Whole-Genome Random Sequencing and Assembly of Haemophilus Influenzae Rd
Author(s): Robert D. Fleischmann, Mark D. Adams, Owen White, Rebecca A. Clayton, Ewen F. Kirkness, Anthony R. Kerlavage, Carol J. Bult, Jean-Francois Tomb, Brian A. Dougherty, Joseph M. Merrick, Keith McKenney, Granger Sutton, Will FitzHugh, Chris Fields, Jeannie D. Gocyne, John Scott, Robert Shirley, Li-Ing Liu, Anna Glodek, Jenny M. Kelley, Janice F. Weidman, Cheryl A. Phillips, Tracy Spriggs, Eva Hedblom, Matthew D. Cotton, Teres ...
Source: *Science*, New Series, Vol. 269, No. 5223 (Jul. 28, 1995), pp. 496-512
Published by: American Association for the Advancement of Science
Stable URL: http://www.jstor.org/stable/2887657
Accessed: 07/10/2008 13:56

RESEARCH ARTICLE

Whole-Genome Random Sequencing and Assembly of *Haemophilus influenzae* Rd

Robert D. Fleischmann, Mark D. Adams, Owen White, Rebecca A. Clayton, Ewen F. Kirkness, Anthony R. Kerlavage, Carol J. Bult, Jean-Francois Tomb, Brian A. Dougherty, Joseph M. Merrick, Keith McKenney, Granger Sutton, Will FitzHugh, Chris Fields,* Jeannine D. Gocayne, John Scott, Robert Shirley, Li-Ing Liu, Anna Glodek, Jenny M. Kelley, Janice F. Weidman, Cheryl A. Phillips, Tracy Spriggs, Eva Hedblom, Matthew D. Cotton, Teresa R. Utterback, Michael C. Hanna, David T. Nguyen, Deborah M. Saudek, Rhonda C. Brandon, Leah D. Fine, Janice L. Fritchman, Joyce L. Fuhrmann, N. S. M. Geoghagen, Cheryl L. Gnehm, Lisa A. McDonald, Keith V. Small, Claire M. Fraser, Hamilton O. Smith, J. Craig Venter†

An approach for genome analysis based on sequencing and assembly of unselected pieces of DNA from the whole chromosome has been applied to obtain the complete nucleotide sequence (1,830,137 base pairs) of the genome from the bacterium *Haemophilus influenzae* Rd. This approach eliminates the need for initial mapping efforts and is therefore applicable to the vast array of microbial species for which genome maps are unavailable. The *H. influenzae* Rd genome sequence (Genome Sequence DataBase accession number L42023) represents the only complete genome sequence from a free-living organism.

A prerequisite to understanding the complete biology of an organism is the determination of its entire genome sequence. Several viral and organellar genomes have been completely sequenced. Bacteriophage ϕX174 [5386 base pairs (bp)] was the first to be sequenced, by Fred Sanger and colleagues in 1977 (*1*). Sanger *et al.* were also the first to use strategy based on random (unselected) pieces of DNA, completing the genome sequence of bacteriophage λ (48,502 bp) with cloned restriction enzyme fragments (*1*). Subsequently, the 229-kb genome of cytomegalovirus (CMV) (*2*), the 192-kb genome of vaccinia (*3*), and the 187-kb mitochondrial and 121-kb chloroplast genomes of *Marchantia polymorpha* (*4*) have been sequenced. The 186-kb genome of variola (smallpox) was the first to be completely sequenced with automated technology (*5*).

At the present time, there are active genome projects for many organisms, including *Drosophila melanogaster* (*6*), *Escherichia coli* (*7*), *Saccharomyces cerevisiae* (*8*), *Bacillus subtilis* (*9*), *Caenorhabditis elegans* (*10*), and *Homo sapiens* (*11*). These projects, as well as viral genome sequencing, have been based primarily on the sequencing of clones usually derived from extensively mapped restriction fragments, or λ or cosmid clones. Despite advances in DNA sequencing technology (*12*) the sequencing of genomes has not progressed beyond clones on the order of the size of λ (~40 kb). This has been primarily because of the lack of sufficient computational approaches that would enable the efficient assembly of a large number (tens of thousands) of independent, random sequences into a single assembly.

The computational methods developed to create assemblies from hundreds of thousands of 300- to 500-bp complementary DNA (cDNA) sequences (*13*) led us to test the hypothesis that segments of DNA several megabases in size, including entire microbial chromosomes, could be sequenced rapidly, accurately, and cost-effectively by applying a shotgun sequencing strategy to whole genomes. With this strategy, a single random DNA fragment library may be prepared, and the ends of a sufficient number of randomly selected fragments may be sequenced and assembled to produce the complete genome. We chose the free-living organism *Haemophilus influenzae* Rd as a pilot project because its genome size (1.8 Mb) is typical among bacteria, its G+C base composition (38 percent) is close to that of human, and a physical clone map did not exist.

Haemophilus influenzae is a small, nonmotile, Gram-negative bacterium whose only natural host is human. Six *H. influenzae* serotype strains (a through f) have been identified on the basis of immunologically distinct capsular polysaccharide antigens. Non-typeable strains also exist and are distinguished by their lack of detectable capsular polysaccharide. They are commensal residents of the upper respiratory mucosa of children and adults and cause otitis media and respiratory tract infections, mostly in children. More serious invasive infection is caused almost exclusively by type b strains, with meningitis producing neurological sequelae in up to 50 percent of affected children. A vaccine based on the type b capsular antigen is now available and has dramatically reduced the incidence of the disease in Europe and North America.

Genome sequencing. The strategy for a shotgun approach to whole genome sequencing is outlined in Table 1. The theory follows from the Lander and Waterman (*14*) application of the equation for the Poisson distribution. The probability that a base is not sequenced is $P_0 = e^{-m}$, where m is the sequence coverage. Thus after 1.83 Mb of sequence has been randomly generated for the *H. influenzae* genome ($m = 1$, 1 × coverage), $P_0 = e^{-1} = 0.37$ and approximately 37 percent of the genome is unsequenced. Fivefold coverage (approximately 9500 clones sequenced from both insert ends and an average sequence read length of 460 bp) yields $P_0 = e^{-5} = 0.0067$, or 0.67 percent unsequenced. If L is genome length and n is the number of random sequence segments done, the total gap length is Le^{-m}, and the average gap size is L/n. Fivefold coverage would leave about 128 gaps averaging about 100 bp in size.

To approximate the random model during actual sequencing, procedures for library construction (*15*) and cloning (*16*) were developed. Genomic DNA from *H. influenzae* Rd strain KW20 (*17*) was mechanically sheared, digested with BAL 31 nuclease to produce blunt ends, and size-fractionated by agarose gel electrophoresis. Mechanical shearing maximizes the randomness of the DNA fragments. Fragments between 1.6 and 2.0 kb in size were excised and recovered. This narrow range was chosen to minimize variation in growth of clones. In addition, we chose this maximum size to minimize the number of complete genes that might be present in a single fragment, and thus might be lost as a result of expression of deleterious gene products. These fragments were ligated to Sma I–cut, phosphatase-treated pUC18 vector, and the ligated products were fractionated on an agarose gel. The linear vector plus insert band was excised and recovered. The ends of the linear recombinant molecules were repaired with T4 polymerase, and the molecules were then ligated into circles. This two-

J.-F. Tomb, B. A. Dougherty, and H. O. Smith are with the Johns Hopkins University School of Medicine, Baltimore, MD 21205, USA. J. M. Merrick is with the State University of New York, Department of Microbiology, Buffalo, NY, 14214, USA. K. McKenney is with the National Institute for Standards and Technology, Gaithersburg, MD 20878, USA. All other authors are with The Institute for Genomic Research (TIGR), Gaithersburg, MD, 20878, USA. The address for TIGR as of 9 September 1995 is 9712 Medical Center Drive, Rockville, MD 20850, USA.

*Present address: The National Center for Genome Resources, Santa Fe, NM, 87505, USA.

†To whom correspondence should be addressed.

stage procedure resulted in a collection of single-insert plasmid recombinants with minimal contamination from double-insert chimeras (<1 percent) or free vector (<3 percent). Because deviation from randomness is most likely to occur during cloning, *E. coli* host cells deficient in all recombination and restriction functions (*18*) were used to prevent rearrangements, deletions, and loss of clones by restriction. Transformed cells were plated directly on antibiotic diffusion plates (*16*) to avoid the usual broth recovery phase that would have allowed multiplication and selection of the most rapidly growing cells and could lead to deviation from randomness. All colonies were used for template preparation regardless of size. Only clones lost because of expression of deleterious gene products would be deleted from the library, resulting in a slight increase in gap number over that expected.

To evaluate the quality of the *H. influenzae* library, sequence data were obtained from ~4000 templates by means of the M13-21 primer. Sequence fragments were assembled with the AUTOASSEMBLER software [Applied Biosystems division of Perkin-Elmer (AB)] after obtaining 1300, 1800, 2500, 3200, and 3800 sequence fragments, and the number of unique assembled base pairs was determined. The data obtained from the assembly of up to 3800 sequence fragments were consistent with a Poisson distribution of fragments with an average "read" length of 460 bp for a genome of 1.9 × 10^6 bp, indicating that the library was essentially random.

Plasmid DNA templates that were double-stranded and of high quality (19,687) were prepared by a method developed in collaboration with Advanced Genetic Technology Corporation (*19*). Plasmids were prepared in a 96-well format for all stages of DNA preparation from bacterial growth through final DNA purification. Template concentration was determined with Hoechst dye and a Millipore Cytofluor 2350. DNA concentrations were not adjusted, but low-yielding templates (<30 ng/μl) were identified where possible and not sequenced. Templates were also prepared from two *H. influenzae* λ genomic libraries (*20*). An amplified library was constructed in vector λ GEM-12 and an unamplified library was constructed in λ DASH II. Both libraries contained inserts in the size range of 15 to 20 kb. Liquid lysates (10 ml) were prepared from selected plaques and templates were prepared on an anion-exchange resin (Qiagen). Sequencing reactions were carried out on plasmid templates by means of a Catalyst LabStation (AB) and PRISM Ready Reaction Dye Primer Cycle Sequencing Kits (AB) for the M13 forward (M13-21) and the M13 reverse (M13RP1) primers (*21*). Dye terminator sequencing reactions were carried out on the λ templates on a Perkin-Elmer 9600 Thermocycler with the Applied Biosystems Prism Ready Reaction Dye Terminator Cycle Sequencing Kits. We used T7 and SP6 primers to sequence the ends of the inserts from the λ GEM-12 library and T7 and T3 primers to sequence the ends of the inserts from the λ DASH II library. Sequencing reactions (28,643) were performed by eight individuals using an average of 14 AB 373 DNA Sequencers per day over a 3-month period. All sequencing reactions were analyzed with the Stretch modification of the AB 373 sequencer. These sequencers were modified to include a heat plate and the height of the laser was reduced. With standard gel plates the "well-to-read" length was increased to 34 cm when standard sequencing plates were used and to 48 cm when 60-cm plates were used. The sequencing reactions in this project were analyzed primarily with a 34-cm well-to-read distance. The overall sequencing success rate was 84 percent for M13-21 sequences, 83 percent for M13RP1 sequences, and 65 percent for dye-terminator reactions. The average usable read length was 485 bp for M13-21 sequences, 444 bp for M13RP1 sequences, and 375 bp for dye-terminator reactions. The high-throughput sequencing phase of the project is summarized in Table 2.

We balanced the desirability of sequencing templates from both ends, in terms of ordering of contigs and reducing the cost of lower total number of templates, against shorter read lengths for sequencing reactions performed with the M13RP1 primer compared to the M13-21 primer. Approximately one-half of the templates were sequenced from both ends. Altogether, 9297 M13RP1 sequencing reactions were done. Random reverse sequencing reactions were done on the basis of successful forward sequencing reactions. Some M13RP1 sequences were obtained in a semidirected fashion; for example, M13-21 sequences pointing outward at the ends of contigs were chosen for M13RP1 sequencing in an effort to specifically order contigs. The semidirected strategy was effective, and clone-based ordering formed an integral part of assembly and gap closure.

In the course of our research on expressed sequence tags (ESTs), we developed a laboratory information management system for a large-scale sequencing laboratory (*22*). The system was designed to automate data flow wherever possible and to reduce user error. It has at its core a series of databases developed with the Sybase relational data management system. The databases store and correlate all information collected during the entire operation from template preparation to final analysis. Although the system was originally designed for EST projects, many of its features were applicable or easily modified for a genomic sequencing project. Because the raw output of the AB 373 sequencers is collected on a Macintosh system and our data management system is based on a Unix system, it was necessary to design and implement multiuser, client-server applications that allow the raw data as well as analysis results to flow seamlessly into the database with a minimum of user effort. To process data collected by the AB 3735, sequence files were first analyzed with FACTURA, an AB program that runs on the Macintosh and is designed for automatic vector sequence removal and end-trimming of sequence files. The Macintosh program ESP, written at The Institute for Genomic Research (TIGR), loaded the feature data extracted from sequence files by FACTURA to the Unix-based *H. influenzae* relational database. Assembly was accom-

Table 1. Whole-genome sequencing strategy.

Stage	Description
Random small insert and large insert library construction	Shear genomic DNA randomly to ~2 kb and 15 to 20 kb, respectively
Library plating	Verify random nature of library and maximize random selection of small insert and large insert clones for template production
High-throughput DNA sequencing	Sequence sufficient number of sequence fragments from both ends for 6× coverage
Assembly	Assemble random sequence fragments and identify repeat regions
Gap closure	
Physical gaps	Order all contigs (fingerprints, peptide links, λ clones, PCR) and provide templates for closure
Sequence gaps	Complete the genome sequence by primer walking
Editing	Inspect the sequence visually and resolve sequence ambiguities, including frameshifts
Annotation	Identify and describe all predicted coding regions (putative identifications, starts and stops, role assignments, operons, regulatory regions)

plished by first retrieving a specified set of sequence files and their associated features by means of STP, another TIGR program, which is an X-windows graphical interface that retrieves sequences from the database with user-defined queries.

TIGR ASSEMBLER is the software component that enabled us to assemble the *H. influenzae* genome. It simultaneously clusters and assembles fragments of the genome. In order to obtain the speed necessary to assemble more than 10^4 fragments, the algorithm builds a table of all 10-bp oligonucleotide subsequences to generate a list of potential sequence fragment overlaps. When TIGR ASSEMBLER is used, a single fragment begins the initial contig; to extend the contig, a candidate fragment is chosen with the best overlap based on oligonucleotide content. The current contig and candidate fragment are aligned by a modified version of the Smith-Waterman (23) algorithm, which provides for optimal gapped alignments. The contig is extended by the fragment only if strict criteria for the quality of the match are met. The match criteria include the minimum length of overlap, the maximum length of an unmatched end, and the minimum percentage match. The algorithm automatically lowers these criteria in regions of minimal coverage and raises them in regions with a possible repetitive element. The number of potential overlaps for each fragment determines which fragments are likely to fall into repetitive elements. Fragments representing the boundaries of repetitive elements and potentially chimeric fragments are often rejected on the basis of partial mismatches at the needs of alignments and excluded from the contig.

TIGR ASSEMBLER was designed to take advantage of clone size information coupled with sequence information from both ends of each template. It enforces the constraint that sequence fragments from two ends of the same template point toward one another in the contig and are located within a certain range of base pairs (definable for each clone on the basis of the insert length or the clone size range for a given library). In order for the assembly process to be successful it was essential that the sequence data be of the highest quality and that sequence fragment lengths be sufficient to span most small repeats. Less than 13 percent of our random sequence fragments were smaller than 400 bp after vector removal and end trimming. Assembly of 24,304 sequence fragments of *H. influenzae* required 30 hours of central processing unit time with the use of one processor on a SPARCenter 2000 containing 512 Mb of RAM. This process resulted in approximately 210 contigs. Because of the high stringency of the TIGR ASSEMBLER, all contigs were searched against each other with GRASTA, which is a modified version of the program FASTA (24). In this way, additional overlaps that enabled compression of the data set into 140 contigs were detected. The location of each fragment in the contigs and extensive information about the consensus sequence itself were loaded into the *H. influenzae* relational database.

Table 2. Summary of features of whole-genome sequencing of *H. influenzae* Rd.

Description	Number
Double-stranded templates	19,687
Forward-sequencing reactions (M13-21 primer)	19,346
Successful (%)	16,240 (84)
Average edited read length (bp)	485
Reverse sequencing reactions (M13RP1 primer)	9,297
Successful (%)	7,744 (83)
Average edited read length (bp)	444
Sequence fragments in random assembly	24,304
Total base pairs	11,631,485
Contigs	140
Physical gap closure	42
PCR	37
Southern analysis	15
λ clones	23
Peptide links	2
Terminator sequencing reactions*	3,530
Successful (%)	2,404 (68)
Average edited read length (bp)	375
Genome size (bp)	1,830,137
G+C content (%)	38
rRNA operons	6
rrnA, rrnC, rrnD (spacer region) (bp)	723
rrnB, rrnE, rrnF (spacer region) (bp)	478
tRNA genes identified	54
Number of predicted coding regions	1,743
Unassigned role (%)	736 (42)
No database match	389
Match hypothetical proteins	347
Assigned role (%)	1,007 (58)
Amino acid metabolism	68 (6.8)
Biosynthesis of cofactors, prosthetic groups, and carriers	54 (5.4)
Cell envelope	84 (8.3)
Cellular processes	53 (5.3)
Central intermediary metabolism	30 (3.0)
Energy metabolism	105 (10.4)
Fatty acid and phospholipid metabolism	25 (2.5)
Purines, pyrimidines, nucleosides and nucleotides	53 (5.3)
Regulatory functions	64 (6.3)
Replication	87 (8.6)
Transcription	27 (2.7)
Translation	141 (14.0)
Transport and binding proteins	123 (12.2)
Other	93 (9.2)

*Includes gap closure, walks on rRNA repeats, random end-sequencing of λ clones for assembly confirmation, and alternative reactions for ambiguity resolution.

After assembly, the relative positions of the 140 contigs were unknown. The program ASM_ALIGN, developed at TIGR, identified clones whose forward and reverse sequencing reactions indicated that they were in different contigs and ordered and displayed these relationships. With this program, the 140 contigs were placed into 42 groups totaling 42 physical gaps (no template DNA for the region) and 98 sequence gaps (template available for gap closure).

Four integrated strategies were developed to order contigs separated by physical gaps. Oligonucleotide primers were designed and synthesized from the end of each contig group. These primers were then available for use in one or more of the strategies outlined below:

1) DNA hybridization (Southern) analysis was done to develop a "fingerprint" for a subset of 72 of the above oligonucleotides. This procedure was based on the supposition that labeled oligonucleotides homologous to the ends of adjacent contigs should hybridize to common DNA restriction fragments, and thus share a similar or identical hybridization pattern or fingerprint (25). Adjacent contigs identified in this manner were targeted for specific PCR reactions.

2) Peptide links were made by searching each contig end with BLASTX (26) against a peptide database. If the ends of two contigs matched the same database sequence appropriately, then the two contigs were tentatively considered to be adjacent.

HI#	Identification	%Sim
0483	ATP Sase F0 β sub (atpF)	79
0481	ATP Sase F1 α sub (atpA)	95
0479	ATP Sase F1 β sub (atpD)	96
0482	ATP Sase F1 δ sub (atpH)	78
0478	ATP Sase F1 ε sub (atpC)	76
0480	ATP Sase F1 γ sub (atpG)	83
1274	ATP Sase sub 3 region prt (atp)	50
	Electron transport	
0885	C-type cytochrome biogenesis prt (copper tolerance) (cycZ)	68
1076	cytochrome oxidase d sub I (cydA)	82
1075	cytochrome oxidase d sub II (cydB)	78
0527	ferredoxin (fdx)	77
0372	ferredoxin (fdx)	84
0191	flavodoxin (fldA)	87
1362	NAD(P) transhydrogenase sub α (pntA)	84
1363	NAD(P) transhydrogenase sub β (pntB)	88
1278	NAD(P)H-flavin oxidoreductase	55
	Entner-Doudoroff	
0047	2-keto-3-deoxy-6-phosphogluconate aldolase (eda)	63
0049	2-keto-3-deoxy-*D*-gluconate kinase (kdgK)	64
	Fermentation	
0499	aldehyde DHase (aldH)	62
0774	butyrate-acetoacetate CoA-Tase sub A (ctfA)	75
0185	glutathione-dependent formaldehyde DHase (gd-faldH)	78
1305	hydrogenase gene region (hypE)	48
1636	phosphoenolpyruvate carboxylase (ppc)	80
0180	pyruvate formate-lyase (pfl)	93
0179	pyruvate formate-lyase activating enzyme (act)	85
1430	short chain alcohol DHase	69
	Gluconeogenesis	
1645	fructose-1,6-bisphosphatase (fbp)	84
0809	phosphoenolpyruvate carboxykinase (pckA)	83
	Glycolysis	
0447	1-phosphofructokinase (fruK)	74
0982	6-phosphofructokinase (pfkA)	84
0932	enolase (eno)	79
0524	fructose-bisphosphate aldolase (fba)	86
1576	glucose-6-P isomerase (pgi)	89
0001	G3PD (gap)	90
0525	phosphoglycerate kinase (pgk)	91
0757	phosphoglyceromutase (gpmA)	75
1573	pyruvate kinase type II (pykA)	87
0678	triosephosphate isomerase (tpiA)	81
	Pentose phosphate pathway	
0553	6-phosphogluconate DHase (gnd)	71
0558	glucose-6-P 1-DHase (G6PD)	65
1023	transketolase 1 (tktA)	88
	Pyruvate dehydrogenase	
1232	dihydrolipoamide acetyltransferase (aceF)	82
0193	dihydrolipoamide acetyltransferase (acoC)	49
1231	lipoamide DHase (lpdA)	92
1233	pyruvate DHase (aceE)	84
	Sugars	
0818	aldose 1-epimerase precursor (mro)	55
0055	*D*-mannonate hydrolase (uxuA)	86
1116	deoxyribose aldolase (deoC)	69
0613	fucokinase (fucK)	65
1012	fuculose-1-P aldolase (fucA)	52
0611	fuculose-1-P aldolase (fucA)	81
0819	galactokinase (galK)	99
0144	glucose kinase (glk)	53
0614	*L*-fucose isomerase (fucI)	85
1025	*L*-ribulose-P 4-epimerase (araD)	82
1108	mal inducer biosyn blocker (malY)	52
0142	*N*-acetylneuraminate lyase (nanA)	61
0505	ribokinase (rbsK)	75
1112	xylose isomerase (xylA)	87
1113	xylulose kinase	50
	TCA cycle	
1662	2-oxoglutarate DHase (sucA)	81
0025	acetate:SH-citrate lyase ligase (AMP)	68
0022	citrate lyase α chain (citF)	86
0023	citrate lyase β chain (citE)	81
0024	citrate lyase γ chain (citD)	72
1661	dihydrolipoamide succinyltransferase (sucB)	84
1398	fumarate hydratase (fumC)	74
1210	malate DHase (mdh)	85
1245	malic acid enzyme	68
1197	succinyl-CoA Sase α sub (sucD)	92
1196	succinyl-CoA Sase β sub (sucC)	80
	Fatty acid and phospholipid metabolism	
1062	(3R)-hydroxymyristol acyl carrier prt dehydrase (fabZ)	85
0734	1-acyl-glycerol-3-P acyltransferase (plsC)	78
0155	3-ketoacyl-acyl carrier prt RDase (fabG)	88
0771	Ac-CoA acetyltransferase (fadA)	80
0406	Ac-CoA carboxylase (accA)	88
0154	acyl carrier prt (acpP)	91
0076	acyl-CoA thioesterase II (tesB)	73
1533	β-ketoacyl-ACP Sase I (fabB)	84
0157	β-ketoacyl-acyl carrier prt Sase III (fabH)	80
0971	biotin carboxyl carrier prt (accB)	83
0972	biotin carboxylase (accC)	91
0919	CDP-diglyceride Sase (cdsA)	67
1325	*D*-3-hydroxydecanoyl-(acyl carrier-prt) dehydratase (fabA)	92
0335	diacylglycerol kinase (dgkA)	72
0426	fatty acid metabolism prt (fadR)	68
0748	glycerol-3-P acyltransferase (plsB)	76
0002	long chain fatty acid CoA ligase	53
0156	malonyl CoA acyl carrier prt transacylase (fabD)	82
0211	phosphatidylglycerophosphate phosphatase B (pgpB)	60
0123	phosphatidylglycerophosphate Sase (pgsA)	83
0160	phosphatidylserine DCase proenzyme (psd)	76
0425	phosphatidylserine Sase (pssA)	71
0689	prt D (hpd)	99
1734	short chain alcohol DHase homolog (envM)	85
1433	USG-1 prt (usg)	54
	Purines, pyrimidines, nucleosides, and nucleotides	
	2'-Deoxyribonucleotide metabolism	
0075	anaerobic ribonucleoside-triphosphate RDase (nrdD)	88
0133	deoxycytidine triphosphate deaminase (dcd)	87
0954	deoxyuridinetriphosphatase (dut)	91
1532	glutaredoxin (grx)	80
1660	ribonucleoside diphosphate RDase ß2 sub	93
1659	ribonucleoside-diphosphate RDase 1 α chain (nrdA)	92
1158	thioredoxin RDase (trxB)	86
0905	thymidylate Sase (thyA)	55
	Nucleotide and nucleoside interconversions	
1077	CTP Sase (pyrG)	90
1299	dGTP triphosphohydrolase (dgt)	58
0132	uridine kinase (udk)	85
	Purine ribonucleotide biosynthesis	
1616	5'-phosphoribosyl-5-amino-4-imidazole carboxylase II (purK)	72
1429	5'-phosphoribosyl-5-aminoimidazole Sase (purM)	87
1743	5'-guanylate kinase (gmk)	82
0349	adenylate kinase (adk)	100
0639	adenylosuccinate lyase (purB)	88
1633	adenylosuccinate Sase (purA)	87
1207	amidoPRTase (purF)	84
0752	formylglycineamide ribonucleotide Sase (purL)	82
1588	formyltetrahydrofolate hydrolase (purU)	85
0222	GMP Sase (guaA)	88
0221	inosine-5'-monophosphate DHase (guaB)	81
0876	nucleoside diphosphate kinase (ndk)	74
0888	phosphoribosylamine-Gly ligase (purD)	85
0887	phosphoribosylaminoimidazole carboxamide formyltransferase (purH)	87
1615	phosphoribosylaminoimidazole carboxylase catalytic sub (purE)	97
1428	phosphoribosylglycinamide formyltransferase (purN)	71
1609	phosphoribosylpyrophosphate Sase (prsA)	91
1726	SAICAR Sase (purC)	55
	Pyrimidine ribonucleotide biosynthesis	
1401	dihydroorotate DHase (pyrD)	77
0272	orotate PRTase (pyrE)	84
1225	orotidine 5'-monophosphate DCase	88
1224	orotidine-5'-monophosphate DCase (pyrF)	79
0459	uracil PRTase (pyrR)	74
	Salvage of nucleosides and nucleotides	
0583	2',3'-cyclic nucleotide 2'-phosphodiesterase (cpdB)	78
1230	adenine PRTase (apt)	83
0551	adenosine tetraphosphatase (apaH)	73
1350	cytidine deaminase (cda)	63
1646	cytidylate kinase (cmk)	77
1219	cytidylate kinase (cmk)	79
0518	purine-nucleoside phosphorylase (deoD)	90
1277	putative ATPase (mrp)	79
0529	thymidine kinase (tdk)	82
1228	uracil PRTase (upp)	94
0280	uridine phosphorylase (udp)	85
0674	xanthine-guanine PRTase	88
0692	xanthine-guanine PRTase	88
	Sugar-nucleotide biosynthesis and conversions	
0206	5'-nucleotidase (ushA)	55
1279	CMP-NeuNAc Sase (siaB)	64
0820	Gal-1-P uridylyltransferase (galT)	100
0812	Glc-P uridylyltransferase (galU)	86
0351	UDP-Glc 4-epimerase (galE)	99
0642	UDP-GlcNAc pyrophosphorylase (glmU)	83
	Regulatory functions	
0604	adenylate cyclase (cyaA)	100
0884	aerobic respiration control prt (arcA)	88
0220	aerobic respiration control sensor prt (arcB)	70
1052	araC-like transcription regulator	48
1209	Arg repressor prt (argR)	81
0236	arsC prt (arsC)	57
0462	ATP-dependent proteinase (lon)	88
0334	ATP:GTP 3'-pyrophosphotransferase (relA)	80
1127	carbon starvation prt (cstA)	54
0813	carbon storage regulator (csrA)	91
0957	cyclic AMP receptor (crp)	100
1200	cys regulon transcriptional activator (cysB)	79
0190	ferric uptake regulation prt (fur)	75
1453	fimbrial transcription regulation repressor (pilB)	53
1455	fimbrial transcription regulation repressor (pilB)	73
1260	folylpolyglutamate-dihydrofolate Sase expression regulator (accD)	83
1425	fumarate (and nitrate) reduction regulatory prt (fnr)	89
0821	galactose operon repressor (galS)	99
0754	glucokinase regulator	56
1194	Gly cleavage system transcriptional activator (gcvA)	69
1009	glycerol-3-P regulon repressor (glpR)	50
0619	glycerol-3-P regulon repressor (glpR)	77
0013	GTP-BP (era)	87
0877	GTP-BP (obg)	71
0571	hydrogen peroxide-inducible activator (oxyR)	86
0615	*L*-fucose operon activator (fucR)	56
0399	lacZ expression regulator (icc)	71
0224	Leu responsive regulatory prt (lrp)	53
1596	Leu responsive regulatory prt (lrp)	87
0749	LexA repressor (lexA)	85
1461	lipooligosaccharide prt (lex2A)	67
1611	maltose regulatory prt sfs1 (sfsA)	71
0294	metF aporepressor (metJ)	93
1473	molybdenum transport system (modD)	52
0199	msbB	67
0763	nadAB transcriptional regulator (nadR)	75
0710	negative regulator of translation (relB)	48
0629	negative rpo regulator (mclA)	63
0267	nitrate sensor prt (narQ)	63
0726	nitrate, nitrite response regulator prt (narP)	79
0337	nitrogen regulatory prt P-II (glnB)	94
1741	penta-P guanosine-3'-pyrophosphohydrolase (spoT)	77
1378	phosphate regulon sensor prt (phoR)	67
1379	phosphate regulon transcriptional regulatory prt (phoB)	72
1635	purine nucleotide synthesis repressor prt (purR)	74
0163	putative murein gene regulator (bolA)	66
0506	rbs repressor (rbsR)	71
0563	regulatory prt (asnC)	81
0893	repressor for cytochrome P450 (Bm3R1)	51
0269	RNA polymerase sigma-32 factor (rpoH)	87
0533	RNA polymerase sigma-70 factor (rpoD)	81
0628	RNA polymerase sigma-E factor (rpoE)	88
1707	sensor prt for basR (basS)	56
1440	stringent starvation prt (sspB)	81
1441	stringent starvation prt A (sspA)	87
1739	trans-activator of metE and metH (metR)	61
0358	transcription activator (tenA)	48
0681	transcriptional activator prt (ilvY)	70
1708	transcriptional regulatory prt (basR)	60
0410	transcriptional regulatory prt (tyrR)	67
0830	Trp repressor (trpR)	67
0054	uxu operon regulator (uxuR)	72
1106	xylose operon regluatory prt (xylR)	75
	Replication	
	Degradation of DNA	
1689	endonuclease III (nth)	92
0249	excinuclease ABC sub A (uvrA)	91
1247	excinuclease ABC sub B (uvrB)	88
0057	excinuclease ABC sub C (uvrC)	80
1377	exodeoxyribonuclease I (sbcB)	75
1321	exodeoxyribonuclease V (recB)	58
0942	exodeoxyribonuclease V (recC)	61
1322	exodeoxyribonuclease V (recD)	59
0041	exonuclease III (xthA)	84
0397	exonuclease VII, large sub (xseA)	74
1214	single-stranded DNA-specific exonuclease (recJ)	77
	DNA replication, restriction, modification, recombination, and repair	
0759	A/G-specific adenine glycosylase (mutY)	75
1226	chromosomal replication initiator (dnaA)	75
0993	chromosomal replication initiator (dnaA)	80
0314	crossover junction endodeoxyribonuclease (ruvC)	88
0209	DNA adenine methylase (dam)	71
1264	DNA gyrase, sub A (gyrA)	85
0567	DNA gyrase, sub B (gyrB)	86
0728	DNA helicase (recQ)	78
1188	DNA helicase II (uvrD)	98
1100	DNA ligase (lig)	80
0654	DNA 3-methyladenine glycosidase I (tagI)	76
0403	DNA mismatch repair prt (mutH)	81
0067	DNA mismatch repair prt (mutL)	67
0707	DNA mismatch repair prt (mutS)	84
0856	DNA polymerase I (polA)	77
0992	DNA polymerase III β sub (dnaN)	80
0923	DNA polymerase III δ sub (holA)	62
0455	DNA polymerase III δ' sub (holB)	57
0137	DNA polymerase III ε sub (dnaQ)	76
0739	DNA polymerase III α chain (dnaE)	86
1397	DNA polymerase III χ sub (holC)	99
0011	DNA polymerase III psi sub (holD)	59
0532	DNA primase (dnaG)	74
1740	DNA recombinase (recG)	80
0070	DNA repair prt (recN)	67
0657	DNA topoisomerase I (topA)	55
0566	dod	93
0062	dosage-dependent dnaK suppressor prt (dksA)	84
0946	formamidopyrimidine-DNA glycosylase (fpg)	75
0582	glucose-inhibited division prt (gidA)	87
0486	glucose-inhibited division prt (gidB)	78
0980	Hin recombinational enhancer BP (fis)	93
0512	HincII endonuclease (HincII)	98
1392	HindIII modification MTase (hindIIIM)	99
1393	HindIII restriction endonuclease (hindIIIR)	100
0313	Holliday junction DNA helicase (ruvA)	80
0312	Holliday junction DNA helicase (ruvB)	90
0676	integrase-recombinase prt (xerC)	74
0309	integrase-recombinase prt (xerD)	85
1313	integration host factor α sub (himA)	83
1221	integration host factor β sub (IHF-β) (himD)	77
0402	methylated-DNA--prt-Cys MTase (dat1)	62
0669	mioC	72
1041	modification methylase HgiDI (MHgiDI)	70
0513	modification methylase HincII (hincIIM)	99
0910	mutator mutT	72
0192	negative modulator of initiation of replication (seqA)	72
0546	primosomal prt n precursor (priB)	100
0339	primosomal prt replication factor (priA)	70
0387	probable ATP-dependent helicase (dinG)	51
0991	DNA, ATP-BP (recF)	76
0332	DNA repair prt (recO)	77
0600	recombinase (recA)	100
0061	recombination prt (rec2)	100
0443	recR prt (recR)	88
0599	regulatory prt (recX)	50
0649	rep helicase (rep)	83
1229	replication prt (dnaX)	70
1574	replicative DNA helicase (dnaB)	83
1040	restriction enzyme (hgiDIR)	64
1172	SAM Sase 2 (metX)	92
1424	shufflon-specific DNA recombinase (rci)	56
0250	single-stranded DNA BP (ssb)	98
1572	site-specific recombinase (rcb)	57
1365	topoisomerase I (topA)	84
0444	topoisomerase III (topB)	79
1529	topoisomerase IV sub A (parC)	85
1528	topoisomerase IV sub B (parE)	89
1258	transcription-repair coupling factor (mfd)	83
0216	type I restriction enzyme ECOK1 specificity prt (hsdS)	59
1287	type I restriction enzyme ECOR124/3 I M (hsdM)	54
0215	type I restriction enzyme ECOR124/3 I M (hsdM)	89
1285	type I restriction enzyme ECOR124/3 R (hsdR)	53
1056	type III restriction-modification ECOP15 enzyme (mod)	56
0018	uracil DNA glycosylase (ung)	80
	Transcription	
	Degradation of RNA	
0218	anticodon nuclease masking-agent (prrD)	86
1733	exoribonuclease II	68
0390	ribonuclease D (rnd)	65
0413	ribonuclease E (rne)	72
0138	ribonuclease H (rnh)	76
1059	ribonuclease HII	83
0014	ribonuclease III (rnc)	80
0273	ribonuclease PH (rph)	88
0999	RNase P (rnpA)	81
0324	RNase T (rnt)	81
	RNA synthesis, modification, and DNA transcription	
0616	ATP-dependent helicase (hepA)	74
0231	ATP-dependent RNA helicase (deaD)	79
0892	ATP-dependent RNA helicase (rhlB)	84
0422	ATP-dependent RNA helicase (srmB)	61
0802	DNA-directed RNA polymerase α chain (rpoA)	97
0515	DNA-directed RNA polymerase β chain (rpoB)	92
0514	DNA-directed RNA polymerase β' chain (rpoC)	91
1304	N utilization substance prt B (nusB)	71
0063	plasmid copy number control prt (pcnB)	73
0229	polynucleotide phosphorylase (pnp)	87
1742	RNA polymerase omega sub (rpoZ)	76
1459	sigma factor (algU)	49
0717	transcription antitermination prt (nusG)	84
1331	transcription elongation factor (greA)	90
0569	transcription elongation factor (greB)	79
1283	transcription factor (nusA)	84
0295	transcription termination factor rho (rho)	95
	Translation	
	Amino acyl tRNA synthetases and tRNA modification	
0814	Ala-tRNA Sase (alaS)	83
1583	Arg-tRNA Sase (argS)	84
1302	Asn-tRNA Sase (asnS)	91
0317	Asp-tRNA Sase (aspS)	85
0708	Cys-tRNA selenium Tase (selA)	76
0078	Cys-tRNA Sase (cysS)	87
1354	Gln-tRNA Sase (glnS)	87
0274	Glu-tRNA Sase (gltX)	84
0927	Gly-tRNA Sase α chain (glyQ)	95
0924	Gly-tRNA Sase β chain (glyS)	82

HI#	Identification	%Sim
0369	His-tRNA Sase (hisS)	79
0962	Ile-tRNA Sase (ileS)	78
0921	Leu-tRNA Sase (leuS)	82
1211	Lys-tRNA Sase (lysU)	84
0836	Lys-tRNA Sase analog (genX)	78
0623	Met-tRNA formyltransferase (fmt)	77
1276	Met-tRNA Sase (metG)	83
0394	peptidyl-tRNA hydrolase (pth)	81
1311	Phe-tRNA Sase α sub (pheS)	82
1312	Phe-tRNA Sase β sub (pheT)	80
0729	Pro-tRNA Sase (proS)	87
1644	pseudouridylate Sase I (hisT)	83
0245	queuosine biosyn prt (queA)	86
0200	selenium metabolism prt (selD)	80
0110	Ser-tRNA Sase (serS)	86
1367	Thr-tRNA Sase (thrS)	86
0202	tRNA (guanine-N1)-MTase (trmD)	93
0848	tRNA (U-5-)-MTase (trmA)	80
0068	tRNA δ(2)-isopentenylpyrophosphate Tase (trpX)	87
1606	tRNA nucleotidyltransferase (cca)	73
0244	tRNA-guanine transglycosylase (tgt)	91
0637	Trp-tRNA Sase (trpS)	86
1610	Tyr-tRNA Sase (tyrS)	73
1391	Val-tRNA Sase (valS)	83
	Degradation of proteins, peptides, and glycopeptides	
0875	aminopeptidase A (pepA)	58
1705	aminopeptidase a/i (pepA)	78
1614	aminopeptidase N (pepN)	76
0816	aminopeptidase P (pepP)	74
0714	ATP-dependent clp protease (clpP)	88
1597	ATP-dependent protease (sms)	92
0715	ATP-dependent protease ATPase sub (clpX)	83
0859	ATP-dependent protease ATP-binding sub (clpB)	89
0419	collagenase (prtC)	53
0150	HflC	78
0990	IgA1 protease (iga1)	100
0247	IgA1 protease (iga1)	57
1324	lon protease (lon)	47
0214	oligopeptidase A (prlC)	85
0675	peptidase D (pepD)	72
0587	peptidase E (pepE)	60
1348	peptidase T (pepT)	71
1259	periplasmic Ser protease Do (htrA)	74
0722	Pro dipeptidase (pepQ)	70
1682	protease (sohB)	74
1541	protease IV (sppA)	64
0151	protease for λ cII repressor (hflK)	73
0530	sialoglycoprotease (gcp)	92
	Nucleoproteins	
0186	DNA-BP	64
1491	DNA-BP (rdgB)	61
1587	DNA-BP H-NS (hns)	65
0430	DNA-BP HU-α	87
	Protein modification and translation factors	
0846	disulfide oxidoreductase (por)	100
0985	DNA processing chain A (dprA)	60
0914	elongation factor EF-Ts (tsf)	85
0578	elongation factor EF-Tu (tufB)	96
0632	elongation factor EF-Tu (tufB)	96
0579	elongation factor G (fusA)	92
0328	elongation factor P (efp)	86
0622	f-Met deformylase (def)	80
0069	Glu-ammonia-ligase adenylyltransferase (glnE)	70
0548	initiation factor IF-1 (infA)	99
1284	initiation factor IF-2 (infB)	85
1318	initiation factor IF-3 (infC)	95
1152	maturation of antibiotic MccB17 (pmbA)	79
1722	Met aminopeptidase (map)	80
0428	oxido-RDase (dsbB)	69
1561	peptide chain release factor 1 (prfA)	88
1212	peptide chain release factor 2 (prfB)	94
1735	peptide chain release factor 3 (prfC)	93
0079	peptidyl-prolyl cis-trans isomerase B (ppiB)	80
0808	ribosome releasing factor (frr)	85
0573	rotamase, peptidyl prolyl cis-trans isomerase (slyD)	73
0699	rotamase, peptidyl prolyl cis-trans isomerase (slyD)	79
0709	translation factor (selB)	65
1213	thiol:disulfide interchange prt (xprA)	67
	Ribosomal proteins: sthesis and modification	
0516	ribosomal prt L1 (rpL1)	93
0640	ribosomal prt L10 (rpL10)	89
0517	ribosomal prt L11 (rpL11)	94
0978	ribosomal prt L11 MTase (prmA)	83
1443	ribosomal prt L13 (rpL13)	96
0788	ribosomal prt L14 (rpL14)	98
0797	ribosomal prt L15 (rpL15)	91
0784	ribosomal prt L16 (rpL16)	96
0803	ribosomal prt L17 (rplQ)	92
0794	ribosomal prt L18 (rpL18)	91
0201	ribosomal prt L19 (rpL19)	98
0780	ribosomal prt L2 (rpL2)	93
1320	ribosomal prt L20 (rpL20)	97
0880	ribosomal prt L21 (rpL21)	86
0782	ribosomal prt L22 (rpL22)	97
0779	ribosomal prt L23 (rpL23)	83
0789	ribosomal prt L24 (rpL24)	86
1630	ribosomal prt L25 (rpL25)	77
0879	ribosomal prt L27 (rpL27)	91
0951	ribosomal prt L28 (rpL28)	95
0785	ribosomal prt L29 (rpL29)	87
0777	ribosomal prt L3 (rpL3)	92
0796	ribosomal prt L30 (rpL30)	86
0758	ribosomal prt L31 (rpL31)	86
0158	ribosomal prt L32 (rpL32)	86
0950	ribosomal prt L33 (rpL33)	91
0998	ribosomal prt L34 (rpL34)	93
1319	ribosomal prt L35 (rpL35)	84
0778	ribosomal prt L4 (rpL4)	93
0790	ribosomal prt L5 (rpL5)	96
0793	ribosomal prt L6 (rpL6)	90
0641	ribosomal prt L7/L12 (rpL7/L12)	92
0544	ribosomal prt L9 (rpL9)	86
1220	ribosomal prt S1 (rpS1)	89
0776	ribosomal prt S10 (rpS10)	99
0800	ribosomal prt S11 (rpS11)	96
0799	ribosomal prt S13 (rpS13)	93
0791	ribosomal prt S14 (rpS14)	95
1328	ribosomal prt S15 (rpS15)	87
1468	ribosomal prt S15 (rpS15)	87
0204	ribosomal prt S16 (rpS16)	85
0786	ribosomal prt S17 (rpS17)	94
0545	ribosomal prt S18 (rpS18)	95
0781	ribosomal prt S19 (rpS19)	98
0913	ribosomal prt S2 (rpS2)	89
0531	ribosomal prt S21 (rpS21)	87
0783	ribosomal prt S3 (rpS3)	93
0801	ribosomal prt S4 (rpS4)	95
0795	ribosomal prt S5 (rpS5)	96
0547	ribosomal prt S6 (rpS6)	87
1531	ribosomal prt S6 modification prt (rimK)	69
0580	ribosomal prt S7 (rpS7)	94
0792	ribosomal prt S8 (rpS8)	91
1442	ribosomal prt S9 (rpS9)	98
0010	ribosomal-prt-Ala acetyltransferase (rimI)	73
0581	streptomycin resistance prt (strA)	100
	Transport and binding proteins	
	Amino acids, peptides and amines	
1177	Arg permease (artM)	80
1178	Arg permease (artQ)	78
1179	Arg-BP (artI)	73
1180	Arg transport ATP-BP artP (artP)	83
0253	biopolymer transport prt (exbB)	99
0252	biopolymer transport prt (exbD)	55
1728	branched chain AA transport system II (braB)	50
0883	*D*-Ala permease (dagA)	65
1187	dipeptide permease (dppB)	79
1186	dipeptide permease (dppC)	83
1185	dipeptide transport ATP-BP (dppD)	84
1184	dipeptide transport ATP-BP (dppF)	87
1079	Gln permease (glnP)	59
1080	Gln-BP (glnH)	48
1530	Glu permease (gltS)	73
0408	Leu-specific transport prt (livG)	55
0226	LIV-II transport system (brnQ)	60
0213	oligopeptide-BP (oppA)	53
1124	oligopeptide-BP (oppA)	69
1123	oligopeptide permease (oppB)	61
1122	oligopeptide permease (oppC)	87
1121	oligopeptide permease ATP-BP (oppD)	85
1120	oligopeptide permease ATP-BP (oppF)	84
1638	peptide permease (sapA)	64
1639	peptide permease (sapB)	64
1640	peptide permease (sapC)	60
1641	peptide permease ATP-BP (sapD)	80
1154	proton Glu symport prt (gltP)	54
0590	putrescine permease (potE)	88
0289	Ser transporter (sdaC)	78
1346	spermidine-putrescine permease (potB)	84
1345	spermidine-putrescine permease (potC)	89
1347	spermidine-putrescine permease ATP-BP (potA)	83
1344	spermidine-putrescine-BP (potD)	72
0498	spermidine-putrescine-BP (potD)	75
0287	Trp-specific permease (mtr)	73
0528	Tyr-specific transport prt (tyrP)	65
0477	Tyr-specific transport prt (tyrP)	68
	Anions	
1691	hydrophilic membrane-bound prt (modC)	75
1692	hydrophobic membrane-bound prt (modB)	85
1381	integral membrane prt (pstA)	78
0354	nitrate transporter ATPase component (nasD)	58
1380	peripheral membrane prt B (pstB)	87
1382	peripheral membrane prt C (pstC)	79
1383	periplasmic phosphate-BP (pstS)	68
1604	phosphate permease	60
	Carbohydrates, organic alcohols, and acids	
0020	2-oxoglutarate/malate translocator	60
0153	Asp transport prt (dcuA)	70
0746	Asp transport prt (dcuA)	70
1110	*D*-xylose transport ATP-BP (xylG)	86
1111	*D*-xylose-BP (rbsB)	88
1712	enzyme I (ptsI)	84
0181	formate transporter	73
0448	fructose permease IIA/FPR component (fruB)	68
0446	fructose permease IIBC component (fruA)	72
0612	fucose operon prt (fucU)	80
1711	Glc phosphotransferase enzyme III (crr)	83
1017	glycerol uptake facilitator prt (glpF)	55
0690	glycerol uptake facilitator prt (glpF)	87
1015	gluconate permease (gntP)	56
0686	glycerol-3-phosphatase transporter (glpT)	79
0502	high affinity ribose transport prt (rbsA)	85
0503	high affinity ribose transport prt (rbsC)	86
0501	high affinity ribose transport prt (rbsD)	78
0610	*L*-fucose permease (fucP)	58
1218	*L*-lactate permease (lctP)	54
1729	lactam utilization prt (lamB)	60
0823	methylgalactoside permease ATP-BP (mglA)	85
0822	methylgalactoside-BP (mglB)	81
0824	methylgalactoside permease (mglC)	90
1690	Na+ and Cl- dependent GABA transporter	53
0736	Na+-dependent noradrenaline transporter	54
0504	periplasmic ribose-BP (rbsB)	87
1713	phosphohistidinoprotein-hexose phosphotransferase (ptsH)	88
0828	potassium channel homolog (kch)	80
1109	ribose permease (xylH)	84
	Cations	
0254	bacterioferritin comigratory prt (bcp)	80
0251	energy transducer (tonB)	98
1272	ferric enterobactin transport ATP-BP (fepC)	51
1470	ferric enterobactin transport ATP-BP (fepC)	55
1466	ferrichrome-iron receptor (fhuA)	49
1385	ferritin like prt (rsgA)	74
1384	ferritin like prt (rsgA)	79
1271	iron(III) dicitrate permease (fecD)	61
0361	iron(III) dicitrate transport ATP-BP (fecE)	56
1035	magnesium and cobalt transport prt (corA)	85
0097	major ferric iron-BP precursor (fbp)	82
1049	mercury transport prt (merT)	54
1050	mercury scavenger prt (merP)	46
0292	mercury scavenger prt (merP)	67
1525	molybdate-BP (modB)	43
0427	Na+,H+ antiporter (nhaB)	87
1107	Na+,H+ antiporter (nhaC)	62
0225	Na+,H+ antiporter 1 (nhaA)	75
0098	periplasmic-BP-dependent iron transport (sfuB)	59
1474	periplasmic-BP-dependent iron transport (sfuC)	58
0911	potassium efflux system (kefC)	66
0290	potassium, copper-transporting ATPase A (copA)	64
1352	sodium, Pro symporter (putP)	79
0625	TRK system potassium uptake prt (trkA)	83
	Nucleosides, purines and pyrimidines	
1087	ribonucleotide transport ATP-BP (mkl)	61
1227	uracil permease (uraA)	62
	Other	
0621	ATP-BP (abc)	87
0060	ATP-dependent translocator (msbA)	100
1619	cystic fibrosis transmembrane conductance regulator	61
0853	heme-binding lpp (dppA)	99
0264	heme-hemopexin-BP (hxuA)	89
1471	hemin permease (hemU)	63
0262	hemin receptor precursor (hemR)	46
1706	high-affinity choline transport prt (betT)	62
0661	lactoferrin-BP (lbpA)	48
0608	Na+, sulfate cotransporter	86
0975	pantothenate permease (panF)	78
0973	transferrin-BP (tfbA)	48
0712	transferrin-BP 1 (tbp1)	49
1565	transferrin-BP 1 (tbp1)	59
0994	transferrin-BP 1 (tbp1)	69
1217	transferrin-BP 1 (tbp1)	80
0635	transferrin-BP 1 (tbp2)	52
0995	transferrin-BP 2 (tbp2)	55
0663	transport ATP-BP (cydD)	54
1157	transport ATP-BP (cydD)	73
	Other categories	
	Adaptations and atypical conditions	
1526	autotrophic growth prt (aut)	61
0071	heat shock prt B253 (grpE)	66
0720	heat shock prt (htpX)	82
1527	heat shock prt B (ibpB)	71
0945	htrA-like prt (htrH)	73
0901	invasion prt (invA)	61
1544	NAD(P)H:menadione oxidoreductase	55
0458	survival prt (surA)	58
0815	universal stress prt (uspA)	87
1251	virulence assoc prt A (vapA)	58
0322	virulence assoc prt C (vapC)	57
0947	virulence assoc prt C (vapC)	61
0450	virulence assoc prt D (vapD)	67
1307	virulence plasmid prt (mlgA)	56
0321	virulence plasmid prt (vagC)	58
	Colicin-related functions	
0382	colicin tolerance prt (tolB)	78
1206	colicin V production prt (cvpA)	79
0384	inner membrane prt (tolR)	79
0385	inner membrane prt (tolQ)	83
1685	outer membrane integrity prt (tolA)	48
0383	outer membrane integrity prt (tolA)	57
	Drug and analog sensitivity	
0895	acriflavine resistance prt (acrB)	55
0300	ampD signalling prt (ampD)	75
1242	bicyclomycin resistance prt (bcr)	69
1623	mercury resistance regulatory prt (merR2)	58
0648	modulator of drug activity (mda66)	75
0897	multidrug resistance prt (emrB)	85
0898	multidrug resistance prt (ermA)	66
0036	multidrug resistance prt (mdl)	51
1462	nodulation prt T (nodT)	46
0549	rRNA (adenosine-N6,N6-)-dimethyltransferase (ksgA)	81
0511	tellurite resistance prt (tehA)	62
1275	tellurite resistance prt (tehB)	71
	Phage-related functions and prophages	
1488	E16 prt (muE16)	53
1503	G prt (muG)	52
1568	G prt (muG)	54
1483	gam prt	74
0411	host factor-I (HF-I) (hfq)	97
1504	I prt (muI)	55
1481	MuB prt (muB)	70
1515	N prt (muN)	52
1516	P prt (muP)	61
1411	terminase sub 1	52
1478	transposase A (muA)	60
	Radiation sensitivity	
0952	DNA repair prt (radC)	72
	Transposon-related functions	
1577	IS1016-V6	61
1329	IS1016-V6	75
1018	IS1016-V6	94
	Other	
1161	15 kD prt (P15)	68
0085	2-hydroxyacid dehydrogenase (ddh)	73
0460	β-lactamase regulatory prt (mazG)	73
0223	chloramphenicol-sensitive prt (rarD)	53
0680	chloramphenicol-sensitive prt (rarD)	55
1670	conjugative transfer co-repressor (finO)	52
0307	δ-1-pyrroline-5-carboxylate RDase (proC)	60
1549	heterocyst maturation prt (devA)	66
1339	embryonic abundant prt, group 3	89
0916	export factor homolog (skp)	76
0937	extragenic suppressor (suhB)	80
0667	glp regulon prt (glpX)	83
1013	glyoxylate-induced prt	58
0497	heat shock prt (hslU)	90
0496	heat shock prt (hslV)	89
1117	ilv-related prt	77
0285	isochorismate Sase (entC)	49
1618	membrane assoc ATPase (cbiO)	53
0461	membrane prt (lapB)	56
1119	membrane prt (lapB)	80
0630	mucoid status locus prt (mucB)	52
0588	*N*-carbamyl-*L*-amino acid amidohydrolase	59
1295	nitrogen fixation prt (nifS)	56
1343	nitrogen fixation prt (nifS)	59
0378	nitrogen fixation prt (nifS)	67
0377	nitrogen fixation prt (nifU)	74
0166	nitrogen fixation prt (rnfE)	48
1686	nitrogen fixation prt (rnfE)	59
0129	nitrogenase C (nifC)	53
1475	nitrogenase C (nifC)	60
1296	partitioning system prt (parB)	68
0171	phenolhydroxylase	57
0368	prt E (gpcE)	94
0556	putative glucose-6-P DHase isozyme (devB)	52
0981	small prt (smpB)	91
1592	spoIIIE prt (spoIIIE)	75
0095	spore germination and vegetative growth prt (gerC2)	55
0896	suppressor prt (msgA)	56
1078	surfactin (sfpo)	78
0357	thiamine-repressed prt (nmt1)	55
0751	toxR regulon (tagD)	64
1407	traN	62
0664	transport ATP-BP (cydC)	52
1156	transport ATP-BP (cydC)	70
1556	vanamycin-resistance prt (vanH)	57

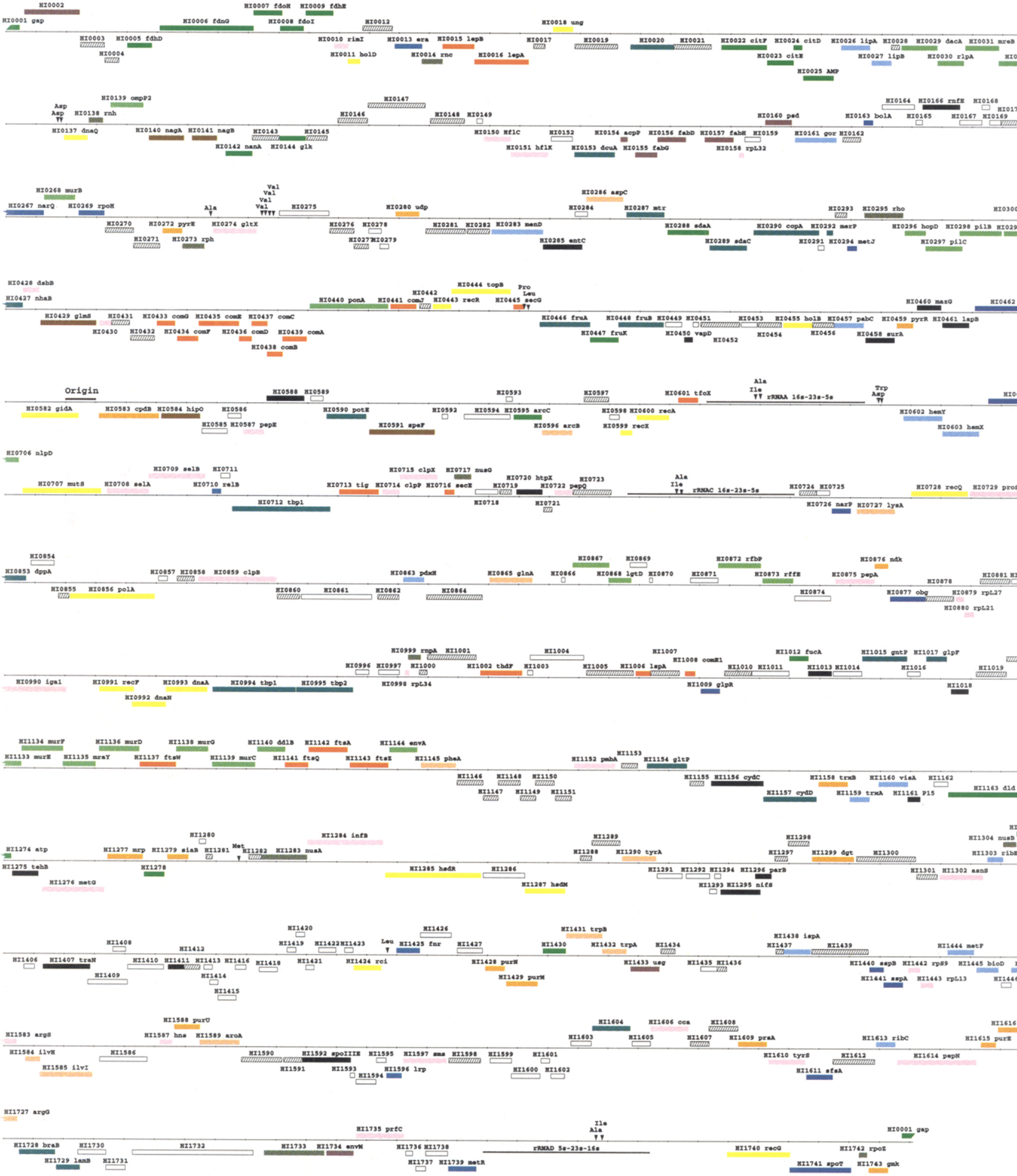
SCIENCE

The Genome of *Haemophilus influenzae* Rd

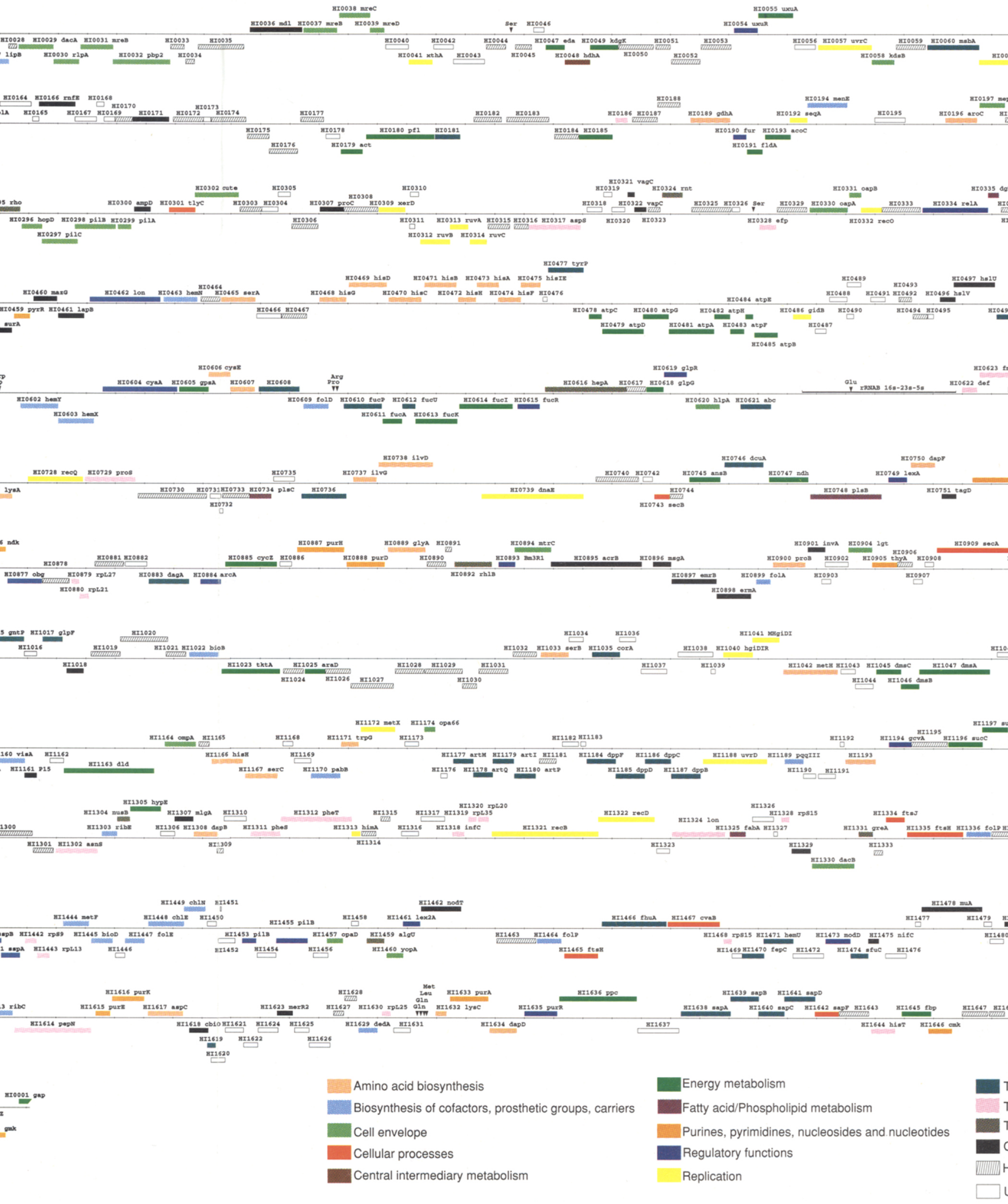
Figure 2. Gene map of the *H. influenzae* Rd genome. Predicted coding regions are shown on each strand. The rRNA and tRNA genes are shown as lines and triangles, respectively. Genes are color-coded by role category as described in the Figure key. Gene identification numbers correspond to those in Table 3. Where possible, three-letter designations are also provided. In the region containing ribosomal proteins HI0782-HI0796 some identification numbers have been omitted because of space limitations. Predicted coding regions with similarity to database sequences designated as hypothetical coding regions are represented as white, cross-hatched rectangles. Predicted coding regions that have no database match are represented as white, unfilled rectangles.

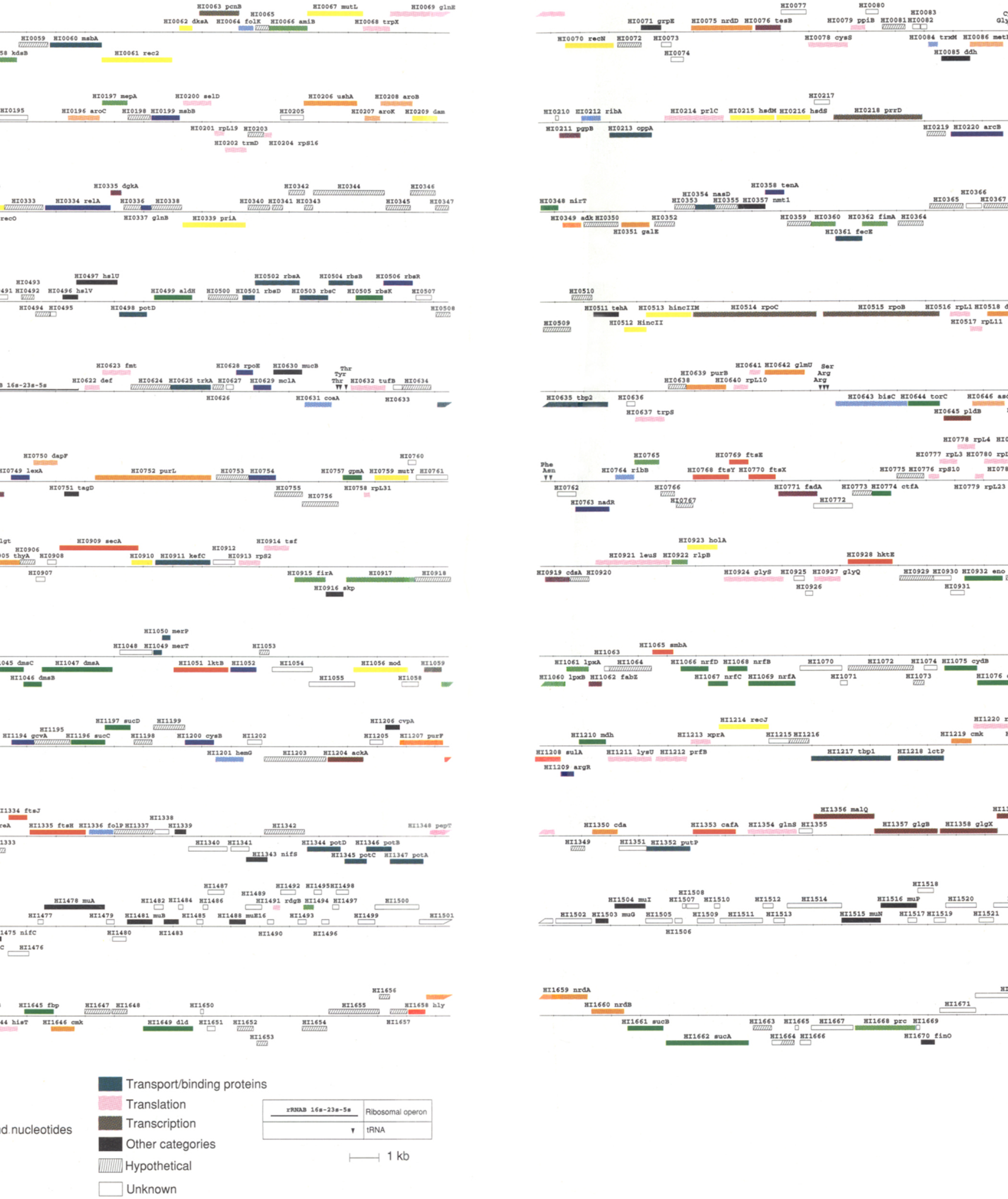
Table 3. Identification of *H. influenzae* genes. Gene identification numbers are listed with the prefix HI in Fig. 3. Each identified gene is listed in its role category [adapted from Riley (*36*)]. The percentage of similarity (Sim) of the best match to the NRBP (as described in the text) is also shown. The amino acid substitution matrix used in the BLAZE analysis is BLOSUM60. An expanded version of this table with additional match information, including species, is available via World Wide Web (URL: http://www.tigr.org/). Abbreviations used: Ac, acetyl; ATase, aminotransferase; BP, binding protein; biosyn, biosynthesis; CoA, coenzyme A; DCase, decarboxylase; DHase, dehydrogenase; DMSO, dimethyl sulfoxide; f-Met, formylmethionine; G3PD, glyceraldehyde-3-phosphate dehydrogenase; GABA, γ-aminobutyric acid; GlcNAc, *N*-acetylglucosamine; LOS, Lipooligosaccharide; lpp, lipoprotein; MTase, methyltransferase; MurNAc, *N*-acetylmuramyl; P, phosphate; prt, protein; PRTase, phosphoribosyltransferase; RDase, reductase; SAM, *S*-adenosylmethionine; Sase, synthase-synthetase; sub, subunit; Tase, transferase. The following hypothetical proteins were matched from the other species as indicated (percent similarity in parentheses after gene identification number): *Alcaligenes eutrophus*: 1053(52); *Anabaena variabilis*: 1349(54); *Bacillus subtilis*: 0115(53), 0259(54), 0355(61), 0404(47), 0415(69), 0416(63), 0417(66), 0454(64), 0456(56), 0522(54), 0687(49), 0775(54), 0959(50), 1083(53), 1203(63), 1627(59), 1647(81), 1648(65), 1654(64); Bacteriophage P22: 1412(54); *Buchnera aphidicola*: 1199(65); *Campylobacter jejuni*: 0560(71); *Chromatium vinosum*: 0105(75); *Clostridium acetobutylicum*: 0773(72); *Clostridium kluyveri*: 0976(48); *Clostridium perfringens*: 0143(58); *Coxiella burnetii*: 1590(74), 1591(50); *Erwinia carotovora*: 1436(72); *Escherichia coli*: 0003(52), 0012(67), 0017(91), 0028(68), 0033(90), 0034(84), 0035(79), 0044(80), 0045(67), 0050(70), 0051(50), 0052(56), 0053(56), 0059(72), 0065(75), 0072(65), 0081(71), 0091(72), 0092(49), 0093(59), 0103(71), 0107(54), 0108(65), 0125(88), 0126(87), 0135(68), 0145(69), 0146(58), 0147(61), 0148(62), 0162(47), 0172(67), 0174(84), 0175(70), 0176(87), 0182(60), 0183(66), 0184(73), 0187(58), 0188(81), 0198(75), 0203(86), 0227(51), 0230(71), 0232(69), 0235(80), 0241(82), 0242(50), 0258(95), 0257(76), 0265(77), 0266(83), 0270(80), 0271(73), 0276(70), 0281(76), 0282(59), 0293(61), 0303(81), 0306(70), 0308(58), 0315(87), 0316(68), 0329(79), 0336(91), 0338(68), 0340(72), 0341(84), 0342(60), 0343(67), 0344(85), 0345(82), 0346(77), 0347(67), 0364(55), 0365(86), 0367(48), 0371(84), 0374(64), 0375(62), 0376(75), 0379(57), 0380(58), 0386(76); 0393(93), 0396(54), 0398(72), 0400(65), 0409(69), 0412(85), 0418(68), 0423(67), 0424(66), 0431(76), 0432(68), 0442(93), 0452(73), 0464(78), 0467(80), 0493(64), 0494(69), 0500(63), 0508(82), 0509(69), 0510(74), 0519(71), 0520(59), 0521(58), 0562(83), 0565(63), 0568(71), 0570(80), 0572(70), 0574(63), 0575(80), 0576(65), 0597(57), 0617(54), 0624(72), 0626(81), 0634(78), 0638(68), 0647(64), 0656(74), 0658(56), 0668(76), 0670(83), 0671(87), 0696(54), 0697(64), 0700(77), 0702(71), 0719(86), 0721(78), 0723(73), 0724(64), 0730(65), 0733(55), 0744(70), 0755(61), 0756(60), 0766(87), 0767(72), 0810(74), 0817(68), 0826(70), 0827(86), 0831(77), 0837(74), 0839(69), 0840(72), 0841(66), 0849(75), 0851(71), 0852(66), 0855(75), 0858(68), 0860(86), 0862(81), 0864(92), 0878(71), 0881(81), 0890(69), 0891(79), 0906(71), 0918(81), 0929(58), 0933(71), 0934(52), 0935(63), 0936(64), 0943(83), 0948(67), 0955(72), 0956(73), 0963(67), 0965(81), 0979(79), 0984(79), 0986(81), 0988(85), 1000(80), 1001(75), 1005(61), 1007(86), 1010(53), 1019(65), 1020(65), 1021(71), 1024(67), 1026(85), 1027(72), 1028(77), 1029(83), 1030(62), 1031(87), 1032(79), 1064(57), 1072(57), 1073(62), 1082(67), 1084(61), 1085(76), 1086(89), 1089(70), 1090(82), 1091(76), 1092(73), 1093(72), 1094(81), 1095(79), 1096(64), 1104(53), 1118(84), 1125(87), 1129(77), 1130(80), 1146(80), 1147(68), 1148(88), 1149(73), 1150(59), 1151(81), 1153(84), 1155(79), 1165(87), 1181(68), 1195(76), 1198(85), 1216(73), 1234(80), 1240(77), 1243(74), 1252(93), 1262(61), 1280(71), 1282(74), 1288(84), 1289(74), 1297(67), 1298(69), 1300(58), 1301(82), 1309(67), 1314(70), 1315(66), 1333(79), 1337(84), 1342(57), 1364(56), 1368(53), 1369(44), 1437(72), 1463(84), 1542(61), 1545(80), 1558(62), 1598(58), 1608(76), 1612(72), 1628(61), 1643(70), 1652(68), 1653(88), 1655(56), 1656(69), 1657(65), 1664(50), 1677(72), 1679(69), 1703(74), 1704(73), 1714(78), 1715(86), 1721(71), 1723(92); *Klebsiella pneumoniae*: 0021(63); *Lactobacillus johnsonii*: 0112(54), 1720(55); *Lactococcus lactis:* 0555(69); *Mycobacterium leprae*: 0004(62), 0019(62), 0136(58), 0260(56), 0694(54), 0740(56), 0920(57), 1663(55); *Mycoplasma hyopneumoniae*: 1281(71); *Pasteurella haemolytica*: 0219(92); *Pseudomonas aeruginosa*: 0090(68), 0177(56); *Rhodobacter capsulatus*: 0170(62), 0672(59), 1439(65), 1683(75), 1684(60), 1688(58); *Salmonella typhimurium*: 0405(51), 0964(67), 1434(76), 1607(51); *Shigella flexneri:* 0277(52); *Streptococcus parasanguis*: 0359(65); *Synechococcus* sp.: 0961(70); *Vibrio parahaemolyticus*: 0323(87), 0325(75); *Vibrio* sp.: 0333(70); *Yersinia enterocolitica*: 0753(69).

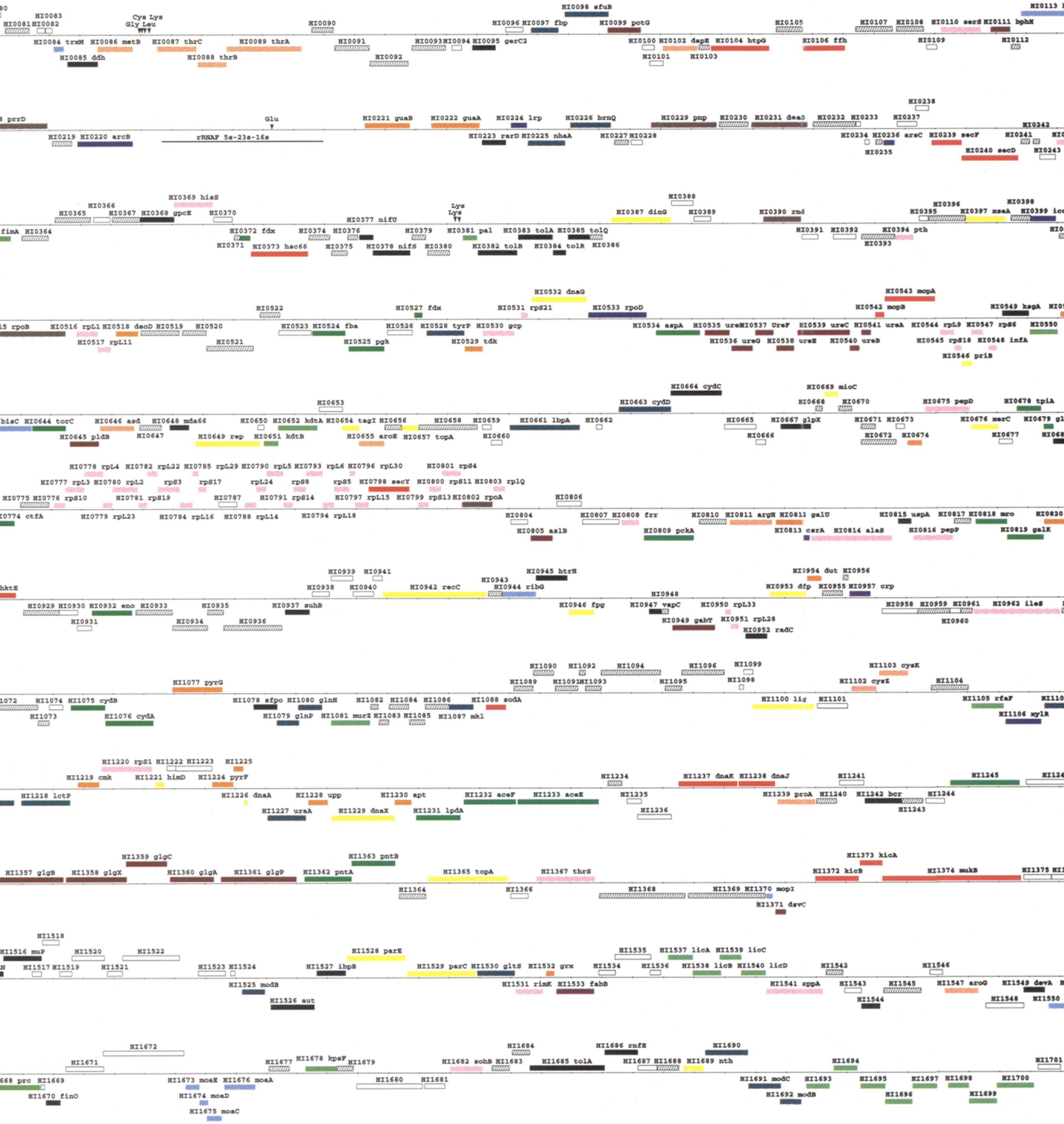
HI#	Identification	%Sim
	Amino acid biosynthesis	
	Aromatic amino acid family	
0970	3-dehydroquinase (aroQ)	83
0208	3-dehydroquinate Sase (aroB)	77
0472	amidotransferase (hisH)	70
1387	anthranilate Sase component I (trpE)	73
1388	anthranilate Sase component II (trpD)	74
1389	anthranilate isomerase (trpC)	75
1171	anthranilate Sase Gln amidotransferase (trpG)	59
0468	ATP PRTase (hisG)	82
1290	chorismate mutase (tyrA)	77
1145	chorismate mutase-prephenate dehydratase (pheA)	75
0196	chorismate Sase (aroC)	88
1547	DAHP Sase (aroG)	84
0607	dehydroquinase shikimate DHase	48
1589	enolpyruvylshikimatephosphateSyn (aroA)	98
1166	Gln amidotransferase (hisH)	61
0469	histidinol dehydrogenase (hisD)	78
0474	hisF cyclase (hisF)	91
0470	histidinol-P ATase (hisC)	77
0471	imidazoleglycerol-P dehydratase (hisB)	81
0475	phosphoribosyl-AMP cyclohydrolase (hisIE)	77
0473	phosphoribosylformimino-5-aminoimidazole caarboximde ribotide isomerase (hisA)	77
0655	shikimate 5-DHase (aroE)	70
0207	shikimic acid kinase I (aroK)	88
1432	Trp Sase α chain (trpA)	73
1431	Trp Sase β chain (trpB)	90
	Aspartate family	
0564	Asn Sase A (asnA)	77
0286	Asp ATase (aspC)	54
1617	Asp ATase (aspC)	79
0646	Asp-semialdehyde DHase (asd)	85
1632	aspartokinase III (lysC)	73
0089	aspartokinase-homoserine DHase (thrA)	77
1042	B12-dependent homocysteine-N5-methyltetrahydrofolate transmethylase (metH)	70
0122	β-cystathionase (metC)	84
0086	cystathionine γ-Sase (metB)	62
1308	dehydrodipicolinate RDase (dapB)	83
0727	diaminopimelate DCase (lysA)	79
0750	diaminopimelate epimerase (dapF)	86
0255	dihydrodipicolinate Sase (dapA)	80
1263	homoserine acetyltransferase (met2)	57
0088	homoserine kinase (thrB)	81
0102	succinyl-diaminopimelate desuccinylase (dapE)	80
1634	tetrahydrodipicolinate *N*-succinyltransferase (dapD)	99
1702	tetrahydropteroyltriglutamate MTase (metE)	68
0087	Thr Sase (thrC)	81
	Branched chain family	
0989	3-isopropylmalate dehydratase (leuD)	86
0987	3-isopropylmalate DHase (leuB)	80
0737	acetohydroxy acid Sase II (ilvG)	79
1585	acetolactate Sase III large chain (ilvI)	84
1584	acetolactate Sase III small chain (ilvH)	85
1193	branched-chain amino acid transaminase	49
0738	dihydroxyacid dehydrase (ilvD)	90
0983	α isopropylmalate Sase (leuA)	100
0682	ketol acid reductoisomerase (ilvC)	90
	Glutamate family	
0811	argininosuccinate lyase (argH)	84
1727	argininosuccinate Sase (argG)	87
0900	γ-glutamyl kinase (proB)	80
1239	γ-glutamyl-P RDase (proA)	79
0865	Gln Sase (glnA)	86
0189	Glu DHase (gdhA)	84
0596	ornithine carbamoyltransferase (arcB)	91
1719	uridylyl Tase (glnD)	68
	Pyruvate family	
1575	Ala racemase, biosynthetic (alr)	75
	Serine family	
1102	Cys Sase (cysZ)	76
1103	Cys Sase (cysK)	84
0465	phosphoglycerate DHase (serA)	84
1167	phosphoserine ATase (serC)	72
1033	phosphoserine phosphatase (serB)	70
0606	Ser acetyltransferase (cysE)	88
0889	Ser hydroxymethyltransferase (glyA)	94
	Biosynthesis of cofactors, prosthetic groups, and carriers	
	Biotin	
1554	7,8-diamino-pelargonic acid ATase (bioA)	74
1553	7-keto-8-aminopelargonic acid Sase (bioF)	56
1551	biotin synthesis prt (bioC)	47
0643	biotin sulfoxide RDase (bisC)	72
1022	biotin Sase (bioB)	78
1550	dethiobiotin Sase (bioD)	60
1445	dethiobiotin Sase (bioD)	62
	Folic acid	
1444	5,10-methylenetetrahydrofolate RDase (metF)	83
0609	5,10-methylenetetrahydrofolate DHase (folD)	82
0064	7,8-dihydro-6-hydroxymethylpterin-pyrophosphokinase (folK)	78
0457	aminodeoxychorismate lyase (pabC)	67
1629	dedA	55
0899	dehydrofolate RDase, type I (folA)	68
1336	dihydropteroate Sase (folP)	71
1464	dihydropteroate Sase (folP)	71
1261	folylpolyglutamate Sase (folC)	68
1447	GTP cyclohydrolase I (folE)	79
1170	p-aminobenzoate Sase (pabB)	54
	Heme and porphyrin	
1160	ferrochelatase (visA)	69
0113	heme utilization prt (hxuC)	46
0263	heme-hemopexin utilization (hxuB)	99
0463	oxygen-independent coproporphyrinogen III oxidase (hemN)	52
0602	protoporphyrinogen oxidase homolog	64
1201	protoporphyrinogen oxidase (hemG)	57
1559	protoporphyrinogen oxidase (hemG)	73
0603	uroporphyrinogen III methylase (hemX)	60
	Lipoate	
0026	lipoate biosyn prt A (lipA)	84
0027	lipoate biosyn prt B (lipB)	84
	Menaquinone and ubiquinone	
0283	2-succinyl-6-hydroxy-2,4-cyclohexadiene-1-carboxylate Sase (menD)	64
0969	4-(2'-carboxyphenyl)-4-oxybutyric acid Sase (menC)	74
1189	coenzyme PQQ synthesis prt III (pqqIII)	49
0968	dihydroxynaphthoic acid Sase (menB)	95
1438	farnesyldiphosphate Sase (ispA)	71
0194	*O*-succinylbenzoate-CoA Sase (menE)	67
	Molybdopterin	
1676	molybdenum biosyn prt A (moaA)	78
1675	molybdenum biosyn prt C (moaC)	89
1370	molybdenum-pterin-BP (mopI)	74
1448	molybdopterin biosyn prt (chlE)	73
0118	molybdopterin biosyn prt (chlN)	53
1449	molybdopterin biosyn prt (chlN)	78
1674	molybdopterin converting factor, sub 1 (moaD)	79
1673	molybdopterin converting factor, sub 2 (moaE)	76
0844	molybdopterin-dinucleotide biosyn prt (mob)	62
	Pantothenate	
0953	pantothenate metabolism flavoprotein (dfp)	77
0631	pantothenate kinase (coaA)	78
	Pyridoxine	
0863	pyridoxamine phosphate oxidase (pdxH)	65
	Riboflavin	
0764	3,4-dihydroxy-2-butanone 4-P Sase (ribB)	83
0212	GTP cyclohydrolase II (ribA)	81
0944	riboflavin biosyn prt (ribG)	76
1613	riboflavin Sase α chain (ribC)	82
1303	riboflavin Sase β chain (ribE)	90
	Thioredoxin, glutaredoxin, and glutathione	
0161	glutathione RDase (gor)	85
1115	thioredoxin (trxA)	59
1159	thioredoxin (trxA)	62
0084	thioredoxin m (trxM)	79
	Cell envelope	
	Membranes, lipoproteins, and porins	
1579	15 kD peptidoglycan-assoc lpp (lpp)	95
0620	28 kD membrane prt (hlpA)	100
0302	apolipoprotein *N*-acyltransferase (cute)	64
0407	hydrophobic membrane prt	61
0360	hydrophobic membrane prt	67
1567	iron-regulated outer membrane prt A (iroA)	51
0693	lpp (hel)	100
0706	lpp (nlpD)	65
0703	lpp B (lppB)	90
0894	membrane fusion prt (mtrC)	54
0401	outer membrane prt P1 (ompP1)	97
0139	outer membrane prt P2 (ompP2)	98
1164	outer membrane prt P5 (ompA)	96
0904	prolipoprotein diacylglyceryl Tase (lgt)	80
0030	rare lpp A (rlpA)	58
0922	rare lpp B (rlpB)	62
	Murein sacculus and peptidoglycan	
1140	*D*-Ala-*D*-Ala ligase (ddlB)	76
1330	*D*-alanyl-*D*-Ala carboxypeptidase (dacB)	68
1138	GlcNAc transferase (murG)	76
1494	MurNAc-*L*-Ala amidase	62
0066	*N*-acetylmuramoyl-*L*-Ala amidase (amiB)	77
0440	penicillin-BP (ponA)	100
1725	penicillin-BP 1B (ponB)	67
0032	penicillin-BP 2 (pbp2)	74
1668	penicillin-BP 3 (prc)	70
0029	penicillin-BP 5 (dacA)	68
0197	penicillin-insensitive murein endopeptidase (mepA)	67
0381	peptidoglycan-assoc outer membrane lpp (pal)	100
1135	phospho-*N*-acetylmuramoyl-pentapeptide-Tase E (mraY)	89
0031	rod shape-determining prt (mreB)	81
0037	rod shape-determining prt (mreB)	90
0038	rod shape-determining prt (mreC)	74
0039	rod shape-determining prt (mreD)	72
0829	soluble lytic murein transglycosylase (slt)	59
1081	UDP-GlcNAc enolpyruvyl Tase (murZ)	85
1139	UDP-MurNAc-Ala ligase (murC)	82
1136	UDP-MurNAc-Ala-*D*-Glu ligase (murD)	74
1134	UDP-MurNAc-pentapeptide Sase (murF)	68
1133	UDP-MurNAc-tripeptide Sase (murE)	73
0268	UDP-NAc-enolpyruvoylglucosamine RDase (murB)	76
	Surface polysaccharides, lipopolysaccharides and antigens	
1557	2-dehydro-3-deoxyphosphooctonate aldolase (kdsA)	92
0652	3-deoxy-*D*-manno-octulosonic-acid Tase (kdtA)	70
1105	ADP-heptose-lps heptosyltransferase II (rfaF)	79
1114	ADP-*L*-glycero-*D*-mannoheptose-6-epimerase (rfaD)	88
0058	CTP:CMP-3-deoxy-*D*-manno-octulosonate-cytidylyl-transferase (kdsB)	82
0868	glycosyl Tase (lgtD)	55
1578	glycosyl Tase (lgtD)	64
1678	kpsF prt (kpsF)	71
1537	lic-1 operon prt (licA)	100
1538	lic-1 operon prt (licB)	99
1539	lic-1 operon prt (licC)	99
1540	lic-1 operon prt (licD)	94
1060	lipid A disaccharide Sase (lpxB)	77
0765	LOS biosyn prt	60
0550	LOS biosyn prt	99
0651	lipopolysaccharide core biosyn prt (kdtB)	76
1700	lsg locus prt 1	100
0867	lsg locus prt 1	83
1699	lsg locus prt 2	99
1698	lsg locus prt 3	97
1697	lsg locus prt 4	98
1696	lsg locus prt 5	98
1695	lsg locus prt 6	99
1694	lsg locus prt 7	98
1693	lsg locus prt 8	99
0261	lipopolysaccharide biosyn prt (opsX)	57
1716	rfe prt	77
1144	UDP-3-*O*-acyl GlcNAc deacetylase (envA)	88
0915	UDP-3-*O*-(R-3-hydroxymyristoyl)-glucosamine *N*-acetyltransferase (firA)	91
1061	UDP-GlcNAc acetyltransferase (lpxA)	79
0873	UDP-GlcNAc epimerase (rffE)	79
0872	undecaprenyl-P Gal-P Tase (rfbP)	75
	Surface structures	
0119	adhesin B precursor (fimA)	48
0362	adhesin B precursor (fimA)	62
0330	cell envelope prt (oapA)	100
0331	opacity assoc prt (oapB)	99
1174	opacity prt (opa66)	59
0414	opacity prt (opa66)	91
1457	opacity prt (opaD)	56
1460	outer membrane adhesin (yopA)	62
0299	pilin biogenesis prt (pilA)	52
0298	pilin biogenesis prt (pilB)	65
0297	pilin biogenesis prt (pilC)	57
0917	protective surface antigen D15	99
	Cellular processes	
	Cell division	
0769	cell division ATP-BP (ftsE)	78
1208	cell division inhibitor (sulA)	56
1142	cell division prt (ftsA)	74
1335	cell division prt (ftsH)	88
1465	cell division prt (ftsH)	88
1334	cell division prt (ftsJ)	90
1131	cell division prt (ftsL)	60
1141	cell division prt (ftsQ)	58
1137	cell division prt (ftsW)	75
0768	cell division prt (ftsY)	81
1143	cell division prt (ftsZ)	83
1374	cell division prt (mukB)	77
1353	cytoplasmic axial filament prt (cafA)	86
0770	cell division membrane prt (ftsX)	70
1065	mukB suppressor prt (smbA)	90
1132	penicillin-BP 3 (ftsI)	71
	Cell killing	
0301	hemolysin (tlyC)	58
1658	hemolysin, 21 kD (hly)	72
1373	killing prt (kicA)	84
1372	killing prt suppressor (kicB)	83
1051	leukotoxin secretion ATP-BP (lktB)	55
	Chaperones	
0373	heat shock cognate prt 66 (hsc66)	82
1238	heat shock prt (dnaJ)	83
1237	heat shock prt 70 (dnaK)	88
0104	heat shock prt C62.5 (htpG)	88
0543	heat shock prt groEL (mopA)	95
0542	heat shock prt groES (mopB)	95
	Detoxification	
0928	catalase (hktE)	99
1088	superoxide dismutase (sodA)	100
1002	thiophene and furan oxidation prt (thdF)	85
	Protein and peptide secretion	
1467	colicin V secretion ATP-BP (cvaB)	56
0016	GTP-binding membrane prt (lepA)	91
1006	lpp signal peptidase (lspA)	72
1642	peptide transport system ATP-BP (sapF)	71
0716	preprotein translocase (secE)	62
0798	preprotein translocase (secY)	87
0240	protein-export membrane prt (secD)	77
0239	protein-export membrane prt (secF)	73
0445	protein-export membrane prt (secG)	81
0743	protein-export prt (secB)	81
0909	preprotein translocase sub (secA)	82
0015	signal peptidase I (lepB)	65
0106	signal recognition particle prt 54 (ffh)	91
0713	trigger factor (tig)	80
0296	type 4 prepilin-like prt specific leader peptidase (hopD)	49
	Transformation	
1008	competence locus E (comE1)	70
0601	tfoX	100
0439	transformation prt (comA)	100
0438	transformation prt (comB)	100
0437	transformation prt (comC)	100
0436	transformation prt (comD)	100
0435	transformation prt (comE)	100
0434	transformation prt (comF)	100
	Central intermediary metabolism	
	Amino sugars	
0140	GlcNAc-6-P deacetylase (nagA)	72
0429	Gln amidotransferase (glmS)	84
0141	glucosamine-6-P deaminase (nagB)	88
	Degradation of polysaccharides	
1356	amylomaltase (malQ)	62
	Other	
0048	7-α-hydroxysteroid DHase (hdhA)	55
1204	acetate kinase (ackA)	84
0949	GABA transaminase (gabT)	56
0111	glutathione Tase (bphH)	57
0691	glycerol kinase (glpK)	89
0584	hippuricase (hipO)	50
0541	urease (ureA)	76
0539	urease α sub (urea amidohydrolase) (ureC)	82
0537	urease accessory prt (UreF)	55
0538	urease prt (ureE)	57
0536	urease prt (ureG)	87
0535	urease prt (ureH)	54
0540	urease sub B (ureB)	77
	Phosphorus compounds	
0695	exopolyphosphatase (ppx)	77
0124	inorganic PPase (ppa)	50
0645	lysophospholipase L2 (pldB)	53
	Polyamine biosynthesis	
0099	nucleotide-BP (potG)	67
0591	ornithine DCase (speF)	80
	Polysaccharides - (cytoplasmic)	
1357	1,4-α-glucan branching enzyme (glgB)	80
1361	α-glucan phosphorylase (glgP)	79
1359	ADP-glucose Sase (glgC)	74
1358	glycogen operon prt (glgX)	68
1360	glycogen Sase (glgA)	71
	Sulfur metabolism	
0805	arylsulfatase regulatory prt (aslB)	67
1371	desulfoviridin γ sub (dsvC)	58
0559	sulfite synthesis pathway prt (cysQ)	56
	Energy metabolism	
	Aerobic	
1163	*D*-lactate DHase (dld)	48
1649	*D*-lactate DHase (dld)	78
0605	glycerol-3-P DHase (gpsA)	81
0747	NADH DHase (ndh)	75
	Amino acids and amines	
0534	aspartase (aspA)	89
0595	carbamate kinase (arcC)	88
0745	*L*-asparaginase II (ansB)	81
0288	*L*-Ser deaminase (sdaA)	83
	Anaerobic	
1047	anaerobic DMSO RDase A (dmsA)	86
1046	anaerobic DMSO RDase B (dmsB)	85
1045	anaerobic DMSO RDase C (dmsC)	65
0644	cytochrome C-type prt (torC)	55
0348	denitrification system component (nirT)	72
0009	formate DHase pathway prt (fdhE)	72
0006	formate DHase (fdnG)	79
0005	formate DHase-N affector (fdhD)	71
0008	formate DHase-O γ sub (fdoI)	72
0007	formate DHase-O, β sub (fdoH)	86
1069	formate-dependent nitrite RDase (nrfA)	75
1068	formate-dependent nitrite RDase (nrfB)	67
1067	formate-dependent nitrite RDase prt Fe-S centers (nrfC)	81
1066	formate-dependent nitrite RDase transmembrane prt (nrfD)	68
0833	fumarate RDase (frdC)	72
0832	fumarate RDase 13 kD hydrophobic prt (frdD)	77
0835	fumarate RDase, flavoprotein sub (frdA)	87
0834	fumarate RDase, iron-sulfur prt (frdB)	85
0685	G3PD, sub A (glpA)	83
0684	G3PD, sub B (glpB)	60
0683	G3PD, sub C (glpC)	76
0679	glpE	63
0618	glpG	65
1390	hydrogenase isoenzymes formation prt (hypC)	82
	ATP-proton motive force interconversion	
0484	ATP Sase C chain (atpE)	82
0485	ATP Sase F0 α sub (atpB)	78

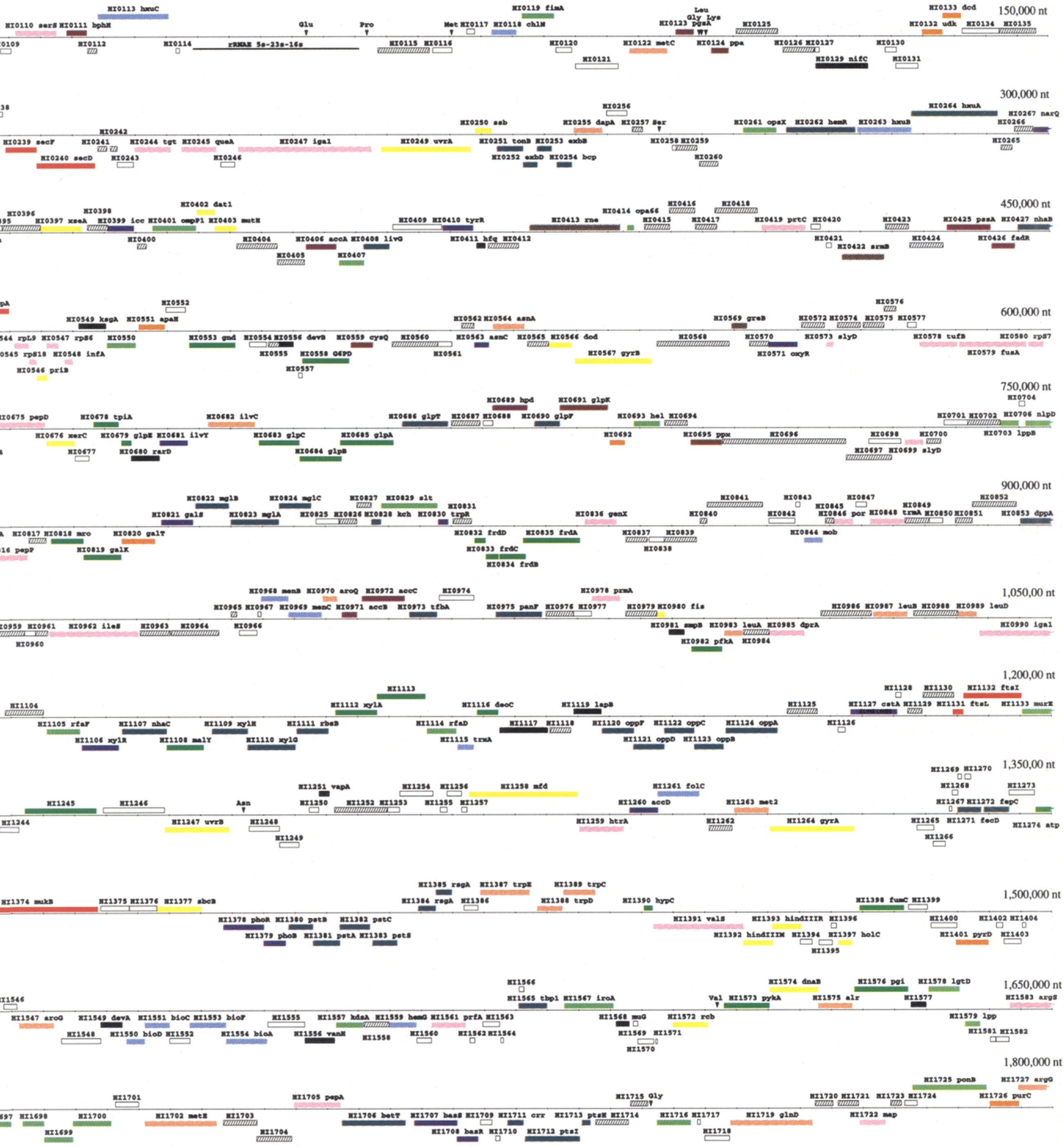


HI0002
HI0001 gap
HI0003
HI0004
HI0005 fdhD
HI0006 fdnG
HI0007 fdoH
HI0008 fdoI
HI0009 fdhE
HI0010 rimI
HI0011 holD
HI0012
HI0013 era
HI0014 rnc
HI0015 lepB
HI0016 lepA
HI0017
HI0018 ung
HI0019
HI0020
HI0021
HI0022 citF
HI0023 citE
HI0024 citD
HI0025 AMP
HI0026 lipA
HI0027 lipB
HI0028
HI0029 dacA
HI0030 rlpA
HI0031 mreB
HI0139 ompP2
Asp
Asp
HI0138 rnh
HI0137 dnaQ
HI0140 nagA
HI0141 nagB
HI0142 nanA
HI0143
HI0144 glk
HI0145
HI0146
HI0147
HI0148
HI0149
HI0150 Hflc
HI0151 hflK
HI0152
HI0153 dcuA
HI0154 acpP
HI0155 fabG
HI0156 fabD
HI0157 fabH
HI0158 rpL32
HI0159
HI0160 psd
HI0161 gor
HI0162
HI0163 bolA
HI0164
HI0165
HI0166 rnfE
HI0167
HI0168
HI0169
HI0268 murB
HI0267 narQ
HI0269 rpoH
HI0270
HI0271
HI0272 pyrE
HI0273 rph
Ala
HI0274 gltX
Val
Val
Val
Val
HI0275
HI0276
HI0277
HI0278
HI0279
HI0280 udp
HI0281
HI0282
HI0283 menD
HI0284
HI0285 entC
HI0286 aspC
HI0287 mtr
HI0288 sdaA
HI0289 sdaC
HI0290 copA
HI0291
HI0292 merP
HI0293
HI0294 metJ
HI0295 rho
HI0296 hopD
HI0297 pilC
HI0298 pilB
HI0300
HI0428 dsbB
HI0427 nhaB
HI0429 glmS
HI0430
HI0431
HI0432
HI0433 comG
HI0434 comF
HI0435 comE
HI0436 comD
HI0437 comC
HI0438 comB
HI0439 comA
HI0440 ponA
HI0441 comJ
HI0442
HI0443 recR
HI0444 topB
HI0445 secG
Pro
Leu
HI0446 fruA
HI0447 fruK
HI0448 fruB
HI0449
HI0450 vapD
HI0451
HI0452
HI0453
HI0454
HI0455 holB
HI0456
HI0457 pabC
HI0458 surA
HI0459 pyrR
HI0460 mazG
HI0461 lapB
HI0462
Origin
HI0582 gidA
HI0583 cpdB
HI0584 hipO
HI0585
HI0586
HI0587 pepE
HI0588
HI0589
HI0590 potE
HI0591 speF
HI0592
HI0593
HI0594
HI0595 arcC
HI0596 arcB
HI0597
HI0598
HI0599 recX
HI0600 recA
HI0601 tfoX
Ala
Ile
rRNAA 16s-23s-5s
Trp
Asp
HI0602 hemY
HI0603 hemX
HI0706 nlpD
HI0707 mutS
HI0708 selA
HI0709 selB
HI0710 relB
HI0711
HI0712 tbp1
HI0713 tig
HI0714 clpP
HI0715 clpX
HI0716 secE
HI0717 nusG
HI0718
HI0719
HI0720 htpX
HI0721
HI0722 pepQ
HI0723
Ala
Ile
rRNAC 16s-23s-5s
HI0724
HI0725
HI0726 narP
HI0727 lysA
HI0728 recQ
HI0729
HI0854
HI0853 dppA
HI0855
HI0856 polA
HI0857
HI0858
HI0859 clpB
HI0860
HI0861
HI0862
HI0863 pdxH
HI0864
HI0865 glnA
HI0866
HI0867
HI0868 lgtD
HI0869
HI0870
HI0871
HI0872 rfbP
HI0873 rffE
HI0874
HI0875 pepA
HI0876 ndk
HI0877 obg
HI0878
HI0879 rpL27
HI0880 rpL21
HI0881
HI0990 iga1
HI0991 recF
HI0992 dnaN
HI0993 dnaA
HI0994 tbp1
HI0995 tbp2
HI0996
HI0997
HI0998 rpL34
HI0999 rnpA
HI1000
HI1001
HI1002 thdF
HI1003
HI1004
HI1005
HI1006 lspA
HI1007
HI1008 comE1
HI1009 glpR
HI1010
HI1011
HI1012 fucA
HI1013
HI1014
HI1015 gntP
HI1016
HI1017 glpF
HI1018
HI1019
HI1134 murF
HI1133 murE
HI1135 mraY
HI1136 murD
HI1137 ftsW
HI1138 murG
HI1139 murC
HI1140 ddlB
HI1141 ftsQ
HI1142 ftsA
HI1143 ftsZ
HI1144 envA
HI1145 pheA
HI1146
HI1147
HI1148
HI1149
HI1150
HI1151
HI1152 pmbA
HI1153
HI1154 gltP
HI1155
HI1156 cydC
HI1157 cydD
HI1158 trxB
HI1159 trxA
HI1160 visA
HI1161 P15
HI1162
HI1163 dld
HI1274 atp
HI1275 tehB
HI1276 metG
HI1277 mrp
HI1278
HI1279 sia8
HI1280
HI1281
Met
HI1282
HI1283 nusA
HI1284 infB
HI1285 hsdR
HI1286
HI1287 hsdM
HI1288
HI1289
HI1290 tyrA
HI1291
HI1292
HI1293
HI1294
HI1295 nifS
HI1296 parB
HI1297
HI1298
HI1299 dgt
HI1300
HI1301
HI1302 asnS
HI1303 ribE
HI1304 nusB
HI1406
HI1407 traN
HI1408
HI1409
HI1410
HI1411
HI1412
HI1413
HI1414
HI1415
HI1416
HI1418
HI1419
HI1420
HI1421
HI1422
HI1423
HI1424 rci
Leu
HI1425 fnr
HI1426
HI1427
HI1428 purN
HI1429 purM
HI1430
HI1431 trpB
HI1432 trpA
HI1433 usg
HI1434
HI1435
HI1436
HI1437
HI1438 ispA
HI1439
HI1440 sspB
HI1441 sspA
HI1442 rpS9
HI1443 rpL13
HI1444 metF
HI1445 bioD
HI1446
HI1583 argS
HI1584 ilvH
HI1585 ilvI
HI1586
HI1587 hns
HI1588 purU
HI1589 aroA
HI1590
HI1591
HI1592 spoIIIE
HI1593
HI1594
HI1595
HI1596 lrp
HI1597 sms
HI1598
HI1599
HI1600
HI1601
HI1602
HI1603
HI1604
HI1605
HI1606 cca
HI1607
HI1608
HI1609 prsA
HI1610 tyrS
HI1611 sfsA
HI1612
HI1613 ribC
HI1614 pepN
HI1615 purE
HI1616
HI1727 argG
HI1728 braB
HI1729 lamB
HI1730
HI1731
HI1732
HI1733
HI1734 envM
HI1735 prfC
HI1736
HI1737
HI1738
HI1739 metR
Ile
Ala
rRNAD 5s-23s-16s
HI1740 recG
HI1741 spoT
HI1742 rpoZ
HI1743 gmk
HI0001 gap


HI0038 mreC
HI0055 uxuA
HI0036 mdl HI0037 mreB HI0039 mreD Ser HI0046 HI0054 uxuR
HI0028 HI0029 dacA HI0031 mreB HI0033 HI0035 HI0040 HI0042 HI0044 HI0047 eda HI0049 kdgK HI0051 HI0053 HI0056 HI0057 uvrC HI0059 HI0060 msbA
lipB HI0030 rlpA HI0032 pbp2 HI0034 HI0041 xthA HI0043 HI0045 HI0048 hdhA HI0050 HI0052 HI0058 kdsB HI00
HI0188 HI0194 menE HI0197 mep
HI0164 HI0166 rnfE HI0168 HI0170 HI0173
lA HI0165 HI0167 HI0169 HI0171 HI0172 HI0174 HI0177 HI0182 HI0183 HI0186 HI0187 HI0189 gdhA HI0192 seqA HI0195 HI0196 aroC HI
HI0175 HI0178 HI0180 pfl HI0181 HI0184 HI0185 HI0190 fur HI0193 acoC
HI0176 HI0179 act HI0191 fldA
HI0321 vagC
HI0302 cute HI0305 HI0310 HI0319 HI0324 rnt HI0331 oapB HI0335 dg
HI0308
5 rho HI0300 ampD HI0301 tlyC HI0303 HI0304 HI0307 proC HI0309 xerD HI0318 HI0322 vapC HI0325 HI0326 Ser HI0329 HI0330 oapA HI0333 HI0334 relA HI0
HI0296 hopD HI0298 pilB HI0299 pilA HI0306 HI0311 HI0313 ruvA HI0315 HI0316 HI0317 aspS HI0320 HI0323 HI0328 efp HI0332 recO HI
HI0297 pilC HI0312 ruvB HI0314 ruvC
HI0477 tyrP
HI0469 hisD HI0471 hisB HI0473 hisA HI0475 hisIE HI0489 HI0497 hslU
HI0464 HI0493
HI0460 mazG HI0462 lon HI0463 hemN HI0465 serA HI0468 hisG HI0470 hisC HI0472 hisH HI0474 hisF HI0476 HI0488 HI0491 HI0492 HI0496 hslV
HI0484 atpE
HI0459 pyrR HI0461 lapB HI0466 HI0467 HI0478 atpC HI0480 atpG HI0482 atpH HI0486 gidB HI0490 HI0494 HI0495 HI049
surA HI0479 atpD HI0481 atpA HI0483 atpF HI0487
HI0485 atpB
HI0606 cysE HI0619 glpR HI0623 fr
HI0604 cyaA HI0605 gpsA HI0607 HI0608 Arg Pro HI0616 hepA HI0617 HI0618 glpG Glu rRNAB 16s-23s-5s HI0622 def
HI0602 hemY HI0609 folD HI0610 fucP HI0612 fucU HI0614 fucI HI0615 fucR HI0620 hlpA HI0621 abc
HI0603 hemX HI0611 fucA HI0613 fucK
HI0738 ilvD HI0746 dcuA HI0750 dapF
HI0728 recQ HI0729 proS HI0735 HI0737 ilvG HI0740 HI0742 HI0745 ansB HI0747 ndh HI0749 lexA
lysA HI0730 HI0731 HI0733 HI0734 plsC HI0736 HI0739 dnaE HI0744 HI0748 plsB HI0751 tagD
HI0732 HI0743 secB
6 ndk HI0887 purH HI0889 glyA HI0891 HI0894 mtrC HI0901 invA HI0904 lgt HI0909 secA
HI0906
HI0878 HI0881 HI0882 HI0885 cycZ HI0886 HI0888 purD HI0890 HI0893 Bm3R1 HI0895 acrB HI0896 msgA HI0900 proB HI0902 HI0905 thyA HI0908
HI0877 obg HI0879 rpL27 HI0883 dagA HI0884 arcA HI0892 rhlB HI0897 emrB HI0899 folA HI0903 HI0907
HI0880 rpL21 HI0898 ermA
5 gntP HI1017 glpF HI1020 HI1034 HI1036 HI1041 MHgiDI
HI1016 HI1019 HI1021 HI1022 bioB HI1032 HI1033 serB HI1035 corA HI1038 HI1040 hgiDIR HI10
HI1018 HI1023 tktA HI1025 araD HI1028 HI1029 HI1031 HI1037 HI1039 HI1042 metH HI1043 HI1045 dmsC HI1047 dmsA
HI1024 HI1026 HI1027 HI1030 HI1044 HI1046 dmsB
HI1172 metX HI1174 opa66 HI1197 su
HI1195
HI1164 ompA HI1165 HI1168 HI1171 trpG HI1173 HI1182 HI1183 HI1192 HI1194 gcvA HI1196 sucC
160 visA HI1162 HI1166 hisH HI1169 HI1177 artM HI1179 artI HI1181 HI1184 dppF HI1186 dppC HI1188 uvrD HI1189 pqqIII HI1193
HI1161 P15 HI1163 dld HI1167 serC HI1170 pabB HI1176 HI1178 artQ HI1180 artP HI1185 dppD HI1187 dppB HI1190 HI1191
HI1305 hypE HI1320 rpL20 HI1326
HI1304 nusB HI1307 mlgA HI1310 HI1312 pheT HI1315 HI1317 HI1319 rpL35 HI1322 recD HI1328 rpS15 HI1334 ftsJ
HI1324 lon
1300 HI1303 ribE HI1306 HI1308 dapB HI1311 pheS HI1313 himA HI1316 HI1318 infC HI1321 recB HI1325 fabA HI1327 HI1331 greA HI1335 ftsH HI1336 folP HI
HI1301 HI1302 asnS HI1309 HI1314 HI1323 HI1329 HI1333
HI1330 dacB
HI1449 chlN HI1451 HI1462 nodT HI1478 muA
HI1444 metF HI1448 chlE HI1450 HI1455 pilB HI1458 HI1461 lex2A HI1466 fhuA HI1467 cvaB HI1477 HI1479 HI
spB HI1442 rpS9 HI1445 bioD HI1447 folE HI1453 pilB HI1457 opaD HI1459 algU HI1463 HI1464 folP HI1468 rpS15 HI1471 hemU HI1473 modD HI1475 nifC HI1480
1 sspA HI1443 rpL13 HI1446 HI1452 HI1454 HI1456 HI1460 yopA HI1465 ftsH HI1469 HI1470 fepC HI1472 HI1474 sfuC HI1476
HI1616 purK HI1628 Met Leu Gln Gln HI1633 purA HI1636 ppc HI1639 sapB HI1641 sapD
3 ribC HI1615 purE HI1617 aspC HI1623 merR2 HI1627 HI1630 rpL25 HI1632 lysC HI1635 purR HI1638 sapA HI1640 sapC HI1642 sapF HI1643 HI1645 fbp HI1647 HI16
HI1614 pepN HI1618 cbiO HI1621 HI1624 HI1625 HI1629 dedA HI1631 HI1634 dapD HI1637 HI1644 hisT HI1646 cmk
HI1619 HI1622 HI1626
HI1620
HI0001 gap
z
gmk
Amino acid biosynthesis
Biosynthesis of cofactors, prosthetic groups, carriers
Cell envelope
Cellular processes
Central intermediary metabolism
Energy metabolism
Fatty acid/Phospholipid metabolism
Purines, pyrimidines, nucleosides and nucleotides
Regulatory functions
Replication
T
T
T
C
H
U


HI0063 pcnB
HI0065
HI0067 mutL
HI0069 glnE
HI0062 dksA
HI0064 folK
HI0066 amiB
HI0068 trpX
HI0059
HI0060 msbA
58 kdsB
HI0061 recQ
HI0077
HI0080
HI0083
HI0071 grpE
HI0075 nrdD
HI0076 tesB
HI0079 ppiB
HI0081
HI0082
HI0070 recN
HI0072
HI0073
HI0078 cysS
HI0084 trxM
HI0086 metI
HI0074
HI0085 ddh
HI0197 mepA
HI0200 selD
HI0206 ushA
HI0208 aroB
HI0195
HI0196 aroC
HI0198
HI0199 msbB
HI0205
HI0207 aroK
HI0209 dam
HI0201 rpL19
HI0203
HI0202 trmD
HI0204 rpS16
HI0217
HI0210
HI0212 ribA
HI0214 prlC
HI0215 hsdM
HI0216 hsdS
HI0218 prrD
HI0211 pgpB
HI0213 oppA
HI0219
HI0220 arcB
HI0335 dgkA
HI0342
HI0344
HI0346
HI0333
HI0334 relA
HI0336
HI0338
HI0340
HI0341
HI0343
HI0345
HI0347
recO
HI0337 glnB
HI0339 priA
HI0358 tenA
HI0354 nasD
HI0366
HI0348 nirT
HI0353
HI0355
HI0357 nmt1
HI0365
HI0367
HI0349 adk
HI0350
HI0352
HI0359
HI0360
HI0362 fimA
HI0364
HI0351 galE
HI0361 fecE
HI0497 hslU
HI0502 rbsA
HI0504 rbsB
HI0506 rbsR
HI0493
491
HI0492
HI0496 hslV
HI0499 aldH
HI0500
HI0501 rbsD
HI0503 rbsC
HI0505 rbsK
HI0507
HI0494
HI0495
HI0498 potD
HI0508
HI0510
HI0511 tehA
HI0513 hincIIM
HI0514 rpoC
HI0515 rpoB
HI0516 rpL1
HI0518
HI0509
HI0512 HincII
HI0517 rpL11
HI0623 fmt
HI0628 rpoE
HI0630 mucB
Thr
Tyr
Thr
3 16s-23s-5s
HI0622 def
HI0624
HI0625 trkA
HI0627
HI0629 mclA
HI0632 tufB
HI0634
HI0626
HI0631 coaA
HI0633
HI0641
HI0642 glmU
Ser
Arg
Arg
HI0639 purB
HI0638
HI0640 rpL10
HI0635 tbp2
HI0636
HI0643 bisC
HI0644 torC
HI0646 asd
HI0637 trpS
HI0645 pldB
HI0750 dapF
HI0760
HI0749 lexA
HI0752 purL
HI0753
HI0754
HI0757 gpmA
HI0759 mutY
HI0761
HI0751 tagD
HI0755
HI0756
HI0758 rpL31
HI0778 rpL4
HI0765
HI0769 ftsE
HI0777 rpL3
HI0780 rpL
Phe
Asn
HI0764 ribB
HI0768 ftsY
HI0770 ftsX
HI0775
HI0776 rpS10
HI078
HI0762
HI0766
HI0771 fadA
HI0773
HI0774 ctfA
HI0779 rpL23
HI0763 nadR
HI0767
HI0772
lgt
HI0909 secA
HI0914 tsf
HI0906
05 thyA
HI0908
HI0910
HI0911 kefC
HI0912
HI0913 rpS2
HI0907
HI0915 firA
HI0917
HI0918
HI0916 skp
HI0923 holA
HI0921 leuS
HI0922 rlpB
HI0928 hktE
HI0919 cdsA
HI0920
HI0924 glyS
HI0925
HI0927 glyQ
HI0929
HI0930
HI0932 eno
HI0926
HI0931
HI1050 merP
HI1048
HI1049 merT
HI1053
045 dmsC
HI1047 dmsA
HI1051 lktB
HI1052
HI1054
HI1056 mod
HI1059
HI1046 dmsB
HI1055
HI1058
HI1065 smbA
HI1063
HI1061 lpxA
HI1064
HI1066 nrfD
HI1068 nrfB
HI1070
HI1072
HI1074
HI1075 cydB
HI1060 lpxB
HI1062 fabZ
HI1067 nrfC
HI1069 nrfA
HI1071
HI1073
HI1076
HI1197 sucD
HI1199
HI1206 cvpA
HI1195
HI1194 gcvA
HI1196 sucC
HI1198
HI1200 cysB
HI1202
HI1205
HI1207 purF
HI1201 hemG
HI1203
HI1204 ackA
HI1214 recJ
HI1220
HI1210 mdh
HI1213 xprA
HI1215
HI1216
HI1219 cmk
HI1208 sulA
HI1211 lysU
HI1212 prfB
HI1217 tbp1
HI1218 lctP
HI1209 argR
HI1334 ftsJ
HI1338
reA
HI1335 ftsH
HI1336 folP
HI1337
HI1339
HI1342
HI1348 pepT
1333
HI1340
HI1341
HI1344 potD
HI1346 potB
HI1343 nifS
HI1345 potC
HI1347 potA
HI1356 malQ
HI13
HI1350 cda
HI1353 cafA
HI1354 glnS
HI1355
HI1357 glgB
HI1358 glgX
HI1349
HI1351
HI1352 putP
HI1487
HI1492
HI1495
HI1498
HI1489
HI1478 muA
HI1482
HI1484
HI1486
HI1491 rdgB
HI1494
HI1497
HI1500
HI1477
HI1479
HI1481 muB
HI1485
HI1488 muE16
HI1493
HI1499
HI1501
1475 nifC
HI1480
HI1483
HI1490
HI1496
C
HI1476
HI1508
HI1518
HI1504 muI
HI1507
HI1510
HI1512
HI1514
HI1516 muP
HI1520
HI1502
HI1503 muG
HI1505
HI1509
HI1511
HI1513
HI1515 muN
HI1517
HI1519
HI1521
HI1506
HI1656
HI1645 fbp
HI1647
HI1648
HI1650
HI1655
HI1658 hly
44 hisT
HI1646 cmk
HI1649 dld
HI1651
HI1652
HI1654
HI1657
HI1653
HI1659 nrdA
HI1
HI1660 nrdB
HI1671
HI1661 sucB
HI1663
HI1665
HI1667
HI1668 prc
HI1669
HI1662 sucA
HI1664
HI1666
HI1670 finO
Transport/binding proteins
Translation
Transcription
d nucleotides
Other categories
Hypothetical
Unknown
rRNAB 16s-23s-5s
Ribosomal operon
tRNA
1 kb


80
HI0083
HI0081 HI0082
Cys Lys
Gly Leu
HI0090
HI0098 sfuB
HI0096 HI0097 fbp
HI0099 potG
HI0105
HI0107 HI0108 HI0110 serS HI0111 bphH
HI0113
HI0084 trxM HI0086 metB HI0087 thrC HI0089 thrA
HI0091
HI0093 HI0094 HI0095 gerC2
HI0100 HI0102 dapE HI0104 htpG
HI0106 ffh
HI0109
HI0112
HI0085 ddh
HI0088 thrB
HI0092
HI0101 HI0103
HI0238
8 prrD
Glu
HI0221 guaB HI0222 guaA HI0224 lrp HI0226 brnQ HI0229 pnp HI0230 HI0231 deaD HI0232 HI0233 HI0237
HI0242
HI0219 HI0220 arcB
rRNAF 5s-23s-16s
HI0223 rarD HI0225 nhaA
HI0227 HI0228
HI0234 HI0236 arsC HI0239 secF HI0241
HI0235
HI0240 secD
HI0243
HI0369 hisS
HI0366
HI0365 HI0367 HI0368 gpcE HI0370
HI0377 nifU
Lys
Lys
HI0388
HI0387 dinG HI0389 HI0390 rnd
HI0396
HI0398
HI0395 HI0397 xseA HI0399 icc
fimA HI0364
HI0372 fdx HI0374 HI0376
HI0379 HI0381 pal HI0383 tolA HI0385 tolQ
HI0371 HI0373 hsc66 HI0375 HI0378 nifS HI0380 HI0382 tolB HI0384 tolR HI0386
HI0391 HI0392 HI0394 pth
HI0393
HI0
HI0532 dnaG
HI0543 mopA
HI0522
HI0527 fdx
HI0531 rpS21
HI0533 rpoD
HI0542 mopB
HI0549 ksgA HI05
15 rpoB HI0516 rpL1 HI0518 deoD HI0519 HI0520
HI0523 HI0524 fba
HI0526 HI0528 tyrP HI0530 gcp
HI0534 aspA HI0535 ureHI0537 UreF HI0539 ureC HI0541 ureA HI0544 rpL9 HI0547 rpS6 HI0550
HI0517 rpL11
HI0521
HI0525 pgk
HI0529 tdk
HI0536 ureG HI0538 ureE HI0540 ureB
HI0545 rpS18 HI0548 infA
HI0546 priB
HI0664 cydC
HI0669 mioC
HI0653
HI0663 cydD
HI0668 HI0670
HI0675 pepD
HI0678 tpiA
bisC HI0644 torC HI0646 asd HI0648 mda66
HI0650 HI0652 kdtA HI0654 tagI HI0656
HI0658 HI0659 HI0661 lbpA HI0662
HI0665 HI0667 glpX
HI0671 HI0673
HI0676 xerC HI0679 gl
HI0645 pldB HI0647
HI0649 rep HI0651 kdtB
HI0655 aroE HI0657 topA
HI0660
HI0666
HI0672 HI0674
HI0677
HI068
HI0778 rpL4 HI0782 rpL22 HI0785 rpL29 HI0790 rpL5 HI0793 rpL6 HI0796 rpL30
HI0801 rpS4
HI0777 rpL3 HI0780 rpL2 rpS3 rpS17 rpL24 rpS8 rpS5 HI0798 secY HI0800 rpS11 HI0803 rplQ
HI0775 HI0776 rpS10 HI0781 rpS19 HI0787 HI0791 rpS14 HI0797 rpL15 HI0799 rpS13 HI0802 rpoA
HI0806
0774 ctfA HI0779 rpL23 HI0784 rpL16 HI0788 rpL14 HI0794 rpL18
HI0804
HI0807 HI0808 frr
HI0810 HI0811 argH HI0812 galU
HI0815 uspA HI0817 HI0818 mro HI0820
HI0805 aslB
HI0809 pckA
HI0813 csrA HI0814 alaS
HI0816 pepP
HI0819 galK
HI0939 HI0941
HI0945 htrH
HI0954 dut HI0956
hktE
HI0938 HI0940
HI0942 recC
HI0943
HI0944 ribG
HI0953 dfp HI0955 HI0957 crp
HI0948
HI0929 HI0930 HI0932 eno HI0933
HI0935
HI0937 suhB
HI0946 fpg
HI0947 vapC
HI0950 rpL33
HI0958 HI0959 HI0961 HI0962 ileS
HI0931
HI0934
HI0936
HI0949 gabT
HI0951 rpL28
HI0952 radC
HI0960
HI1090 HI1092 HI1094 HI1096 HI1099
HI1103 cysK
HI1089 HI1091HI1093 HI1095 HI1098
HI1102 cysZ
HI1104
HI1077 pyrG
1072 HI1074 HI1075 cydB
HI1078 sfpo HI1080 glnH HI1082 HI1084 HI1086 HI1088 sodA
HI1100 lig HI1101
HI1105 rfaF HI110
HI1073
HI1076 cydA
HI1079 glnP HI1081 murZ HI1083 HI1085 HI1087 mkl
HI1106 xylR
HI1220 rpS1 HI1222 HI1223 HI1225
HI1219 cmk HI1221 himD HI1224 pyrF
HI1234
HI1237 dnaK HI1238 dnaJ
HI1241
HI1245
HI124
HI1218 lctP
HI1226 dnaA HI1228 upp HI1230 apt HI1232 aceF HI1233 aceE HI1235
HI1239 proA HI1240 HI1242 bcr HI1244
HI1227 uraA HI1229 dnaX HI1231 lpdA
HI1236
HI1243
HI1359 glgC
HI1363 pntB
HI1373 kicA
HI1357 glgB HI1358 glgX HI1360 glgA HI1361 glgP HI1362 pntA HI1365 topA HI1367 thrS
HI1372 kicB HI1374 mukB HI1375 HI1
HI1364 HI1366 HI1368 HI1369 HI1370 mopI
HI1371 dsvC
HI1518
HI1516 muP HI1520 HI1522
HI1528 parE
HI1535 HI1537 licA HI1539 licC
N HI1517 HI1519 HI1521 HI1523 HI1524 HI1527 ibpB HI1529 parC HI1530 gltS HI1532 grx HI1534 HI1536 HI1538 licB HI1540 licD
HI1542 HI1546
HI1525 modB
HI1531 rimK HI1533 fabB
HI1541 sppA HI1543 HI1545 HI1547 aroG HI1549 devA H
HI1526 aut
HI1544
HI1548 HI1550
HI1672
HI1684
HI1686 rnfE
HI1690
HI1671
HI1677 HI1678 kpsF HI1679
HI1682 sohB HI1683 HI1685 tolA
HI1687 HI1688 HI1689 nth
HI1694
HI1701
668 prc HI1669
HI1673 moaE HI1676 moaA
HI1680 HI1681
HI1691 modC HI1693 HI1695 HI1697 HI1698 HI1700
HI1670 finO
HI1674 moaD
HI1675 moaC
HI1692 modB
HI1696
HI1699


HI0113 hxuC
HI0119 fimA
Leu
Gly Lys
HI0133 dcd
150,000 nt
HI0110 serS HI0111 bphH
Glu
Pro
Met HI0117 HI0118 chlN
HI0123 pgsA
HI0125
HI0132 udk HI0134 HI0135
0109
HI0112
HI0114
rRNAE 5s-23s-16s
HI0115 HI0116
HI0120
HI0122 metC
HI0124 ppa
HI0126 HI0127
HI0130
HI0121
HI0129 nifc
HI0131
38
HI0256
HI0264 hxuA
300,000 nt
HI0267 narQ
HI0250 ssb
HI0255 dapA HI0257 Ser
HI0261 opsX HI0262 hemR HI0263 hxuB
HI0266
HI0242
HI0239 secF
HI0241
HI0244 tgt HI0245 queA
HI0247 igal
HI0249 uvrA
HI0251 tonB HI0253 exbB
HI0258 HI0259
HI0265
HI0240 secD
HI0243
HI0246
HI0252 exbD HI0254 bcp
HI0260
HI0402 dat1
HI0416
HI0418
450,000 nt
HI0396
HI0398
HI0414 opa66
95 HI0397 xseA HI0399 icc HI0401 ompP1 HI0403 mutH
HI0409 HI0410 tyrR
HI0413 rne
HI0415
HI0417
HI0419 prtC HI0420
HI0423
HI0425 psaA HI0427 nhaB
HI0400
HI0404
HI0406 accA HI0408 livG
HI0411 hfq HI0412
HI0421
HI0424
HI0426 fadR
HI0405
HI0407
HI0422 srmB
pA
HI0552
HI0576
600,000 nt
HI0549 ksgA HI0551 apaH
HI0562 HI0564 asnA
HI0569 greB
HI0572 HI0574 HI0575 HI0577
544 rpL9 HI0547 rpS6
HI0550
HI0553 gnd
HI0554 HI0556 devB
HI0559 cysQ
HI0560
HI0563 asnC
HI0565 HI0566 dod
HI0568
HI0570
HI0573 slyD
HI0578 tufB
HI0580 rpS7
0545 rpS18 HI0548 infA
HI0555
HI0558 G6PD
HI0561
HI0567 gyrB
HI0571 oxyR
HI0579 fusA
HI0546 priB
HI0557
HI0689 hpd
HI0691 glpK
750,000 nt
HI0704
HI0675 pepD
HI0678 tpiA
HI0682 ilvC
HI0686 glpT HI0687 HI0688 HI0690 glpF
HI0693 hel HI0694
HI0701 HI0702 HI0706 nlpD
HI0676 xerC HI0679 glpE HI0681 ilvY
HI0683 glpC
HI0685 glpA
HI0692
HI0695 ppx
HI0696
HI0698
HI0700
HI0703 lppB
HI0677
HI0680 rarD
HI0684 glpB
HI0697 HI0699 slyD
HI0822 mglB
HI0824 mglC
HI0827 HI0829 slt
HI0841
HI0843
HI0847
HI0852
900,000 nt
HI0831
HI0845
HI0849
HI0821 galS
HI0823 mglA
HI0825 HI0826 HI0828 kch HI0830 trpR
HI0836 genX
HI0840
HI0842
HI0846 por HI0848 trmA HI0850 HI0851
HI0853 dppA
HI0817 HI0818 mro
HI0820 galT
HI0832 frdD
HI0835 frdA
HI0837
HI0839
HI0844 mob
16 pepP
HI0819 galK
HI0833 frdC
HI0838
HI0834 frdB
HI0968 menB HI0970 aroQ HI0972 accC
HI0974
HI0978 prmA
1,050,00 nt
HI0965 HI0967 HI0969 menC HI0971 accB HI0973 tfbA
HI0975 panF HI0976 HI0977
HI0979 HI0980 fis
HI0986 HI0987 leuB HI0988 HI0989 leuD
0959 HI0961
HI0962 ileS
HI0963 HI0964
HI0966
HI0981 smpB HI0983 leuA HI0985 dprA
HI0990 igal
HI0960
HI0982 pfkA HI0984
HI1113
HI1128
HI1130
HI1132 ftsI
1,200,00 nt
HI1104
HI1112 xylA
HI1116 deoC
HI1119 lapB
HI1125
HI1127 cstA HI1129 HI1131 ftsL HI1133 murE
HI1105 rfaF
HI1107 nhaC
HI1109 xylH
HI1111 rbsB
HI1114 rfaD
HI1117 HI1118
HI1120 oppF HI1122 oppC
HI1124 oppA
HI1126
HI1106 xylR
HI1108 malY
HI1110 xylG
HI1115 trxA
HI1121 oppD HI1123 oppB
HI1269 HI1270
1,350,00 nt
HI1251 vapA
HI1254 HI1256
HI1258 mfd
HI1261 folC
HI1268
HI1273
HI1245
HI1246
Asn
HI1250
HI1252 HI1253
HI1255 HI1257
HI1260 accD
HI1263 met2
HI1267 HI1272 fepC
HI1244
HI1247 uvrB
HI1248
HI1259 htrA
HI1262
HI1264 gyrA
HI1265 HI1271 fecD
HI1274 atp
HI1249
HI1266
HI1385 rsgA HI1387 trpE
HI1389 trpC
1,500,000 nt
HI1374 mukB
HI1375 HI1376 HI1377 sbcB
HI1384 rsgA HI1386
HI1388 trpD
HI1390 hypC
HI1398 fumC HI1399
HI1378 phoR HI1380 pstB HI1382 pstC
HI1391 valS
HI1393 hindIIIR HI1396
HI1400
HI1402 HI1404
HI1379 phoB HI1381 pstA HI1383 pstS
HI1392 hindIIIM HI1394 HI1397 holC
HI1401 pyrD HI1403
HI1395
HI1566
HI1574 dnaB
HI1576 pgi
HI1578 lgtD
1,650,000 nt
HI1546
HI1565 tbp1 HI1567 iroA
Val HI1573 pykA
HI1575 alr
HI1577
HI1583 argS
HI1547 aroG
HI1549 devA HI1551 bioC HI1553 bioF
HI1555
HI1557 kdsA HI1559 hemG HI1561 prfA HI1563
HI1568 mug
HI1572 rcb
HI1579 lpp
HI1548
HI1550 bioD HI1552
HI1554 bioA
HI1556 vanH
HI1558
HI1560
HI1562 HI1564
HI1569 HI1571
HI1581 HI1582
HI1570
1,800,000 nt
HI1725 ponB
HI1727 argG
HI1701
HI1705 pepA
HI1715 Gly
HI1720 HI1721 HI1723 HI1724
HI1726 purC
697 HI1698
HI1700
HI1702 metE
HI1703
HI1706 betT HI1707 basS HI1709 HI1711 crr HI1713 ptsH HI1714
HI1716 HI1717
HI1719 glnD
HI1722 map
HI1699
HI1704
HI1708 basR HI1710 HI1712 ptsI
HI1718

3) The two λ libraries constructed from *H. influenzae* genomic DNA were probed with oligonucleotides designed from the ends of contig groups (*27*). The positive plaques were then used to prepare templates, and the sequence was determined from each end of the λ clone insert. These sequence fragments were searched with GRASTA against a database of all contigs. Two contigs that matched the sequence from the opposite ends of the same λ clone were ordered. The λ clone then provided the template for closure of the sequence gap between the adjacent contigs.

4) To confirm the order of contigs found by the other approaches and establish the order of the remaining contigs, we performed amplifications by polymerase chain reaction (PCR), both standard and long range (XL) (*28*). Although a PCR reaction was done for essentially every combination of physical gap ends, techniques such as DNA fingerprinting, database matching, and the probing of large insert clones were particularly valuable in ordering contigs adjacent to each other and reducing the number of combinatorial PCRs necessary to achieve complete gap closure. Use of these strategies to an even greater extent in future genome projects will increase the overall efficiency of complete genome closure. In the program ASM_ALIGN Southern analysis data, identification of peptide links, forward and reverse sequence data from λ clones, and PCR data are used to establish the relative order of the contigs separated by physical gaps. The number of physical gaps ordered and closed by each of these techniques is summarized in Table 2.

Lambda clones were a central feature for completion of the genome sequence and assembly. It was probable that some fragments of the *H. influenzae* genome would be nonclonable in a high copy plasmid because they would produce deleterious proteins in the *E. coli* host cells. Lytic λ clones would provide DNA for these segments because such genes would not inhibit plaque production. Furthermore, sequence information from the ends of 15- to 20-kb clones is particularly suitable for gap closure and providing general confirmation of genome assembly. Because of their size, they would be likely to span any physical gap. Approximately 100 random plaques were picked from the amplified λ library, templates were prepared, and sequence information was obtained from each end. These sequences were searched (GRASTA) against the contigs and linked in the database to their appropriate contig, thus providing a scaffolding of λ clones that contributed additional support to the accuracy of the genome assembly (Fig. 1). In addition to confirmation of the contig structure, the λ clones provided closure for 23 physical gaps. Approximately 78 percent of the genome was covered by λ clones.

The λ clones were particularly useful for solving repeat structures. All repeat structures identified in the genome were small enough to be spanned by a single clone from the random insert library, except for the six ribosomal RNA (rRNA) operons and one repeat (two copies) that was 5340 bp in length. The ability to distinguish and assemble the six rRNA operons of *H. influenzae* (each containing in order 16S, 23S, and 5S subunit genes) was a test of our overall strategy to sequence and assemble a complex genome that might contain a significant number of repeat regions. The high degree of sequence similarity and the length of the six operons caused the assembly process to cluster all the underlying sequences into a few indistinguishable contigs. To determine the correct placement of the operons in the sequence, unique sequences were identified at the 5S ends. Oligonucleotide primers were designed from these six flanking regions and used to probe the two λ libraries. For five of the six rRNA operons at least one positive plaque was identified that completely spanned the rRNA operon and contained uniquely identifying flanking sequence at the 16S and 5S ends. These plaques provided the templates for obtaining the sequence for these rRNA operons. For *rrnA* a plaque was identified that contained the particular 5S end and terminated in the 16S end. The 16S end of *rrnA* was obtained by PCR from *H. influenzae* Rd genomic DNA.

An additional confirmation of the global structure of the assembled circular genome was obtained by comparing a computer-

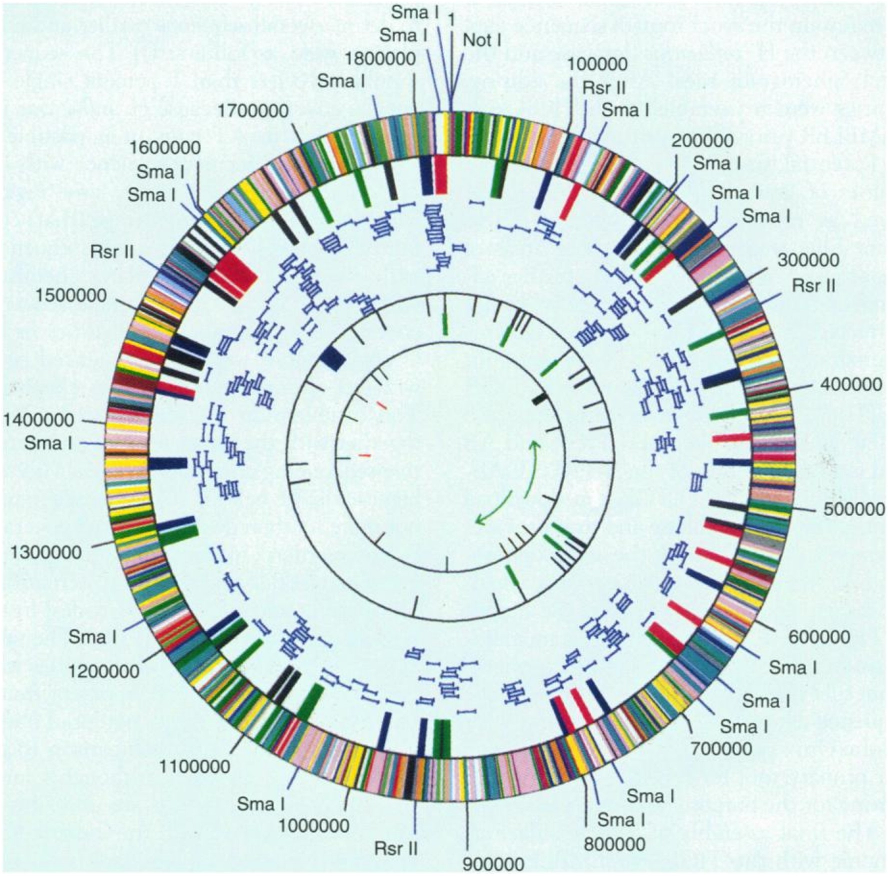


Fig. 1. A circular representation of the *H. influenzae* Rd chromosome illustrating the location of each predicted coding region containing a database match as well as selected global features of the genome. Outer perimeter: The location of the unique Not I restriction site (designated as nucleotide 1), the Rsr II sites, and the Sma I sites. Outer concentric circle: Coding regions for which a gene identification was made. Each coding region location is classified as to role according to the color code in Fig. 2. Second concentric circle: Regions of high G+C content (>42 percent, red; >40 percent, blue) and high A+T content (>66 percent, black; >64 percent, green). Third concentric circle: Coverage by λ clones (blue). More than 300 λ clones were sequenced from each end to confirm the overall structure of the genome and identify the six ribosomal operons. Fourth concentric circle: The locations of the six ribosomal operons (green), the tRNAs (black) and the cryptic mu-like prophage (blue). Fifth concentric circle: Simple tandem repeats. The locations of the following repeats are shown: CTGGCT, GTCT, ATT, AATGGC, TTGA, TTGG, TTTA, TTATC ,TGAC, TCGTC, AACC, TTGC, CAAT, CCAA. The putative origin of replication is illustrated by the outward pointing arrows (green) originating near base 603,000. Two potential termination sequences are shown near the opposite midpoint of the circle (red).

generated restriction map based on the assembled sequence for the endonucleases Apa I, Sma I, and Rsr II with the predicted physical map of Lee *et al.* (*29*). The restriction fragments from the sequence-derived map matched those from the physical map in size and relative order (Fig. 1).

At the same time that the final gap filling process occurred, each contig was edited visually by reassembling overlapping 10-kb sections of contigs by means of the AB AUTOASSEMBLER and the Fast Data Finder hardware. AUTOASSEMBLER provides a graphical interface to electropherogram data for editing. The electropherogram data was used to assign the most likely base at each position. Where a discrepancy could not be resolved or a clear assignment made, the automatic base calls were initially left unchanged. Individual sequence changes were written to the electropherogram files and a program was designed (CRASH) to maintain the synchrony of sequence data between the *H. influenzae* database and the electropherogram files. After the editing, contigs were reassembled with TIGR ASSEMBLER prior to annotation.

Potential frameshifts identified in the course of annotating the genome were saved as reports in the database. These frameshifts were used to indicate areas of the sequence that might require further editing or sequencing. Frameshifts were not corrected for cases in which clear electropherogram data disagreed with a frameshift. Frameshift editing was done with TIGR EDITOR. This program was developed as a collaborative effort between TIGR and AB and is a modification of the AB AUTOASSEMBLER. TIGR EDITOR can download contigs from the database and thus provides a graphical interface to the electropherogram for the purpose of editing data associated with the aligned sequence file output of TIGR ASSEMBLER. The program maintains synchrony between the electropherogram files on the Macintosh system and the sequence data in the *H. influenzae* database on the Unix system. TIGR EDITOR is now our primary tool for sequence viewing and editing for the purpose of genome assembly.

The final assembly of the *H. influenzae* genome with the TIGR ASSEMBLER was precluded by the rRNA and other repeat regions, and was accomplished by means of COMB_ASM (a program written at TIGR) that splices together contigs on the basis of short sequence overlaps.

Throughout the project, we paid particular attention to the accuracy of the sequence generated and included various quality control measures. In particular, we constructed random small and large insert libraries (as described above), used strict criteria for excluding any single sequence in which more than 3 percent of the nucleotides could not be identified with certainty, determined that there was no vector contamination in each sequence, and rejected chimeric sequences from the assembly process. The most important measure of the sequence accuracy is the correct assembly of the 1.8-Mb genome. Any deviation from inclusion of only high-quality sequences would have resulted in an inability to assemble the final genome. In addition, the use of the large insert λ clones confirmed the accuracy of the final assembly. Our finding that the restriction map of the *H. influenzae* Rd genome based on our sequence data is in complete agreement with that previously published (*29*) further confirms the accuracy of the assembly.

As a consequence of our shotgun approach, we reached an average of more than sixfold redundancy across the genome, although there are some regions in which the coverage is lower. The criteria that we used to define overall sequence quality and completion were as follows: (i) The sequence should have less than 1 percent single sequence coverage. Because *H. influenzae* is a genome rich in AT pairs, it is possible to obtain a highly accurate sequence with single-pass coverage. However, any regions with single sequence coverage that contained ambiguities were again sequenced with an alternative sequencing chemistry. (ii) Areas with more than single sequence coverage that contained ambiguities or G–C compressions were also sequenced again with an alternative sequencing chemistry. The combination of sequence redundancy together with the application of an alternative sequencing chemistry in areas with ambiguities is, we believe at least as accurate, if not more so, than double-stranded coverage. By these criteria we have reduced the number of nucleotide ambiguities [International Union of Biochemistry (IUB) codes] in the sequence to less than 1 in 19,000. The same approaches used to resolve ambiguities were also applied to areas where apparent frameshifts were indicated. Sixty potential frameshifts were identified by comparison to entries in peptide databases. Although some of these potential frameshifts are undoubtedly real, others may reflect the hundreds of frameshifts present in GenBank sequences from public databases (*30*). They may also represent biologically significant phenomena such as insertions or deletions in insertion elements, or in tandem repeats often associated with virulence genes (*31*).

We also considered comparison of our sequence to existing GenBank *H. influenzae* Rd sequences as a method for evaluating sequence accuracy as reported for yeast chromosome VIII (*32*). Unlike yeast, only a limited number of *H. influenzae* sequences are in GenBank (38 *H. influenzae* Rd accessions) and these are not necessarily of high accuracy. The results of such a comparison show that our sequence is 99.67 percent identical overall to those GenBank sequences annotated as *H. influenzae* Rd. Two problems were apparent with this type of comparison. Sequences could differ because of strain variation, which is poorly annotated in the GenBank entries. It is also difficult to evaluate the significance of differences as the accuracy of the GenBank entries was impossible to assess. We compared GenBank accession M86702 (*strA* resistance gene) to our sequence and found the identity to be 94.7 percent over 545 bp. There are 24 single base pair mismatches relative to our sequence as well as an insertion and a deletion. Comparison of our sequence to GenBank accession L23824 (adenylate cyclase) shows a 99.7 percent match over 2960 bp. There are nine single base pair mismatches and one insertion. In this case the mismatches all fall in the noncoding flanking regions. While we cannot speak to the accuracy of these GenBank sequences, we are very confident of our sequences in these regions because of the 3× to 9× coverage with high-quality sequence data. Thus, a comparison of our sequence to sequences in GenBank annotated as *H. influenzae* Rd is not a meaningful way to evaluate the accuracy of the sequence.

Although it is extremely difficult to assess sequence accuracy, we wanted to provide an approximation of accuracy based on frequency of shifts in open reading frames, unresolved ambiguities, overall quality of raw data, and fold coverage. We estimate our error rate to be between 1 base in 5000 and 1 base in 10,000.

We also attempted to estimate the cost of the complete sequencing of the genome. Reagent and labor costs for construction of small insert and λ libraries, template preparation and sequencing, gap closure, sequence confirmation, annotation, and preparation for publication were summed and divided by the genome length. Sequencing projects that require up front mapping should include the cost of construction of the clone maps for sequencing. Not included were costs associated with development of technology and software that will be used for future sequencing projects. The estimated direct cost was 48 cents per finished base pair. Because of the techniques developed during this project any future genomes of this size should cost less.

Data and software availability. The *H. influenzae* genome sequence has been deposited in the Genome Sequence DataBase (GSDB) with the accession number L42023 and is termed version 1.0. The nucleotide sequence and peptide translation of each predicted coding region with identified start and stop codons have also been accessioned by GSDB. We consider annotation, accuracy checking, and error resolution to be ongoing tasks. As outlined above, there are predicted coding regions with potential frameshift errors in the sequence. As these are resolved, they will be deposited with GSDB. We also expect the annotation of the sequence to increase over time and be updated in GSDB.

Additional data are available on our World Wide Web site (http://www.tigr.org). An expanded version of Table 3 has links to the database accessions that were used to identify the predicted coding regions, additional sequence similarity data, and coordinates of the predicted coding regions. The alignments between the predicted coding regions and the database sequences are also available. The data can also be queried by gene identification number, putative identification, matching accession, and role. The entire sequence and the sequences of all predicted coding regions and their translations, including those having frameshifts, are also available. This Web site will be maintained as an up-to-date source of *H. influenzae* genome sequence data, and we encourage the scientific community to forward their results for inclusion (with proper attribution) at this site.

The software developed at TIGR that is described in the article is still under development. However, TIGR will work with other genome centers to make its software available upon request.

Genome analysis. We have attempted to predict all of the coding regions and identify genes, transfer RNAs (tRNAs) and rRNAs, as well as other features of the DNA sequence (such as repeats, regulatory sites, replication origin sites, and nucleotide composition), with the realization that biochemical and biological conformation of many of these will be an ongoing task. We include a description of some of the most obvious sequence features.

The *H. influenzae* Rd genome is a circular chromosome of 1,830,137 bp. The overall G+C nucleotide content is approximately 38 percent (A, 31 percent; C, 19 percent; G, 19 percent; T, 31 percent). The G+C content of the genome was examined with several window lengths to look for global structural features. With a window of 5000 bp, the G+C content is relatively even except for seven large regions rich in G+C and several regions rich in A+T (Fig. 1). The G+C–rich regions correspond to six rRNA operons and a cryptic mu-like prophage. Genes for several proteins similar to proteins encoded by bacteriophage mu are located at approximately position 1.56 to 1.59 Mbp of the genome. This area of the genome has a markedly higher G+C content than average for *H. influenzae* (~50 percent G+C compared to ~38 percent for the rest of the genome).

The minimal origin of replication (*oriC*) in *E. coli* is a 245-bp region defined by three copies of a 13-bp repeat at one end (sites for initial DNA unwinding) and four copies of a 9-bp repeat (sites for DnaA binding, the first step in replication) at the other (*33*). An approximately 280-bp sequence containing structures similar to the three 13-bp and four 9-bp repeats defines the putative origin of replication in *H. influenzae* Rd. This region lies between sets of ribosomal operons *rrnF, rrnE, rrnD* and *rrnA, rrnB, rrnC*. These two groups of ribosomal operons are transcribed in opposite directions and the placement of the origin is consistent with their polarity for transcription. Termination of *E. coli* replication is marked by two 23-bp termination sequences located ~100 kb on either side of the midway point at which the two replication forks meet. Two potential termination sequences sharing a 10-bp core sequence with the *E. coli* termination sequence were identified in *H. influenzae*. These two regions are offset approximately 100 kb from a point approximately 180° opposite of the proposed origin of *H. influenzae* replication.

Six rRNA operons were identified. Each contains three subunits and a variable spacer region in the order: 16S subunit—spacer region—23S subunit—5S subunit. The subunit lengths are 1539, 2653, and 116 bp, respectively. The G+C content of the three ribosomal subunits (50 percent) is higher than that of the genome as a whole. The G+C content of the spacer region (38 percent) is consistent with the remainder of the genome. The nucleotide sequence of the three rRNA subunits is completely identical in all six ribosomal operons. The rRNA operons can be grouped into two classes based on the spacer region between the 16S and 23S sequences. The shorter of the two spacer regions is 478 bp (*rrnb, rrnE*, and *rrnF*) and contains the gene for $\mathrm{tRNA^{Glu}}$. The longer spacer is 723 bp (*rrnA, rrnC*, and *rrnD*) and contains the genes for $\mathrm{tRNA^{Ile}}$ and $\mathrm{tRNA^{Ala}}$. The two sets of spacer regions are also completely identical across each group of three operons. Other tRNA genes are present at the 16S and 5S ends of two of the rRNA operons. The genes for $\mathrm{tRNA^{Arg}}$, $\mathrm{tRNA^{His}}$, and $\mathrm{tRNA^{Pro}}$ are located at the 16S end of *rrnE* while the genes for $\mathrm{tRNA^{Trp}}$ and $\mathrm{tRNA^{Asp}}$ are located at the 5S end of *rrnA*.

The predicted coding regions were initially defined by evaluating their coding potential with the program GENEMARK (*34*) based on codon frequency matrices derived from 122 *H. influenzae* coding sequences in GenBank. The predicted coding region sequences (plus 300 bp of flanking sequence) were used in searches against a database of nonredundant bacterial proteins (NRBP) created specifically for the annotation. Redundancy was removed from NRBP at two stages. All DNA coding sequences were extracted from GenBank (release 85), and sequences from the same species were searched against each other. Sequences having more than 97 percent identity over regions longer than 100 nucleotides were combined. In addition, the sequences were translated and used in protein comparisons with all sequences in Swiss-Prot (release 30). Sequences belonging to the same species and having more than 98 percent similarity over 33 amino acids were combined. NRBP is composed of 21,445 sequences extracted from 23,751 GenBank sequences and 11,183 Swiss-Prot sequences from 1099 different species.

A total of 1743 predicted coding regions was identified. Searches of the predicted coding regions for *H. influenzae* were performed against NRBP with BLAZE (*35*) run on a Maspar MP-2 massively parallel computer with 4096 microprocessors. BLAZE translates the query DNA sequence in the three plus-strand reading frames and identifies the protein sequences that match the query. The protein-protein matches were aligned with PRAZE, a modified Smith-Waterman (*23*) algorithm. In cases where insertions or deletions in the DNA sequence produced a potential frameshift, the alignment algorithm started with protein regions of maximum similarity and extended the alignment to the same database match in alternative frames by means of the 300-bp flanking region. Unidentified predicted coding regions and the remaining intergenic sequences were searched against a dataset of all available peptide sequences from Swiss-Prot, the Protein Information Resource (PIR), and GenBank. Identification of operon structures is expected to be facilitated by experimental determination of promoter and termination sites.

Each putatively identified *H. influenzae* gene was assigned to one of 102 biological role categories adapted from Riley (*36*). Assignments were made by linking the protein sequence of the predicted coding regions with the Swiss-Prot sequences in the Riley database. Of the 1743 predicted coding regions, 736 have no role assignment. Of these, no database match was found for 389, while 347 matched "hypothetical proteins" in the database. Role assignments were made for 1007 of the predicted coding regions. Each of the 102 role categories was grouped into one of 14 broader role categories (Table 2). A compilation of all the predicted coding regions, their identifiers, a three-letter gene identifier, and percent similarity are presented in Table 3 (fold-out). An annotated complete genome map of *H. influenzae* Rd is presented in Fig. 2 (fold-out). The map places each predicted coding region on the *H. influenzae* chromosome, indicates its direction of transcription and color codes its role assignment. Role assignments are also represented in Fig. 1.

A survey of the genes and their chromosomal organization in *H. influenzae* Rd makes possible a description of the metabolic processes *H. influenzae* requires for survival as a free-living organism, the nutritional requirements for its growth in the laboratory, and the characteristics that make it different from other organisms specifically as they relate to its pathogenicity and virulence. The genome would be expected to have complete complements of certain classes of genes known to be essential for life. For example, there is a one-to-one correspondence of published *E. coli* ribosomal protein sequences to potential homologs in the *H. influenzae* database. Likewise, as shown in Table 3, an aminoacyl tRNA synthetase is present in the genome for each amino acid. Finally, the location of tRNA genes was mapped onto the genome. There are 54 identified tRNA genes, including representatives of all 20 amino acids.

In order to survive as a free-living organism, *H. influenzae* must produce energy in the form of ATP via fermentation or electron transport. As a facultative anaerobe, *H. influenzae* Rd is known to ferment glucose, fructose, galactose, ribose, xylose, and fucose (*37*). As indicated by the genes identified in Table 3, transport systems are available for the uptake of these sugars by the phosphoenolypyruvate-phosphotransferase system (PTS), and by non-PTS mechanisms. Genes that specify the common phosphate-carriers enzyme I and Hpr (*ptsI* and *ptsH*) of the PTS system were identified as well as the glucose-specific *crr* gene. We have not, however, identified the gene-encoding, membrane-bound, glucose-specific enzyme II. The latter enzyme is required for transport of glucose by the PTS system. A complete PTS system for fructose was identified.

Genes encoding the complete glycolytic pathway and for the production of fermentative end products were identified. Also identified were genes encoding functional anaerobic electron transport systems that depend on inorganic electron acceptors such as nitrates, nitrites, and dimethyl sulfoxide. Genes encoding three enzymes of the tricarboxylic acid (TCA) cycle appear to be absent from the genome. Citrate synthase, isocitrate dehydrogenase, and aconitase were not found by searching the predicted coding regions or by using the *E. coli* enzymes as peptide queries against the entire genome in translation. This provides an explanation for the large amount of glutamate (1 g/liter) that is required in defined culture media (*38*). Glutamate can be directed into the TCA cycle by conversion to α-ketoglutarate by glutamate dehydrogenase. In the absence of a complete TCA cycle, glutamate presumably serves as the source of carbon for biosynthesis of amino acids from precursors that branch from the TCA cycle. Functional electron transport systems that depend on oxygen as a terminal electron acceptor are available for the production of adenosine triphosphate.

Previously unanswered questions regarding pathogenicity and virulence can be addressed by examining certain classes of genes such as adhesins and the lipo-oligosaccharide biogenesis genes. Moxon and coworkers (*31*) have obtained evidence that a number of these virulence-related genes contain tandem tetramer repeats that undergo frequent addition and deletion of one or more repeat units during replication such that the reading frame of the gene is changed and its expression thereby altered. It is now possible, by means of the complete genome sequence, to locate all such tandem repeat tracts (Fig. 2) and to begin to determine their roles in phase variation of such potential virulence genes.

Haemophilus influenzae Rd has a highly efficient, DNA transformation system. The DNA uptake sequence site, 5′ AAGTGCGGT, present in multiple copies in the genome, is necessary for efficient DNA uptake (*39*). It is now possible to locate all of these sites and describe their distribution with respect to genic and intergenic regions (*40*). Fifteen genes involved in transformation have already been described and sequenced (*41*). Six of the genes, *comA* to *comF*, comprise an operon that is under positive control by a 22-bp, palindromic, competence regulatory element (CRE) located approximately one helix turn upstream of the promoter. It is now feasible to locate additional copies of CRE in the genome and discover potential transformation genes under CRE control (*42*). In addition, other global regulatory elements may be discovered with an ease not previously possible.

Table 4. Two-component systems in *H. influenzae* Rd. ID, identity; Sim, similarity.

Identification number	Location	Best match*	Id (%)	Sim (%)	Length (bp)
		Sensors			
HI0220	239,378	*arcB*	39.5	63.9	200
HI0267	299,541	*narQ*	38.1	68.0	562
HI1707	1,781,143	*basS*	27.7	51.5	250
HI1378	1,475,017	*phoR*	38.1	61.6	280
		Regulators			
HI0726	777,934	*narP*	59.3	77.0	209
HI0837	887,011	*cpxR*	51.9	73.0	229
HI0884	936,624	*arcA*	77.2	87.8	236
HI1379	1,475,502	*phoB*	52.9	71.4	228
HI1708	1,781,799	*basR*	43.5	59.3	219

*In all cases, the best match was to a gene of *E. coli*.

One well-described system for gene regulation in bacteria is the "two-component" system composed of a sensor molecule that detects an environmental signal and a regulator molecule that is phosphorylated by the activated form of the sensor. The regulator protein is generally a transcription factor that, when activated by the sensor, turns on or off expression of a specific set of genes. It has been estimated that *E. coli* harbors 40 sensor-regulator pairs (*43*). The *H. influenzae* genome was searched with representative proteins from each family of sensor and regulator proteins with TBLASTN and TFASTA. Four sensor and five regulator proteins were identified with similarity to proteins from other species (Table 4). There appears to be a corresponding sensor for each regulator protein except CpxR. Searches with the CpxA protein from *E. coli* identified three of the four sensors listed in Table 4, but no additional significant matches were found. It is possible that the sequence similarity is low enough to be undetectable with TFASTA. All of the regulator proteins present fall into the OmpR subclass (*43*). No representatives of the NtrC class of regulators were found. This class of proteins interacts directly with the sigma-54 subunit of RNA polymerase, which is absent from *H. influenzae*, and which plays a major role in the regulation of a large number of operons in *E. coli* and other enterobacteria. The absence of the Ntr network in *H. influenzae* suggests significant differences in the regulatory processes between these two groups of organisms.

Some of the most interesting questions that can be answered by a complete genome sequence relate to the genes or pathways that are absent. The nonpathogenic *H. influenzae* Rd strain varies significantly from the pathogenic serotype b strains. Many of the differences between these two strains appear in factors affecting infectivity. For example, we have found that the eight genes that make up the fimbrial gene cluster (*44*) involved in adhesion of bacteria to host cells are absent in the Rd strain. The *pepN* and *purE* genes, which flank the fimbrial cluster in *H. influenzae* type b strains,

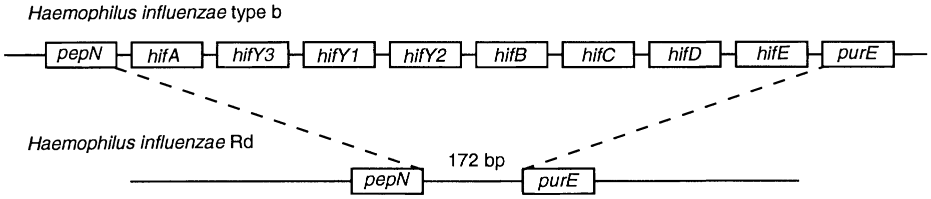


Fig. 3. A comparison of the region of the *H. influenzae* chromosome containing the eight genes of the fimbrial gene cluster present in *H. influenzae* type b and the same region in *H. influenzae* Rd. The region is flanked by *pepN* and *purE* in both organisms. However, in the noninfectious Rd strain the eight genes of the fimbrial gene cluster have been excised. A 172-bp spacer region is located in this region in the Rd strain and continues to be flanked by the *pepN* and *purE* genes.

are adjacent to one another in the Rd strain (Fig. 3), suggesting that the entire fimbrial cluster was excised.

On a broader level, we determined which *E. coli* proteins are not in *H. influenzae* by taking advantage of a nonredundant set of protein-coding genes from *E. coli*, namely the University of Wisconsin Genome Project contigs in GenBank: 1216 predicted protein sequences from GenBank accessions D10483, L10328, U00006, U00039, U14003, and U18997 (*45*). The minimum threshold for matches was set so that even weak matches would be scored as positive, thereby giving a minimal estimate of the *E. coli* genes not present in *H. influenzae*. We used TBLASTN to search each of the *E. coli* proteins against the complete genome. All BLAST scores greater than 100 were considered matches. Altogether 627 *E. coli* proteins matched at least one region of the *H. influenzae* genome and 589 proteins did not. The 589 nonmatching proteins were examined and found to contain a disproportionate number of hypothetical proteins from *E. coli*. Sixty-eight percent of the identified *E. coli* proteins were matched by an *H. influenzae* sequence whereas only 38 percent of the hypothetical proteins were matched. Proteins are annotated as hypothetical on the basis of a lack of matches with any other known proteins (*45*). At least two potential explanations can be offered for the overrepresentation of hypothetical proteins among those without matches: (i) some of the hypothetical proteins are not, in fact, translated (at least in the annotated frame), or (ii) these are *E. coli*–specific proteins that are unlikely to be found in any species except those most closely related to *E. coli*, for example, *Salmonella typhimurium*.

A total of 389 predicted coding regions did not display significant similarity with a six-frame translation of GenBank release 87. These unidentified coding regions were compared to one another with FASTA. Two previously unidentified gene families were identified. Two predicted coding regions without database matches (HI0589 and HI0850) share 75 percent identity over almost their entire lengths (139 and 143 amino acid residues respectively). A second pair of predicted coding regions (HI1555 and HI1548) encode proteins that share 30 percent identity over almost their entire lengths (394 and 417 amino acids respectively). These similarities suggest that there may be previously unidentified gene families present in these regions.

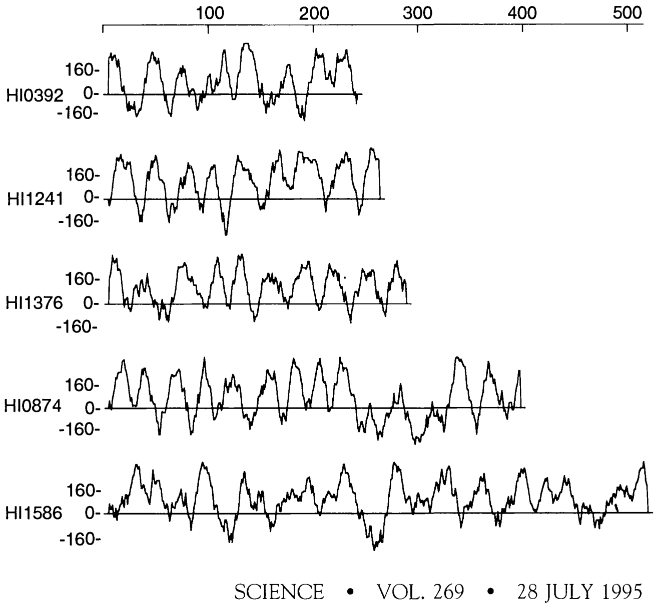


Fig. 4. Hydrophobicity analysis of five potential channel proteins. The amino acid sequences of five predicted coding regions that do not display similarity with known peptide sequences (GenBank release 87), each exhibit multiple hydrophobic domains that are characteristic of channel-forming proteins. The predicted coding region sequences were analyzed by the Kyte-Doolittle algorithm (*46*) (with a range of 11 residues) with the GENEWORKS software package (Intelligenetics).

Another analysis that can be applied to the unidentified coding regions is hydropathy analysis, which indicates the patterns of potential membrane-spanning domains that are often conserved between members of receptor and transporter gene families, even in the absence of significant amino acid identity. The five best examples of unidentified predicted coding regions that display potential transmembrane domains with a periodic pattern that is characteristic of membrane-bound channel proteins are shown in Fig. 4. Such information can be used to focus on specific aspects of cellular function that are affected by targeted deletion or mutation of these genes.

We have learned some important lessons concerning overall strategy from the *H. influenzae* sequencing project that should reduce the effort required for future bacterial genome sequencing projects. For example, the small insert library and the large insert library should be constructed and end-sequenced concurrently. It is essential that the sequence fragments used for the assembly are of the highest quality. The sequences should be rigorously checked for vector contamination. Although it is important that sequence read lengths be long enough to span most small repeats, they must also be highly accurate. Our raw sequence data contained on average less than 1.5 percent uncertainties. The use of high quality individual sequence fragments and a rigorous assembly algorithm essentially eliminated difficulty with achieving closure. The success of whole genome shotgun sequencing offers the potential to accelerate research in a number of areas. Comparative genomics could be advanced by the availability of an increased number of complete genomes from a variety of prokaryotes and eukaryotes. Knowledge of the complete genomes of pathogenic organisms could lead to new vaccines. Information obtained from the genomes of particular organisms could have industrial applications. Finally, this strategy has potential to facilitate the sequencing of the human genome.

REFERENCES AND NOTES

1. F. Sanger *et al.*, *Nature* **246**, 687 (1977); F. Sanger, A. R. Coulson, G. F. Hong, D. F. Hill, G. B. Petersen, *J. Mol. Biol.* **162**, 729 (1982).
2. A. T. Bankier *et al.*, *DNA Seq.* **2**, 1 (1991).
3. S. J. Goebel *et al.*, *Virology* **179**, 247 (1990).
4. K. Oda *et al.*, *J. Mol. Biol.* **223**, 1 (1992); K. Ohyama *et al.*, *Nature* **322**, 572 (1986).
5. R. F. Massung *et al.*, *Nature* **366**, 748 (1993).
6. D. L. Hartl and M. J. Palazzolo, *Genome Research in Molecular Medicine and Virology*, K. W. Adolph, Ed. (Academic Press, Orlando, FL, 1993), pp. 115–129.
7. H. J. Sofia *et al.*, *Nucleic Acids Res.* **22**, 2576 (1994).
8. J. Levy, *Yeast* **10**, 1689 (1994).
9. P. Glaser *et al.*, *Mol. Microbiol.* **10**, 371 (1993).
10. J. Sulston *et al.*, *Nature* **356**, 37 (1992).
11. W. F. Bodmer, *Rev. Invest. Clin.* (suppl., pp. 3–5) (1994).
12. M. D. Adams, C. Fields, J. C. Venter, Eds. *Automat-*

ed DNA Sequencing and Analysis (Academic Press, San Diego, CA, 1994).

13. M. D. Adams *et al.*, *Science* **252**, 1651 (1991); M. D. Adams *et al.*, *Nature* **355**, 632 (1992); M. D. Adams *et al.*, *ibid.*, in press.
14. E. S. Lander and M. S. Waterman, *Genomics* **2**, 231 (1988).
15. *Haemophilus influenzae* Rd KW20 DNA was prepared by extraction with phenol. A mixture (3.3 ml) containing 600 μg of DNA, 300 mM sodium acetate, 10 mM tris-HCl, 1 mM Na-EDTA, and 30 percent glycerol was sonicated (Branson Model 450 Sonicator) at the lowest energy setting for 1 minute at 0°C with a 3-mm probe. The DNA was precipitated in ethanol and redissolved in 500 μl of tris-EDTA (TE) buffer to create blunt ends; a 100-μl portion was digested for 10 minutes at 30°C in 200 μl of BAL 31 buffer with 5 units of BAL 31 nuclease (New England BioLabs). The DNA was extracted with phenol, precipitated in ethanol, redissolved in 100 μl of TE buffer, and fractionated on a 1.0 percent low melting agarose gel. A fraction (1.6 to 2.0 kb) was excised, extracted with phenol, and redissolved in 20 μl of TE buffer. A two-step ligation procedure was used to produce a plasmid library in which 97 percent of the recombinants contained inserts, of which >99 percent were single inserts. The first ligation mixture (50 μl) contained 2 μg of DNA fragments, 2 μg of Sma I + bacterial alkaline phosphatase pUC18 DNA (Pharmacia), and 10 units of T4 ligase (Gibco/BRL), and incubation was at 14°C for 4 hours. After extraction with phenol and ethanol precipitation, the DNA was dissolved in 20 μl of TE buffer and separated by electrophoresis on a 1.0 percent low melting agarose gel. A ladder of ethidium bromide–stained linearized DNA bands, identified by size as insert (i), vector (v), v+i, v+2i, v+3i, and so on, was visualized by 360-nm ultraviolet light, and the v+i DNA was excised and recovered in 20 μl of TE. The v+i DNA was blunt-ended by T4 polymerase treatment for 5 minutes at 37°C in a reaction mixture (50 μl) containing the linearized v+i fragments four deoxynucleotide triphosphates (dNTPs) (500 μM each) and 9 units of T4 polymerase (New England BioLabs) under buffer conditions recommended by the supplier. After phenol extraction and ethanol precipitation, the repaired v+i linear pieces were dissolved in 20 μl of TE. The final ligation to produce circles was carried out in a 50-μl reaction containing 5 μl of v+i DNA and 5 units of T4 ligase at 14°C overnight. The reaction mixture was heated for 10 minutes at 70°C and stored at −20°C.
16. A 100-μl portion of Epicurian Coli SURE 2 Supercompetent Cells (Stratagene 200152) was thawed on ice and transferred to a chilled Falcon 2059 tube on ice. A 1.7-μl volume of 1.42 M β-mercaptoethanol was added to the cells to a final concentration of 25 mM. Cells were incubated on ice for 10 minutes. A 1-μl sample of the final ligation mix was added to the cells and incubated on ice for 30 minutes. The cells were heat-treated for 30 seconds at 42°C and placed back on ice for 2 minutes. The outgrowth period in liquid culture was omitted to minimize the preferential growth of any given transformed cell. Instead, the transformed cells were plated directly on a nutrient rich SOB plate containing a 5-ml bottom layer of SOB agar (1.5 percent SOB agar consisted of 20 g of tryptone, 5 g of yeast extract, 0.5 g of NaCl, and 1.5 percent Difco agar/liter). The 5-ml bottom layer was supplemented with 0.4 ml of ampicillin (50 mg/ml) per 100 ml of SOB agar. The 15-ml top layer of SOB agar was supplemented with 1 ml of X-gal (2 percent), 1 ml of $MgCl_2$ (1 M), and 1 ml of $MgSO_4$ (1 M) per 100 ml of SOB agar. The 15-ml top layer was poured just before plating. Our titer was approximately 100 colonies per 10-μl aliquot of transformation.
17. K. W. Wilcox and H. O. Smith, *J. Bact.* **122**, 443 (1975).
18. A. Greener, *Strategies* **3**, 5 (1990).
19. T. R. Utterback *et al.*, in preparation.
20. For the unamplified λ library, *H. influenzae* Rd KW20 DNA (>100 kb) was partially digested in a reaction mixture (200 μl) containing 50 μg of DNA, 1× Sau3A I buffer, and 20 units of Sau3A I for 6 minutes at 23°C. The digested DNA was extracted with phenol and fractionated on a 0.5 percent low melting agarose gel at 2 V/cm for 7 hours. Fragments from 15 to 25 kb were excised and recovered in a final volume of 6 μl. We used 1 μl of fragments with 1 μl of DASHII vector (Strategene) in the recommended ligation reaction. One microliter of the ligation mixture was used per packaging reaction as recommended in the protocol with the Gigapack II XL Packaging Extract (Stratagene, 227711). Phage were plated directly without amplification from the packaging mixture (after dilution with 500 μl of recommended SM buffer and treatment with chloroform). [SM buffer contains (per liter) 5.8 g of NaCl, 2 g of $MgSO_4 \cdot H_2O$, 50 ml of 1 M tris-HCl, pH7.5, and 5 ml of a 2 percent solution of gelatin.] The yield was about 2.5×10^3 plaque-forming units (PFU) per microliter. The amplified library was prepared essentially as above except the λ GEM-12 vector was used. After packaging, about 3.5×10^4 PFU were plated on the restrictive NM539 host. The lysate was harvested in 2 ml of SM buffer and stored frozen in 7 percent dimethyl sulfoxide. The phage titer was approximately 1×10^9 PFU/ml.
21. M. D. Adams, *et al.*, *Nature* **368**, 474 (1994).
22. A. R. Kerlavage *et al.*, *Proceedings of the Twenty-Sixth Annual Hawaii International Conference on System Science* (IEEE Computer Society Press, Washington, DC, 1993), p. 585; A. R. Kerlavage *et al.*, *IEEE Computers in Medicine and Biology* (IEEE, Computer Society Press, Washington, DC, in press).
23. M. S. Waterman, *Methods Enzymol.* **164**, 765 (1988).
24. W. Pearson and D. Lipman, *Proc. Natl. Acad. Sci. U.S.A.* **85**, 2444 (1988).
25. Oligonucleotides were labeled by combining 50 pmol of each 20-mer and 250 mCi of $[\gamma\text{-}^{32}P]$ adenosine triphosphate and T4 polynucleotide kinase. The labeled oligonucleotides were purified with Sephadex G-25 superfine (Pharmacia). A portion containing 10^7 counts per minute of each was used in a Southern hybridization analysis of *H. influenzae* Rd chromosomal DNA digested with one frequently cleaving endonuclease (Ase I) and five less-frequent ones (Bgl II, Eco RI, Pst I, Xba I, and Pvu II). The DNA from each digest was fractionated on a 0.7 percent agarose gel and transferred to nylon (Nytran Plus) membranes (Schleicher & Schuell). Hybridization was carried out for 16 hours at 40°C. To remove nonspecific signals, we sequentially washed each blot at room temperature with increasingly stringent conditions up to 0.1× saline sodium citrate and 0.5 percent SDS. Blots were exposed to a PhosphorImager cassette (Molecular Dynamics) for several hours; hybridization patterns were compared visually.
26. S. Altschul *et al.*, *J. Mol. Biol.* **215**, 403 (1990).
27. E. F. Kirkness *et al.*, *Genomics* **10**, 985 (1991).
28. Standard amplification by polymerase chain reaction (PCR) was performed in the following manner. Each reaction (57 μl) contained a 37-μl mixture of 16.5 μl of H_2O, 3 μl of 25 mM $MgCl_2$, 8 μl of a dNTP mix (1.25 mM each dNTP), 4.5 μl of 10× PCR core buffer II (Perkin-Elmer N808-0009), and 25 ng of *H. influenzae* Rd KW20 genomic DNA. The appropriate two primers (4 μl, 3.2 pmol/μl) were added to each reaction. A preliminary incubation (hotstart) was performed at 95°C for 5 minutes followed by a 75°C hold. During the holding period, Amplitaq DNA polymerase (Perkin-Elmer N801-0060, 0.3 μl in 4.3 μl of H_2O, 0.5 μl of 10× PCR core buffer II) was added to each reaction. The PCR profile was 25 cycles of 94°C for 45 seconds, then denature; 55°C for 1 minute, then aneal; 72°C for 3 minutes, then extension. All reactions were performed in a 96-well format on a Perkin-Elmer GeneAmp PCR System 9600. Long-range PCR was performed as follows: Each reaction contained a 35.2-μl mixture of 12.0 μl of H_2O, 2.2 μl of 25 mM magnesium acetate, 4 μl of a dNTP mixture (200 μM final concentration), 12.0 μl of 3.3× PCR buffer, and 25 ng of *H. influenzae* Rd KW20 genomic DNA. The appropriate two primers (5 μl, 3.2 pmol/μl) were added to each reaction. A preliminary incubation (hot start) was performed at 94°C for 1 minute. Then r*Tth* polymerase (Perkin-Elmer N808-0180) (4 units per reaction) in 2.8 μl of 3.3× PCR buffer II was added to each reaction. The PCR profile was 18 cycles of 94°C for 15 seconds, denature; 62°C for 8 minutes, anneal and extend followed by 12 cycles 94°C for 15 seconds, denature; 62°C for 8 minutes (increase 15 per cycle), anneal and extend; and 72°C for 10 minutes, final extension. All reactions were done in a 96-well format on a Perkin-Elmer GeneAmp PCR System 9600.
29. J. J. Lee, H. O. Smith, R. R. Redfield, *J. Bacteriol.* **171**, 3016 (1989).
30. J. M. Claverie, *J. Mol. Biol.* **234**, 1140 (1993).
31. J. N. Weiser *et al.*, *Cell* **59**, 657 (1989).
32. M. Johnston *et al., Science* **265**, 2077 (1994).
33. B. Lewin, Ed., *Genes V* (Oxford Univ. Press, New York, 1994), chaps. 18 and 19.
34. M. Borodovsky and J. McIninch, *Comp. Chem.* **17**, 123 (1993). In the GeneMark program second-order phased Markov chain models were used; it was trained on 188,572 bp of protein coding sequence and 33,118 bp of noncoding sequence as annotated in GenBank *H. influenzae* entries. It was shown that the second-order program is the most accurate given the size of the training set. The accuracy level was assessed by a cross-validation procedure with a set of 96-bp nonoverlapping fragments derived from the same sets of sequences. With the use of a threshold of 0.5, coding fragments were identified correctly in 91.2 percent of the cases; noncoding fragments were identified correctly in 93.3 percent of the cases.
35. D. Brutlag *et al.*, *ibid.*, p. 203. The BLOSUM 60–amino acid substitution matrix was used in all protein-protein comparisons [S. Henikoff and J. G. Henikoff, *Proc. Natl. Acad. Sci. U.S.A.* **89**, 10915 (1992)].
36. M. Riley, *Microbiol. Rev.* **57**, 862 (1993).
37. I. R. Dorocicz *et al.*, *J. Bacteriol.* **175**, 7142 (1993); B. Dougherty, unpublished results.
38. R. D. Klein and G. H. Luginbuhl, *J. Gen. Microbiol.* **113**, 409 (1979).
39. D. B. Danner *et al.*, *Gene* **11**, 311 (1980); D. B. Danner *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **79**, 2393 (1982); M. E. Kahn and H. O. Smith, *J. Membr. Biol.* **138**, 155 (1984).
40. H. O. Smith *et al.*, *Science* **269**, 538 (1995).
41. R. R. Redfield, *J. Bacteriol.* **173**, 5612 (1991); M. S. Chandler, *Proc. Natl. Acad. Sci. U.S.A.* **89**, 1616 (1992); R. Barouki and H. O. Smith, *J. Bacteriol.* **163**, 629 (1985); J.-F. Tomb, H. El-Hajji, H. O. Smith, *Gene* **104**, 1 (1991); J.-F. Tomb, *Proc. Natl. Acad. Sci. U.S.A.* **89**, 10252 (1992).
42. J.-F. Tomb, unpublished results.
43. L. M. Albright, E. Huala, F. M. Ausubel, *Annu. Rev. Genet.* **23**, 311 (1989); J. S. Parkinson and E. C. Kofoid, *Am. Rev. Genet.* **26**, 71 (1992).
44. M. S. vanHam, L. vanAlphen, F. R. Mooi, J. P. VanPattern, *Mol. Microbiol.* **13**, 673 (1994).
45. T. Yura *et al.*, *Nucleic Acids Res.* **20**, 3305 (1992); V. Burland *et al.*, *Genomics* **16**, 551 (1993).
46. J. Kyte and R. F. Doolittle, *J. Mol. Biol.* **157**, 105 (1982).

47. Supported in part by a core grant from Human Genome Sciences and an American Cancer Society grant (NP-838C) (to H.O.S.). Reagents for sequencing reactions and the synthesis of the oligonucleotides were a gift from the Applied Biosystems Division of Perkin-Elmer. We thank T. Burcham of Applied Biosystems for his contribution in the development of the TIGR EDITOR software; M. Riley, Marine Biological Laboratory, Woods Hole, for making her *E. coli* database available; M. Borodovsky and W. Hayes, School of Biology, Georgia Institute of Technology for providing and tuning the GeneMark software for use with *H. influenzae*; and J. Kelley, T. Dixon, and V. Sapiro for their excellent computer system support. H.O.S. is an American Cancer Society research professor.

16 May 1995; accepted 28 June 1995