The Complete Genome Sequence of *Escherichia coli* K-12

Frederick R. Blattner,* Guy Plunkett III,* Craig A. Bloch, Nicole T. Perna, Valerie Burland, Monica Riley, Julio Collado-Vides, Jeremy D. Glasner, Christopher K. Rode, George F. Mayhew, Jason Gregor, Nelson Wayne Davis, Heather A. Kirkpatrick, Michael A. Goeden, Debra J. Rose, Bob Mau, Ying Shao

The 4,639,221–base pair sequence of *Escherichia coli* K-12 is presented. Of 4288 protein-coding genes annotated, 38 percent have no attributed function. Comparison with five other sequenced microbes reveals ubiquitous as well as narrowly distributed gene families; many families of similar genes within *E. coli* are also evident. The largest family of paralogous proteins contains 80 ABC transporters. The genome as a whole is strikingly organized with respect to the local direction of replication; guanines, oligonucleotides possibly related to replication and recombination, and most genes are so oriented. The genome also contains insertion sequence (IS) elements, phage remnants, and many other patches of unusual composition indicating genome plasticity through horizontal transfer.

Because of its extraordinary position as a preferred model in biochemical genetics, molecular biology, and biotechnology, *E. coli* K-12 was the earliest organism to be suggested as a candidate for whole genome sequencing (*1*, *2*). The availability of the complete sequence of *E. coli* should stimulate further research toward a more complete understanding of this important experimental, medical, and industrial organism. Since the inception of the *E. coli* project, six other complete genomes have become publicly available (*3*). Genome sequences, especially those of well-studied experimental organisms, help to integrate a vast resource of biological knowledge and serve as a guide for further experimentation. Availability of the complete set of genes also enables global approaches to biological function in living cells (*4*) and has led to new ways of looking at the evolutionary history of bacteria (*5*).

Escherichia coli is an important component of the biosphere. It colonizes the lower gut of animals, and, as a facultative anaerobe, survives when released to the natural environment, allowing widespread dissemination to new hosts (*6*). Pathogenic *E. coli* strains are responsible for infections of the enteric, urinary, pulmonary, and nervous systems. We chose strain MG1655 as the representative to sequence because it has been maintained as a laboratory strain with minimal genetic manipulation, having only been cured of the temperate bacteriophage lambda and F plasmid by ultraviolet light and acridine orange, respectively (*7*). We now know that these treatments resulted in a frameshift mutation at the end of *rph*, causing low expression of the downstream gene *pyrE* and, in turn, a pyrimidine starvation phenotype (*8*). In addition, a mutation in *ilvG* disrupts one of the isoleucine-valine biosynthesis pathways in all K-12 isolates (*9*). Finally, almost all K-12 derivatives, including MG1655, carry the *rfb-50* mutation, where an IS5 insertion results in the absence of O-antigen synthesis in the lipopolysaccharide (*10*). It will be interesting to compare strain MG1655 with the K-12 strain W3110, which has been carried through more experimental treatments and is being sequenced in Japan (*11*).

F. R. Blattner, G. Plunkett III, N. T. Perna, J. D. Glasner, G. F. Mayhew, J. Gregor, N. W. Davis, H. A. Kirkpatrick, M. A. Goeden, D. J. Rose, B. Mau, and Y. Shao are at the Laboratory of Genetics, University of Wisconsin–Madison, 445 Henry Mall, Madison, WI 53706, USA. C. A. Bloch and C. K. Rode are in the Department of Pediatrics, University of Michigan School of Medicine, 1150 West Medical Center Drive, Ann Arbor, MI 48105, USA. V. Burland is at FMC Bioproducts, 191 Thomaston Street, Rockland, ME 04841, USA. M. Riley is at Marine Biological Laboratories, Woods Hole, MA 02543, USA. J. Collado-Vides is at the Centro de Investigación sobre Fijación de Nitrógeno, Universidad Nacional Autónoma de México, Cuernavaca A.P. 565-A, Morelos 62100, México.

*To whom correspondence should be addressed. E-mail: ecoli@genetics.wisc.edu

Sequencing Strategy

Sequencing was carried out in sections, with steadily improving technical approaches. The M13 Janus shotgun strategy proved to be the most efficient strategy for data collection and closure. It involved initial random sequencing at a four- to fivefold redundancy in the Janus vector (*12*), followed by computerized selection of templates to be resequenced from the opposite end, followed by limited primer walking.

The first 1.92 Mb (*13*, *14*), positions 2,686,777 to 4,639,221 [in base pairs (bp)], was sequenced from our overlapping set of 15- to 20-kb MG1655 lambda clones (*15*) by means of radioactive chemistry and was deposited in GenBank between 1992 and 1995. Subsequently, we switched to dye-terminator fluorescence sequencing (Applied Biosystems). In addition to greater speed and lower cost, this new technology avoided electrophoretic compression artifacts, which, owing to its 50.8% G+C content, occur in practically every gene of *E. coli*. For the next segment (positions 2,475,719 to 2,690,160), we obtained DNA for sequencing by the popout plasmid approach (*16*), in which nonoverlapping segments were excised directly from the chromosome in circular form, gel-purified, and shotgunned for sequencing. The largest portion of the genome (positions 22,551 to 2,497,976) was sequenced from M13 Janus shotguns prepared from 11 I–Sce I fragments of ~250 kb (*17*). Among the many advantages of the I–Sce I method are the ability to select the size of fragment to be shotgunned, elimination of redundant sequencing at the borders between segments, and the reliability inherent in sequencing DNA without intermediate cloning steps. Because the DNA is never amplified, genes that might be deleterious when present in multicopy form are not subject to rearrangements or deletions. Each I–Sce I fragment shotgun contained 15 to 30% random clones from elsewhere in the genome, which apparently arose from randomly sheared genomic fragments comigrating in the pulsed-field gel.

The final stages entailed special attention to problem areas. The region between positions 0 and 22,551 did not yield a suitable I–Sce I fragment, so three lambda clones were selected to finally complete the genome. One of them was found to contain a deletion and had to be finished by shotgun sequencing of a long-range polymerase chain reaction (PCR) fragment (*18*). Other areas of the genome were also resequenced in this way. In total, long-range PCR (*18*) was used to close 36.9 kb of gaps, with amplimers used directly as sequencing templates or as source material for shotguns.

The completed sequence was deposited in GenBank on 16 January 1997; in that sequence 168 ambiguity codes reflected uncertainties in the original determination. While this manuscript was in review, additional PCR sequencing was undertaken to resolve all of these ambiguous residues, and the affected annotations were updated accordingly.

Annotation

Annotation is an ongoing task whose goal is to make the genome sequence more useful by correlating it with other knowledge. Specifically, we attempted to (i) identify genes, operons, regulatory sites, mobile genetic elements, and repetitive sequences in the genome; (ii) assign or suggest functions where possible; and (iii) relate the *E. coli* sequence to other organisms, especially those for which complete genome sequences are available. Currently, the annotation includes 4288 actual and proposed protein-coding genes, and one-third of these genes are well characterized. Postulation of genes in uncharacterized base sequences was surprisingly difficult. They were selected from among the numerous available open reading frames (ORFs) on the basis of codon usage statistics, sequence searches versus SWISS-PROT release 34, Link's database of NH_2-terminal peptide sequences from *E. coli*, computer prediction of signal peptides, upstream matches to the Shine-Delgarno ribosome binding site, and other information including personal communications from colleagues (*19*). Assignment of NH_2-termini posed special problems because most ORFs contain multiple in-frame start codons. In the absence of other information, we generally selected the ORF with the longest possible NH_2-terminus. This method preserves the most coding information for analysis, but it may not reflect the situation in vivo.

Functions of previously known *E. coli* proteins were collected from the GenProtEC (*20*) and EcoCyc (*21*) databases. The function of new translated sequences was imputed from sequence similarity (*22*). Each gene (including stable RNA genes) in the sequence was assigned a unique numeric identifier beginning with a lowercase "b"; when no name has been assigned to a given gene, it is referred to by this number. A specific physiological role was assigned if most of the hits were for a specific function such as alcohol dehydrogenase, but if the substrates varied among the hits, the common denominator (for example, permease or kinase) was assigned to the ORF, substrate specificity unknown. If less specificity was found among the hits, a general function was assigned to an ORF when a majority of the hits were for one type of function, such as a permease or a class of enzymes. When the functions of the hit sequences were varied and there was no solid agreement even for type of function, or when only one sequence was hit, no function was assigned to the query ORF and it was counted among the unknowns.

The average distance between *E. coli* genes is 118 bp. The 70 intergenic regions larger than 600 bp were reevaluated for the presence of ORFs (Geneplot, DNASTAR Inc.) and searched against the entire GenBank database for DNA sequence (BLASTN) and protein coding (BLASTX) features (*23*). Closer inspection revealed that 15 of these regions contain previously unannotated ORFs, which in most cases were overlooked because of their small size. An additional 11 intergenic regions contain sequence features such as long untranslated leader sequences [for example, *oppA* messenger RNA (mRNA) extends ~500 bp upstream of the start codon (*24*)] or well-characterized control regions [for example, the *araFGH* operon control region (*25*)]. The remaining 44 large intergenic regions fall into three general classes: putative gene regulatory regions, large repetitive sequences, and unknowns.

Genes separated by more than 600 bp are likely to contain independent regulatory sequences. Twenty-nine large intergenic regions contain sequences suggestive of regulatory functions, including 21 with predicted regulatory protein binding sites. There are 13 regions between divergently transcribed ORFs, and 11 of these have at least one predicted promoter for each ORF (2 have only one predicted promoter). The 16 regions between ORFs transcribed in the same direction contain at least one predicted promoter for the downstream ORF, and several contain a terminator for the upstream ORF. Seven of the large intergenic regions, including the largest region overall (1730 bp), consist of repeated sequences such as REP or LDR, as described below. Seven intergenic regions larger than 600 bp have no predicted regulatory or coding functions. Five of these regions contain sequences that could encode proteins of at least 50 amino acids, although codon usage patterns for these ORFs suggest that they are not expressed. It is likely that these regions contain additional, as yet undiscovered, functions such as binding sites for additional regulatory proteins.

We searched for promoter and protein binding site sequences upstream of 2436 genes. This includes all genes except those that are less than 70 bp from the 3′ end of an adjacent gene transcribed in the same direction. We limited our search to the 250 bp upstream of the predicted translational start sites because *E. coli* promoters are typically found within this region. We also searched for other potential regulatory sites within the 400-bp segments upstream of genes. This search was based on an exhaustive collection of known functional sites for 56 transcriptional regulatory proteins. More detail about these methods is available elsewhere (*26*).

The codon adaptation index (CAI) was calculated for each ORF according to the method of Sharp and Li (*27*). The CAI measures the extent to which codon usage agrees with an *E. coli* reference set from highly expressed genes. CAI is a predictor of the extent of gene expression. This is attributed to correspondence with iso-accepting tRNA abundance of *E. coli* and optimal (intermediate) codon-anticodon interaction energy (*28*). Genes with exceptionally low CAI values may be recent horizontal transfers that still reflect the optimal codon usage or mutational spectrum of their previous host (*29*). We identified clusters of four or more adjacent genes with low CAI values (<0.25) and also identified all genes in the lower 10th percentile of CAI observed in this genome.

The annotated sequence (accession number U00096) is available at the National Center for Biotechnology Information (NCBI) through the Entrez Genomes division, GenBank, and the BLAST databases. Our FTP site (ftp.genetics.wisc.edu) will maintain an updated version of the sequence as additional annotations or corrections are made; the version discussed here is M49.

Overview of the Sequence

The genome of *E. coli*, diagrammed in Fig. 1, consists of 4,639,221 bp of circular duplex DNA (*30*). Both base pair and minute scales are shown; base pair 1 was assigned in an apparently featureless region between genes *lasT* and *thrL*. Protein-coding genes account for 87.8% of the genome, 0.8% encodes stable RNAs, and 0.7% consists of noncoding repeats, leaving ~11% for regulatory and other functions. A radial plot shows *E. coli*'s local similarity to sequenced bacteriophage genes. The polar coordinate plot of CAI is designed to highlight regions of the genome with unusual codon usage, which may signify recent immigration by horizontal transfer. Some gene clusters with low CAI values correspond to known cryptic prophages, and others point to possible locations of additional horizontally acquired elements.

The origin and terminus of replication divide the genome into oppositely replicated halves, which we term replichores. Replichore 1, which is replicated clockwise, has the presented strand of *E. coli* as its leading strand; in replichore 2 the complementary

strand is the leading one. Many features of *E. coli* are oriented with respect to replication. All seven ribosomal RNA (rRNA) operons, and 53 of 86 tRNA genes, are expressed in the direction of replication (Fig. 1). Approximately 55% of protein-coding genes are also aligned with the direction of replication, confirming an early observation of Brewer (*31*).

Compositional organization of the genome. Several authors (*32*, *33*), in analyzing a variety of systems, have commented on base compositional asymmetries correlated with the direction of replication. For *E. coli*, the leading strands of both replichores have significantly ($P < 0.001$) greater abundance of G (26.22%) than its complementary partner C (24.58%) or the alternative pair A (24.52%) or T (24.69%). Lobry (*33*) plotted G-C skew for a 1.6-Mb section of *E. coli* surrounding the origin and summarized the data by codon position and gene direction. We extended this G-C skew analysis to the entire *E. coli* genome (Fig. 2), observing the same sharp transition at the terminus that he reported at the origin. These clear trends in base compositional skew apply to genes in both orientations, to intergenic regions, and to all codon positions (Table 1), supporting the idea advanced by Lobry, Perna, and Wu (*32*, *33*) that leading and lagging strands are subject to differential mutation as the result of

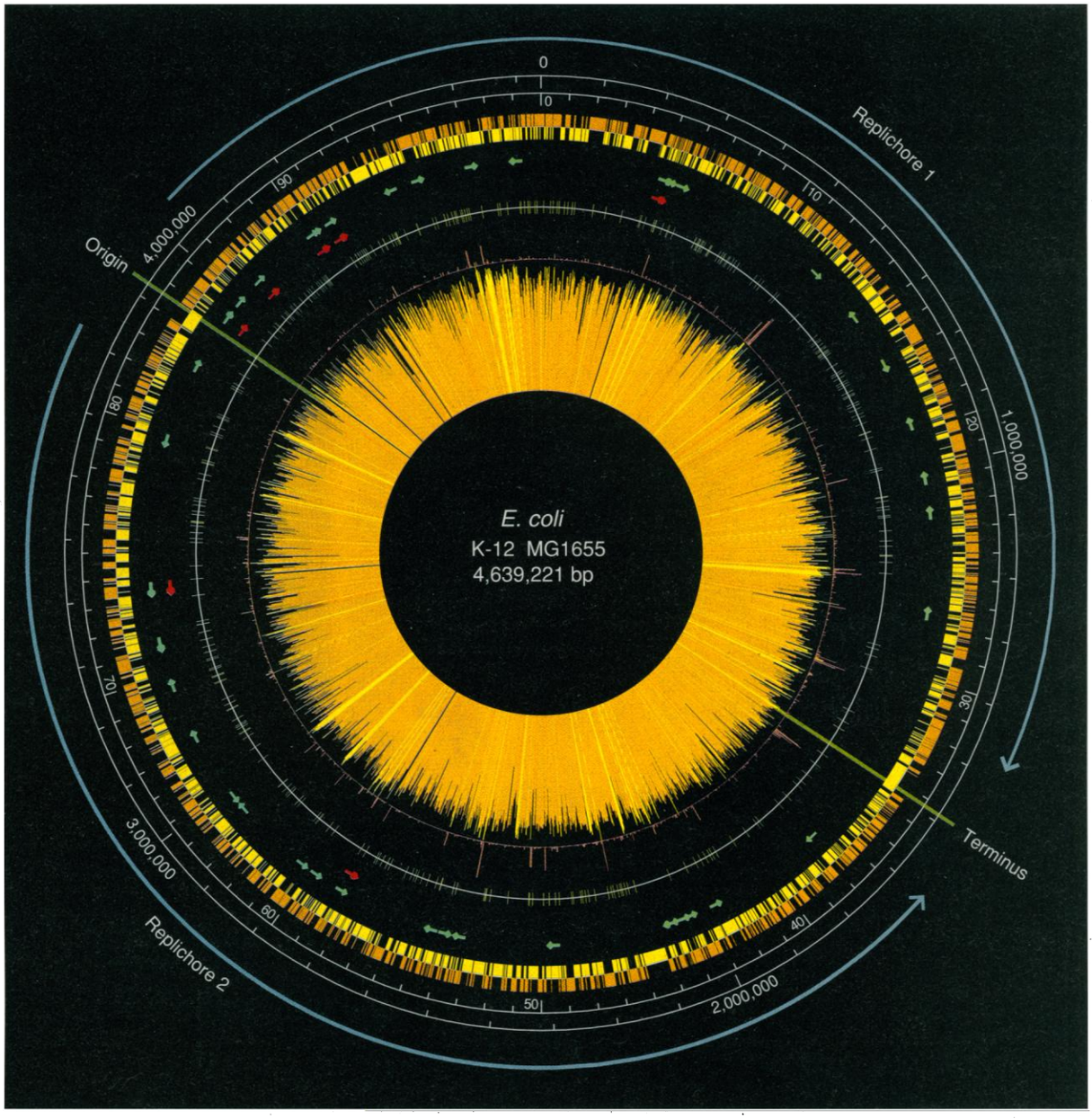


Fig. 1. The overall structure of the *E. coli* genome. The origin and terminus of replication are shown as green lines, with blue arrows indicating replichores 1 and 2. A scale indicates the coordinates both in base pairs and in minutes (actually centisomes, or 100 equal intervals of the DNA). The distribution of genes is depicted on two outer rings: The orange boxes are genes located on the presented strand, and the yellow boxes are genes on the opposite strand. Red arrows show the location and direction of transcription of rRNA genes, and tRNA genes are shown as green arrows. The next circle illustrates the positions of REP sequences around the genome as radial tick marks. The central orange sunburst is a histogram of inverse CAI (1 – CAI), in which long yellow rays represent clusters of low (<0.25) CAI. The CAI plot is enclosed by a ring indicating similarities between previously described bacteriophage proteins and the proteins encoded by the complete *E. coli* genome; the similarity is plotted as described in Fig. 3 for the complete genome comparisons.

asymmetry inherent to the DNA replication mechanism. This, combined with natural selection, leads to an observed base distribution that depends in part on the mutational pattern and in part on selection. Hence, intergenic regions and third positions in *E. coli* are more skewed than first and second positions, and the net G-rich tendency of the leading strand relative to the lagging one is seen in both first and second codon positions, despite strong and sometimes opposite codon usage preferences at those positions.

Replication, recombination, and skew. We carried out further analyses of *E. coli* by constructing a reference sequence composed of the leading strands of each replichore concatenated at a novel joint, and we examined this sequence for oligonucleotide distribution. The most frequent oligomers in this leading strand (for example, octam-

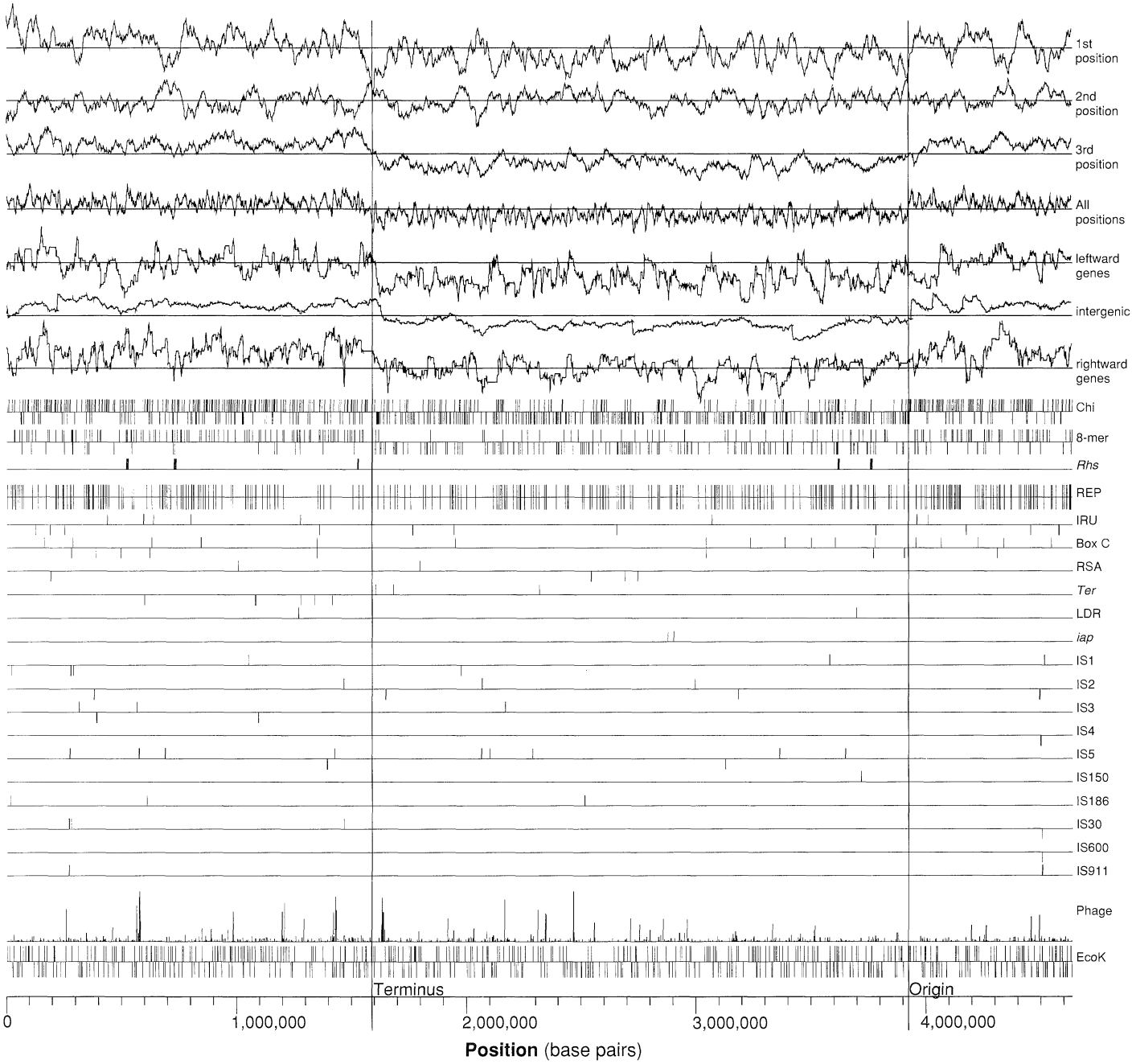


Fig. 2. Base composition is not randomly distributed in the genome. G-C skew [(G – C)/(G + C)] is plotted as a 10-kb window average for one strand of the entire *E. coli* genome. Skew plots for the three codon positions are presented separately; leftward genes, rightward genes, and non–protein-coding regions are shown in lines 5, 6, and 7. The two horizontal lines below the skew plots show the distribution of two highly skewed octamer sequences, GCTGGTGG (Chi) and GCAGGGCG (8-mer). Tick marks indicate the position of each copy of a sequence in the complete genome and are vertically offset to indicate the strand containing the sequence. The next 18 horizontal lines correspond to distinct classes of repetitive elements. The penultimate line contains a histogram showing the similarity (the product of the percent of each protein in the pairwise alignment and the percent amino acid identity across the aligned region) of known phage proteins to the proteins encoded by the complete *E. coli* genome. The last line indicates the position and orientation of the EcoK restriction-modification site AACNNNNNNGTGC (N, any nucleotide). Two vertical lines through the plots show the location of the origin and terminus of replication.

ers; Table 2) form a family containing the trimer CTG, often within the pentamer GCTGG, as also noticed by Karlin and co-workers (*34*). We note that the DnaG primase-binding site includes (or is) the sequence CTG, with T being the template for the first base of the RNA primer of Okazaki fragments (*35*). Although there is no direct proof implicating these sequences in discontinuous replication, their spacing is consistent with Okazaki fragment sizes and their distribution is skewed toward the leading strand, as expected. Although the skews are significant, the most frequent octamers on the leading strand are overrepresented on the lagging strand as well. Although leading strand replication is highly asymmetric in vitro, both leading and lagging strands are reported to replicate discontinuously in vivo (*36*). The high abundance of these proposed DnaG primase-binding sites on both strands supports a model in which both strands are replicated discontinuously. The associated skews imply that the leading strand has fewer sites for Okazaki initiation.

The recombinational hotspot Chi (*37, 38*), the third most abundant octamer of the leading strand, also contains the proposed DnaG primase-binding site. In fact, none of the frequent octamers differs from Chi by changes known to inactivate the recombinational activity of Chi (*37, 38*). Hence, it is possible that other members of the family may display Chi activity. As noted earlier (*14*), the Chi site is markedly skewed toward the leading strand. One must skip to the 251st most frequent octamer, GCAGGGCG, to locate a higher skew (57%) than that of the Chi site (50%) (Fig. 2, lines 8 and 9). Chi sites are implicated in RecBCD-mediated recombination (*37*), and as part of this process it is supposed that single-stranded DNA intermediates having Chi at the 3′ end are formed, which then invade the recipient chromosome to form a "D loop." This implies the existence of a Chi site on the displaced strand. If the CTG of Chi is also a primase binding site, Okazaki initiation at Chi could facilitate strand assimilation by branch migration. Kuzminov has recently proposed a role for Chi sites in the recombinational repair of collapsed replication forks, which may explain their extreme skew (*38*), but a secondary role as a primase-binding site may be sufficient to explain this bias.

Rare tetramer CTAG. It is well known that the palindromic tetramer CTAG is extremely rare in *E. coli*, with an abundance 5% of that predicted from the base composition. Various explanations have been offered (*39, 40*). In Table 3 we have analyzed its distribution in various subsets of the genome. Clearly, the rarity of CTAG is most pronounced in protein-coding regions. Its occurrence is considerably higher in intergenic DNA, but it is surprisingly abundant in genes coding for structural RNAs, especially in that minuscule portion of the genome that codes for tRNAs. Danchin and co-workers (*40*) have hypothesized that CTAG may "kink" DNA and thereby interfere with function. It is also possible that some peculiar folding behavior of CUAG in RNA might interfere with mRNA function while having no negative effect on stable RNA species.

Newly Proposed Genes and Previously Mapped Genes

Six new tRNA genes. In this study we discovered six new tRNA genes. Four of the genes—*valZ*, *lysY*, *lysZ*, and *lysQ* (positions 780,291 to 780,875)—are part of the *lysT* operon and consist of a duplicate of *valT* and three duplicates of *lysW*. The other two genes form single-gene transcriptional units: *asnW* (positions 2,056,049 to 2,056,124) is a duplicate copy of *asnT*, and *ileY* (positions 2,783,782 to 2,783,857) is a near copy of *ileX*, differing in a single compensating base pair change in the aminoacyl stem of the tRNA (C6·G67 in *ileX*, A6·T67 in *ileY*).

An operon for degradation of aromatic compounds. Six *E. coli* enzymes are known to constitute a pathway for the degradation of aromatic compounds such as phenylpropionate, but only two of the genes have been previously identified, *mhpB* and *mhpE* (*41*). On the basis of similarity searches, we are confident that there is an operon that starts with the monooxygenase gene *mhpA*, followed by the known dioxygenase gene *mhpB*, the hydrolase gene *mhpC*, the hydratase gene *mhpD*, the dehydrogenase gene *mhpF*, and the known gene *mhpE* coding for 4-hydroxy-2-oxovalerate aldolase. All the genes (positions 367,835 to 373,095; b0347 to b0352) are in the same order as the enzymes of the pathway. We propose that the next gene upstream (positions 366,811 to 367,758;

Table 1. G-C skew for each of the three codon positions, calculated separately for the coding strand of 2357 forward genes (whose coding strand is the leading strand) and 1929 backward genes (whose coding strand is the lagging strand). The net skew attributable to replication direction is the difference between the values for the forward and the backward genes divided by 2.

Position	Forward genes	Backward genes	Average G-C skew	Net G-C skew attributable to replication direction
1	19.41	16.08	17.74	1.66
2	−9.34	−11.79	−10.57	1.22
3	7.99	−0.48	3.75	4.23
Average	6.02	1.27	3.64	2.37

Table 2. Frequent octamers and their skew. The 24 most frequent octamers are ranked by frequency of occurrence on the leading strand (octamers with the same frequency of occurrence are ordered alphabetically). Frequent octamers that are reverse complements of frequent octamers are identified by their rank (in parentheses) beside that of their complement. All primary sequences are aligned by the CTG trimer. The average spacing for nonoverlapping sequences from this list is on the order of 1.3 kb. The percent skew is $100 \times (f - f')/(f + f')$, where f is the frequency of an octamer and f' that of its reverse complement.

Rank	Octamer	Skew (%)	Count
1 (9)	cgCTGgcg	15.6	867
2 (16)	ggcgCTGg	19.6	826
3 (= Chi)	gCTGgtgg	50.8	761
4 (17)	gCTGgcgg	13.1	719
5 (11)	tgCTGgcg	9.4	719
6	gcgCTGgc	17.2	691
7	tggcgCTG	15.4	677
8 (24)	gCTGgcgc	12.6	659
10	cgCTGgtg	27.0	617
12	CTGgcggc	16.2	589
13	CTGgcgca	13.2	575
14	gCTGgcga	9.4	570
15	TGgcggcg	19.3	561
18	aaCTGgcg	12.5	543
19	gCTGgaag	11.0	538
20	CTGgcgcg	15.4	524
21	gcgCTGga	16.4	519
22	CTGgcgaa	14.9	515
23	tgCTGgtg	29.1	515

Table 3. Distribution of CTAG sequences.

Category of DNA	CTAG count	Average spacing
All *E. coli*	886	7161
Protein-coding sequence	569	7159
TAG terminators	67	
REP sequences	4	6144
All non–protein-coding sequences	317	1782
Regulatory regions	251	1999
rRNA genes	46	697
tRNA genes	13	514
10Sa RNA (*ssrA*)	2	233
RNase P M1 RNA (*rnpB*)	1	377
Expected from base composition	18,101	256

b0346) may be the regulator for this pathway, because this sequence is similar to a number of transcriptional regulators.

A second operon for degradation of aromatic compounds. We have found a previously unrecognized set of *E. coli* genes (positions 2,667,052 to 2,671,269) that resemble *Pseudomonas* genes for the degradation of the aromatic compounds toluene, benzene, and biphenyl (*42*). The first three genes (b2538 to b2540) encode the α and β subunits and the ferredoxin component of the 1,2-dioxygenase that opens the rings and oxidizes carbons 1 and 2. The gene encoding the last component of the dioxygenase, the ferredoxin reductase (b2542), is separated from the first three genes by another ORF (b2541). The product of this ORF resembles the enzyme dihydro-1,2-diol dehydrogenase, which acts on the product of the dioxygenase to generate catechol. This proposed operon is preceded by a divergently transcribed ORF (b2537) resembling a number of transcriptional regulators, which may be involved in the regulation of the genes. We do not know the substrate for this operon, or whether it has enzymes with sufficiently broad specificity to use several related substrates. It is also not clear how catechol might be further metabolized. In *Pseudomonas* catechol is normally metabolized by either an ortho or meta pathway, and *E. coli* has some very distant sequence similarities to some of the meta pathway enzymes, especially to the penultimate step. In addition, MhpB can metabolize catechol at a slow rate (*41*). Further research will be needed to determine whether this is a physiologically significant pathway, and if so, under what conditions.

Flagellar operons nearly identical to those of Salmonella. *Escherichia coli* has an array of 14 flagellar synthesis genes (b1070 to b1083), only two of which have been previously reported: *flgM* and *flgL*. One additional gene is involved with initiation of filament assembly: *flgN*, which precedes *flgM*, a negative regulator of flagellin synthesis. In the region between *flgM* and *flgL*, we identified homologs of the *Salmonella typhimurium flgA* (basal-body P-ring formation), *flgB* (putative flagellar basal-body formation protein), *flgC* (putative flagellar basal-body formation protein), *flgD* (basal-body rod modification protein), *flgE* (flagellar hook protein), *flgF* (putative flagellar basal-body formation protein), *flgG* (flagellar basal-body formation protein), *flgH* (flagellar L-ring protein precursor), *flgI* (flagellar P-ring protein precursor), *flgJ* (flagellar protein), and *flgK* (flagellar hook-associated protein 1) genes. The gene arrangement of this cluster (positions 1,128,637 to 1,140,209) is identical to that of the cluster at 26.5 centisomes on the *Salmonella* chromosome. In fact, the entire flagellar systems of *E. coli* and *S. typhimurium* are essentially identical in most respects, with the current organization of genes predating the divergence of these two species (*43*). Two additional genes (b1068 and b1069), preceding the *flg* genes, show strong similarity (81% and 94% identity, respectively, as well as near-equal length) to *mviM* and *mviN*, two *Salmonella* virulence factors (*43*). Homologs of both *mviM* and *mviN* also have been identified in *Haemophilus* (*3*).

Open reading frames and gene function class assignments. Figure 3 is a detailed graphical presentation of the genome showing the arrangement of putative and known genes, operons, promoters, and protein binding sites. Of the 4288 ORFs annotated in the sequence, 1853 are previously described genes. (A complete listing of *E. coli* ORFs is available at www.genetics.wisc.edu/ and is likely to change as functional data accumulate.) The distribution of start codons is as follows: ATG, 3542; GTG, 612; and TTG, 130. There is also one ATT and possibly a CTG (*44*). The distribution of translation termination codons is as follows: TAA, 2705; TGA, 1257; and TAG, 326. We assigned 405 genes with the start codon overlapping the preceding stop, distributed as follows: ATGA, 224; TAATG, 98; TGATG, 48; GTGA, 28; TAGTG, 4; and TTGA, 3. The most common overlap in phage lambda is also ATGA (*45*).

The 4288 ORFs were searched for matches to the Link database of peptides excised from two-dimensional gels (*19*). These searches confirmed the expression of 30 hypothetical ORFs. In addition to the 194 Link sequences annotated in SWISS-PROT release 34, our searches identified nine NH_2-terminal sequences corresponding to *dsbA*, b2548, *gcvT*, *glpQ*, *trpB*, *ydfG*, *ygaG*, *ygiN*, and *yifE*.

The longest ORF encodes a 2383–amino acid protein of unknown function, resembling several bacterial attaching and effacing proteins and invasins—virulence factors in pathogenic strains of *E. coli* and other enteric bacteria (*46*). The average ORF size is 317 amino acids; there are four ORFs in the range 1500 to 1700 amino acids, 51 in the range 1000 to 1500 amino acids, and 381 that are smaller than 100 amino acids. In general, it was difficult to assign small ORFs unless they exhibited typical *E. coli* codon usage or had been characterized biochemically (for example, leader peptides).

Two complementary catalogs were devised originally to classify functions of *E. coli* gene products, one for broad functions of the gene product (for example, enzyme, regulator, or transport protein) and another for specific physiological roles in the cell (*47*). A simplified composite system was devised to represent *E. coli* gene products ranging from precisely known to loosely attributed functions in Fig. 3. Table 4 summarizes the functional class assignments used to classify each ORF. Pending the location of the coding sequences for 383 known *E. coli* proteins that are not yet associated with ORFs, nearly 40% of the

Table 4. Distribution of *E. coli* proteins among 22 functional groups (simplified schema).

Functional class	Number	Percent of total
Regulatory function	45	1.05
Putative regulatory proteins	133	3.10
Cell structure	182	4.24
Putative membrane proteins	13	0.30
Putative structural proteins	42	0.98
Phage, transposons, plasmids	87	2.03
Transport and binding proteins	281	6.55
Putative transport proteins	146	3.40
Energy metabolism	243	5.67
DNA replication, recombination, modification, and repair	115	2.68
Transcription, RNA synthesis, metabolism, and modification	55	1.28
Translation, posttranslational protein modification	182	4.24
Cell processes (including adaptation, protection)	188	4.38
Biosynthesis of cofactors, prosthetic groups, and carriers	103	2.40
Putative chaperones	9	0.21
Nucleotide biosynthesis and metabolism	58	1.35
Amino acid biosynthesis and metabolism	131	3.06
Fatty acid and phospholipid metabolism	48	1.12
Carbon compound catabolism	130	3.03
Central intermediary metabolism	188	4.38
Putative enzymes	251	5.85
Other known genes (gene product or phenotype known)	26	0.61
Hypothetical, unclassified, unknown	1632	38.06
Total	4288	100.00*

*Total of these rounded values is 99.97%.

ORFs are completely uncharacterized. This is similar to the proportion of unassigned ORFs in other recently sequenced bacterial genomes: *Haemophilus influenzae* (43%), *Synechocystis* sp. (45%), and *Mycoplasma genitalium* (32%) (3).

The largest well-defined functional group consists of 281 transport and binding proteins, and there are an additional 146 putative transport and binding proteins. In contrast, 123 transport proteins have been identified in *Haemophilus* and 34 in *Mycoplasma* (3). Whether this difference reflects a larger number of substrates to transport, greater specificity of particular transporters, or greater redundancy in *E. coli* is not yet clear. In sharp contrast, the number of proteins involved in translation is similar for *E. coli* (182), *Haemophilus* (141), and *Mycoplasma* (101).

On the basis of 1827 characterized *E. coli* proteins, Riley and Labedan (48) described 75 pairs of isozymes, or multiple enzymes with identical or nearly identical function. An additional 11 groups of potentially redundant enzymes have been identified among the newly sequenced ORFs. Although sequence similarity and functional overlap are not synonymous, these highly conserved proteins [point accepted mutations per 100 residues (PAM) < 110] are likely to carry out the same physiological function.

We have not yet attempted to represent proteins with multiple roles that depend on physiological circumstances. On the basis of our present knowledge, one-fourth of the cell's resources are devoted to small-molecule metabolism and about one-eighth to large-molecule metabolism, and at least one-fifth of the cell's resources are associated with cell structure and processes. Of course, this distribution may be altered when the specific functions of the remaining 40% of the gene products become known.

Homology between E. coli *proteins and the other sequenced genomes.* Figure 3 also presents comparisons of the 4288 *E. coli* proteins with data from five other complete genomes (3), representing the three major kingdoms. There are two components to the significance of each database hit: the degree of similarity between the aligned proteins, and the amounts of the two proteins that are alignable. In Fig. 3, we have plotted a simple index that takes both components into account.

To provide a preliminary estimate of the number of orthologous sequences shared by *E. coli* and each of these other complete genomes, we counted only matches including at least 60% of both proteins in an alignment with at least 30% identity. Each protein from another species was permitted only one match to an *E. coli* protein. The largest number of matches to *E. coli* is found in the *Haemophilus influenzae* genome (1.83 Mb encoding 1703 proteins with 1130 hits to *E. coli* proteins). *Haemophilus*, like *E. coli*, is a member of the gamma subdivision proteobacteria, making it the most closely related complete genome available for consideration (49). We also compared two additional eubacterial genomes: *Synechocystis* sp. (3.6 Mb, 3168 proteins, 675 hits) and *Mycoplasma genitalium* (0.58 Mb, 468 proteins, 158 hits). All four eubacteria have 111 proteins in common.

The numbers of matches across kingdoms in the archeon *Methanococcus jannaschii* (1.6 Mb, 1738 proteins, 231 hits) and the eukaryote *Saccharomyces cerevisiae* (12.1 Mb, 5885 proteins, 254 hits) are remarkably similar to each other. However, according to our significance criteria, only 16 proteins are conserved among all six taxa; they are largely translation proteins, including seven ribosomal proteins and two aminoacyl synthetases. One is classified as a hypothetical ORF in *E. coli*, *Saccharomyces*, and *Methanococcus*, but is described as a putative O-sialoglycoprotein endopeptidase in both *Haemophilus* and *Mycoplasma* on the basis of similarity to a *Pasteurella haemolytica* protein (50).

Nearly 60% of *E. coli* proteins have no match in any other complete genome considered. These may represent the subset of proteins specific to enterobacterial or *E. coli* processes as well as insertion elements and phage with restricted host range. The 629 proteins shared exclusively by *Haemophilus* and *E. coli* include new genes acquired in this lineage. The 292 proteins common to *E. coli* and just one of the other four species are indicative of numerous gene losses over the course of genome evolution. This preliminary analysis of similarity among sequences of complete genomes provides many avenues for further study.

Similarity among E. coli *proteins.* Also presented in Fig. 3 is a comparison of all the proteins of *E. coli* with each other. These can be divided into families defined by sequence relatedness (5). A paralogous family is generally composed of proteins within a single species with similar, though not necessarily identical, functions. We define putative paralogs as ORFs that share at least 30% sequence identity over more than 60% of their lengths. The similarity index for the best putative paralog of each gene is plotted in Fig. 3. Many *E. coli* proteins—1345—have at least one paralogous sequence in the genome. The relative size of a gene family for each protein is also shown in Fig. 3. The largest number of significant hits to a single protein (b1917) was 37. This protein is a member of the largest family of paralogous proteins in *E. coli*, the ABC transporter proteins. Riley and Labedan (5) compiled a list of 54 ABC transporters among *E. coli* proteins, and analysis of the proteins from the complete genome reveals an additional 26 members of this family. Determination of the number of independent paralogous groups requires a careful examination of all the matches to a particular protein, followed by inspection of all hits to proteins contained within the initial list of matches (5), and will require further analysis.

Many proteins are members of paralogous gene families and have significant matches in other species. It will be difficult, if not impossible, to unambiguously determine the relation between similar genes in different species when the level of divergence between orthologous genes approaches the level of divergence among paralogs within a species. The genes in all genomes are derived from a set of unique ancestral genes present in a progenitor of all extant organisms. Upon duplication of an ancestral gene, copies of the gene may be subsequently lost through natural selection or simply by a neutral stochastic process. Alternately, the copies may be retained as redundant systems for executing the original biological function, or they may diverge, with one or both copies giving rise to a novel function. This process of duplication and divergence, along with the occasional transfer of genes between strains and species, gives rise to the present contents of a genome (51). Characterization of all *E. coli* paralogous groups and comparison with groups from other species will allow examination of the evolutionary events surrounding protein diversification.

Operons, promoters, and protein binding sites. Operons, promoters, and regulatory protein binding sites are shown in Fig. 3. A total of 2584 predicted and known operons are represented. Of 2192 predicted operons, a surprisingly high 73% have only one gene, 16.6% have two genes, 4.6% have three genes, and 6% have four or more genes. All of them have at least one promoter, either known or predicted. Of 2405 operon regions with predicted promoters, 68% contain one promoter, 20% contain two promoters, and 12% contain three or more promoters. Regulatory sites are described in 603 regions corresponding to 16% of operon regions and 10% of interoperonic regions. We estimate that our search included representatives of 15 to 25% of the total number of different regulatory binding proteins in *E. coli*, including sites that are recognized by global regulators of transcription (for example, sites bound by the cyclic AMP receptor protein, CRP). Within the regions with predicted sites, 89.2% are regulated by one protein, 8.4% by two proteins, and 2.4% by three or more proteins. In 81.2% of these regions

only one site was found, 12.2% have two sites, and 6.6% have three or more sites. These numbers are more or less consistent with the distribution of regulatory sites among a set of promoters where transcriptional regulation has been well studied. In this collection of 132 promoters, 73% are regulated by one protein, and 43% contain only one site for the binding of a regulator (*52*). A number of *E. coli* genes are part of known operons (Fig. 3, red arrows).

Repeated sequences. A number of repeated sequences have been characterized in the *E. coli* genome (*53*). The number and distribution of these sequences in the whole genome are summarized in Fig. 2. The largest repeated sequences in *E. coli* K-12 are the five *Rhs* elements (all previously described), which are 5.7 to 9.6 kb in length and together comprise 0.8% of the genome. They have no known function, although strain comparisons suggest they may be mobile elements. The ~40-bp palindromic sequences variously referred to as REP, BIME, or PU constitute the largest class of repeats. They are often found as tandem copies, alternating in orientation, in complexes called REP elements. We have located 581 such sequences, in 314 REP elements containing from 1 to 12 tandem copies (see also Fig. 1). These elements account for 0.54% of the genome and are of unknown origin and function. These can be subdivided into distinct classes, as described by Bachellier *et al.* (*53*). Of the other known small dispersed repeats, we find four new IRU (or ERIC) elements, for a total of 19; four new copies of Box C, for a total of 33; and only the previously described six copies of RSA. The distribution of some of these repeated sequences may not be totally random; for example, Box C is absent over a 1-Mbp span in replichore 2.

Another repeated sequence found in the *E. coli* genome is the *Ter* sequence, which acts as a one-way gate or valve to block the progression of the DNA replication fork such that replication starting from the origin is prevented from progressing beyond the terminus marked by the *dif* site (*54*). François *et al.* (55) identified 10 different chromosomal fragments with homology to an oligomeric *TerA* probe, but only seven *Ter* sequences (*TerA* through *Ter*G) have been identified to date. We found two new copies of the 11-bp *Ter* core sequence TGTTGTAACTA, both of which are located and oriented as expected relative to *dif*.

The sequence named LDR (*11*) occurs as three tandem copies at positions 1,268,308 to 1,269,848; a lone fourth copy, shorter and diverged from the consensus of the other copies, is located at positions 3,697,525 to 3,697,888. In the region between positions 2,875,665 and 2,902,430, a 29-bp sequence called the *iap* repeat is found in three clusters of 14, 2, and 7 copies, for a total of 23 copies (*53, 56*). No additional copies of either of these sequences are found in the rest of the genome.

Insertion sequences. The chromosome of *E. coli* K-12 contains a number of autonomously transposable elements that are implicated in the generation of many spontaneous mutations—not only by insertional inactivation, but also by deletions, duplications, and inversions. Estimates have been made as to the IS element set present in *E. coli* K-12 when originally isolated (*57*). The IS elements' map positions are shown in Fig 2. There are two multicomponent clusters. At positions 269,430 to 271,751, there is an IS911-related sequence (65% match), which we term IS911A, interrupted by a copy of IS30. At positions 4,504,683 to 4,507,369, there is a more faithful copy of IS911 (designated IS911B), which is also interrupted by a copy of IS30 as well as by a piece of IS600. This is the only IS600-related sequence in the genome. We did not find the copy of IS629 that had been suspected from hybridization studies (*58*).

Cryptic prophage and phage remnants. As originally isolated, *E. coli* K-12 carried bacteriophage lambda plus the defective lambdoid prophages DLP12, Rac, and Qin, the element e14, and the recently described CP4-57 (59). Defective, or cryptic, prophages have lost some functions essential for lytic growth and the production of infectious particles, but still retain other functional phage genes. They can rescue mutations in related infecting bacteriophages by recombining with them to generate viable hybrids. Figure 2 shows a histogram plot presenting all sequence matches to the phage proteins in SWISS-PROT. In addition to clarifying the structure of the known prophages, we identified three new cryptic prophages. Moreover, we found numerous instances of isolated genes that are similar to bacteriophage genes. We call these single genes "phage remnants" to distinguish them from the larger cryptic prophages. Although this implies a phage origin—the last vestiges of a cryptic prophage ravaged by deletions—these genes may actually be homologs encoded by both a bacteriophage and its host, with no ready indication as to which genome was the original carrier.

We determined the precise endpoints of e14 in MG1655 (positions 1,195,432 and 1,210,646), including terminal 11-bp direct repeats, from the published excised element and e14-free chromosome sequences (GenBank accession numbers M19693 and M19683). The 1829-bp Pin invertable P-region of e14 is in the (–) orientation in this sequence. The precise boundaries of the other lambdoid prophage remain to be annotated. The "cryptic P4" phage CP4-57 (59) is located at 57 minutes, where it is inserted into the stable RNA gene *ssrA*. The junction sequences (59) allowed us to identify the extended attL and attR sequences and to define the endpoints of the prophage (positions 2,753,956 and 2,776,007); our earlier report (GenBank accession number U36840) that attR was deleted in MG1655 was a misinterpretation.

We have discovered two new cryptic prophages, seemingly related to CP4-57, which we name CP4-6 and CP4-44 after their minute positions. The three CP4 prophage are organized similarly and encode several similar proteins, although they do not share the same attachment sites. We infer that CP4-6 is integrated into tRNA gene *thrW* (*60*) because the 3′ end of *thrW* is duplicated 34,242 bp downstream adjacent to b0281, a homolog of several integrases. This prophage (positions 262,122 to 296,489) includes *argF*, a known "duplicate" gene in the arginine biosynthesis pathway that has been suggested to have been acquired through a transposition event (*61*). It also includes the IS911A complex, a partial IS30 copy, two copies of IS1, and one copy of IS5. CP4-44 is less well defined (approximate endpoints at positions 2,064,181 and 2,077,053) and we suspect that insertion of the IS5 at its left end may have been accompanied by a deletion of part of the prophage; although it shares other ORFs with CP4-6 and CP4-57, it has no candidate integrase or associated direct repeats that might be att sites.

A third new cryptic prophage is located in the *eut* operon. Its presumptive integrase (b2442) resembles that of phiR-73, Sf6, and the CP4 family, but no other ORFs suggest its inclusion in the CP4 group. The endpoints of the element (positions 2,556,711 and 2,563,508) were defined by comparison with the sequence of *Salmonella typhimurium*, from which the element is missing (*62*). The 8-bp direct repeat TCAGGAAG at the ends is present as a single copy in *Salmonella*. The W3110 sequence from the Japanese group (http://mol.genes.nig.ac.jp/ecoli/) is missing this element, which, in light of the K-12 pedigree, suggests that this element is able to excise.

Conclusion

Although the determination of the complete *E. coli* sequence has required almost 6 years, this represents only the beginning of our understanding. Further research will be required to determine the precise functions for all of the genes by global tran-

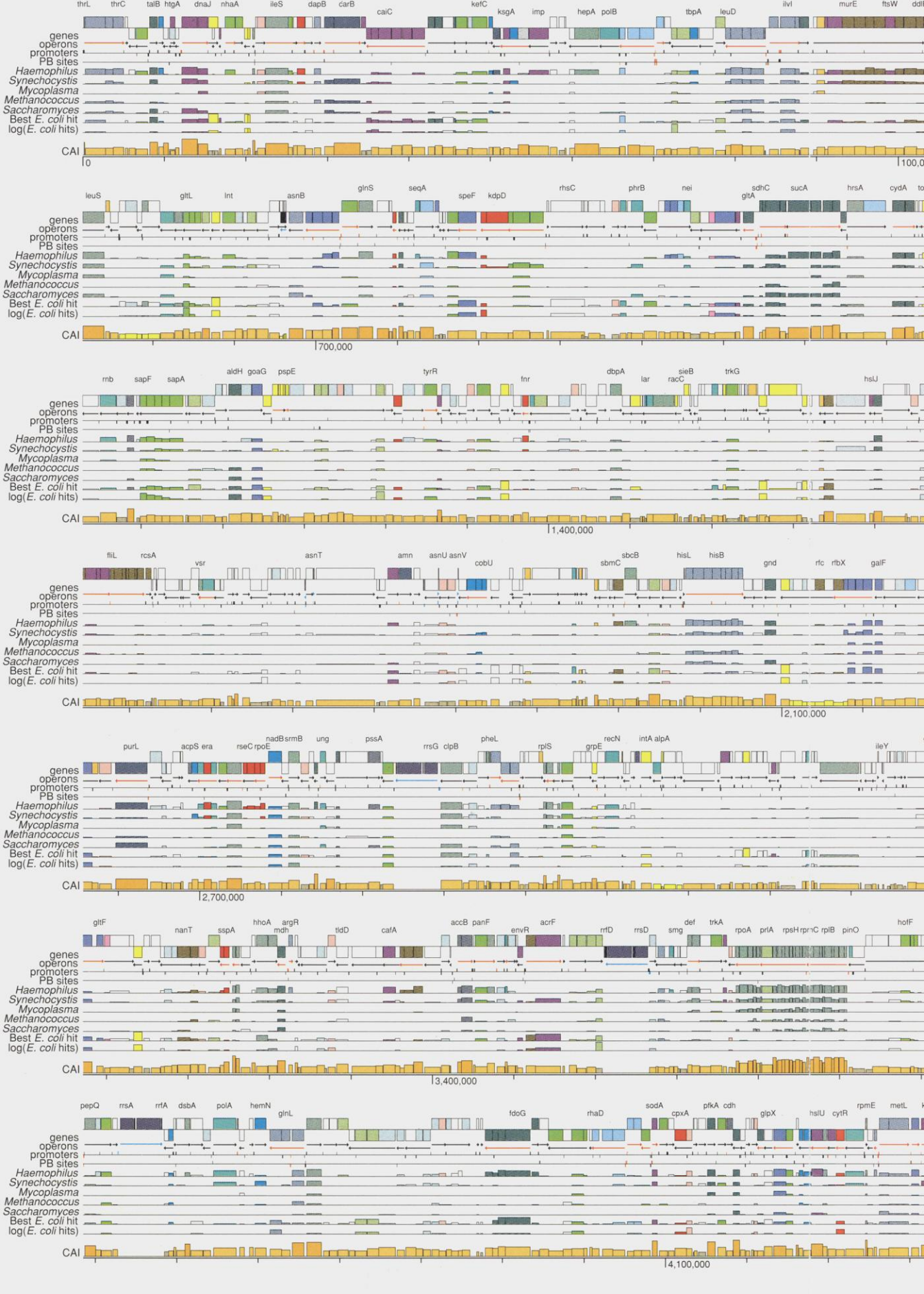
scriptional analysis, phenotypic analysis of mutants, and analysis of biochemical and catalytic properties of the expressed proteins. Another fruitful avenue for exploration will lie in whole genome comparisons—both with related pathogens to identify those genes that confer unique detrimental or beneficial properties, and with other microbial genomes to ascertain evolutionary relations.

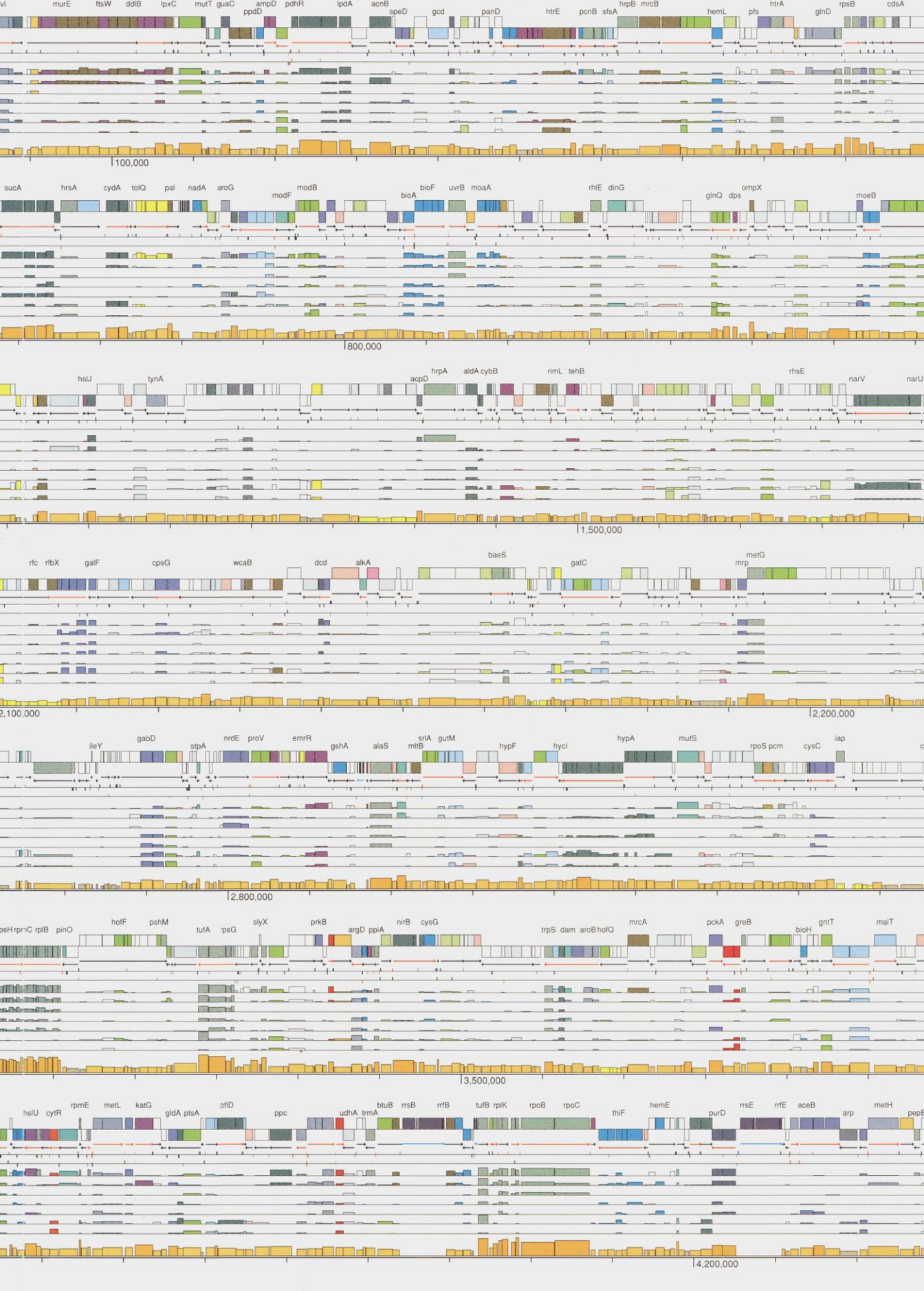
REFERENCES AND NOTES

1. F. R. Blattner, *Science* **222**, 719 (1983). *Escherichia coli* has been the subject of extensive monographs, the most recent of which is (*2*).
2. Escherichia coli *and* Salmonella *Cellular and Molecular Biology*, F. C. Neidhardt *et al.*, Eds. (ASM Press, Washington, DC, 1996).
3. The publicly available complete genome sequences are those of *Haemophilus influenzae* Rd [R. D. Fleischmann *et al.*, *Science* **269**, 496 (1995)], *Mycoplasma genitalium* [C. M. Fraser *et al.*, *ibid.* **270**, 397 (1995)], *Methanococcus jannaschii* [C. J. Bult *et al.*, *ibid.* **273**, 1058 (1996)], *Mycoplasma pneumoniae* [R. Himmelreich *et al.*, *Nucleic Acids Res.* **24**, 4420 (1996)], *Synechocystis* sp. strain PCC6803 [T. Kaneko *et al.*, *DNA Res.* **3**, 109 (1996)], and *Saccharomyces cerevisiae* [A. Goffeau *et al.*, *Science* **274**, 546 (1996)].
4. S.-E. Chuang, D. L. Daniels, F. R. Blattner, *J. Bacteriol.* **175**, 2026 (1993); D. J. Lockart *et al.*, *Nature Biotechnol.* **14**, 1675 (1996).
5. M. Riley and B. Labedan, *J. Mol. Biol.* **269**, 1 (1997).
6. F. C. Neidhardt, in (*2*), vol. 2, pp. 1–3.
7. B. Bachmann, in (*2*), vol. 2, pp. 2460–2488.
8. K. F. Jensen, *J. Bacteriol.* **175**, 3401 (1993).
9. R. P. Lawther *et al.*, *ibid.* **149**, 294 (1982).
10. D. Liu and P. R. Reeves, *Microbiology* **140**, 49 (1994).
11. T. Yura *et al.*, *Nucleic Acids Res.* **20**, 3305 (1992); N. Fujita, H. Mori, T. Yura, A. Ishihama, *ibid.* **22**, 1637 (1994); T. Oshima *et al.*, *DNA Res.* **3**, 137 (1996); H. Aiba *et al.*, *ibid.*, p. 363; T. Itoh *et al.*, *ibid.*, p. 379.
12. V. Burland, D. L. Daniels, G. Plunkett III, F. R. Blattner, *Nucleic Acids Res.* **21**, 3385 (1993).
13. Six segments of the genome were sequenced using radioactive chemistry (*14*) [D. L. Daniels, G. Plunkett III, V. Burland, F. R. Blattner, *Science* **257**, 771 (1992); G. Plunkett III, V. Burland, D. L. Daniels, F. R. Blattner, *Nucleic Acids Res.* **21**, 3391 (1993); F. R. Blattner, V. Burland, G. Plunkett III, H. J. Sofia, D. L. Daniels, *ibid.*, p. 5408; H. J. Sofia, V. Burland, D. L. Daniels, G. Plunkett III, F. R. Blattner, *ibid.* **22**, 2576 (1994); V. Burland, G. Plunkett III, H. J. Sofia, D. L. Daniels, F. R. Blattner, *ibid.* **23**, 2105 (1995)]. We determined experimentally that deoxyinosine triphosphate (dITP) is the most effective analog for resolving G-C compressions, although it also causes premature termination. With radioactive sequencing, a dITP sequence lane must be run in addition to, rather than in place of, a deoxyguanosine triphosphate (dGTP) run. For efficiency in the areas of *E. coli* we sequenced radioactively, tiling software was used to select a minimal set of M13 clones for resequencing with dITP after the bulk of the assembly had been completed with dGTP. On the other hand, because prematurely terminated chains are not labeled by the fluorophore with dye-terminator fluorescent sequencing, dITP can substitute totally for dGTP and can be used for all routine data collection.
14. V. Burland, G. Plunkett III, D. L. Daniels, F. R. Blattner, *Genomics* **16**, 551 (1993).
15. D. L. Daniels, in *The Bacterial Chromosome*, K. Drlica and M. Riley, Eds. (American Society for Microbiology, Washington, DC, 1990), pp. 43–51. It was often necessary to resequence overlapping regions between adjacent clones, and screening to remove lambda vector sequences before sequencing was costly. Occasionally we found deleted, mismapped, or chimeric lambda clones that created unexpected gaps in genome coverage.
16. Although the 1-μg yield of popout plasmid [G. Pósfai *et al.*, *Nucleic Acids Res.* **22**, 2392 (1994)] was low for early shotgun protocols, the assemblies were successful when supplemented with lambda clone and long-range PCR data. The main problem with extending this approach was the need to specifically engineer each popout plasmid by insertional recombination into the host.
17. I–Sce I is a site-specific intron-encoded homing endonuclease from yeast [A. Perrin, M. Buckle, B. Dujon, *EMBO J.* **12**, 2939 (1993)], whose 18-bp nonpalindromic recognition site is absent from *E. coli* (C. A. Bloch and C. K. Rode, unpublished data). Single I–Sce I sites were introduced into MG1655 on a transposable element to produce a mapped collection of strains, each with a unique I–Sce I site [C. K. Rode, V. H. Obreque, C. A. Bloch, *Gene* **166**, 1 (1995); C. A. Bloch, C. K. Rode, V. H. Obreque, J. Mahillon, *Biochem. Biophys. Res. Commun.* **223**, 104 (1996)]. P1 transduction was used to combine sites in pairs, permitting isolation of I–Sce I fragments as single bands by pulsed-field gel electrophoresis. Sequencing confirmed the expected nine-base overlap between adjacent fragments. Although the background contamination for entire I–Sce I fragment shotguns ranged from 15 to 30%, we occasionally observed individual preparative gels that seemed to have <5% background, as assessed from gel images. We therefore suspect that improvements in gel handling and electrophoretic conditions could improve the overall quality of the fragment preparations.
18. V. Burland, F. P. Curtis, N. Kusukawa, *Biotechniques* **21**, 142 (1996).
19. Codon usage statistics [M. Borodovsky and J. McininCh, *Comput. Chem.* **17**, 123 (1993); M. Gribskov, J. Devereux, R. R. Burgess, *Nucleic Acids Res.* **12**, 539 (1984)] were graphically displayed by means of the program Geneplot (DNASTAR). Protein searches were to SWISS-PROT release 34 [A. Bairoch and R. Apweiler, *ibid.* **24**, 21 (1996)]. The Link database is described in A. J. Link, thesis, Harvard University (1994). Signal peptide searches used an unpublished BASIC program written by F.R.B. Predictions for ribosomal binding sites were provided by W. S. Hayes and M. Borodovsky (personal communication).
20. M. Riley, *Nucleic Acids Res.* **25**, 51 (1997).
21. P. Karp, M. Riley, S. M. Paley, A. Pellegrini-Toole, M. Krummenacker, *ibid.*, p. 43.
22. Similarity searches were conducted using both the DeCypher II hardware-software system (Time Logic Inc., Incline Village, NV) and the PepPepSearch program of the Darwin suite at Zurich, http://cbrg.inf.ethz.ch/ [G. H. Gonnet, M. A. Cohen, S. A. Benner, *Science* **256**, 1443 (1992)]. PepPepSearch returns up to 30 hit sequences per query, and returns each pairwise alignment and the corresponding PAM scores. For most of the cases, only matches with PAM < 200 were used. See B. Labedan and M. Riley, *Mol. Biol. Evol.* **12**, 980 (1995).
23. S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman, *J. Mol. Biol.* **215**, 403 (1990).
24. K. Kashiwagi, Y. Yamaguchi, Y. Sakai, H. Kobayashi, K. Igarashi, *J. Biol. Chem.* **265**, 8387 (1990).
25. Y. Lu, C. Flaherty, W. Hendrickson, *ibid.* **267**, 24848 (1992).
26. Using the database of 392 known operons that we have localized in the genome sequence, we first predicted operons on the basis of the functional class conservation within genes of an operon. This gives a better prediction (68% positive prediction) than the method of predicting operons on the basis of the distance of genes inside operons versus the distance between operons (59% positive prediction). We predicted 2281 operons by functional class conservation and predicted the remainder with unclassified genes, using 50 bp as the distance criterion. The strategy found to give the highest number of positive promoter predictions (~40% when tested with an independent set of known promoters) involves an initial search with a pair of weight matrices, one for the −10 region and one for the −35 region. Candidate promoters using a low threshold of matches and 15 to 21 bp between −10 and −35 are saved. A subset of best candidates are selected on the basis of a context measure that compares alternative candidates within a given region of 200 bp upstream of each ORF. This includes a weight preference for candidates located closer to the beginning of the gene. The method can find zero, one, or several promoters in a single region. Inside operons, we only saved promoters where regulatory sites were also found. Regulatory sites were searched with a combined weight matrix (when at least three sequences are known) and a string search that allows a fixed number of mismatches for each regulatory site. To avoid overrepresentation of particular sites, we adjusted the number of allowed mismatches such that the number of predicted sites did not exceed 10 times the number of known sites for a given regulatory protein [D. A. Rosenblueth, D. Thieffry, A. M. Huerta, H. Salgado, J. Collado-Vides, *Comput. Appl. Biosci.* **12**, 415 (1997)].
27. P. M. Sharp and W. H. Li, *Nucleic Acids Res.* **15**, 1281 (1987).
28. H. Grosjean and W. Fiers, *Gene* **18**, 199 (1982); T. Ikemura, *Mol. Biol. Evol.* **2**, 13 (1985).
29. C. Médigue, T. Rouxel, P. Vigier, A. Henaut, A. Danchin, *J. Mol. Biol.* **222**, 851 (1991).
30. The zero reference (0/100, formerly 0/60) of the map was originally defined as the position of the first marker (*thr*) transferred by *E. coli* Hfr H, which was used in genetic mapping by interrupted mating, and a convention has arisen of using the first residue of the *thrA* gene as residue 1. However, this results in placing the regulatory region of the *thr* operon at the opposite end of the 4.6-Mb sequence from the operon itself. We therefore defined nucleotide 1 as the A residue 189 nucleotides upstream of the initiation codon for *thrL*, the first gene on the genetic map. We did not detect any feature spanning this point.
31. B. J. Brewer, in *The Bacterial Chromosome*, K. Drlica and M. Riley, Eds. (American Society for Microbiology, Washington, DC, 1990), pp. 61–83.
32. C.-I. Wu and N. Maeda, *Nature* **327**, 169 (1987); N. T. Perna and T. D. Kocher, *J. Mol. Evol.* **41**, 353 (1995).
33. J. R. Lobry, *Mol. Biol. Evol.* **13**, 660 (1996); *Science* **272**, 745 (1996).
34. L. R. Cardon, C. Burge, G. A. Schachtel, B. E. Blaisdell, S. Karlin, *Nucleic Acids Res.* **21**, 3875 (1993); B. E. Blaisdell, K. E. Rudd, A. Matin, S. Karlin, *J. Mol. Biol.* **229**, 833 (1993).
35. K. Yoda, H. Yasuda, X. W. Xiang, T. Okazaki, *Nucleic Acids Res.* **16**, 6531 (1988); H. Hiasa *et al.*, *Gene* **84**, 9 (1989); K. Yoda and T. Okazaki, *Mol. Gen. Genet.* **227**, 1 (1991); J. R. Swart and M. A. Griep, *J. Biol. Chem.* **268**, 12970 (1993).
36. T.-C. V. Wang and S.-H. Chen, *Biochem. Biophys. Res. Commun.* **184**, 1496 (1992); *ibid.* **198**, 844 (1994).
37. The major recombination pathway in *E. coli* is the RecBCD pathway, so called because of the central involvement of the enzyme encoded by the *recBCD* genes. For a review of RecBCD-mediated recombination, see F. Stahl and R. Myers, *J. Hered.* **86**, 327 (1995); see also (*38*). For a review of recombination-deficient variants of Chi, see D. W. Schultz, J. Swindle, G. R. Smith, *J. Mol. Biol.* **146**, 275 (1981).
38. A. Kuzminov, *Mol. Microbiol.* **16**, 373 (1995).
39. C. Burge, A. M. Campbell, S. Karlin, *Proc. Natl. Acad. Sci. U.S.A.* **89**, 1358 (1992); M. McClelland and A. S. Bhagwat, *Nature* **355**, 595 (1992); A. S. Bhagwat and M. McClelland, *Nucleic Acids Res.* **20**, 1663 (1992); R. Merkl, M. Kroger, P. Rice, H. J. Fritz, *ibid.*, p. 1657; S. Karlin and L. R. Cardon, *Annu. Rev. Microbiol.* **48**, 619 (1994).
40. C. Médigue, A. Viari, A. Hénaut, A. Danchin, *Mol. Microbiol.* **5**, 2629 (1991).
41. R. P. Burlingame, L. Wyman, P. J. Chapman, *J. Bacteriol.* **168**, 55 (1986); T. D. H. Bugg, *Biochim. Biophys. Acta* **1202**, 258 (1993); E. Spence, M. Kawamukai, J. Sanvoisin, H. Braven, T. Bugg, *J. Bacteriol.* **178**, 5249 (1996).
42. H. M. Tan, H. Y. Tang, C. L. Joannou, N. H. Abdel-Wahab, J. R. Manson, *Gene* **130**, 33 (1993).
43. R. M. Macnab, in (*2*), vol. 2, pp. 123–145; M.

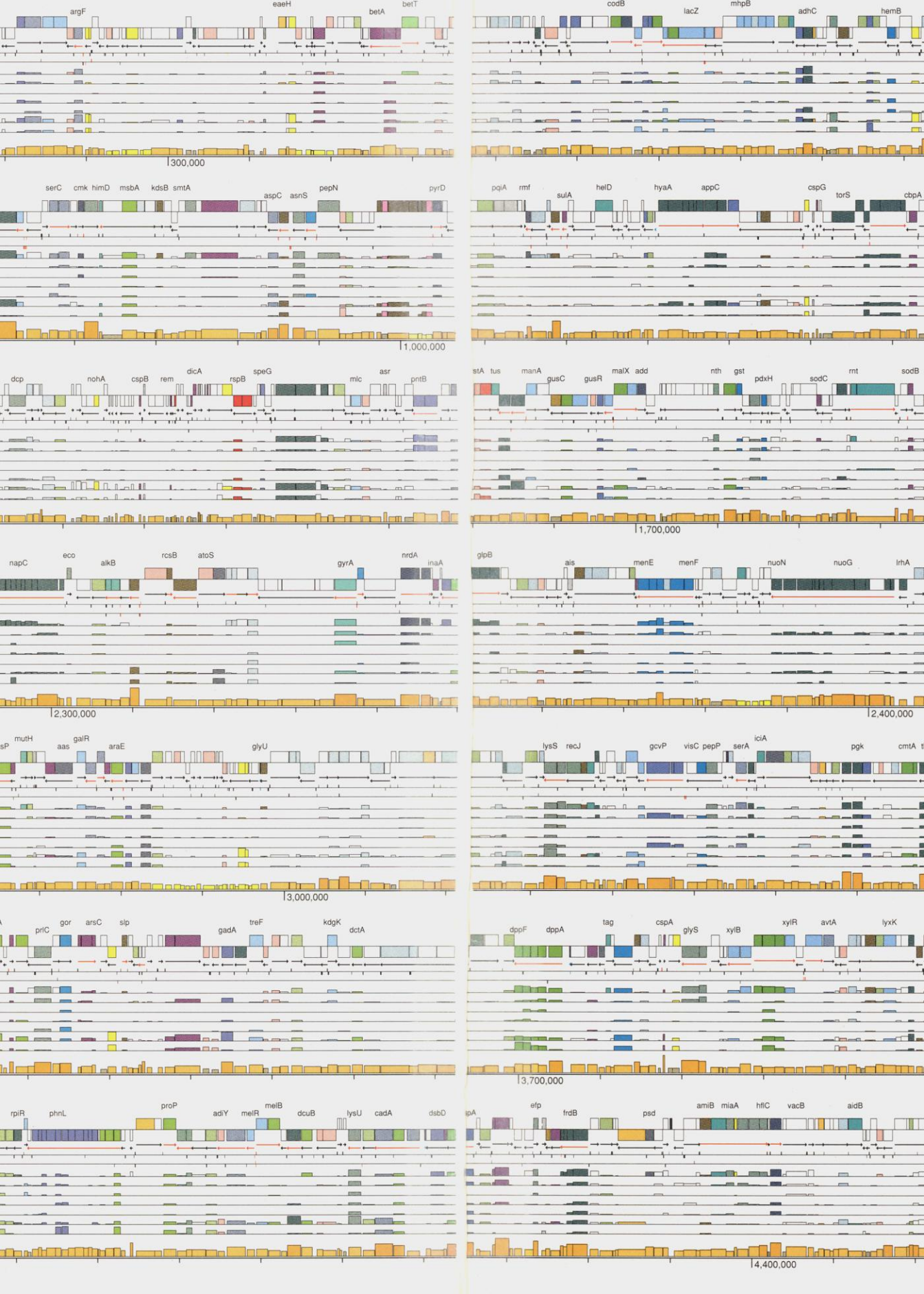
Homma, D. J. DeRosier, R. M. Macnab, *J. Mol. Biol.* **213**, 819 (1990); K. Ohnishi, Y. Ohto, S. Aizawa, R. M. Macnab, T. Iino, *J. Bacteriol.* **176**, 2272 (1994); For a discussion of *mviM* and *mviN*, see K. Kutsukake, T. Okada, T. Yokoseki, T. Iino, *Gene* **143**, 49 (1994).
44. For a discussion of ATT start in *infC*, see C. Sacerdot *et al.*, *EMBO J.* **1**, 311 (1982); for a discussion of CTG start in *htgA*, see D. Missiakas, C. Georgopoulos, S. Raina, *J. Bacteriol.* **175**, 2613 (1993).
45. D. L. Daniels, F. Sanger, A. R. Coulson, *Cold Spring Harbor Symp. Quant. Biol.* **47**, 1009 (1983); F. Sanger, A. R. Coulson, G. F. Hong, D. F. Hill, G. B. Petersen, *J. Mol. Biol.* **162**, 729 (1982).
46. A number of bacterial proteins have been implicated in mediating the invasion of host cells by pathogens. Attaching and effacing proteins are involved in eliciting an extensive rearrangement of host cell actin by enteropathogenic *E. coli* strains, whereas invasins are bacterial surface proteins that provoke the endocytic uptake of *Yersinia* and *Salmonella* spp. by host cells. For an overview of bacterial pathogenesis, including virulence factors, see A. A. Salyers and D. D. Whitt, *Bacterial Pathogenesis: A Molecular Approach* (ASM Press, Washington, DC, 1994).
47. M. Riley, *Microbiol. Rev.* **57**, 862 (1993).
48. _____ and B. Labedan, in (*2*), vol. 2, pp. 2118–2202.
49. Relations among these eubacteria are estimated by a rRNA phylogeny [G. J. Olsen, C. R. Woese, R. Overbeek, *J. Bacteriol.* **176**, 1 (1994)]. A previous estimate of 1128 *Haemophilus influenzae* orthologs among 75% of the complete *E. coli* genome [R. L. Tatusov *et al.*, *Curr. Biol.* **6**, 279 (1996)] is based on less restrictive criteria and includes sequences with as little as 18% identity.
50. K. M. Abdullah, R. Y. Lo, A. Mellors, *J. Bacteriol.* **173**, 5597 (1991).
51. S. Ohno, *Evolution by Gene Duplication* (Springer-Verlag, Berlin, 1970).
52. J. D. Gralla and J. Collado-Vides, in (*2*), vol. 1, pp. 1232–1244.
53. S. Bachellier, E. Gilson, M. Hofnung, C. W. Hill, in (*2*), vol. 2, pp. 2012–2040.
54. T. M. Hill, in (*2*), vol. 2, pp. 1602–1612.
55. V. François, J. Louarn, J.-M. Louarn, *Mol. Microbiol.* **3**, 995 (1989).
56. A. M. Nakata, M. Amemura, K. Makino, *J. Bacteriol.* **171**, 3553 (1989).
57. R. C. Deonier, in (*2*), vol. 2, pp. 2000–2011.
58. S. Matsutani and E. Ohtsubo, *Gene* **127**, 111 (1993).
59. For a review of K-12 prophage, see A. M. Campbell, in (*2*), vol. 2, pp. 2041–2046. CP4-57 is described in D. M. Retallack, L. L. Johnson, D. I. Friedman, *J. Bacteriol.* **176**, 2082 (1994); J. E. Kirby, J. E. Trempy, S. Gottesman, *ibid.*, p. 2068.
60. P22 [D. F. Lindsey, C. Martinez, J. R. Walker, *J. Bacteriol.* **174**, 3834 (1992)] and a phage from a clinical isolate [D. Lim, *Mol. Microbiol.* **6**, 3531 (1992)] also integrate into *thrW*.
61. F. Van Vliet, A. Boyen, N. Glansdorff, *Ann. Inst. Pasteur Microbiol.* **139**, 493 (1988).
62. E. Kofoid and J. Roth, personal communication.
63. This is Laboratory of Genetics paper 3487. We thank the entire *E. coli* community for their support, encouragement, and sharing of data, and especially D. L. Daniels and N. Peterson, who were present at the creation. We also thank R. Straussburg and M. Guyer, our program administrators; R. R. Burgess and M. Sussman for critical reading of the manuscript; M. Borodovsky and W. S. Hayes for application of a new version of the GeneMark program to the analysis of the sequence; K. Rudd for his Ecoseq7 melds of GenBank data; J. Mahillon for providing I–Sce I strains; J. Roth and E. Kofoid for unpublished *Salmonella* data; the Japanese group under H. Mori and T. Horiuchi for cooperative competition; G. Pósfai and W. Szybalski for the popout strains; S. Baldwin, C. Allex, N. Manola, G. Bouriakov, and J. Schroeder of DNASTAR for extraordinary software; A. Huerta, H. Salgado, and D. Thieffry for help with promoter, operon, and regulatory site identification; T. Thiesen for Postscript illustrations; H. Kijenski, G. Peyrot, P. Soni, G. Diarra, E. Grotbeck, T. Forsythe, M. Maguire, M. Federle, S. Subramanian, and K. Kadner for excellent technical work; and 169 University of Wisconsin undergraduates who participated over the last decade. Supported by NIH grants P01 HG01428 (from the Human Genome Project) and S10 RR10379 (for ABI machines from the National Center for Research Resources–Biomedical Research Support Shared Instrumentation Grant). We thank IBM for the gift of workstations, the State of Wisconsin for remodeling support, and especially SmithKline Beecham Pharmaceuticals and Genome Therapeutics Corp. for financial support of the annotation of this sequence. N.P. is an NSF fellow in molecular evolution.

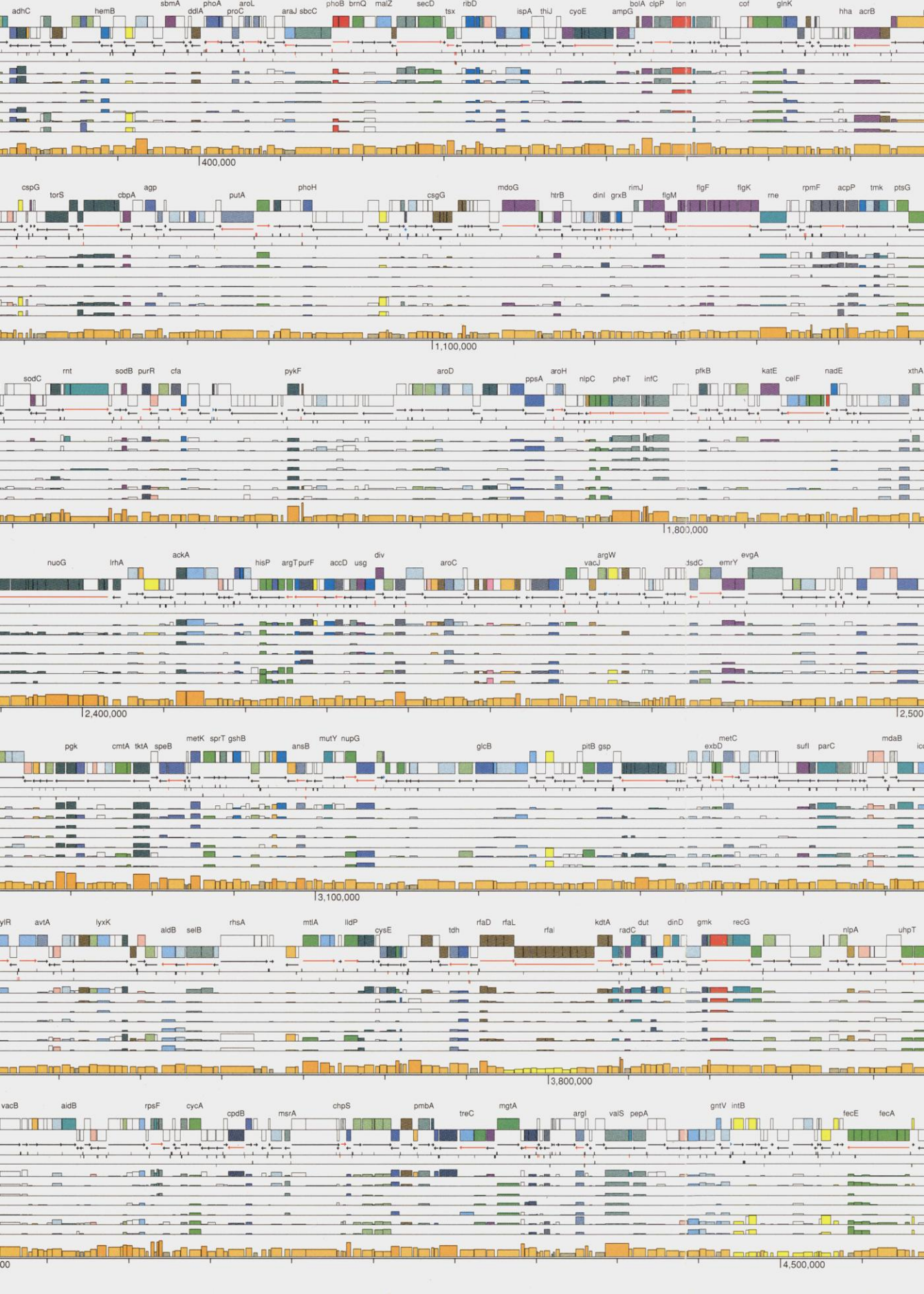
Fig. 3 (foldout). Map of the complete *E. coli* sequence, its features and similarities to proteins from five other complete genome sequences, proceeding from left to right in 42 tiers. The top line shows each gene or hypothetical gene, color-coded to represent its known or predicted function as assigned on the basis of biochemical and genetic data. Genes are vertically offset to indicate their direction of transcription. Space permitting, names of previously described *E. coli* genes are indicated above the line. The second line contains arrows indicating documented (red) and predicted (black) operons. Documented operons encoding stable RNAs are blue. Line 3, below the operons, contains tick marks showing the position of documented (red), predicted (black), and stable RNA (blue) promoter sequences. Line 4 consists of tick marks showing the position of documented (red) and predicted (black) protein binding sites. Lines 5 to 9 are histograms showing the results of alignments between *E. coli* proteins and the products encoded by five other complete genomes. The height of each bar is a simple index of similarity: the product of the percent of each protein in the pairwise alignment and the percent amino acid identity across the aligned region. Line 10 indicates similarity among proteins in *E. coli* in the same fashion. Line 11 histograms show the logarithm of the number of proteins in the *E. coli* genome that match a particular protein. Line 12 in each tier is a histogram that indicates the CAI of each ORF. Genes with intermediate CAI values are shown in orange, genes with high CAI values (>90th percentile) are a darker shade of orange, genes with low CAI values (<10th percentile) are light brown, and clusters of four or more genes with low CAI values (<0.25) are yellow. The final line in each tier is a scale showing position (in base pairs).

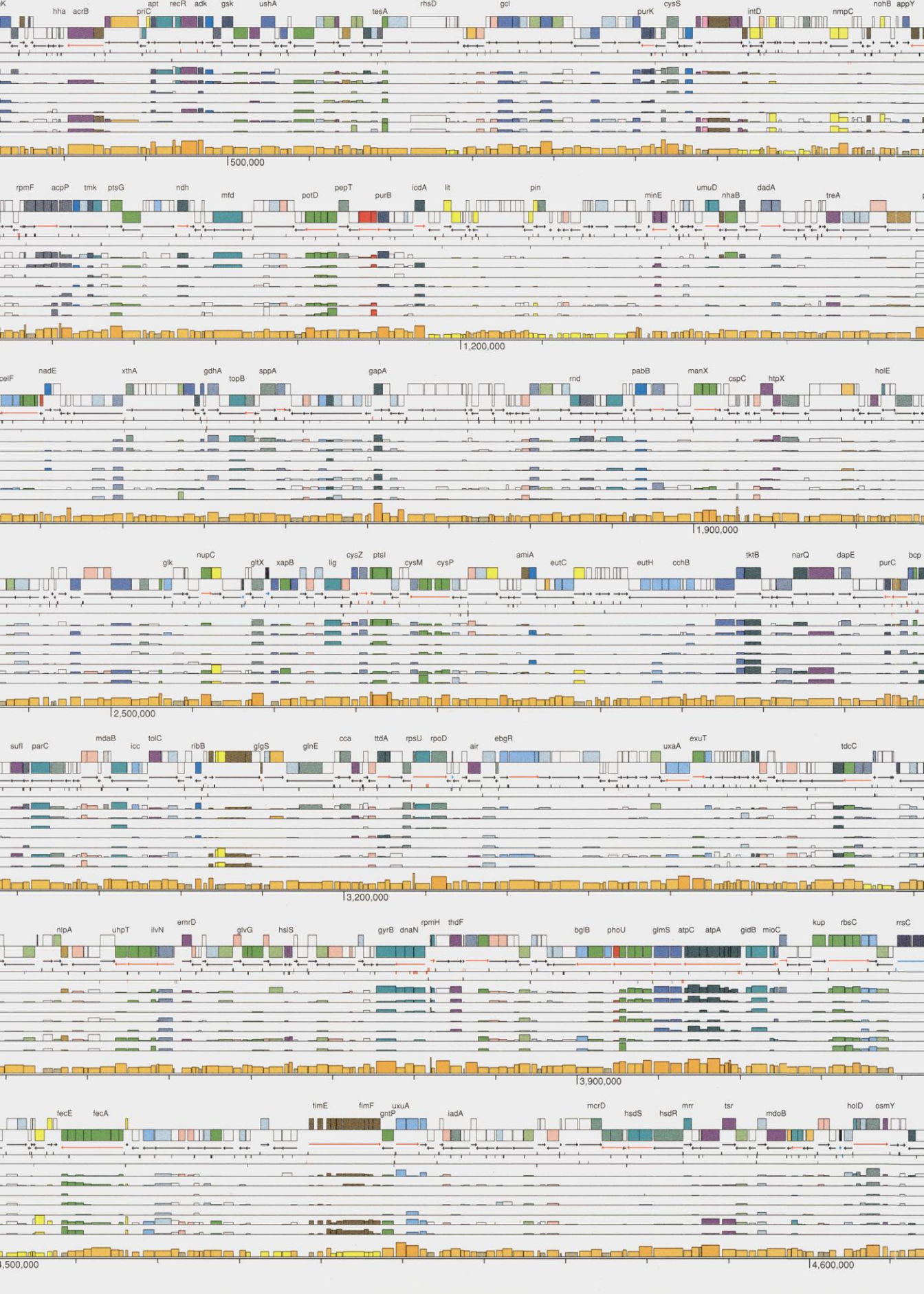

genes
operons
promoters
PB sites
Haemophilus
Synechocystis
Mycoplasma
Methanococcus
Saccharomyces
Best E. coli hit
log(E. coli hits)
CAI
0
700,000
1,400,000
2,100,000
2,700,000
3,400,000
4,100,000

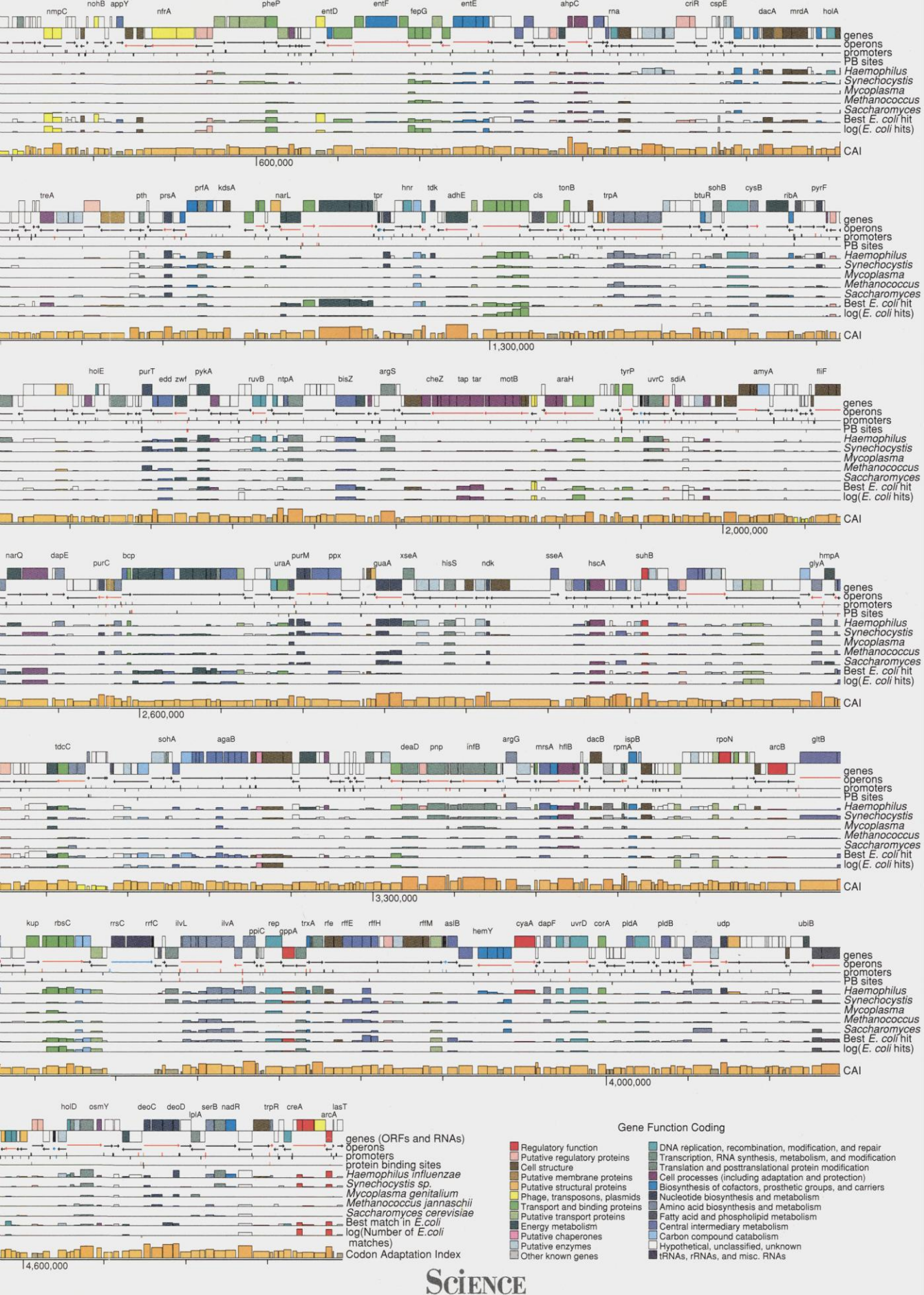

murE ftsW ddlB lpxC mutT guaC ampD ppdD pdhR lpdA acnB speD gcd panD htrE pcnB sfsA hrpB mrcB hemL pfs htrA glnD rpsB cdsA
100,000
sucA hrsA cydA tolQ pal nadA aroG modF modB bioA bioF uvrB moaA rhlE dinG glnQ dps ompX moeB
800,000
hslJ tynA acpD hrpA aldA cybB rimL tehB rhsE narV narU
1,500,000
rfc rfbX galF cpsG wcaB dcd alkA baeS gatC metG mrp
2,100,000
2,200,000
ileY gabD stpA nrdE proV emrR gshA alaS srlA mltB gutM hypF hycI hypA mutS rpoS pcm cysC iap
2,800,000
hotF pshM tufA rpsG slyX prkB argD ppiA nirB cysG trpS dam aroB hofQ mrcA pckA greB bioH gntT malT
3,500,000
hslU cytR rpmE metL katG gldA ptsA pflD ppc udhA trmA btuB rrsB rrfB tufB rplK rpoB rpoC thiF hemE purD rrsE rrfE aceB arp metH
4,200,000


glnD rpsB cdsA hlpA rnhB accA mesJ cutF rcsF abc rrsH rrfH dniR dnaQ gmhA mbhA prfH gpt crl proB argF
200,000
moeB dacC mdaA potF grxA artM poxB aqpZ cspD clpA aat trxB lrp lolA dmsA pflA serC cmk himD msbA
900,000
rhsE narV narU fdnG sfcA osmC xasA hipA uxaB marR dcp nohA
1,600,000
bglX dld cdd mglC galS cirA nfo fruK bcr rplY narP ccmF napC eco alkB
2,200,000
2,300,000
cysC iap cysH eno relA barA syd sdaC exo gcvA mltA argA ptr recC thyA ptsP mutH aas galR araE
2,900,000
bioH gntT malT glpR glgP glgC asd gntK ggt livF livH rpoH nikA rhsB pitA prlC gor arsC slp
3,600,000
aceB arp metH pepE lysC pgi xylI lamB ubiC plsB lexA qor alr uvrA ssb acs nrfA gltP fdhF rpiR phnL
4,300,000


argF
eaeH
betA
betT
codB
lacZ
mhpB
adhC
hemB
300,000
serC
cmk
himD
msbA
kdsB
smtA
aspC
asnS
pepN
pyrD
pqiA
rmf
sulA
helD
hyaA
appC
cspG
torS
cbpA
1,000,000
dcp
nohA
cspB
rem
dicA
rspB
speG
mlc
asr
pntB
stA
tus
manA
gusC
gusR
malX
add
nth
gst
pdxH
sodC
rnt
sodB
1,700,000
napC
eco
alkB
rcsB
atoS
gyrA
nrdA
inaA
glpB
ais
menE
menF
nuoN
nuoG
lrhA
2,300,000
2,400,000
mutH
galR
aas
araE
glyU
lysS
recJ
gcvP
visC
pepP
serA
iciA
pgk
cmtA
3,000,000
prlC
gor
arsC
slp
gadA
treF
kdgK
dctA
dppF
dppA
tag
cspA
glyS
xylB
xylR
avtA
lyxK
3,700,000
rpiR
phnL
proP
adiY
melR
melB
dcuB
lysU
cadA
dsbD
efp
frdB
psd
amiB
miaA
hflC
vacB
aidB
4,400,000


adhC
hemB
sbmA
ddlA
phoA
proC
aroL
araJ
sbcC
phoB
brnQ
malZ
secD
tsx
ribD
ispA
thiJ
cyoE
ampG
bolA
clpP
lon
cof
glnK
hha
acrB
400,000
cspG
torS
cbpA
agp
putA
phoH
csgG
mdoG
htrB
dinI
grxB
rimJ
figM
figF
figK
rne
rpmF
acpP
tmk
ptsG
1,100,000
sodC
rnt
sodB
purR
cfa
pykF
aroD
ppsA
aroH
nlpC
pheT
infC
pfkB
katE
celF
nadE
xthA
1,800,000
nuoG
lrhA
ackA
hisP
argT
purF
accD
usg
div
aroC
argW
vacJ
dsdC
emrY
evgA
2,400,000
2,500
pgk
cmtA
tktA
speB
metK
sprT
gshB
ansB
mutY
nupG
glcB
pitB
gsp
metC
exbD
sufI
parC
mdaB
3,100,000
ylR
avtA
lyxK
aldB
selB
rhsA
mtlA
lldP
cysE
tdh
rfaD
rfaL
rfaI
kdtA
radC
dut
dinD
gmk
recG
nlpA
uhpT
3,800,000
vacB
aidB
rpsF
cycA
cpdB
msrA
chpS
pmbA
treC
mgtA
argI
valS
pepA
gntV
intB
fecE
fecA
00
4,500,000


hha acrB
apt recR adk gsk ushA
priC
tesA
rhsD
gcl
purK
cysS
intD
nmpC
nohB appY
500,000
rpmF acpP tmk ptsG
ndh
mfd
potD
pepT
purB
icdA lit
pin
minE
umuD nhaB
dadA
treA
1,200,000
celF
nadE
xthA
gdhA topB
sppA
gapA
rnd
pabB
manX
cspC
htpX
holE
1,900,000
glk
nupC
gltX xapB
lig
cysZ ptsI
cysM cysP
amiA
eutC
eutH cchB
tktB
narQ
dapE
purC
bcp
2,500,000
sufI parC
mdaB
icc tolC
ribB
glgS
glnE
cca
ttdA
rpsU rpoD
air
ebgR
uxaA
exuT
tdcC
3,200,000
nlpA
uhpT
ilvN
emrD
glvG
hslS
gyrB dnaN
rpmH thdF
bglB
phoU
glmS atpC atpA
gidB mioC
kup
rbsC
rrsC
3,900,000
fecE fecA
fimE
fimF
uxuA
gntP
iadA
mcrD
hsdS hsdR
mrr
tsr
mdoB
holD osmY
4,500,000
4,600,000


nmpC nohB appY nfrA pheP entD entF fepG entE ahpC rna criR cspE dacA mrdA holA
genes
operons
promoters
PB sites
Haemophilus
Synechocystis
Mycoplasma
Methanococcus
Saccharomyces
Best E. coli hit
log(E. coli hits)
CAI
600,000
treA pth prsA prfA kdsA narL tpr hnr tdk adhE cls tonB trpA btuR sohB cysB ribA pyrF
1,300,000
holE purT edd zwf pykA ruvB ntpA bisZ argS cheZ tap tar motB araH tyrP uvrC sdiA amyA fliF
2,000,000
narQ dapE purC bcp uraA purM ppx guaA xseA hisS ndk sseA hscA suhB glyA hmpA
2,600,000
tdcC sohA agaB deaD pnp infB argG mrsA hflB dacB rpmA ispB rpoN arcB gltB
3,300,000
kup rbsC rrsC rrfC ilvL ilvA ppiC rep gppA trxA rfe rffE rffH rffM aslB hemY cyaA dapF uvrD corA pldA pldB udp ubiB
4,000,000
holD osmY deoC deoD lplA serB nadR trpR creA arcA lasT
genes (ORFs and RNAs)
operons
promoters
protein binding sites
Haemophilus influenzae
Synechocystis sp.
Mycoplasma genitalium
Methanococcus jannaschii
Saccharomyces cerevisiae
Best match in E.coli
log(Number of E.coli matches)
Codon Adaptation Index
4,600,000
Gene Function Coding
Regulatory function
Putative regulatory proteins
Cell structure
Putative membrane proteins
Putative structural proteins
Phage, transposons, plasmids
Transport and binding proteins
Putative transport proteins
Energy metabolism
Putative chaperones
Putative enzymes
Other known genes
DNA replication, recombination, modification, and repair
Transcription, RNA synthesis, metabolism, and modification
Translation and posttranslational protein modification
Cell processes (including adaptation and protection)
Biosynthesis of cofactors, prosthetic groups, and carriers
Nucleotide biosynthesis and metabolism
Amino acid biosynthesis and metabolism
Fatty acid and phospholipid metabolism
Central intermediary metabolism
Carbon compound catabolism
Hypothetical, unclassified, unknown
tRNAs, rRNAs, and misc. RNAs

SCIENCE