica, K. and Riley, M. (eds.), *The Bacterial Chromosome*. American Society for Microbiology, Washington, D.C., pp. 43–52.
17. Borodovsky, M. and McIninch, J. (1993) *Computers & Chemistry* **17**, 123–133.
18. Bachmann, B.J. (1990) *Microbiol. Rev.* **54**, 130–197.
19. Connerton, I.F., Fincham, J.R.S., Sandeman, R.A. and Hynes, M.J. (1990) *Mol. Micro.* **4**, 451–460.
20. Beckman, D.L., Trawick, D.R. and Kranz, R.G. (1992) *Genes and Dev.* **6**, 268–283.
21. Conlin, C.A., Håkansson, K., Liljas, A. and Miller, C.G. (1993) *J. Bacteriol.*, in press.
22. Wu, G., Williams, H.D., Zamanian, M., Gibson, F. and Poole, R.K. (1992) *J. Gen. Micro.* **138**, 2101–2112.
23. Darwin, A., Hussain, H., Griffiths, L., Grove, J., Sambongi, Y., Busby, S. and Cole, J. (1993) *Mol. Micro.* **9**, 1255–1263.
24. Wijsman, H.J.W. (1972) *Genet. Res.* **20**, 269–277.
25. Wild, J., Hennig, J., Lobočka, M., Walczak, W. and Kłopotowski, T. (1985) *Mol. Gen. Genet.* **198**, 315–322.
26. Galakatos, N.G., Daub, E., Botstein, D. and Walsh, C.T. (1986) *Biochemistry* **25**, 3255–3260.
27. Wasserman, S.A., Daub, E., Grisafi, P., Botstein, D. and Walsh, C.T. (1984) *Biochemistry* **23**, 5182–5187.
28. Lilley, L., Stamford, N.P.J., Vasudevan, S.G. and Dixon, N.E. (1993) *Gene* **129**, 9–16.
29. Mallinder, P.R., Pritchard, A. and Moir, A. (1992) *Gene* **110**, 9–16.
30. Sprenger, G.A., Hammer, B.A., Johnson, E.A. and Lin E.C.C. (1989) *J. Gen. Micro.* **135**, 1255–1262.
31. Jin, R.Z., Tang, J.T.-C. and Lin, E.C.C. (1983) *J. Mol. Evol.* **19**, 429–436.
32. Meadow, N.D., Fox, D.K. and Roseman, S. (1990) *Annu. Rev. Biochem.* **59**, 497–542.
33. Wu, L.-F., Tomich, J.M., and Saier, Jr., M.H. (1990) *J. Mol. Biol.* **213**, 687–703.
34. Knappe, J. and Sawers, G. (1990) *FEMS Microbiol. Rev.* **75**, 383–398.
35. Kano, Y. and Imamoto, F. (unpublished) GenBank accession D12624.
36. Schaffer, S., Hantke, K. and Braun, V. (1985) *Mol. Gen. Genet.* **200**, 100–113.
37. Davison, J., Brunel, F., Phanopoulos, A., Prozzi, D. and Terpstra, P. (1992) *Gene* **114**, 19–24.
38. Heller, K., Mann, B.J., Kadner, R.J. (1985) *J. Bacteriol.* **161**, 896–903.
39. Szmecman, S. and Hofnung, M. (1975) *J. Bacteriol.* **124**, 112–118.
40. Kellerman, O. and Szmecman, S. (1974) *Eur. J. Biochem.* **47**, 139–149.
41. Olsen, G.J., Overbeek, R., Larsen, N., Marsh, T.L., McCaughey, M.J., Maciukenas, M.A., Kuan, W.-M., Macke, T.J., Xing, Y., Woese, C.R. (1992) *Nucleic Acids Res.* **20**, 2199–2200.
42. Chuang, S.-E. (1993) Ph.D. Thesis, University of Wisconsin.
43. 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.
44. Gøberdhan, P.D., Rudd, K.E., Morgan, M. K., Bayat, H. and Ames, G.F.-L. (1992) *J. Bacteriol.* **174**, 4583–4593.
45. Gilson, E., Saurin, W., Perrin, D., Bachellier, S. and Hofnung, M. (1991) *Nucleic Acids Res.* **19**, 1375–1383.
46. Vander Horn, P.B., Backstrom A.D., Stewart, V. and Begley, T.P. (1993) *J. Bacteriol.* **175**, 982–992.
47. Johnson, J.L., Wuebbens, M.M. and Rajagopalan, K.V. (1989) *J. Biol. Chem.* **264**, 13440–13447.
48. Stoker, K., Reijnders, W.N.M., Oltmann, L.F. and Stouthamer, A.H. (1989) *J. Bacteriol.* **171**, 4448–4456.
49. Chopra, A.K., Peterson, J.W., and Prasad, R. (1991) *Biochim. Biophys. Acta.* **1129**, 115–118.
50. Galinier, A., Bleicher, F., Negre, D., Perriere, G., Duclos, B., Cozzone, A.J. and Cortay, J.C. (1991) *Gene* **97**, 149–150.
51. M. Schleyer and E.P. Bakker, *J. Bacteriol.* **175**, 6925–6931.
52. Benson, S.A., Hall, M.N. and Silhavy, T.J. (1985) *Ann. Rev. Biochem.* **54**, 101–134.
53. Stan-Lotter, H., Clarke, D.M. and Bragg, P.D. (1986) *FEBS Lett.* **197**, 121–124.
54. Jensen, K.J. (1993) *J. Bacteriol.* **175**, 3401–3407.
55. Berlyn, M. and Letovsky, S. (1992) *Nucleic Acids Res.* **20**, 6143–6151.
56. Fuchs, R. (1991) *Comput. Applic. Biosci.* **7**, 105–106.
57. Schroeder, J.L. and Blattner, F.R. (1982) *Nucleic Acids Res.* **10**, 69–73.