MICROBIOLOGY AND MOLECULAR BIOLOGY REVIEWS, Sept. 1998, p. 985–1019 Vol. 62, No. 3
1092-2172/98/$04.00+0

Linkage Map of *Escherichia coli* K-12, Edition 10: The Physical Map

KENNETH E. RUDD

Department of Biochemistry and Molecular Biology, University of Miami School of Medicine, Miami, Florida 33101-6129

INTRODUCTION

EcoMap10, the physical map of edition 10 of the *Escherichia coli* K-12 linkage map, is a map of restriction sites and genomic positions of a set of bacteriophage lambda clones and includes a graphic representation of EcoGene10, a refined annotation of the *Escherichia coli* genome sequence. The previous version, EcoMap7, was published as part of edition 9 of the *Escherichia coli* K-12 linkage map (4). A brief description of EcoMap10 is provided here, and a more detailed description of EcoMap10 and the EcoGene10 data set will be published separately (16). The most significant change in EcoMap construction is that it is now based upon the complete genome sequence of *E. coli* K-12 strain MG1655 version M52 (4,639,221 bp) as determined by Blattner et al. (5). EcoMap10 features, including the predicted restriction sites, Kohara clone alignments, protein coding regions, gene and open reading frame (ORF) designations, insertion sequence (IS) elements, and repetitive extragenic palindrome (REP) clusters, have all been derived by using version M52 of the MG1655 DNA sequence (GenBank/EMBL/DDBJ accession no., U00096). The tables and references in the traditional map of edition 10 of the *Escherichia coli* K-12 linkage map (3) also apply to the genes displayed in the physical map.

EcoMap10

Restriction Enzyme Recognition Sites

The recognition sites for the eight restriction enzymes used to create the whole genome restriction map of Kohara et al. (10) are predicted from the genomic DNA sequence. These 6-bp recognition sites are mapped at the position of their first base pair. Although any set of restriction sites can now be used to create a restriction map of the entire chromosome, this set of sites was used in order to retain continuity with the original Kohara/Isono genomic restriction map and previous EcoMap versions. This set of commonly used enzymes provides a convenient pattern of restriction sites and includes a wide range in the number of predicted recognition sites in the MG1655 genome: *Bam*HI, 495; *Kpn*I, 516; *Hin*dIII, 556; *Eco*RI, 645; *Pst*I, 958; *Pvu*II, 1,778; *Bgl*I, 1,919; and *Eco*RV, 2,040. The expected number of 6-bp restriction enzyme recognition sites in a randomly generated DNA sequence of this length and composition would be 1,133. The mean number of predicted sites for this set of eight enzymes is 1,113.

Kohara/Isono Miniset Clones

The Kohara/Isono miniset is a widely used collection of ordered *E. coli* bacteriophage lambda clones derived from strain *E. coli* K-12 W3110 (10). Four hundred and seventy-three of the original 476 miniset clones have been aligned to EcoMap10. Seven of the clones were split into two portions labeled A and B because they crossed the 0-min point, the IN(*rrnD-rrnE*)1 inversion endpoints, or the sites of a duplication and translocation of the *tdc* region specific to the Kohara/Isono version of W3110, as previously described (4, 11, 14, 17, 19, 21). One hundred and eighty-six of the clones are present in GenBank/EMBL/DDBJ as individual sequence entries, and these clones are precisely aligned to the genomic DNA sequence since their chromosomal DNA inserts have been sequenced (1, 9, 13, 22). The remaining clones were positioned by using the gel electrophoresis-derived restriction enzyme map of Kohara et al. (10) as previously described (14, 17). These clones are referred to as "unsequenced" because there are no individual GenBank/EMBL/DDBJ records available for them, even though many of them may in fact have already been sequenced. When additional information about the remaining clones becomes available, this information will be incorporated into the EcoMap alignments. Most of the miniset clones are *Sau*3A partial restriction fragments cloned into the *Bam*HI site of lambda EMBL4, and no attempt was made to align the ends of the unsequenced clones to specific *Sau*3A sites in the genomic sequence. Twenty-four of the miniset clones depicted in EcoMap10 are *Eco*RI partial fragments cloned into the *Eco*RI site of lambda 2001, identified by clone names that begin with the designations E1 to E25. Fourteen of these have GenBank/EMBL/DDBJ entries and have terminal *Eco*RI restriction sites in the database entries that are all aligned to *Eco*RI sites in the genomic DNA sequence. The alignments of the 10 unsequenced *Eco*RI clones were manually adjusted so that their ends align to *Eco*RI restriction sites in the genomic sequence. The orientations depicted for the Kohara/Isono clone inserts indicate that the right arm of lambda is to the

* Mailing address: Department of Biochemistry and Molecular Biology (R-629), University of Miami School of Medicine, P.O. Box 016129, Miami, FL 33101-6129. Phone: (305) 243-6055. Fax: (305) 243-3955. E-mail: rudd@ecogene.med.miami.edu.

right of the insert's restriction map as depicted in EcoMap10 (positive orientation, rightward arrow) or to the left (negative orientation, leftward arrow).

Caution must be taken if EcoMap10 is used as the source of a restriction map for the Kohara/Isono miniset clone since the miniset was derived from W3110. In addition to the rare occurrence of DNA sequence errors and strain-specific DNA sequence polymorphisms that might lead to minor restriction map differences, there are major differences due to genome rearrangements (noted above) and the W3110-specific IS elements (see below) (reviewed in reference 6). Solutions to this problem include using the DNA sequence database entries for the sequenced clone subset or using the original Kohara/Isono W3110 restriction map (10) for the unsequenced subset of clones. In either case, the experimental verification of critical restriction sites is recommended.

IS Elements and REP Clusters

IS and REP (also called PU) elements are repeated DNA sequences and major extragenic features of the *E. coli* chromosome (2, 6). The positions of the IS elements present in MG1655 are determined by searching the complete genomic MG1655 DNA sequence with representative IS family member sequences. The positions of the W3110-specific IS element insertion points are determined from the sequenced W3110 clones whenever possible or estimated from the physical mapping data as previously described (14). The orientations of the IS elements indicate the direction of transcription of the transposase gene, as previously described (4, 6). The IS*5* family element orientations were depicted incorrectly in EcoMap7 (4), and this error has been corrected in EcoMap10. Three putative IS-related sequences of unknown origin were identified and are temporarily designated ISX (2793.3 kb), ISY (2714.1 kb), and ISZ (1293.8 kb). The IS-encoded genes are not considered *E. coli* genes in EcoGene, and it is the full length of the IS element that is represented, not the coding regions contained within them.

REP elements have been postulated to have a variety of RNA- and DNA-related functions, but the stabilization of mRNA is the only firmly established function (2). The positions of individual REP elements were determined by a variety of pattern searches, as will be described elsewhere (16). This approach identified nearly all previously reported REP elements (2) and was used to locate new REP elements. The few REP elements identified earlier that were missed by this approach were annotated manually. Individual REP elements occur in intergenic REP clusters, also called bacterial interspersed mosaic elements (BIMEs) containing from 1 to 12 REP elements interspersed with other small conserved sequences (2, 8). Three hundred and fifty-five REP clusters (BIMEs) containing a total of 697 individual REP elements were identified. Particular attention was given to the detection of a class of REP-like putative bidirectional transcription terminators referred to as PU* or Y* (2, 7). A total of 108 individual Y* elements are included in the REP tabulation (16). Y* elements can also be found as subsequences of a number of other REP elements, but these overlapping Y* elements are not counted separately in the REP tabulation. The serially numbered REP clusters (BIMEs) identified in the MG1655 genome sequence are denoted R1 to R355 directly under the restriction map portion of EcoMap10 along with the minute position labels.

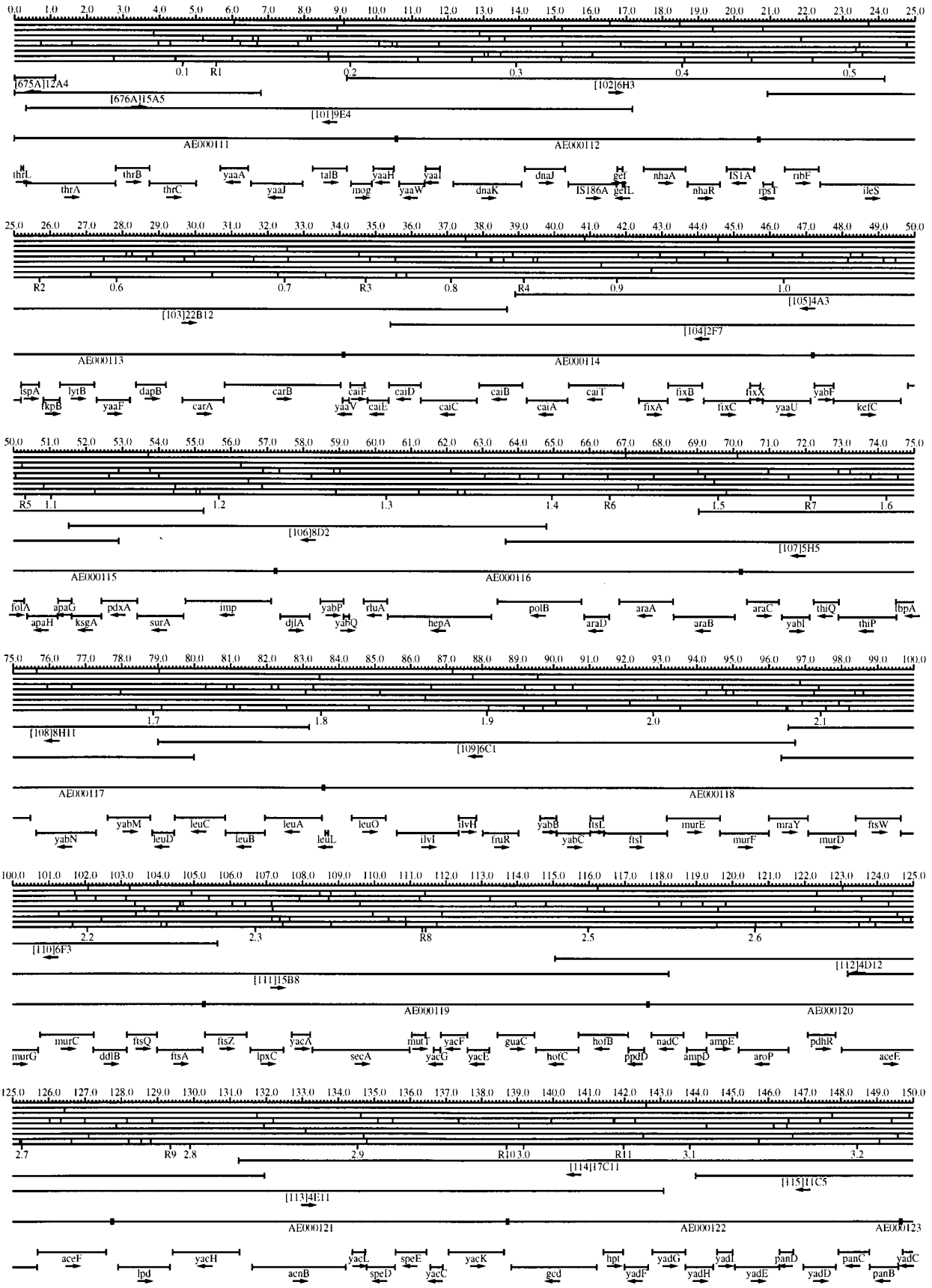
Genes and ORFs

A detailed description of the annotation of genes and functionally uncharacterized ORFs in EcoGene10 is presented in a separate publication (16). The entire genome sequence annotation of protein coding regions has been reviewed, and revisions have been made to approximately 15% of them. The most frequent revisions were the choice of an alternative translation start site, although sequences encoding small proteins were added and deleted from the set of coding regions as well. These two areas were acknowledged as difficult aspects of protein coding region annotation (5), and the EcoGene annotation should be thought of as one view of the *E. coli* K-12 genome. Producing a set of predicted protein sequences as accurately as possible was the goal of the reannotation effort, but experimental verification is the only way to establish the coding regions definitively. Published experimental data was used to establish gene intervals as much as possible. Anyone wishing to communicate additional prepublication information directly is encouraged to do so, especially if he or she has no objection to the information being made publicly available in the EcoGene and SWISS-PROT databases as a personal communication. The *E. coli* genome sequence annotation refinement has been a close collaboration with the curator of the SWISS-PROT database, Amos Bairoch.

Partial or frameshifted ORFs and genes are marked in Fig. 1 (see the figure legend). In most cases, but not all, the presence of a frameshift or deletion is based on sequence analysis alone and thus should be considered a prediction. It is not known if any particular putative frameshift or deletion is the result of a DNA sequencing error, a cloning artifact, an adaptation to the laboratory environment, natural evolutionary pressure, or pseudogene formation. Errors introduced during the reannotation process are also possible, and everyone is encouraged to contact this author or SWISS-PROT if he or she thinks an error has been made; we will take appropriate steps to update our databases. These sequence-based frameshift predictions should assist in the experimental determination of the source of the frameshifts.

The traditional and physical maps of edition 10 of the *Esch-*

FIG. 1. EcoMap10, a DNA sequence-derived map depicting restriction sites, Kohara/Isono clones, genes, ORFs, IS elements, and REP clusters of the *E. coli* K-12 chromosome. The derivation of this map from the complete genome sequence of *E. coli* K-12 strain MG1655 is briefly described in the text. The map depicts sites for eight restriction enzymes (top line to bottom line: *Bam*HI, *Hin*dIII, *Eco*RI, *Eco*RV, *Bgl*I, *Kpn*I, *Pst*I, and *Pvu*II). Above the restriction map are position coordinates in kilobases; immediately below the map are minute coordinates (in 0.1-min increments). Also immediately below the map are the designations R1 to R355 referring to the 355 serially numbered REP clusters, placed at the genomic position of the base pair at their left ends. Some minute designations were omitted as they overlapped with the REP serial numbers, but the tick marks for these unlabeled 0.1-minute positions are present, and their values can be easily determined from the flanking minute values. The first set of spanning lines below the map represent the genomic positions and clone insert orientations of the Kohara miniset clones. Those Kohara miniset clone W3110 chromosomal DNA inserts that have been completely sequenced are additionally labeled with their GenBank/EMBL/DDBJ accession numbers, D90699 to D90892 (1, 9, 13, 22). The second set of spanning lines, labeled with database accession numbers AE000111 to AE000510, represent the locations of the GenBank/EMBL/DDBJ complete-genome MG1655 sequence entries of Blattner et al. (5). The third set of spanning lines depict the positions and orientations of the genes, ORFs, and IS elements that constitute EcoGene10. An asterisk following a gene or ORF name indicates that a frameshift or in-frame stop codon that prevents the EcoGene10 representation of the coding region from being translated is present in the genome sequence. A prime indicates a partial EcoGene entry, i.e., a deletion or IS element insertion is predicted to have disrupted the ancestral complete gene, ORF, or IS element. This figure was created by using the PrintMap Postscript drawing program, which implements the Plasmid Description Language developed by Craig Werner (18).



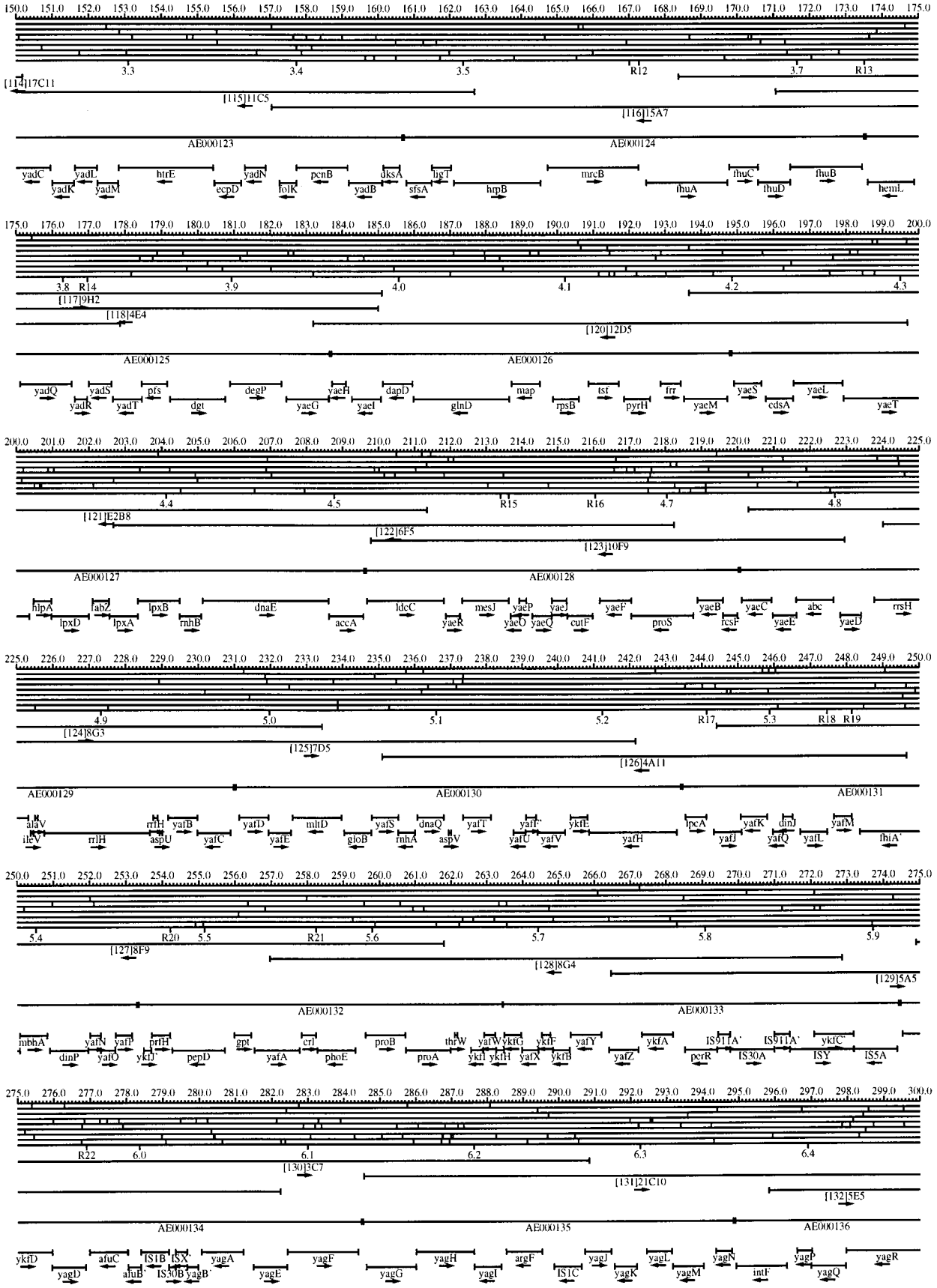


FIG. 1—*Continued.*

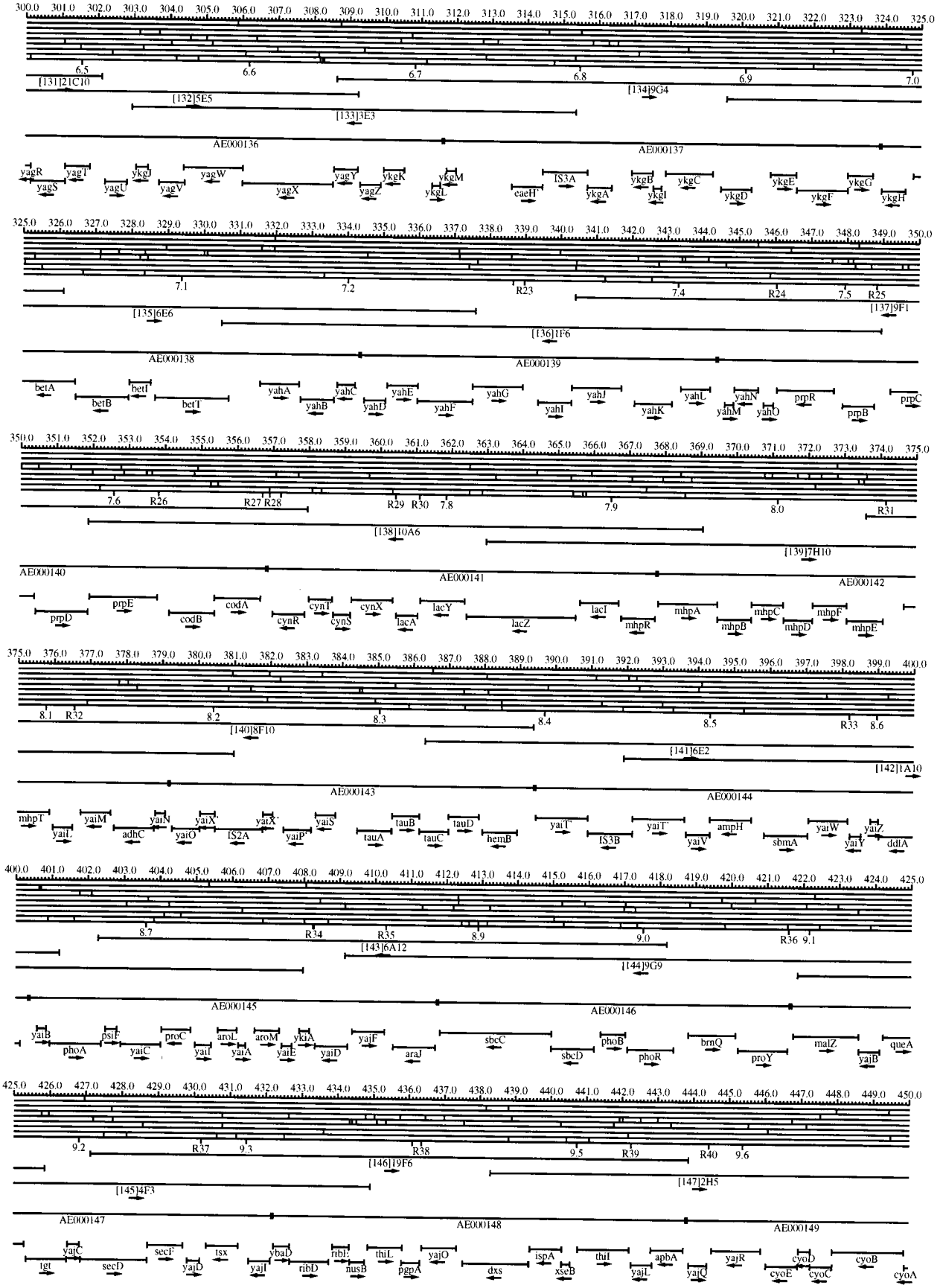


FIG. 1—*Continued.*

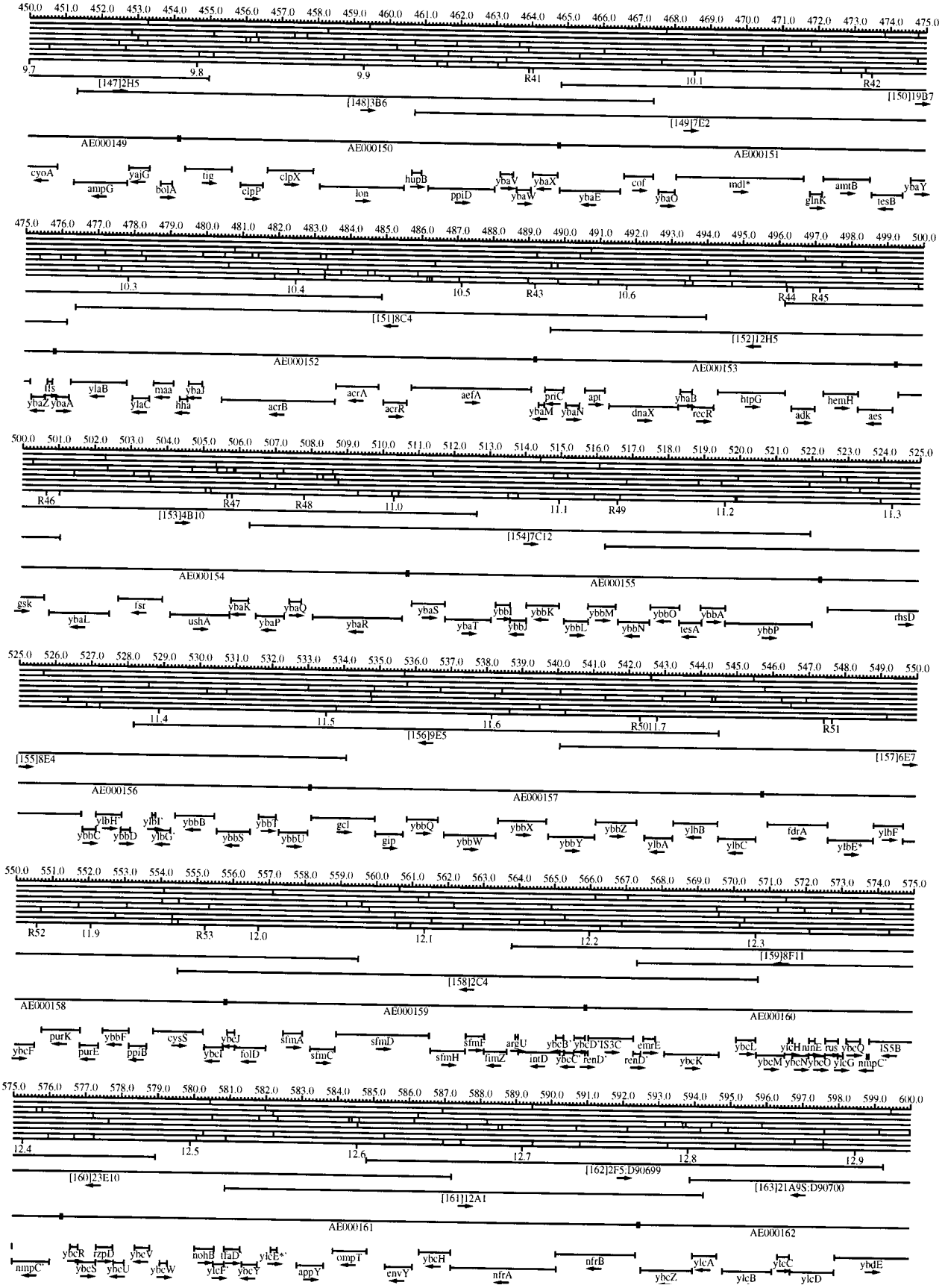


FIG. 1—*Continued.*

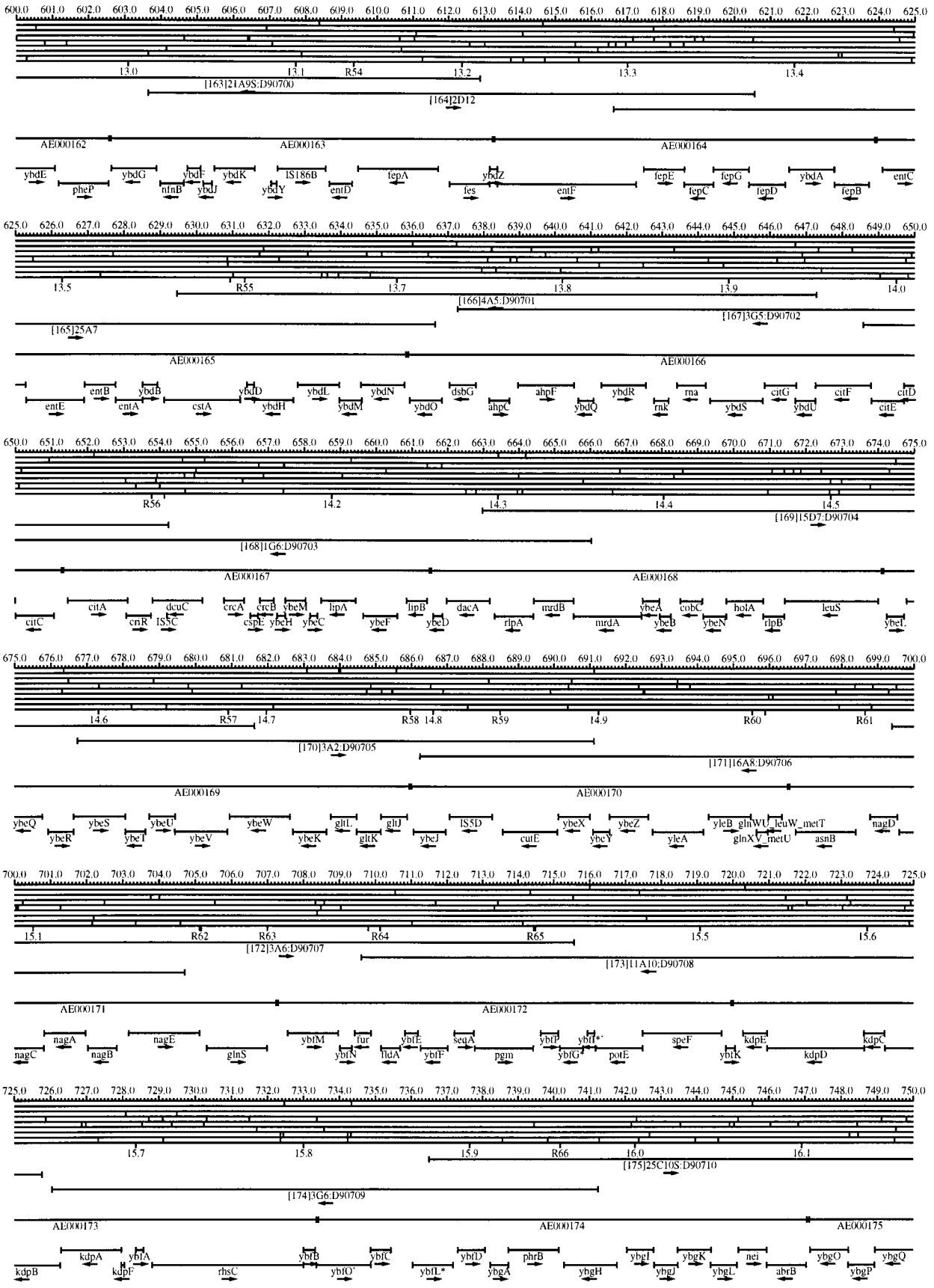


FIG. 1—*Continued.*

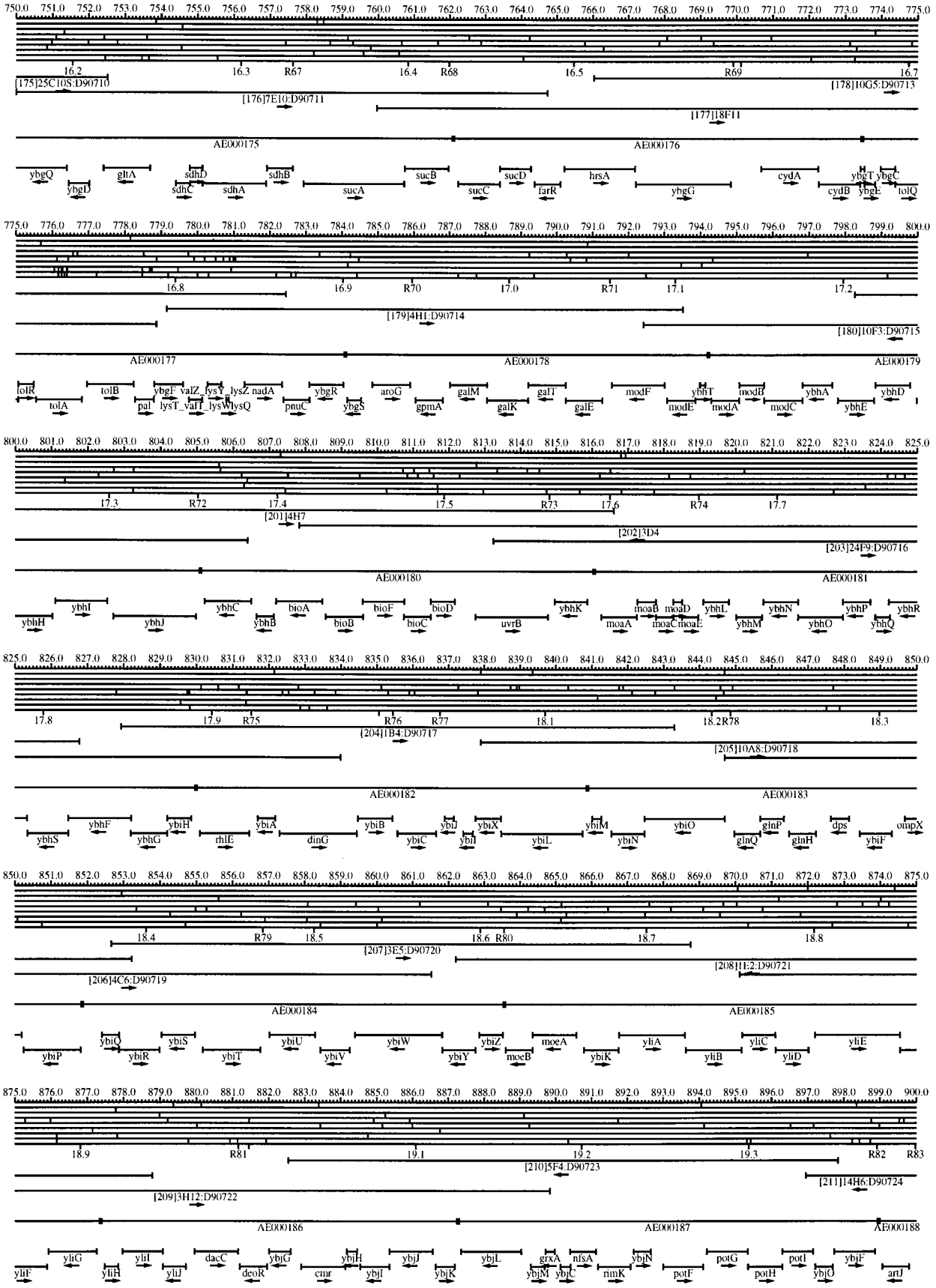


FIG. 1—*Continued.*

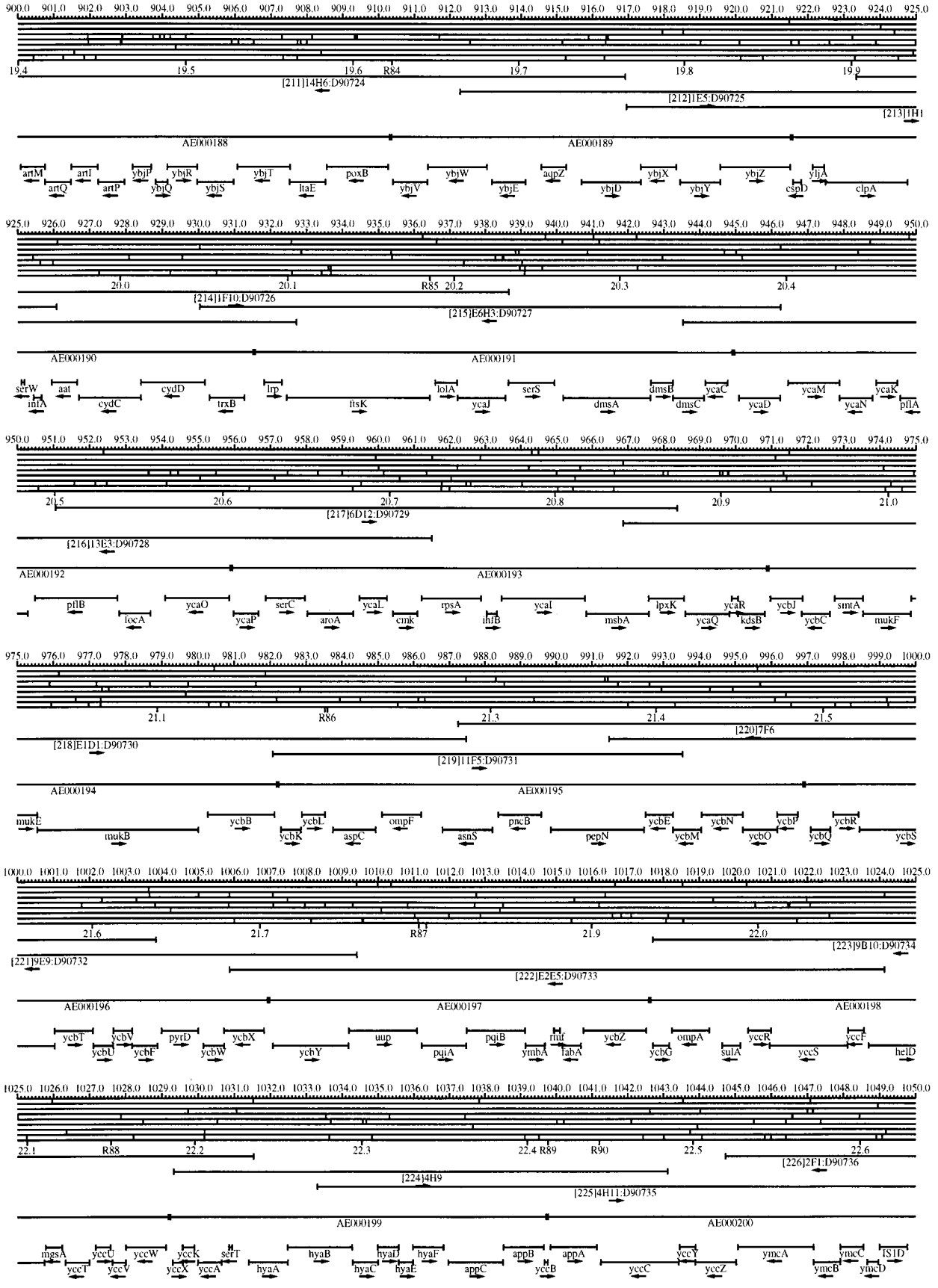


FIG. 1—*Continued.*

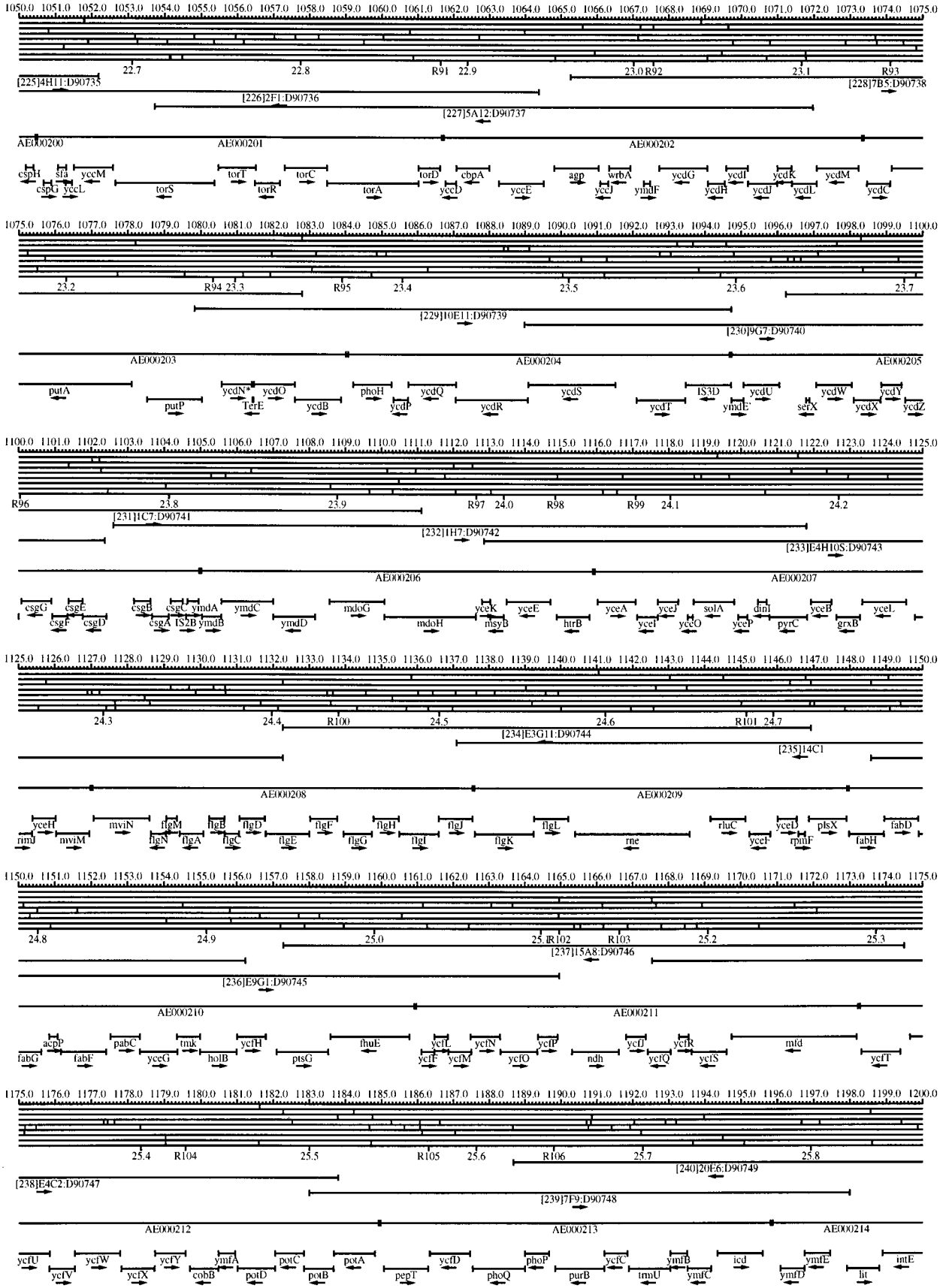


FIG. 1—*Continued.*

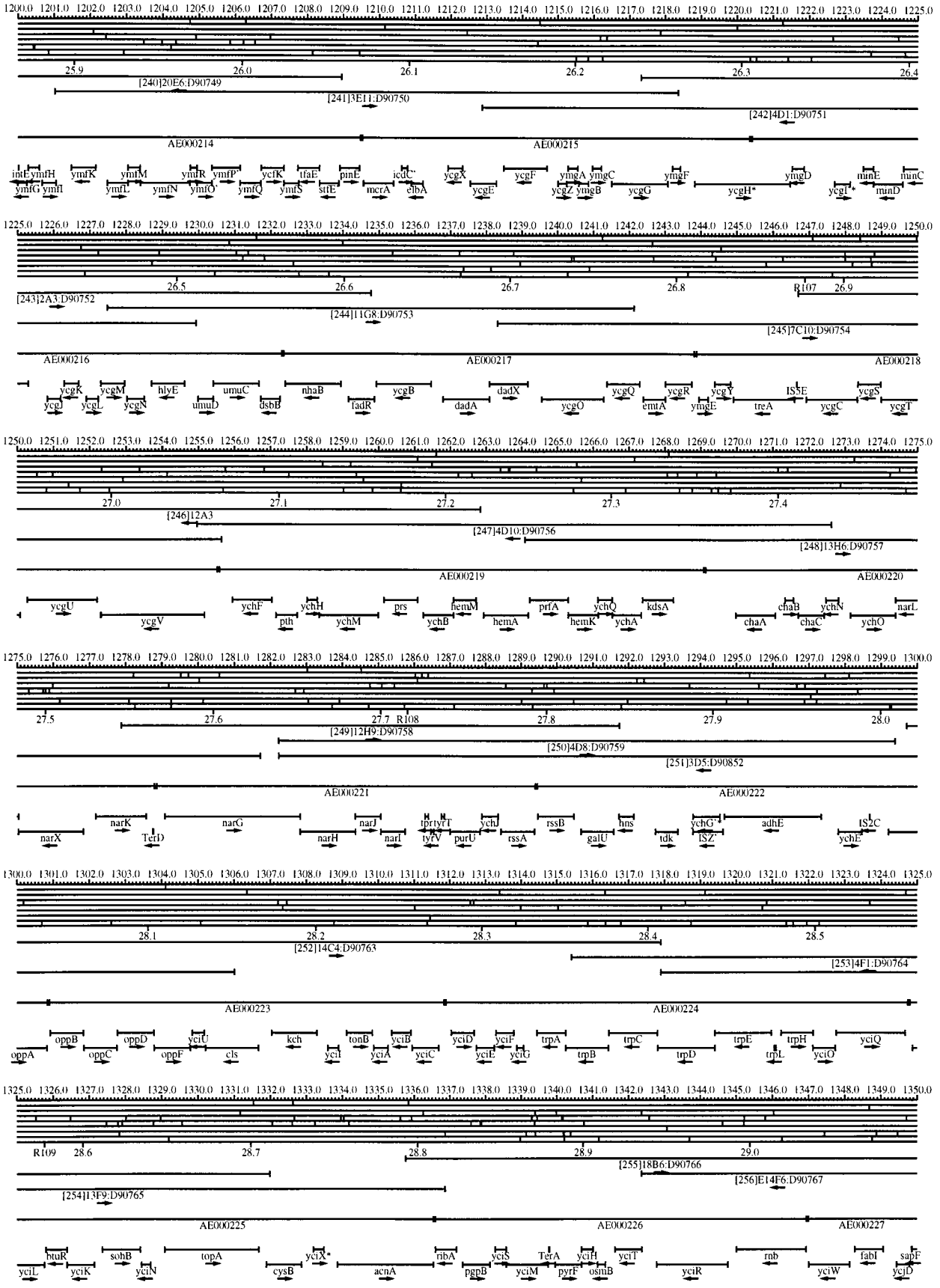


FIG. 1—*Continued.*

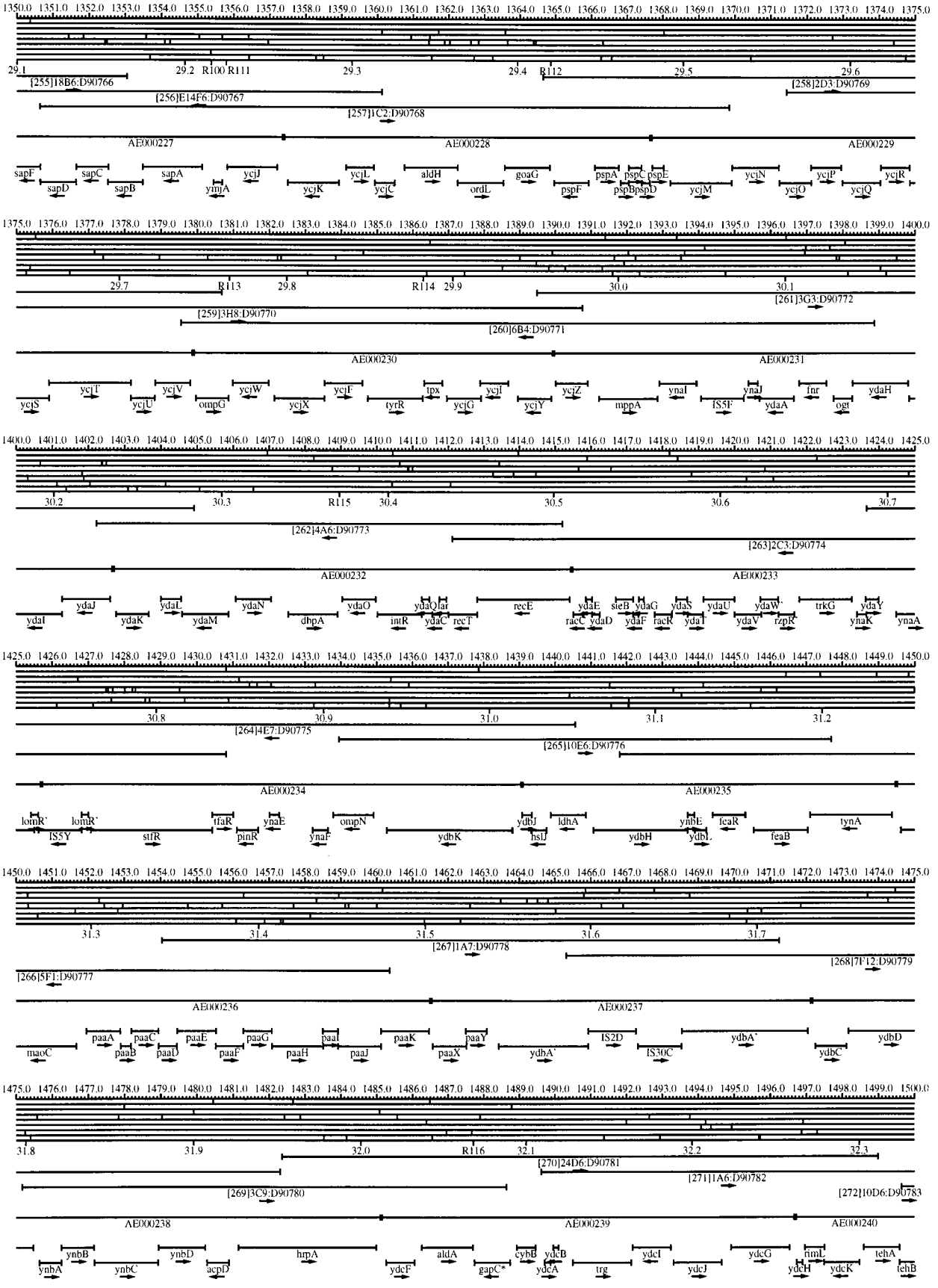


FIG. 1—*Continued.*

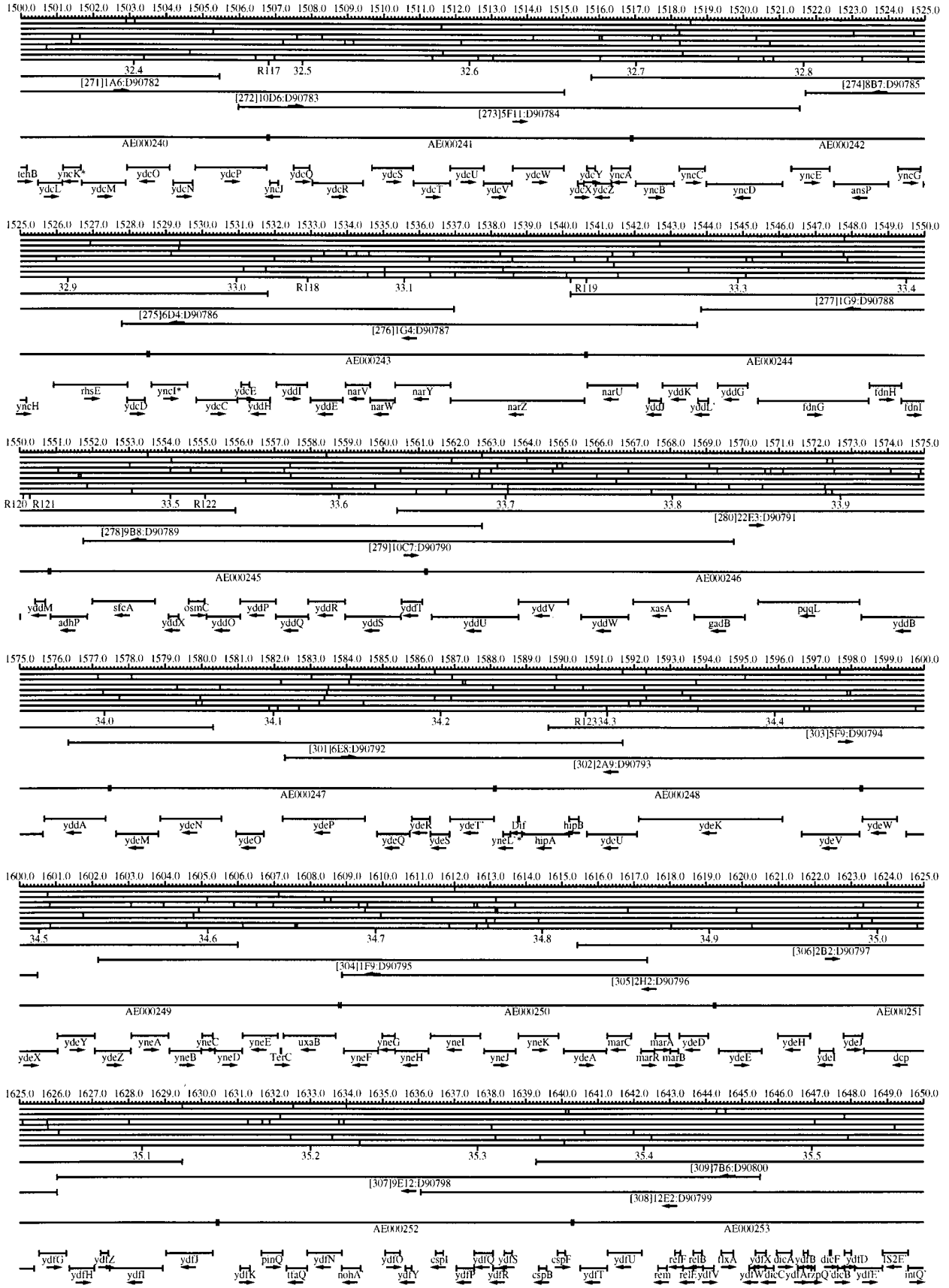


FIG. 1—*Continued.*

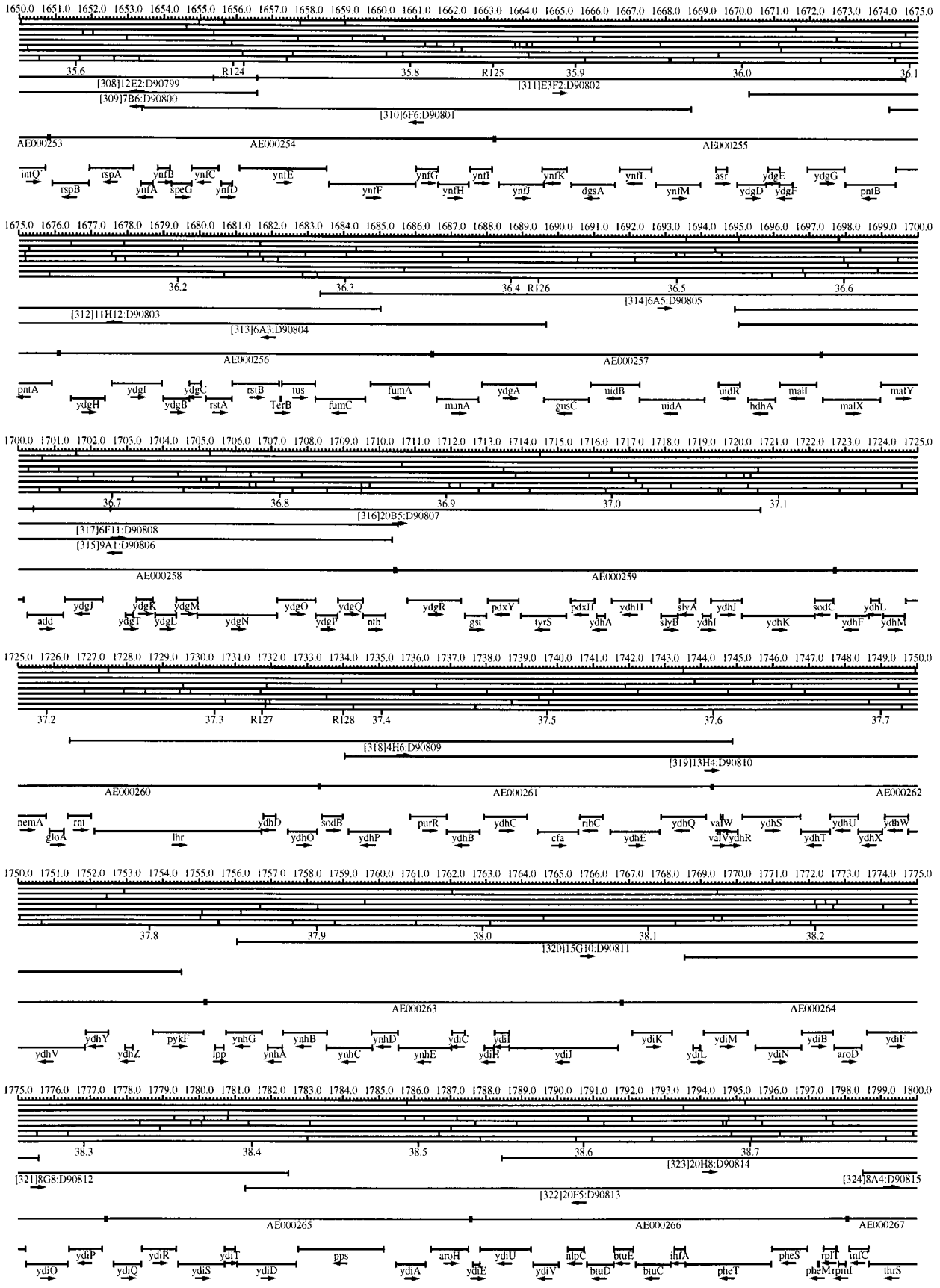


FIG. 1—*Continued.*

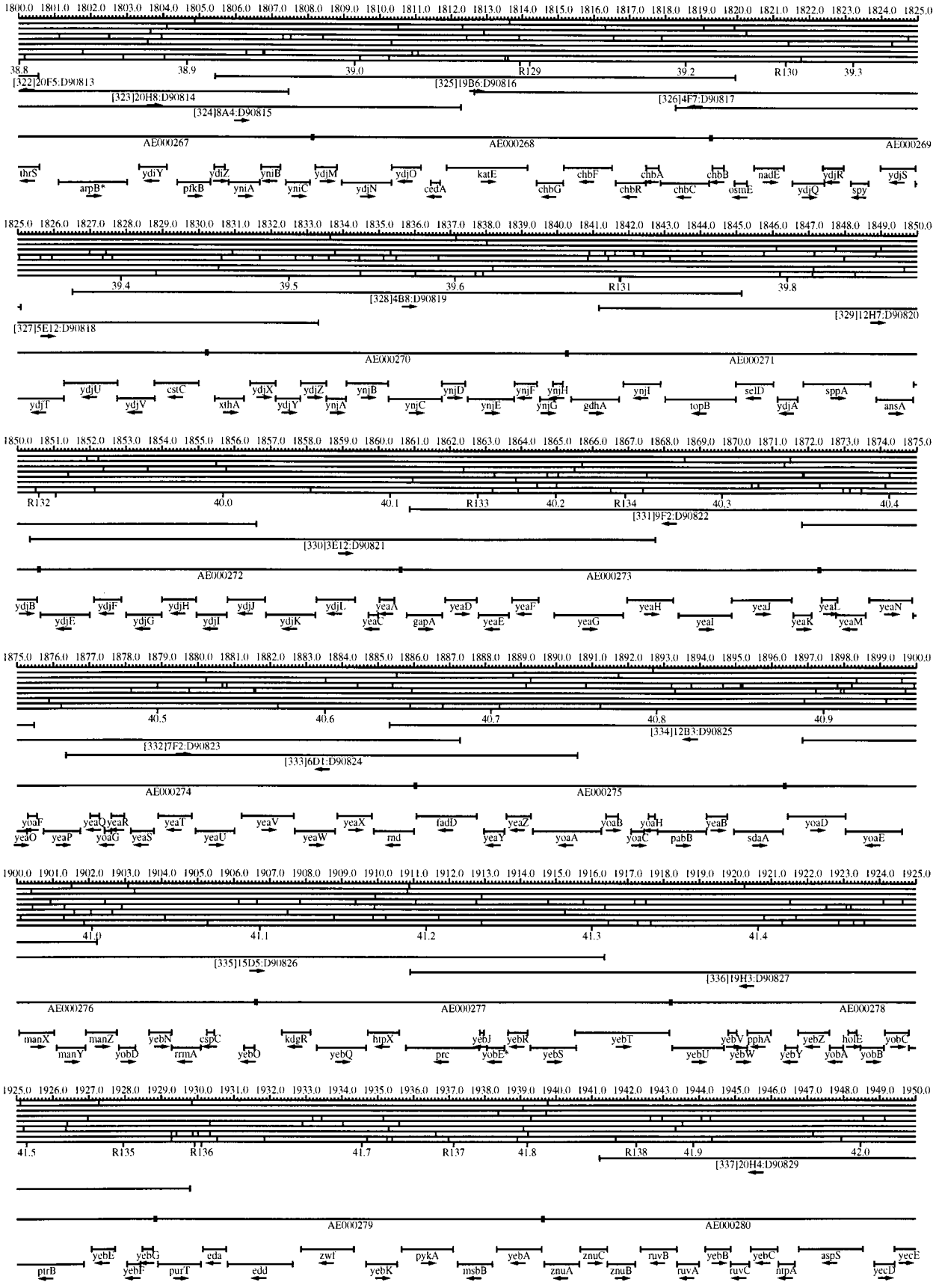


FIG. 1—*Continued.*

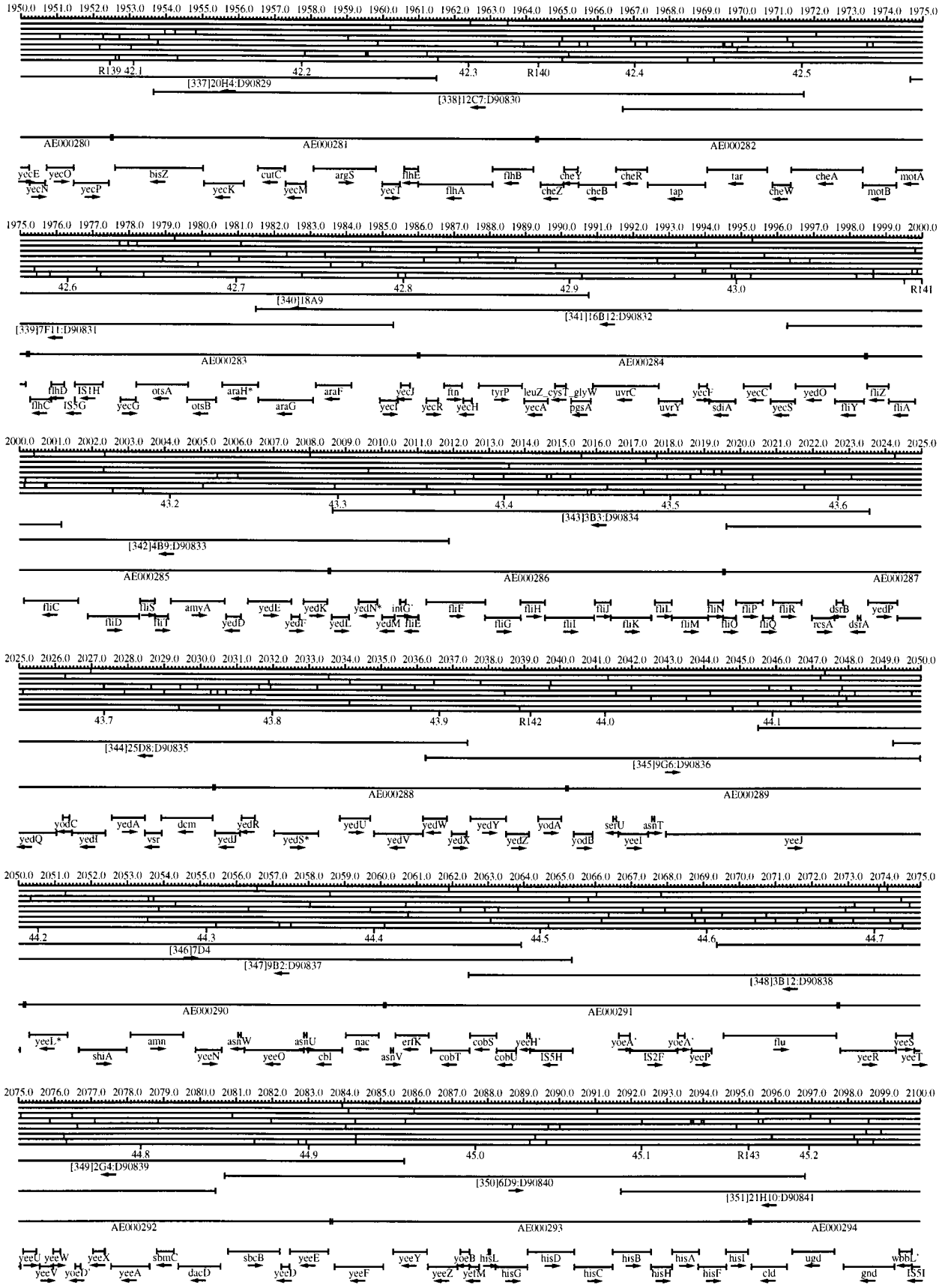


FIG. 1—*Continued.*

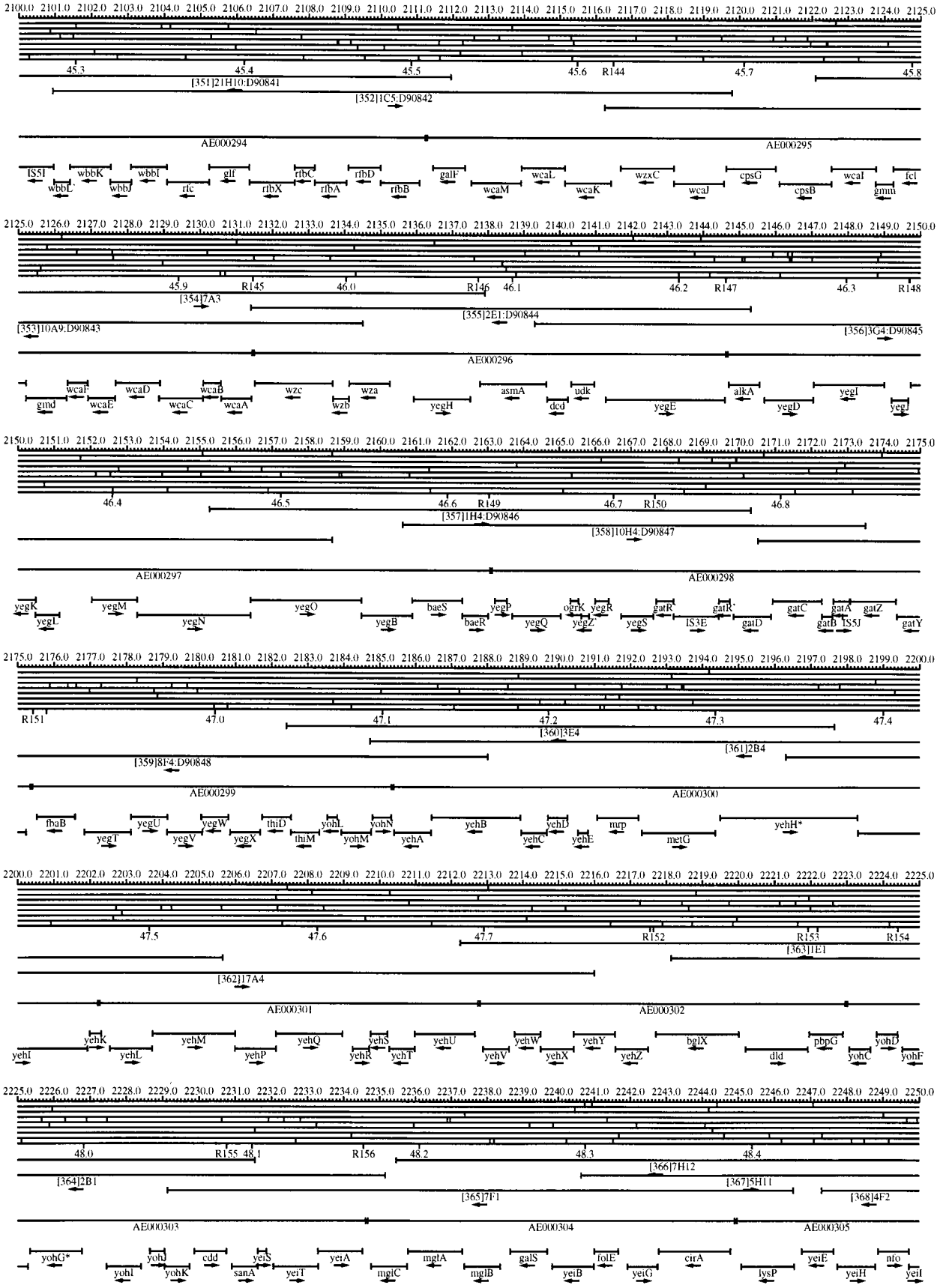


FIG. 1—*Continued.*

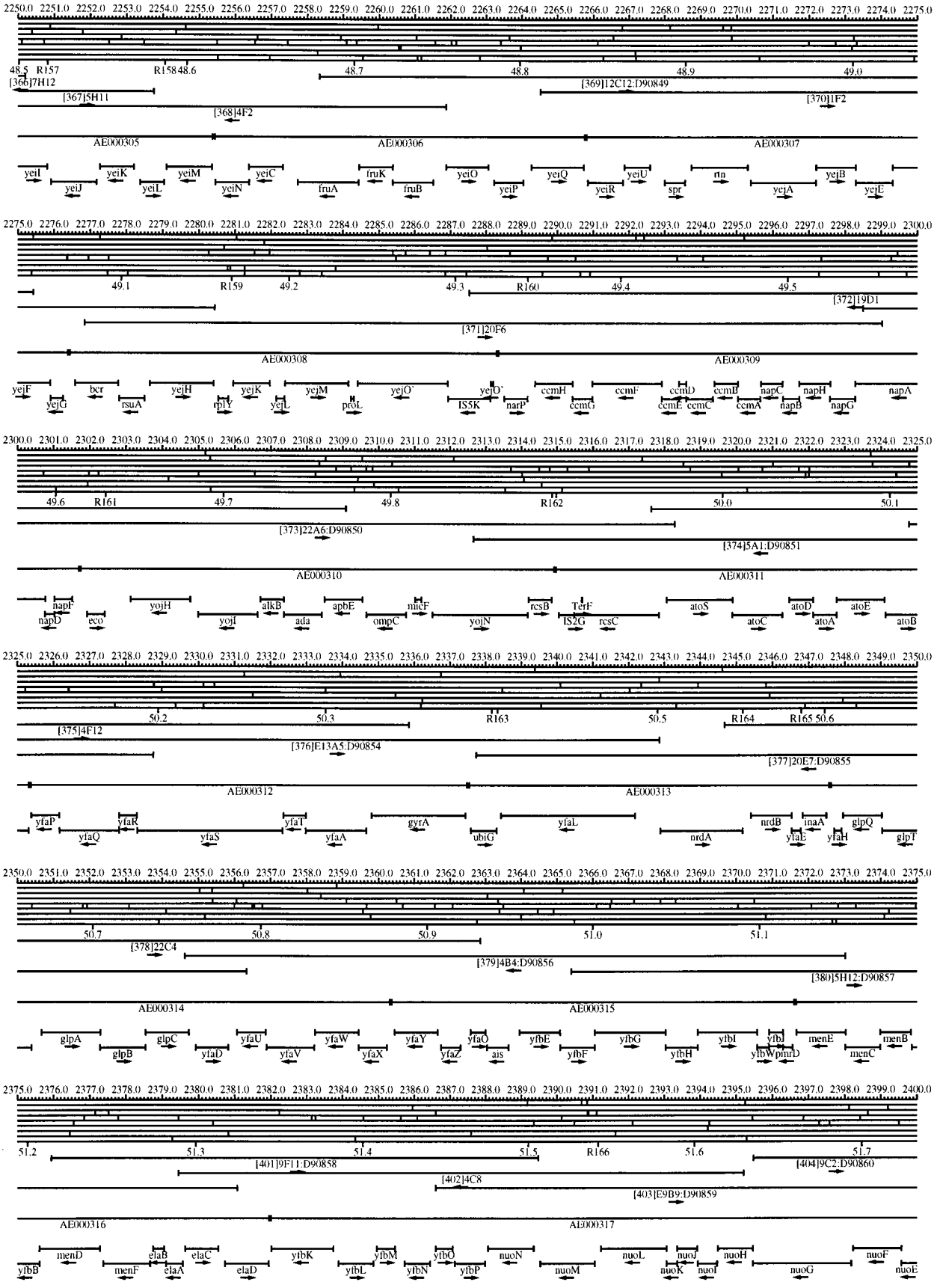


FIG. 1—*Continued.*

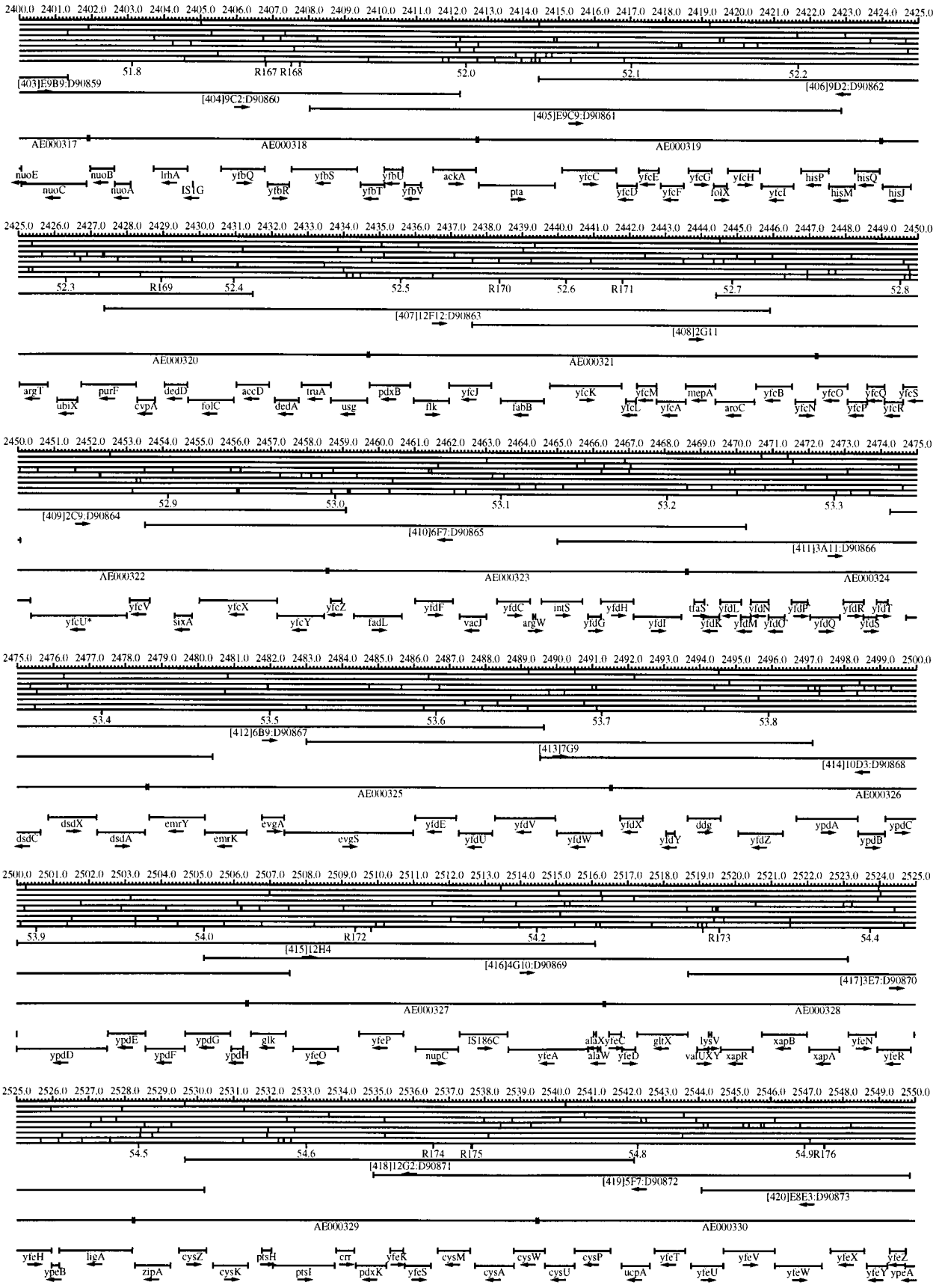


FIG. 1—*Continued.*

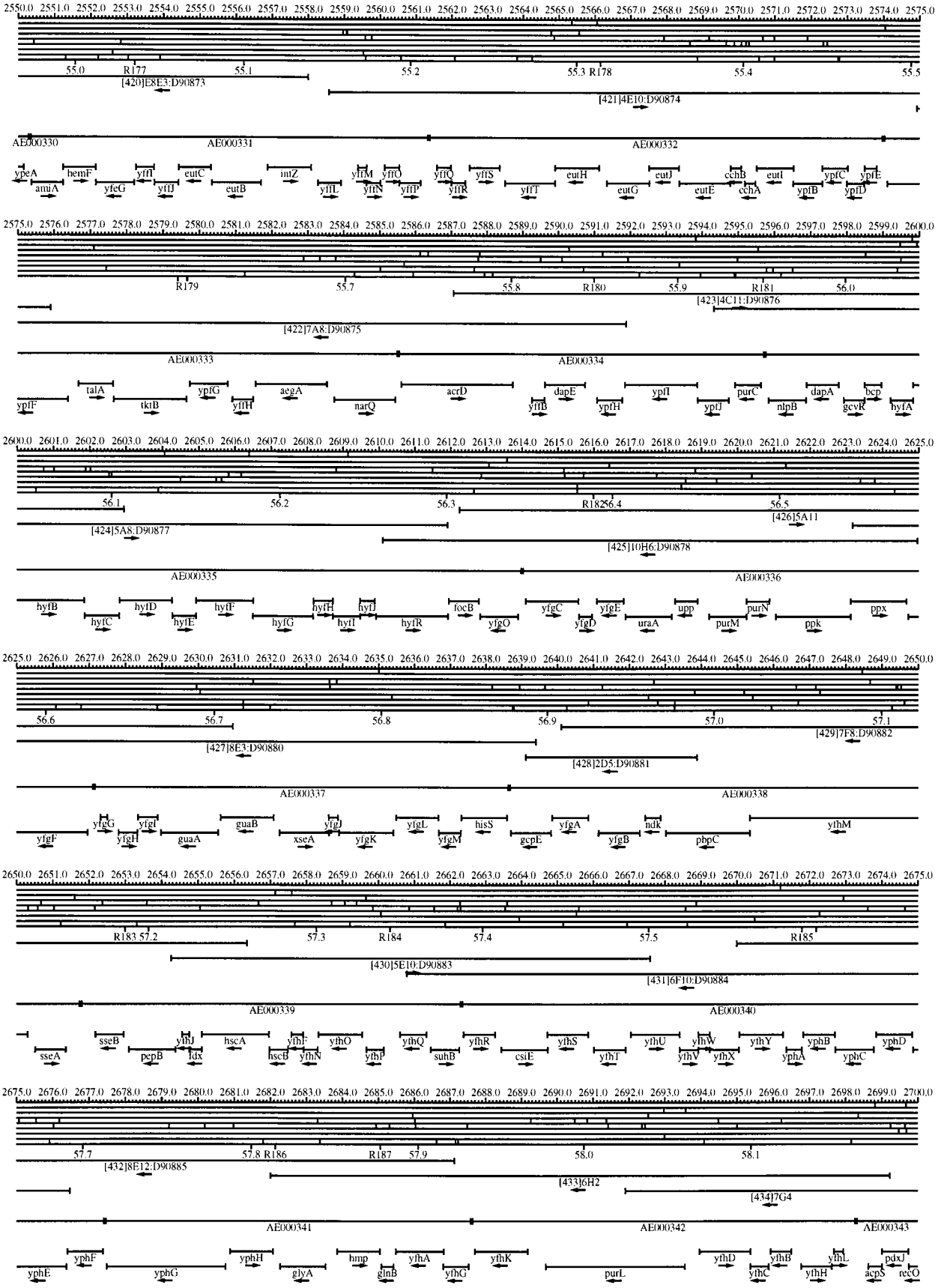


FIG. 1—*Continued.*

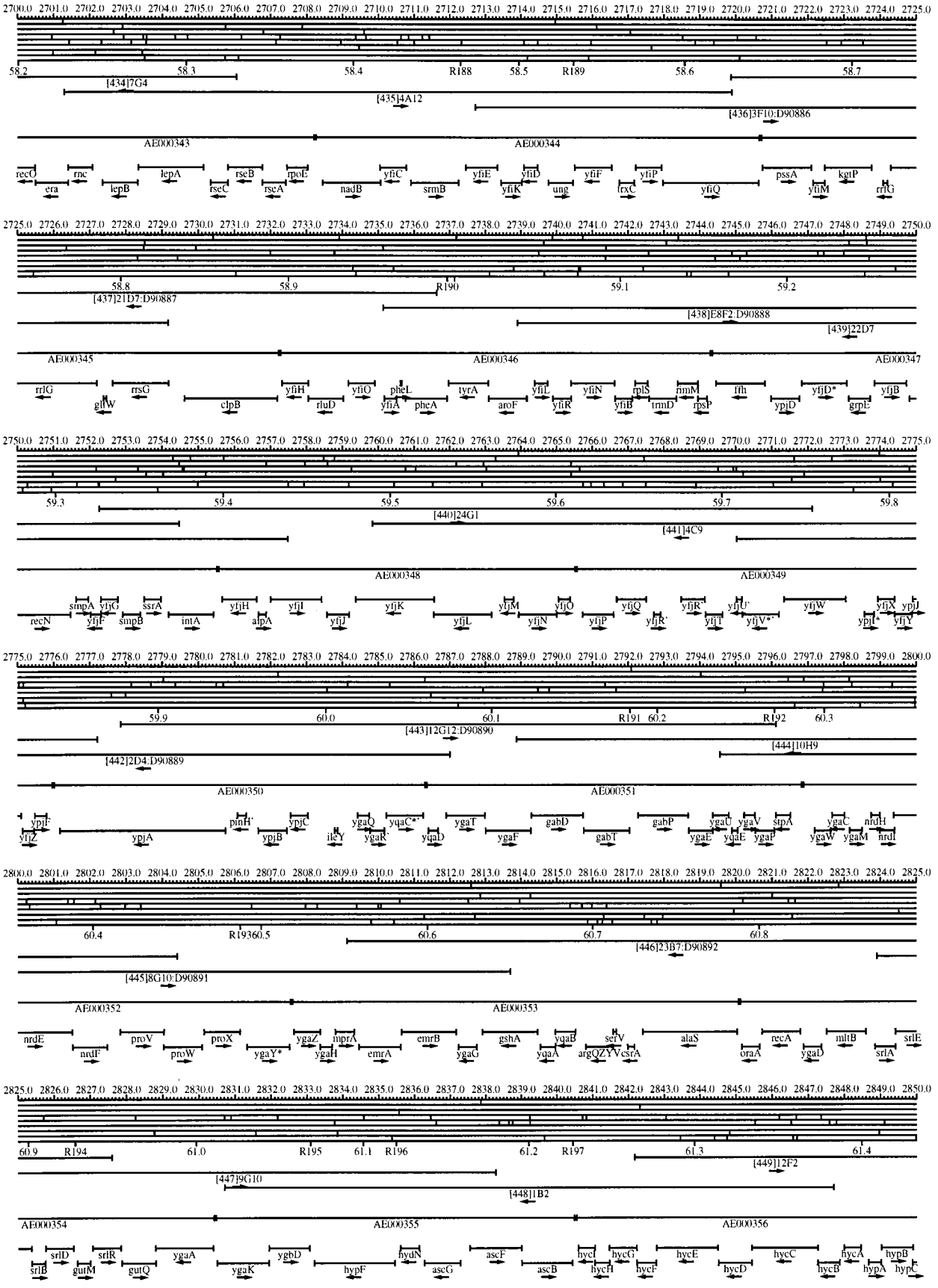


FIG. 1—*Continued.*

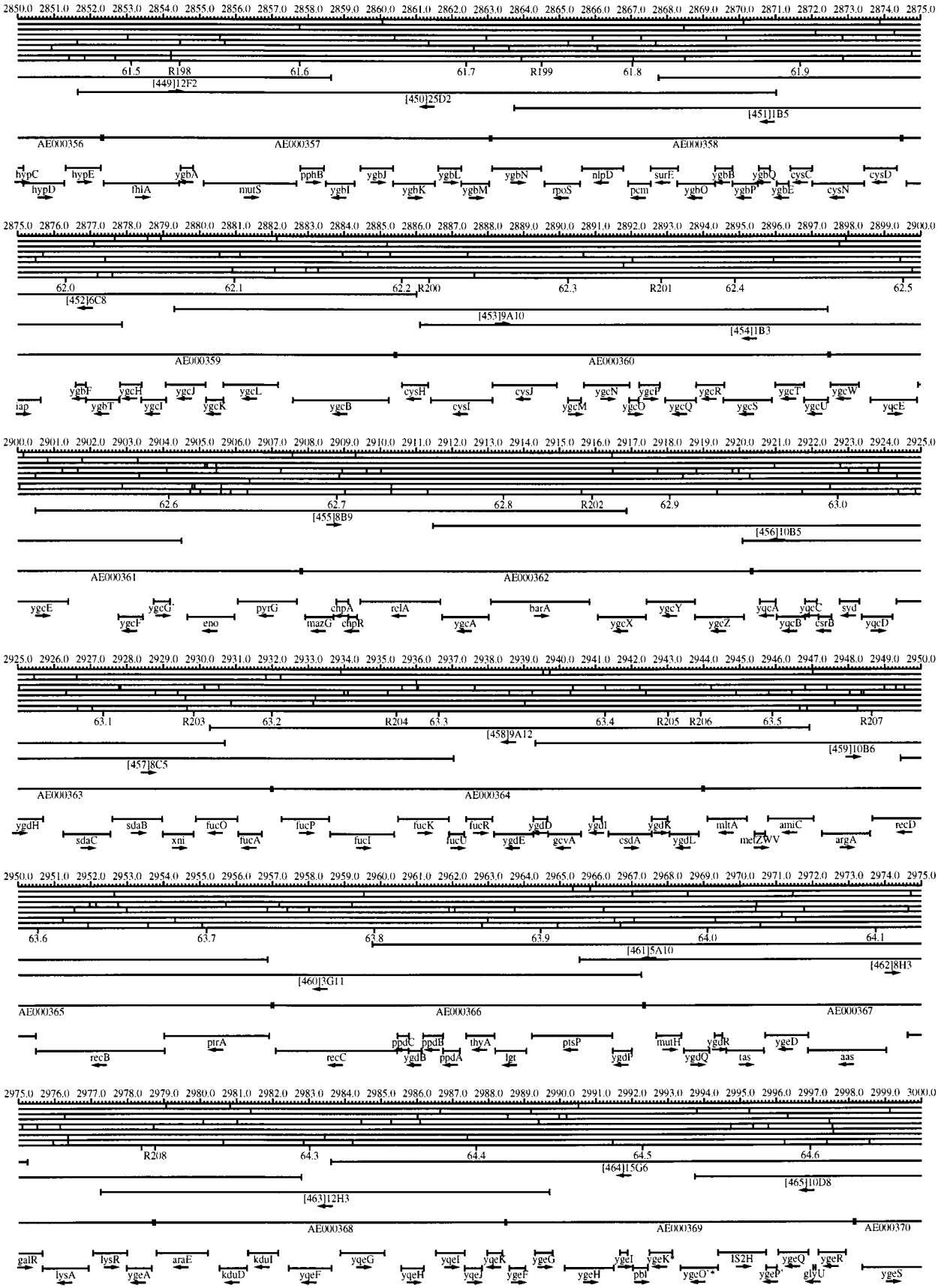


FIG. 1—*Continued.*

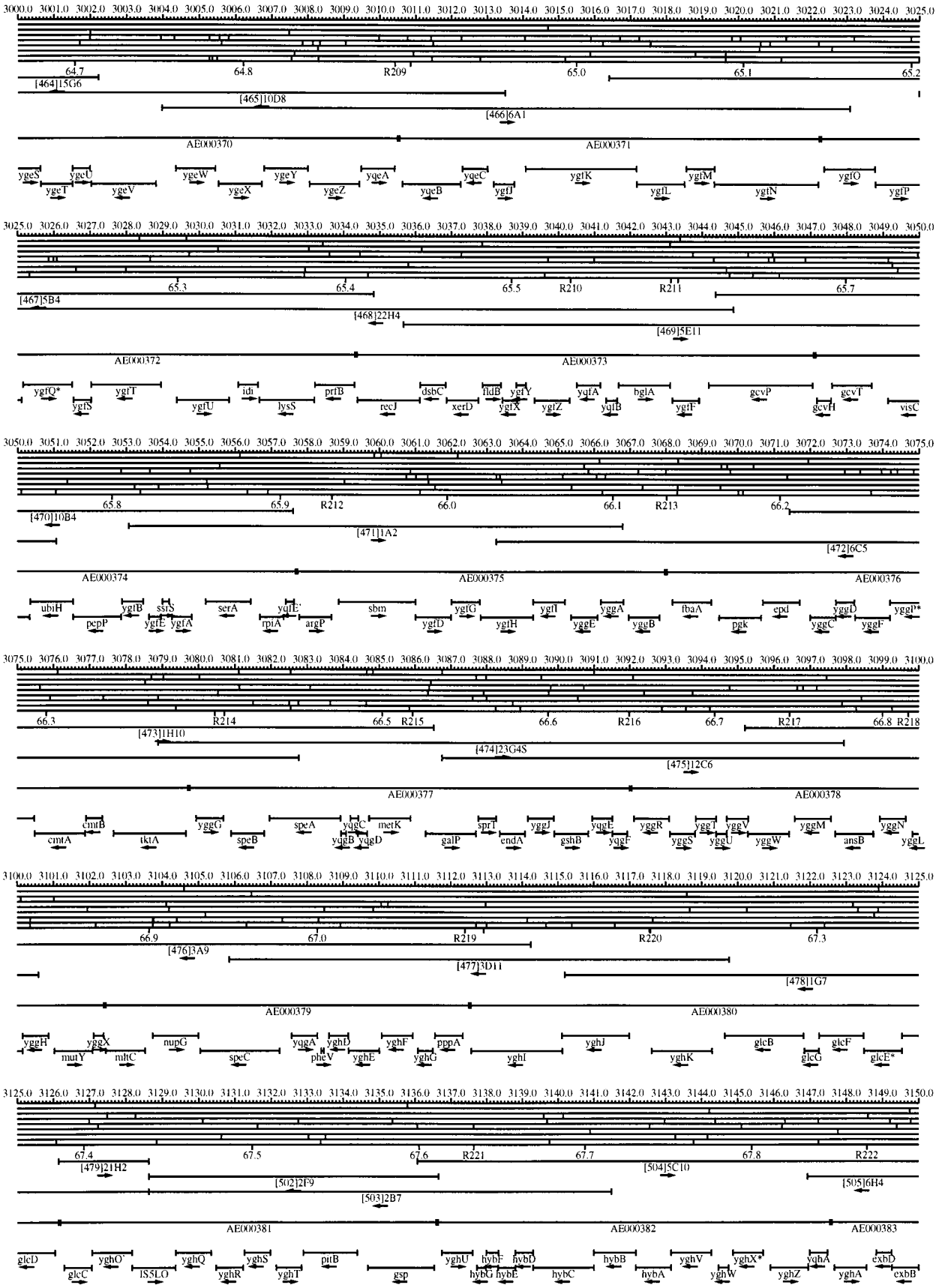


FIG. 1—*Continued.*

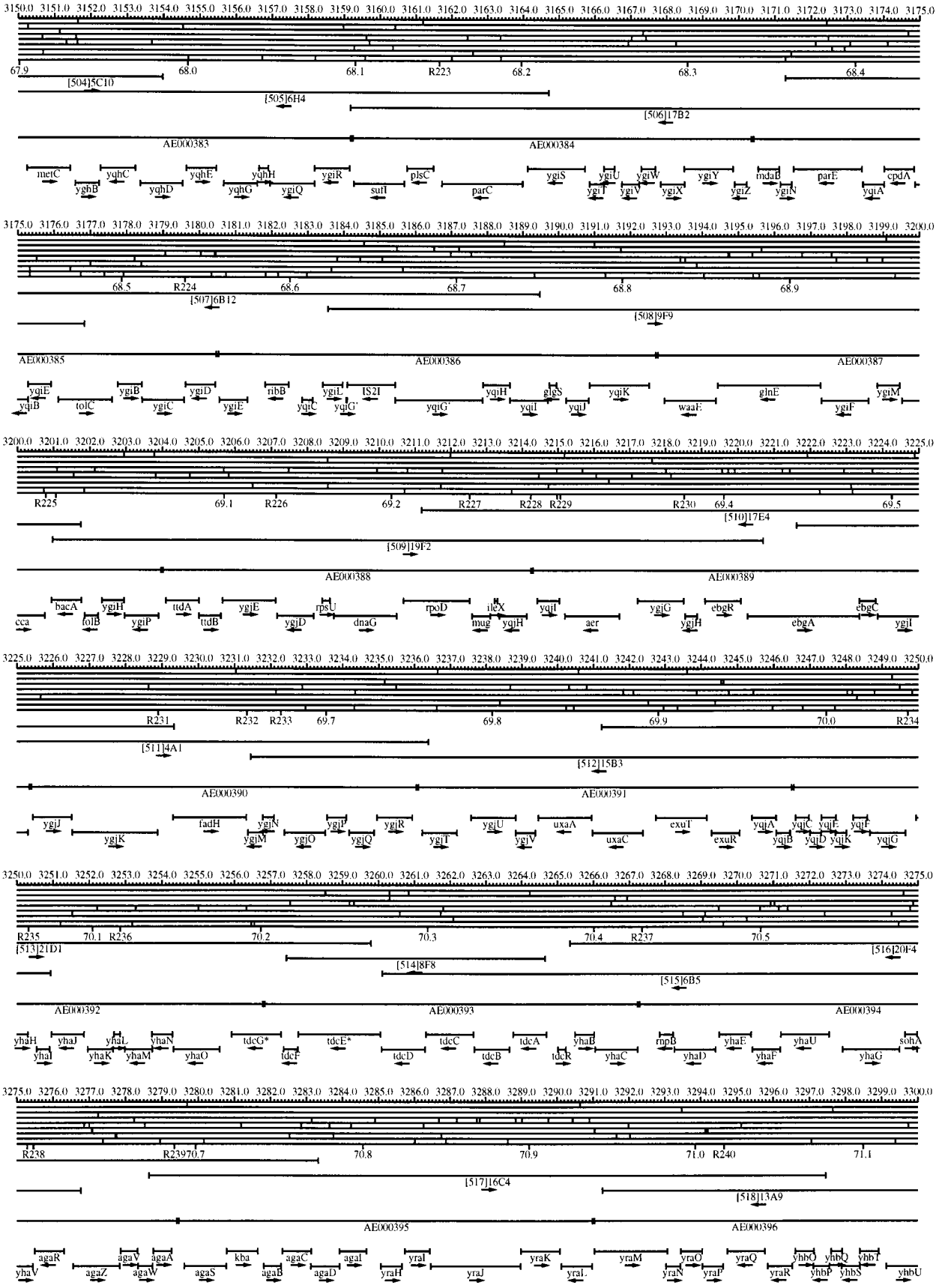


FIG. 1—*Continued.*

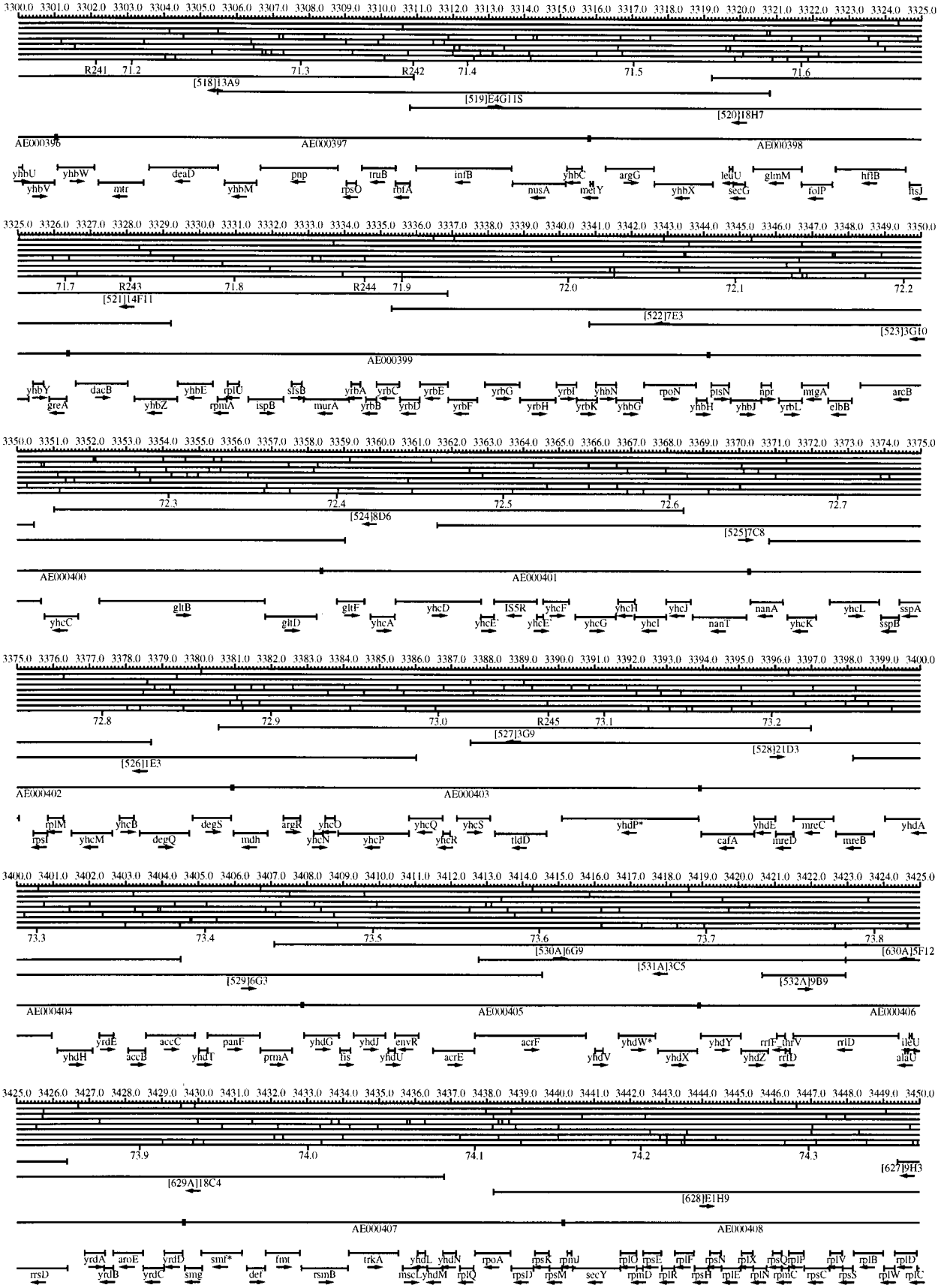


FIG. 1—*Continued.*

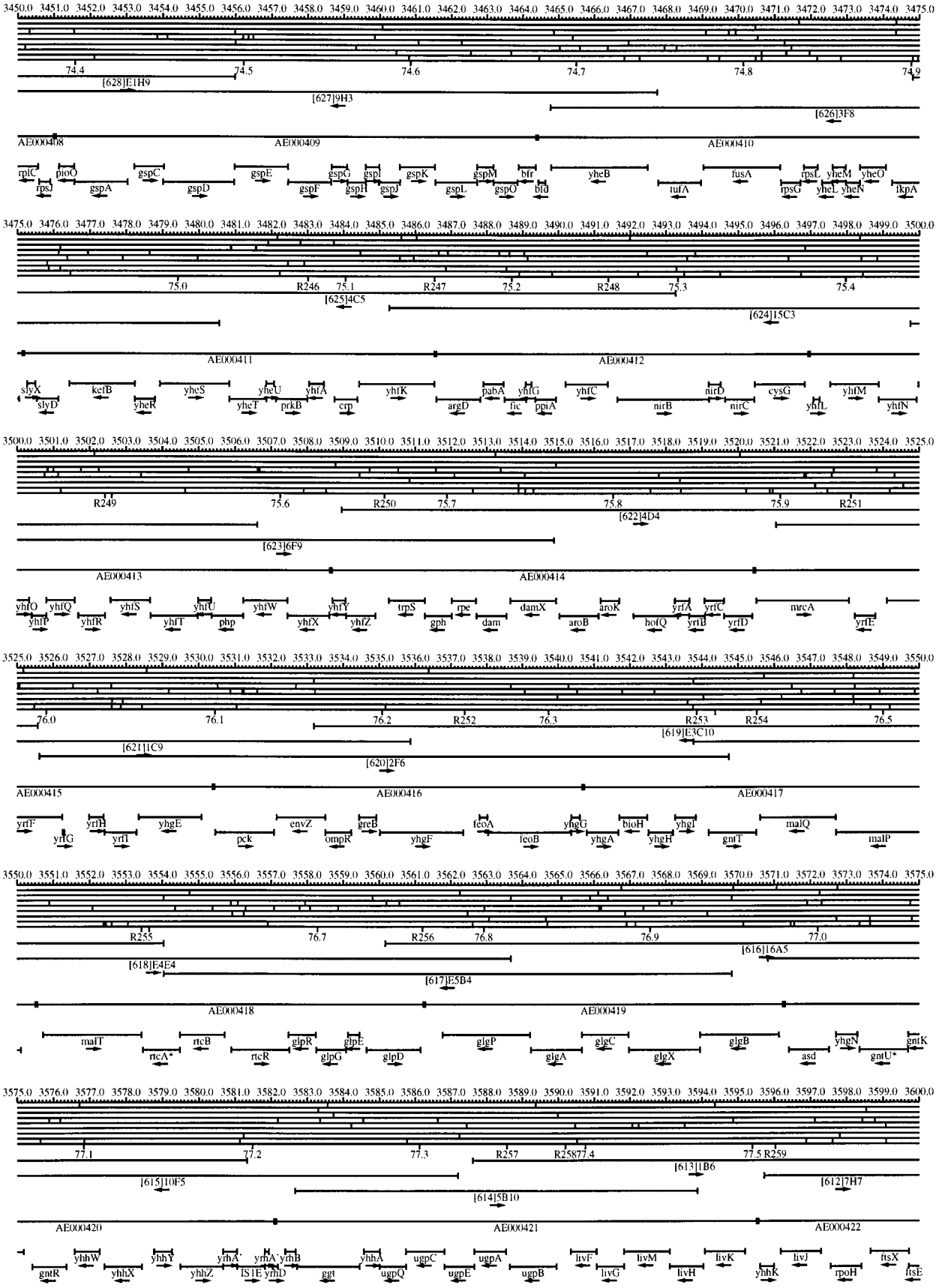


FIG. 1—*Continued.*

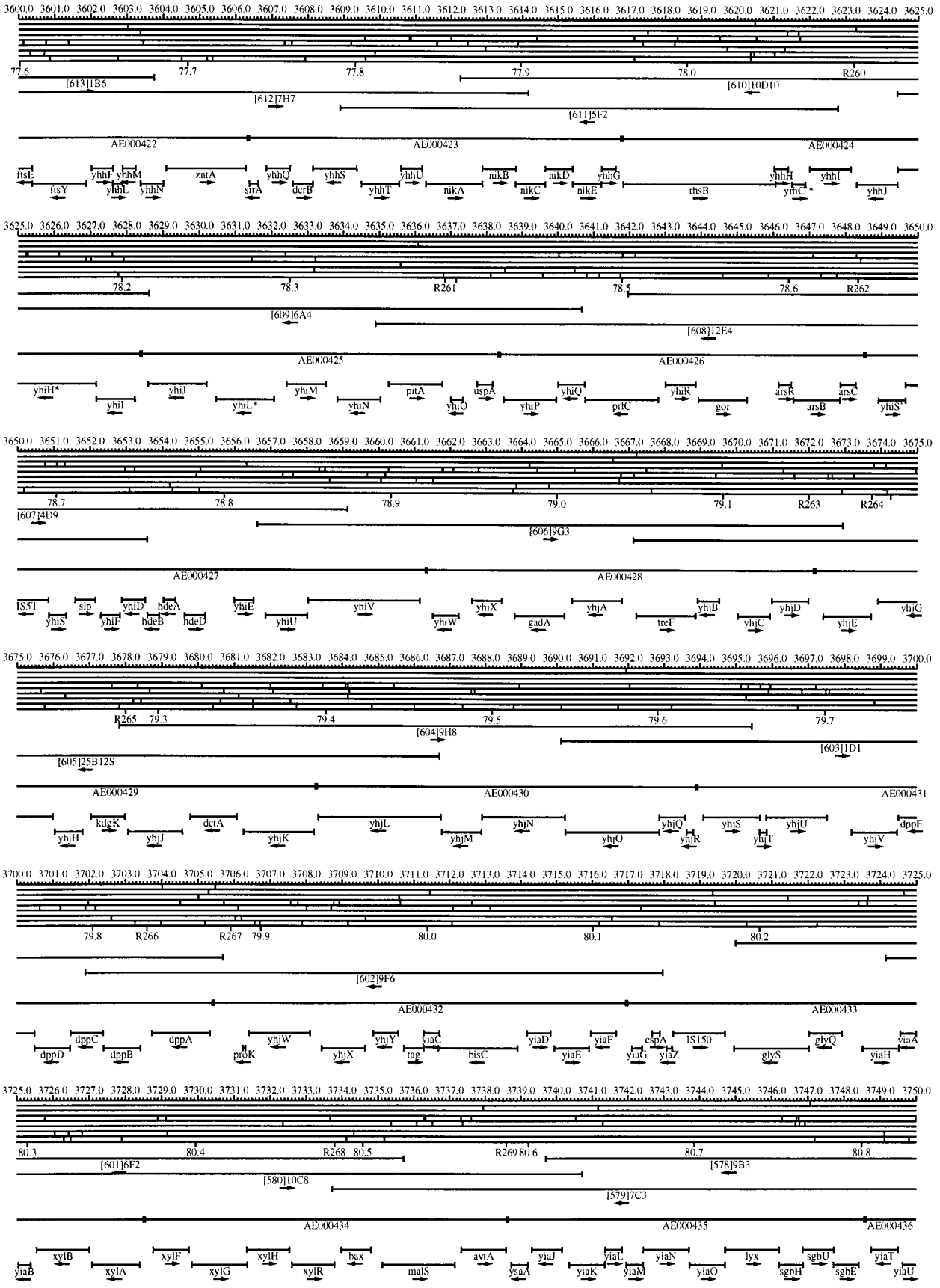


FIG. 1—*Continued.*

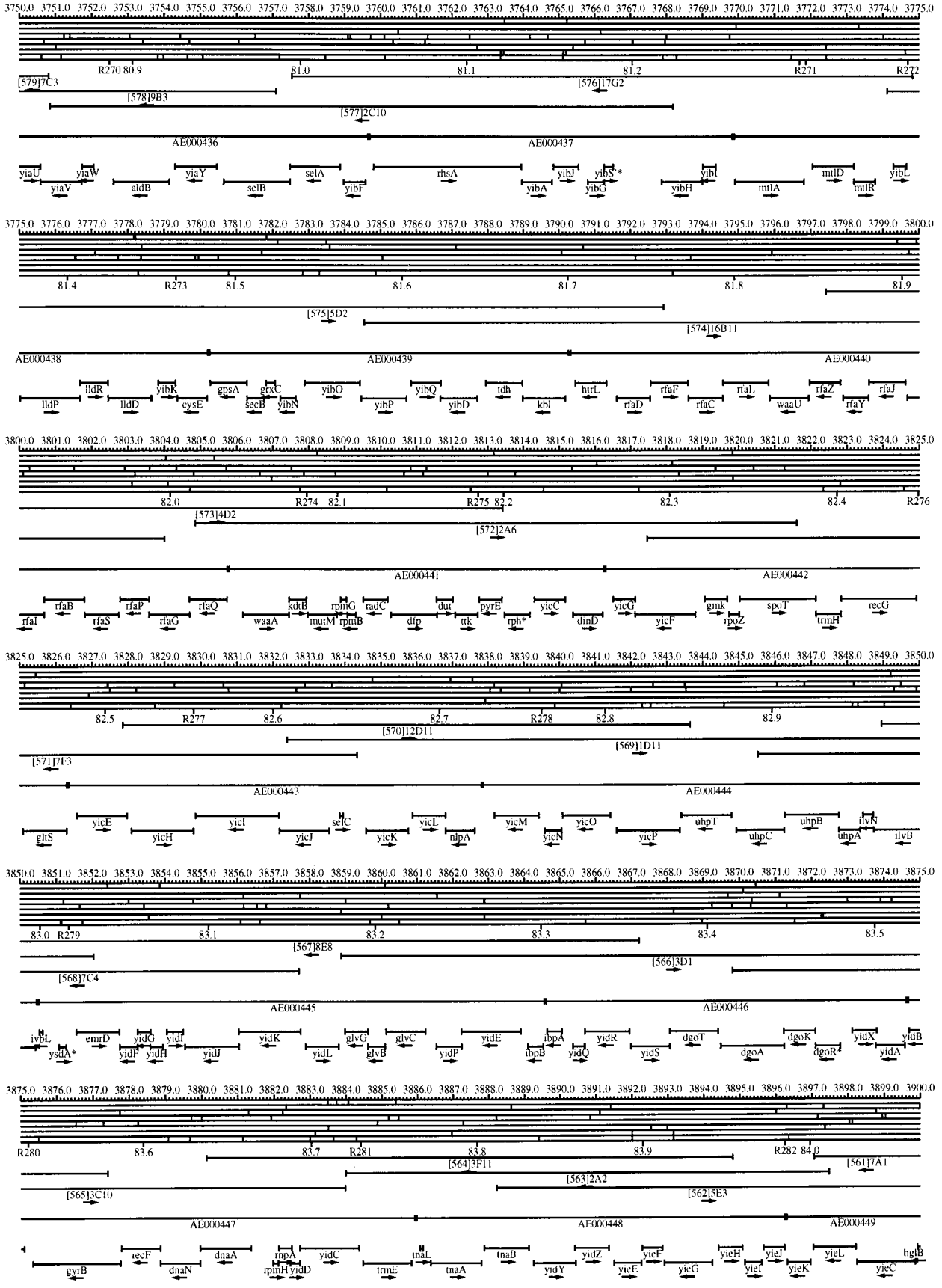


FIG. 1—*Continued.*

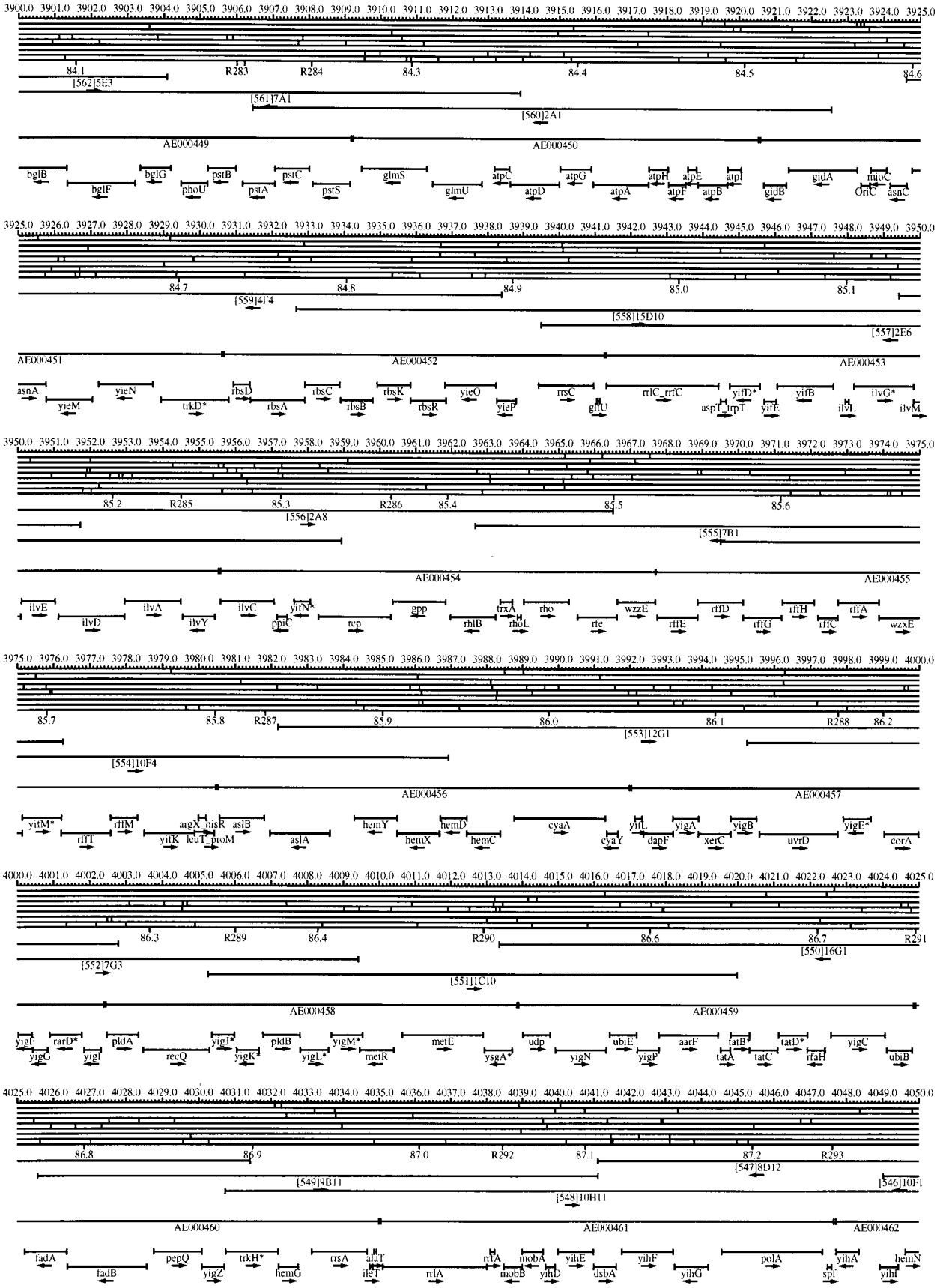


FIG. 1—*Continued.*

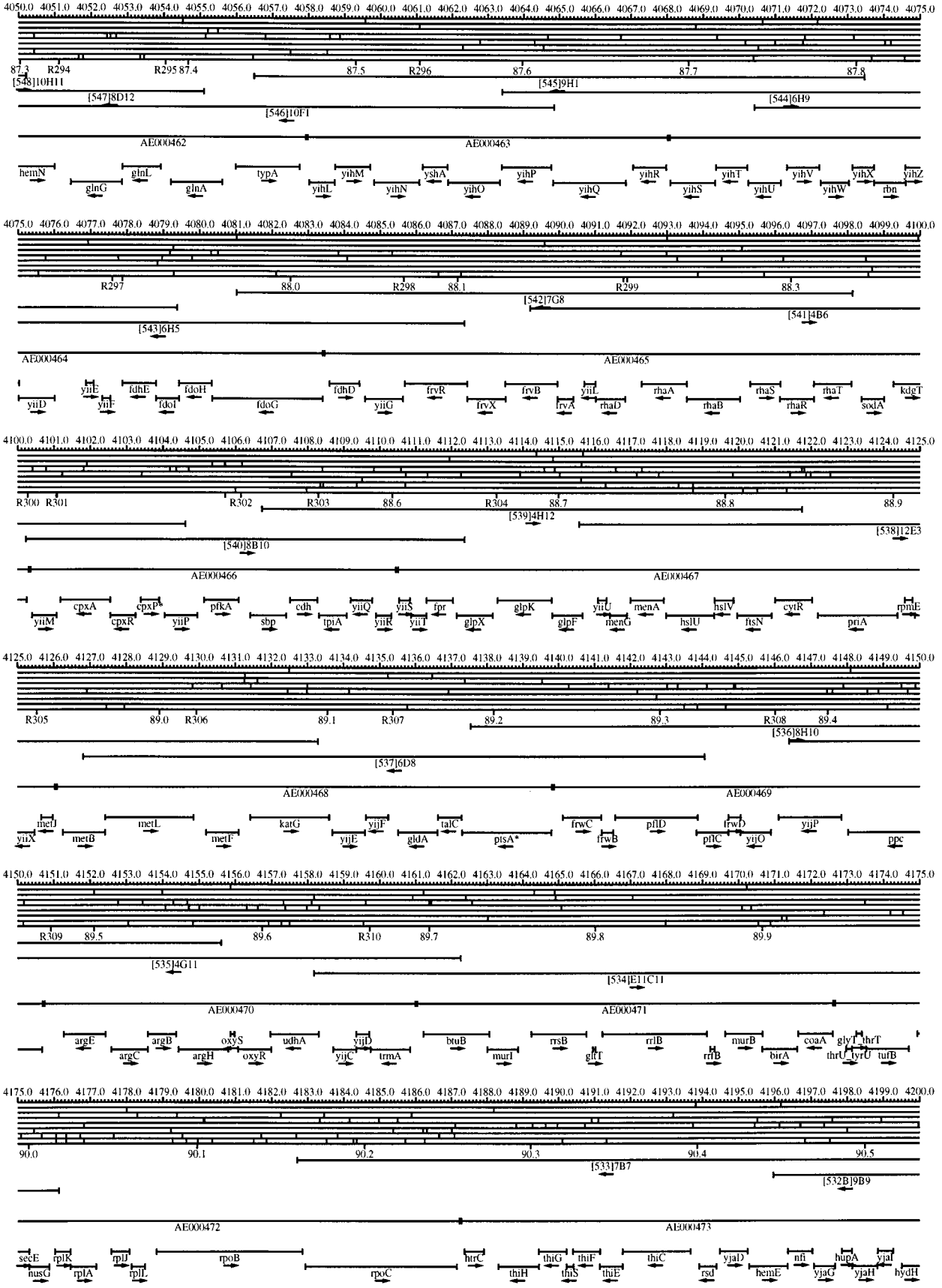


FIG. 1—*Continued.*

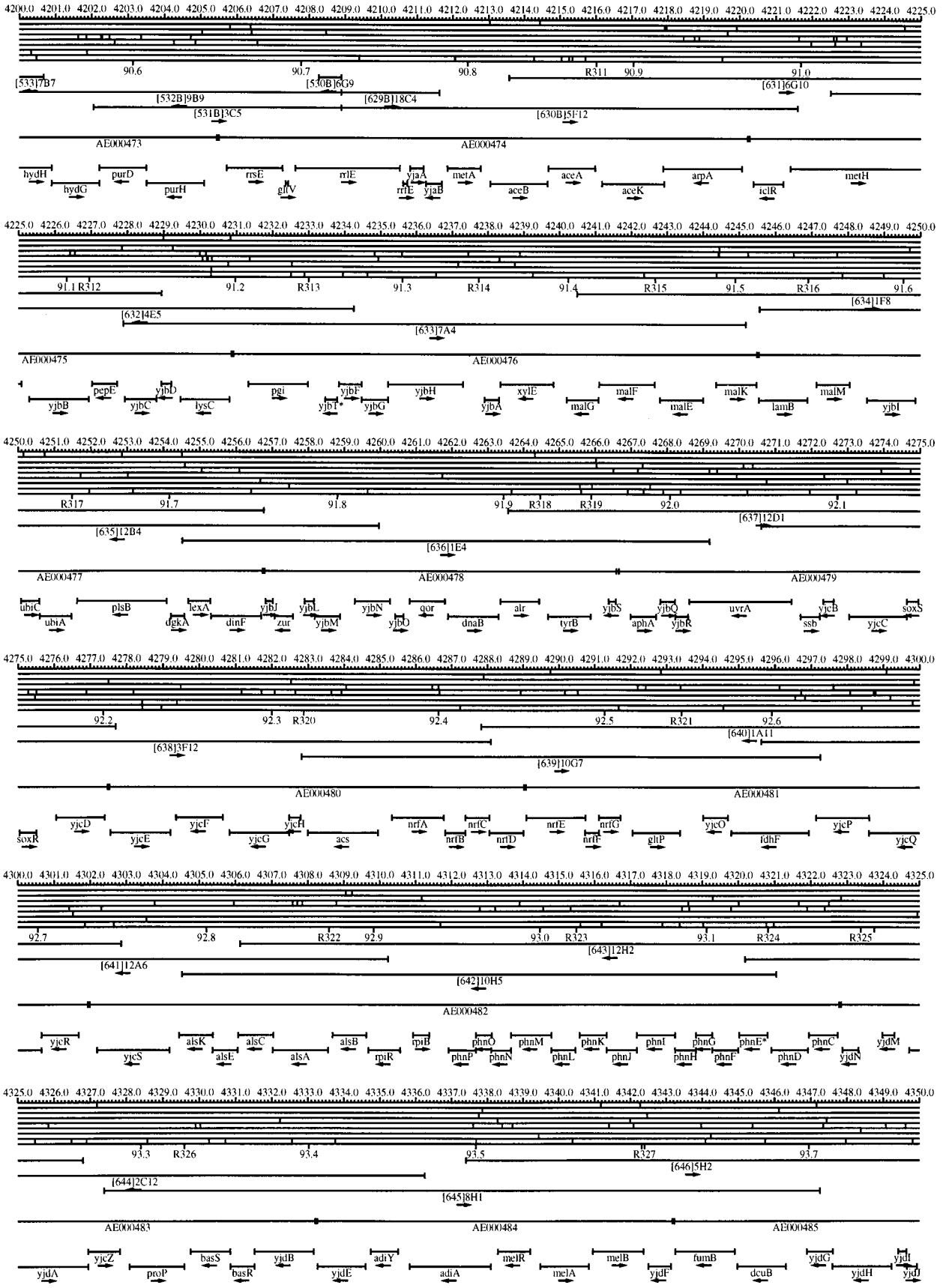


FIG. 1—*Continued.*

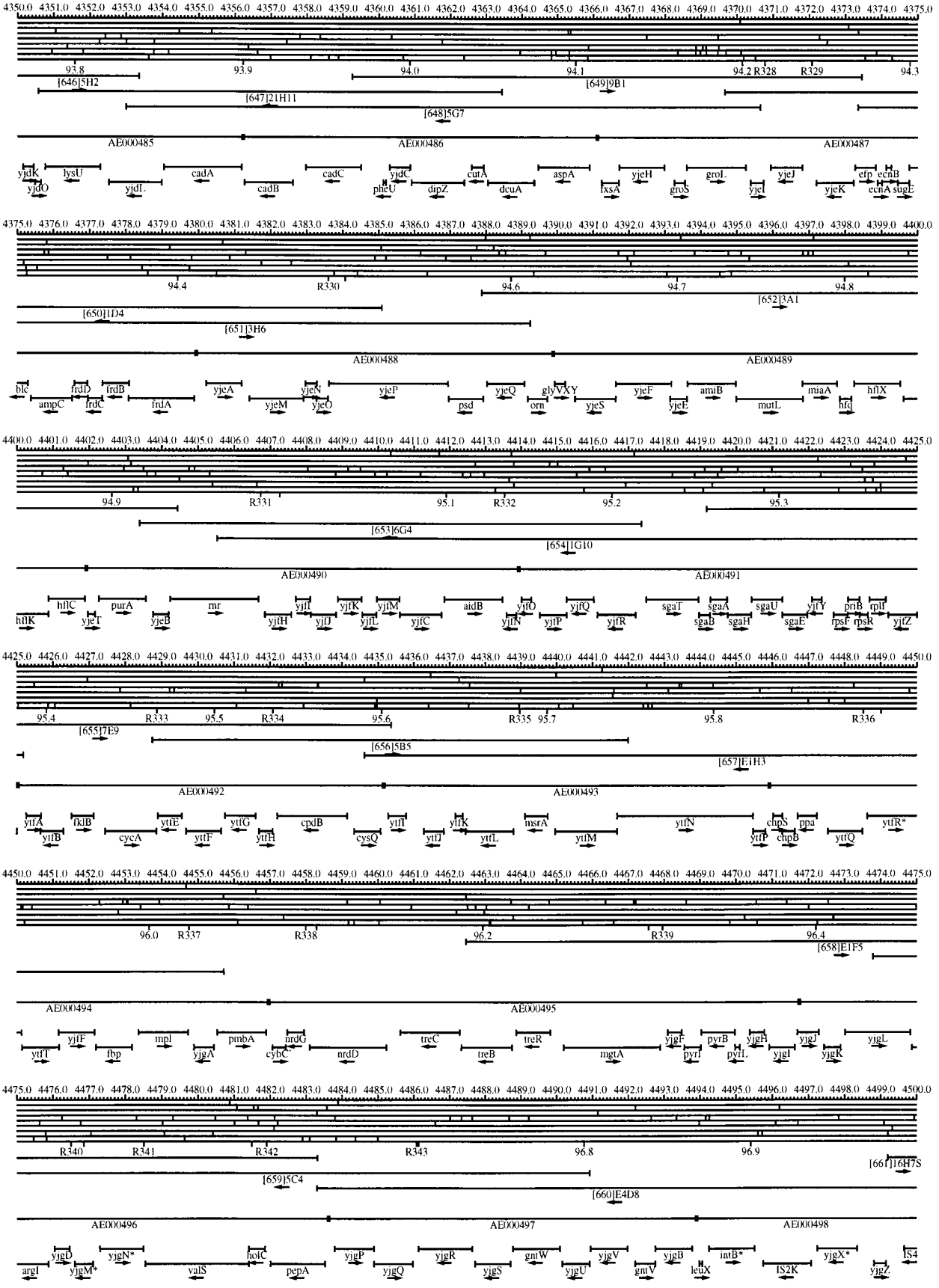
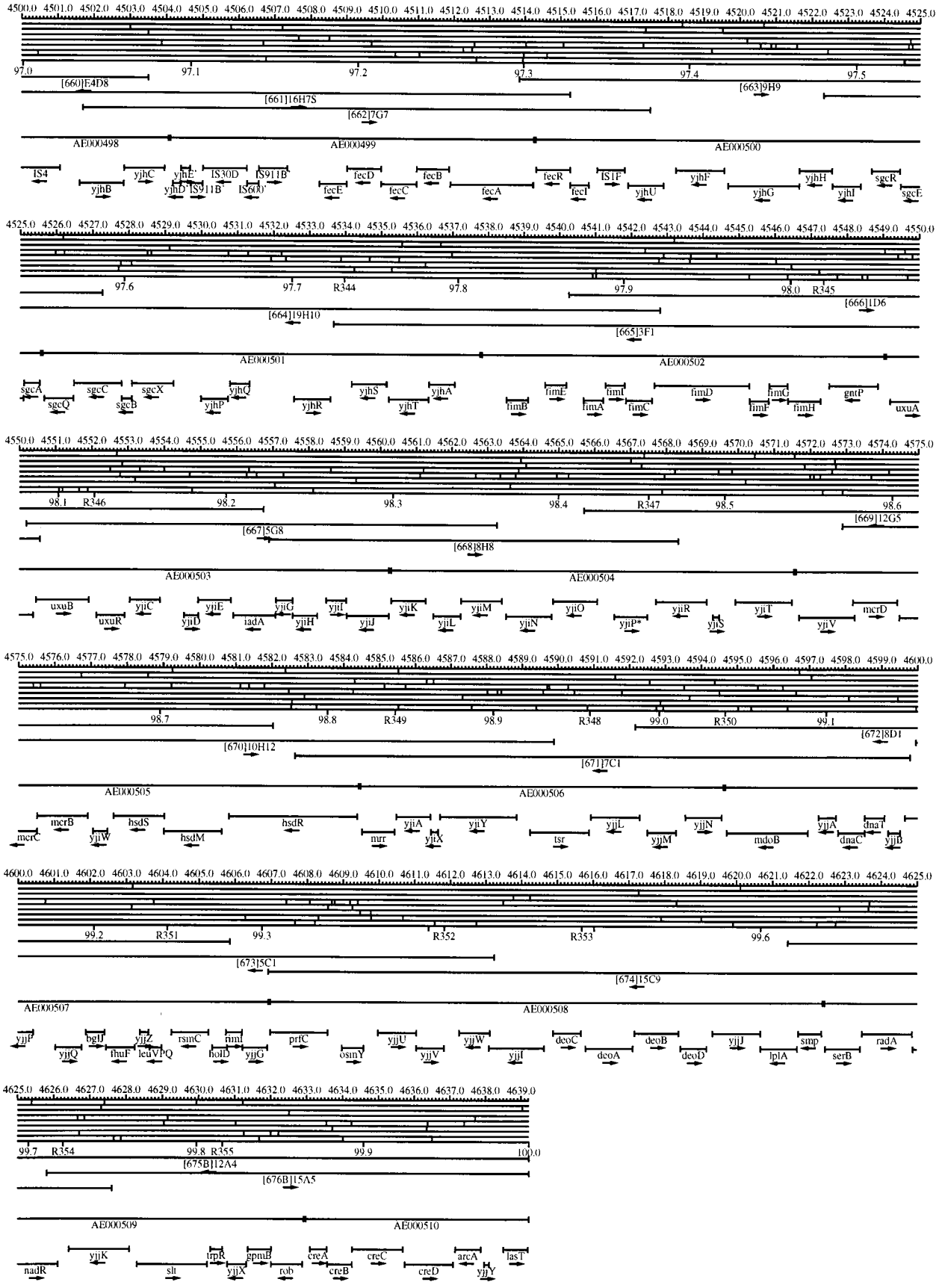


FIG. 1—*Continued.*



erichia coli K-12 linkage map are closely correlated. When there is a choice of several synonyms to use as the primary gene name, the physical map uses the same primary gene name as the traditional map. The primary names of genes not yet in the *E. coli* Genetic Stock Center (CGSC) database are considered provisional primary gene names. When choosing names for genes that are being functionally characterized for the first time, gene names already present in the database at the CGSC or the EcoGene database should be avoided. Guidance on naming and renaming genes is given in the paper containing the traditional map (3).

For the cases in which no standard-format gene name was assigned to a functionally uncharacterized gene or ORF, a systematic ORF nomenclature, the "y" naming system, was used to generate a provisional name (4, 15, 20). The first three letters of a "y" name are based on the map position of an ORF at the time the name was assigned. Similar to the "z" naming system for transposon insertions, ya[a to j]A to Z designates ORFs in the 0- to 10-min region of the chromosome, yb[a to j]A to Z designates ORFs in the 11- to 20-min region, and so on. The fourth letters (A to Z) can be assigned in any order within the 1-min interval. If all 26 names in any 1-min interval are exhausted, a new second letter is assigned to generate another 26 possibilities; additional ORFs after yaaZ would be ykaA, ykaB, and so on; additional ORFs after ybaZ would be ylaA, ylaB, and so on. The "y" names are not reused if a "y" ORF is given a new gene name or if an ORF becomes defunct, e.g., if a frameshift correction fuses two adjacent ORFs. Map locations provide a convenient and systematic method for naming ORFs, and the "y" names can guide one to an approximate map position. However, to avoid unnecessary renaming the "y" name of an ORF is not changed if a map revision moves it into an adjacent minute interval. The "y" names are now assigned to all the functionally uncharacterized, unnamed ORFs in EcoGene10. Once a new function is established for an *E. coli* gene, the provisional "y" name should be abandoned and a new gene name should be chosen.

Information concerning the availability of the EcoMap10 and EcoGene10 electronic datasets in various formats, including the Colibri database management program (12), can be obtained at http://cesspit.med.miami.edu. Additional information about the genes and ORFs in EcoGene10 is contained in SWISS-PROT records (http://www.expasy.ch/sprot) that can be accessed by using the names that are depicted on EcoMap10 and that are indexed in a master file (http://www.expasy.ch/cgi-bin/lists?ecoli.txt.

ACKNOWLEDGMENTS

This work was supported by funds made available to K.E.R. from a Lucille P. Markey Charitable Trust grant to the Department of Biochemistry and Molecular Biology at the University of Miami School of Medicine.

I am especially indebted to Amos Bairoch for his dedication to *E. coli*, his enthusiastic support of EcoGene, and for the many gene discoveries, literature citations, and protein sequence refinements that he has shared with me since the beginning of the EcoGene project. I thank Yuji Kohara and Katsumi Isono for providing the miniset of *E. coli* lambda clones, for allowing me to freely redistribute them, and for providing the original individual restriction maps of each miniset clone in electronic format. My collaboration with Mary Berlyn of the CGSC has been an essential component of the EcoMap/EcoGene project, and I am grateful for her patience and kindness throughout our data sharing and map coordination effort. I would also like to thank and acknowledge Gabrielle Redfern, Yuhong Zuo, Webb Miller, Karl Sirotkin, Craig Werner, Gerald Bouffard, Bobby Baum, Mark Borodovsky, Nir Hus, Rick Mitchell, Valerie Wasinger, Peter Maxwell, Ian Humphery-Smith, Ivan Moszer, and Antoine Danchin variously for general assistance, programming support, and helpful comments as well as for their continuing friendship. I acknowledge this work as a being derived from the many scientific contributions of the entire *E. coli* research community and extend my sincere gratitude to this community for their contributions. I gratefully acknowledge Fred Blattner and his colleagues for the complete MG1655 sequence and all the members of the Japanese research consortium who participated in the sequencing of the W3110 genome.

REFERENCES

1. **Aiba, H., T. Baba, K. Hayashi, T. Inada, K. Isono, T. Itoh, H. Kasai, K. Kashimoto, S. Kimura, M. Kitakawa, M. Kitagawa, K. Makino, T. Miki, K. Mizobuchi, H. Mori, T. Mori, K. Motomura, S. Nakade, Y. Nakamura, H. Nashimoto, Y. Nishio, T. Oshima, N. Saito, G. Sampei, Y. Seki, S. Sivasunddaram, H. Tagami, J. Takeda, K. Takemoto, Y. Takeuchi, C. Wada, Y. Yamamoto, and T. Horiuchi.** 1996. A 570-kb DNA sequence of the *Escherichia coli* K-12 genome corresponding to the 28.0-40.1 min region on the linkage map. DNA Res. **3:**363–377.
2. **Bachellier, S., E. Gilson, M. Hofnung, and C. W. Hill.** 1996. Repeated sequences, p. 2012–2040. *In* F. C. Neidhardt, R. Curtiss III, J. L. Ingraham, E. C. C. Lin, K. B. Low, B. Magasanik, W. S. Reznikoff, M. Riley, M. Schaechter, and H. E. Umbarger (ed.), *Escherichia coli* and *Salmonella*: cellular and molecular biology, 2nd ed. ASM Press, Washington, D.C.
3. **Berlyn, M. B.** 1998. Linkage map of *Escherichia coli* K-12, edition 10: the traditional map. Microbiol. Mol. Biol. Rev. **62:**814–984.
4. **Berlyn, M. B., K. B. Low, and K. E. Rudd.** 1996. Integrated linkage map of *Escherichia coli* K-12, edition 9, p. 1715–1902. *In* F. C. Neidhardt, R. Curtiss III, J. L. Ingraham, E. C. C. Lin, K. B. Low, B. Magasanik, W. S. Reznikoff, M. Riley, M. Schaechter, and H. E. Umbarger (ed.), *Escherichia coli* and *Salmonella*: cellular and molecular biology, 2nd ed. ASM Press, Washington, D.C.
5. **Blattner, F. R., G. Plunkett, C. A. Bloch, N. T. Perna, V. Burland, M. Riley, J. Collado-Vides, J. D. Glasner, C. K. Rode, G. F. Mayhew, J. Gregor, N. W. Davis, H. A. Kirkpatrick, M. A. Goeden, D. J. Rose, B. Mau, and Y. Shao.** 1997. The complete genome sequence of *Escherichia coli* K-12. Science **277:**1453–1462.
6. **Deonier, R. C.** 1996. Native insertion sequence elements: locations, distributions, and sequence relationships, p. 2000–2011. *In* F. C. Neidhardt, R. Curtiss III, J. L. Ingraham, E. C. C. Lin, K. B. Low, B. Magasanik, W. S. Reznikoff, M. Riley, M. Schaechter, and H. E. Umbarger (ed.), *Escherichia coli* and *Salmonella*: cellular and molecular biology, 2nd ed. ASM Press, Washington, D.C.
7. **Gilson, E., J. P. Rousset, J. M. Clement, and M. Hofnung.** 1986. A subfamily of *E. coli* palindromic units implicated in transcription termination? Ann. Inst. Pasteur Microbiol. **137B:**259–270.
8. **Gilson, E., W. Saurin, D. Perrin, S. Bachellier, and M. Hofnung.** 1991. The BIME family of bacterial highly repetitive sequences. Res. Microbiol. **142:** 217–222.
9. **Itoh, T., H. Aiba, T. Baba, K. Hayashi, T. Inada, K. Isono, H. Kasai, S. Kimura, M. Kitakawa, M. Kitagawa, K. Makino, T. Miki, K. Mizobuchi, H. Mori, T. Mori, K. Motomura, S. Nakade, Y. Nakamura, H. Nashimoto, Y. Nishio, T. Oshima, N. Sato, G. Sampei, Y. Seki, S. Sivasunddaram, H. Tagami, J. Takeda, K. Takemoto, C. Wada, Y. Yamamoto, and T. Horiuchi.** 1996. A 460-kb DNA sequence of the *Escherichia coli* K-12 genome corresponding to the 40.1-50.0 min region on the linkage map. DNA Res. **3:**379–392.
10. **Kohara, Y., K. Akiyama, and K. Isono.** 1987. The physical map of the whole *E. coli* chromosome: application of a new strategy for rapid analysis and sorting of a large genomic library. Cell **50:**495–508.
11. **Komine, Y., and H. Inokuchi.** 1991. Precise mapping of the *rnpB* gene encoding the RNA component of RNase P in *Escherichia coli* K-12. J. Bacteriol. **173:**1813–1816.
12. **Medigue, C., A. Viari, A. Henaut, and A. Danchin.** 1993. Colibri: a functional data base for the *Escherichia coli* genome. Microbiol. Rev. **57:**623–654.
13. **Oshima, T., H. Aiba, T. Baba, K. Fujita, K. Hayashi, A. Honjo, K. Ikemoto, T. Inada, T. Itoh, M. Kajihara, K. Kanai, K. Kashimoto, S. Kimura, M. Kitagawa, K. Makino, S. Masuda, T. Miki, K. Mizobuchi, H. Mori, K. Motomura, Y. Nakamura, H. Nashimoto, Y. Nishio, N. Saito, G. Sampei, Y. Seki, H. Tagami, K. Takemoto, C. Wada, Y. Yamamoto, M. Yano, and T. Horiuchi.** 1996. A 718-kb DNA sequence of the *Escherichia coli* K-12 genome corresponding to the 12.7-28.0 min region on the linkage map. DNA Res. **3:**137–155.
14. **Rudd, K. E.** 1992. Alignment of *E. coli* DNA sequences to a revised, integrated genomic restriction map, p. 2.3–2.4.3. *In* J. Miller (ed.), A short course in bacterial genetics: a laboratory manual and handbook for *Escherichia coli* and related bacteria, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y.
15. **Rudd, K. E.** 1993. Maps, genes, sequences, and computers: an *Escherichia coli* case study. ASM News **59:**335–341.
16. **Rudd, K. E., M. K. B. Berlyn, Y. Zuo, A. Danchin, I. Moszer, N. Hus, R. Mitchell, and A. Bairoch.** Submitted for publication.
17. **Rudd, K. E., W. Miller, J. Ostell, and D. A. Benson.** 1990. Alignment of

Escherichia coli K12 DNA sequences to a genomic restriction map. Nucleic Acids Res. **18:**313–321.
18. **Rudd, K. E., W. Miller, C. Werner, J. Ostell, C. Tolstoshev, and S. G. Satterfield.** 1991. Mapping sequenced *E. coli* genes by computer: software, strategies and examples. Nucleic Acids Res. **19:**637–647.
19. **Schweizer, H. P., and P. Datta.** 1990. Physical map location of the *tdc* operon of *Escherichia coli*. J. Bacteriol. **172:**2825.
20. **Stewart, A.** 1995. Genetic nomenclature guide including information on genomic databases. Elsevier, New York, N.Y.
21. **Umeda, M., and E. Ohtsubo.** 1990. Mapping of insertion element IS5 in the *Escherichia coli* K-12 chromosome. Chromosomal rearrangements mediated by IS5. J. Mol. Biol. **213:**229–237.
22. **Yamamoto, Y., H. Aiba, T. Baba, K. Hayashi, T. Inada, K. Isono, T. Itoh, S. Kimura, M. Kitagawa, K. Makino, T. Miki, N. Mitsuhashi, K. Mizobuchi, H. Mori, S. Nakade, Y. Nakamura, H. Nashimoto, T. Oshima, S. Oyama, N. Saito, G. Sampei, Y. Satoh, S. Sivasundaram, H. Tagami, H. Takahashi, J. Takeda, K. Takemoto, K. Uehara, C. Wada, S. Yamagata, and T. Horiuchi.** 1997. Construction of a contiguous 874-kb sequence of the *Escherichia coli*-K12 genome corresponding to 50.0-68.8 min on the linkage map and analysis of its sequence features. DNA Res. **4:**91–113.