Patterns of Genome Organization in Bacteria

Frederick R. Blattner *et al.* (*1*), when describing the complete sequence of the *Escherichia coli* chromosome, correlated an overall DNA property, "GC skew" [the quantity (G − C)/(G + C) averaged over a sliding window of arbitrary length 10 kb] with the direction of DNA replication. GC skew for replichore 1 (rightwards from the origin on the presented strand) oscillates considerably, yet remains almost entirely positive for its entire length, while replichore 2 shows the opposite behavior. Kunst *et al.* (*2*) did not present such an analysis for the sequence of the *Bacillus subtilis* chromosome, but did note that the GC skew changes sign at the origin, an observation made earlier by Lobry (*3*), who documented it for the replication origins of *E. coli*, *Haemophilus influenzae*, *B. subtilis*, and *Mycoplasma genitalium* for the terminus of *H. influenzae*.

In contrast to GC skew, which is a derivative function of the base composition along a DNA sequence, we have computed three integral functions of the sequences of nine complete prokaryotic genomes (Table 1). Composite graphs for three of these genomes are presented (Fig. 1), and the remainder are available on a linked website. We define "purine-excess" as the sum of all purines minus the sum of all pyrimidines encountered in a walk along the sequence up to the point plotted (*4*). "Keto excess" is the same function calculated for the keto bases (GT) minus the amino bases (AC), and "coding-strand excess" is the sum of all nucleotides encountered along the sequence that are in coding sequences, minus those that have *complements* (on the opposite strand that are in coding sequences; bases in noncoding regions add zero to this sum. Correlations between purine excess and coding-strand reveal nonrandom patterns, the most striking of which is the clear correlation between purine excess and the origins and termini of DNA replication (Fig. 1). In *every case* where independent information is available, the minimum in the purine-excess curve corresponds to the origin (Table 1). We suggest that this regularity may hold for most prokaryotic genomes. Conversely, the maxima of the purine-excess curves (Fig. 1) correlate strongly with known or suspected replication termini (*5*). Keto-excess curves reflect the same correlation, although for most genomes the minima and maxima (thus, predicted origins and termini) are not as sharply defined as for the purine-excess functions. *H. influenzae* represents a notable exception to this rule (for example, the keto-excess curve in Fig. 1B).

Other genome features stand out in these graphs. The relatively smooth, featureless curve for *E. coli* contrasts with the much rougher patterns displayed by *H. influenzae* and *Synechocystis* PCC6803 (see linked website for data). This likely reflects a greater tendency of the latter organisms to take up foreign DNA and integrate it into the chromosome (*6, 7*), a point supported by the correlation of the density of DNA-uptake sequences in *H. influenzae* (*6*) with many of the inflection points of the purine-excess curve (*8*). Likewise, the sites of m prophage integration in *H. influenzae* cluster most densely around the pronounced minimum in the purine-excess curve adjacent to the terminus (Fig. 1B). The larger megaplasmid (pNGR234a) of *Rhizobium* sp. NGR234 also displays similar behavior (*8*), in keeping with its recognized characteristics as a "transposon trap" (*9*).

Examination of the relationship between base-composition and coding asymmetries at the whole-genome level shows close parallels between coding-strand and purine excess for seven out of nine genomes. *E. coli* shows typical behavior (Fig. 1A). *H. influenzae* and *Synechocystis* display much weaker correlations on this scale. At a finer level of detail, there are substantial correlations between these functions for all the genomes we studied, but the results for the two archaebacteria, *M. jannaschii* and *M. thermoautotrophicum*, are particularly striking (Fig. 1C), showing strong correspondence between coding-strand and purine excess.

What forces might give rise to the long-range patterns of strand asymmetry in bacterial genomes? There is a prominent correlation between purine excess and replication direction, which suggests as an explanation asymmetrical errors in DNA synthesis. In the absence of transpositions and insertions, a bias favors accumulation of purines in the leading strand. However, this contradicts expectations that *lagging* strand synthesis should be more error-prone (*10*), and thus that most purine substitutions (the principal cause of transversions) should occur there. Francino and Ochman (*11*) have argued, on the other hand, that *transcriptional* effects can account for DNA strand asymmetry because transcription-coupled repair will remove the most frequent types of DNA damage (deaminated cytosines and pyrimidine-dimers), thereby reducing harmful mutations. This only occurs on the transcribed (that is, template) strand, which therefore will become pyrimidine-rich. In addition, the template strand is significantly protected against DNA damage during transcription, whereas the coding strand is exposed. Under this model, evolutionary selection should increase the less mutationally vulnerable purine content of the coding strand.

M. genitalium conforms to the predictions of the transcription-coupled repair model particularly well: in replichore 1, 85% of the ORFs correspond to the presented (purine-rich) strand up to the putative terminus (maximum in the purine-excess curve). For the other replichore, 77% of the ORFs occur in the complementary strand. In *E. coli*, strand preference is less pronounced: only 55% of the genes are aligned with the replication direction (*1*). However, Francino has analyzed the codon adaptation index (CAI), a measure strongly associated with the extent of gene expression in *E. coli*, and finds that 74% of the genes with CAI ≥ 0.5 and 84% of those with CAI ≥ 0.6 are situated on the leading strand (*11*), that is, with the direction of transcription the same as replication (*12*). In addition to favoring transcriptional repair, a major advantage to this arrangement is that head-on collisions between replication and transcription complexes will be reduced (*13*).

Functions like those described here promise to be revealing tools for whole-

Table 1. Completely-sequenced bacterial genomes analyzed for base and coding asymmetries and their origins and termini of replication.

Species	Length (Mbp)	Origin (bp)	Ref.	Terminus (bp)	Ref.
Escherichia coli	4.64	3,923,500	(*1*)	1,588,800	(*18*)
Bacillus subtilis	4.21	1	(*2*)	2,017,000	(*2*)
Mycoplasma pneumoniae	0.82	205,000	(*19*)	n.a.*	
Mycoplasma genitalium	0.58	1	(*3, 20*)	n.a.	
Helicobacter pylori	1.67	1	(*21*)	n.a.	
Haemophilus influenzae	1.83	603,000	(*3, 16*)	1,518,000	(*3, 16*)
Synechocystis PCC6803	3.57	1,351,000?	(*22*)	n.a.	
Methanococcus jannaschii	1.66	n.a.		n.a.	
Methanobacterium thermoautotrophicum	1.75	n.a.		n.a.	

*Data not available.

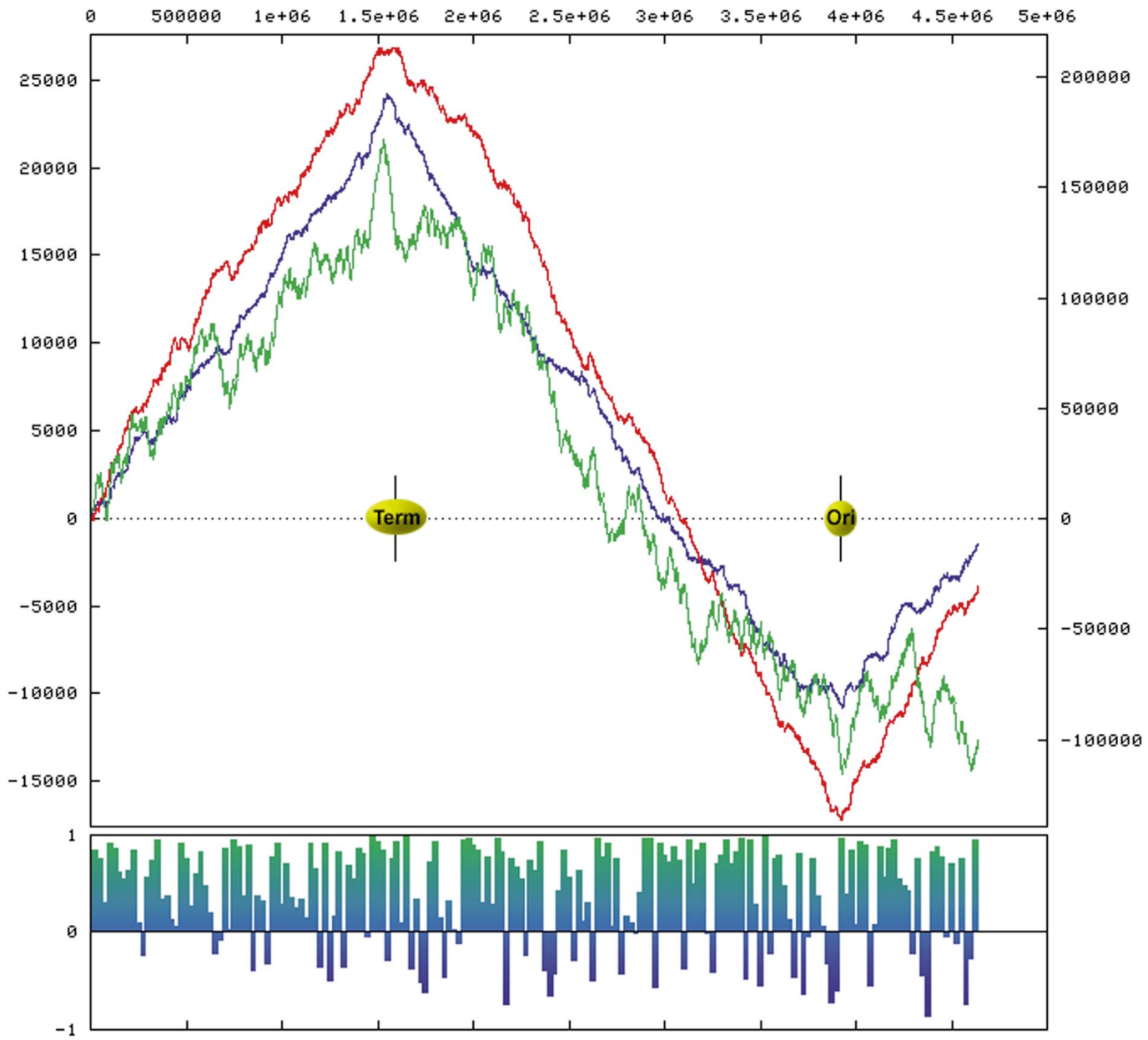


Fig. 1. Purine excess (blue curves), keto excess (red curves), and coding-strand excess (green curves) for the complete genomes of (**A**) *E. coli (1)*, (**B**) *H. influenzae* (*16*), and (**C**) *M. jannaschii* (*17*). Known origins and termini of replication are marked. Abscissa represents the genomic sequence position from the beginning to the end of the genome; left ordinate represents the count of purine and keto excesses; right ordinate represents the Watson coding-strand excess count at a given position. Green histograms across the bottom of each graph display the correlation coefficients between purine excess and coding-strand excess for 25 kb windows. Click on each image to enlarge. Graphs of six additional genomes (Table 1) can be viewed on the Web at http://bmerc-www.bu.edu/genomeplot/

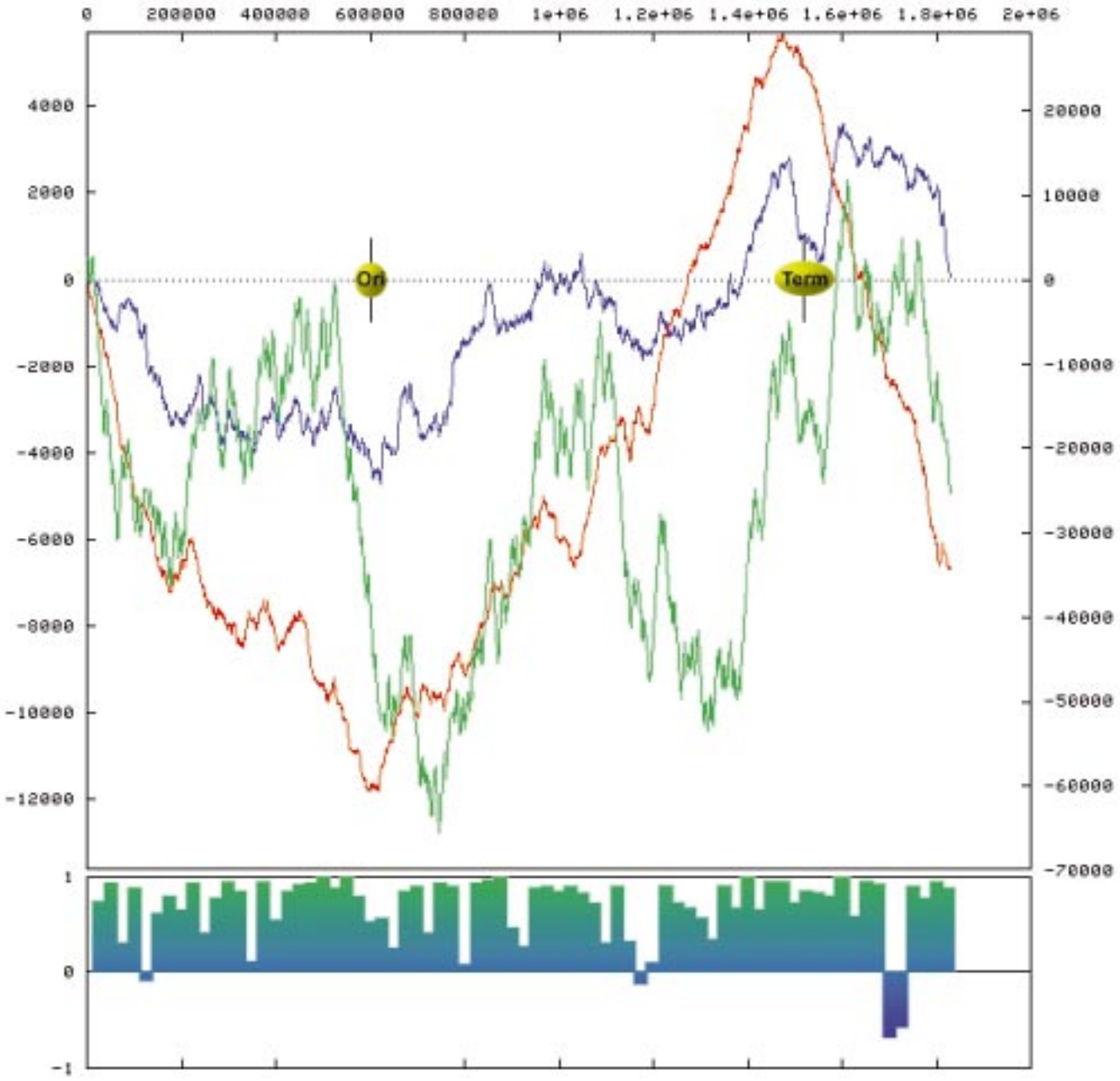
genome analysis (*4*). For example, in the absence of any other information, the global minimum of the purine excess locates the probable origin of replication, and its maximum is the likely terminus for prokaryotic genomes. Similar regularities may emerge from the impending deluge of eukaryotic DNA sequences. We have already shown that the patterns of purine-excess plots correlate well with phylogenetic position for mitochondrial DNAs (*14*), and graphs of yeast shows ORFs as tending to display positively sloped purine-excess curves (*15*).

James M. Freeman
Biomolecular Engineering Research Center, Boston University, Boston, MA 02215, USA
Thomas N. Plasterer
Department of Pharmacology, Boston University
Temple F. Smith
Biomolecular Engineering Research Center, Boston University
Scott C. Mohr
Department of Chemistry, Boston University

REFERENCES AND NOTES

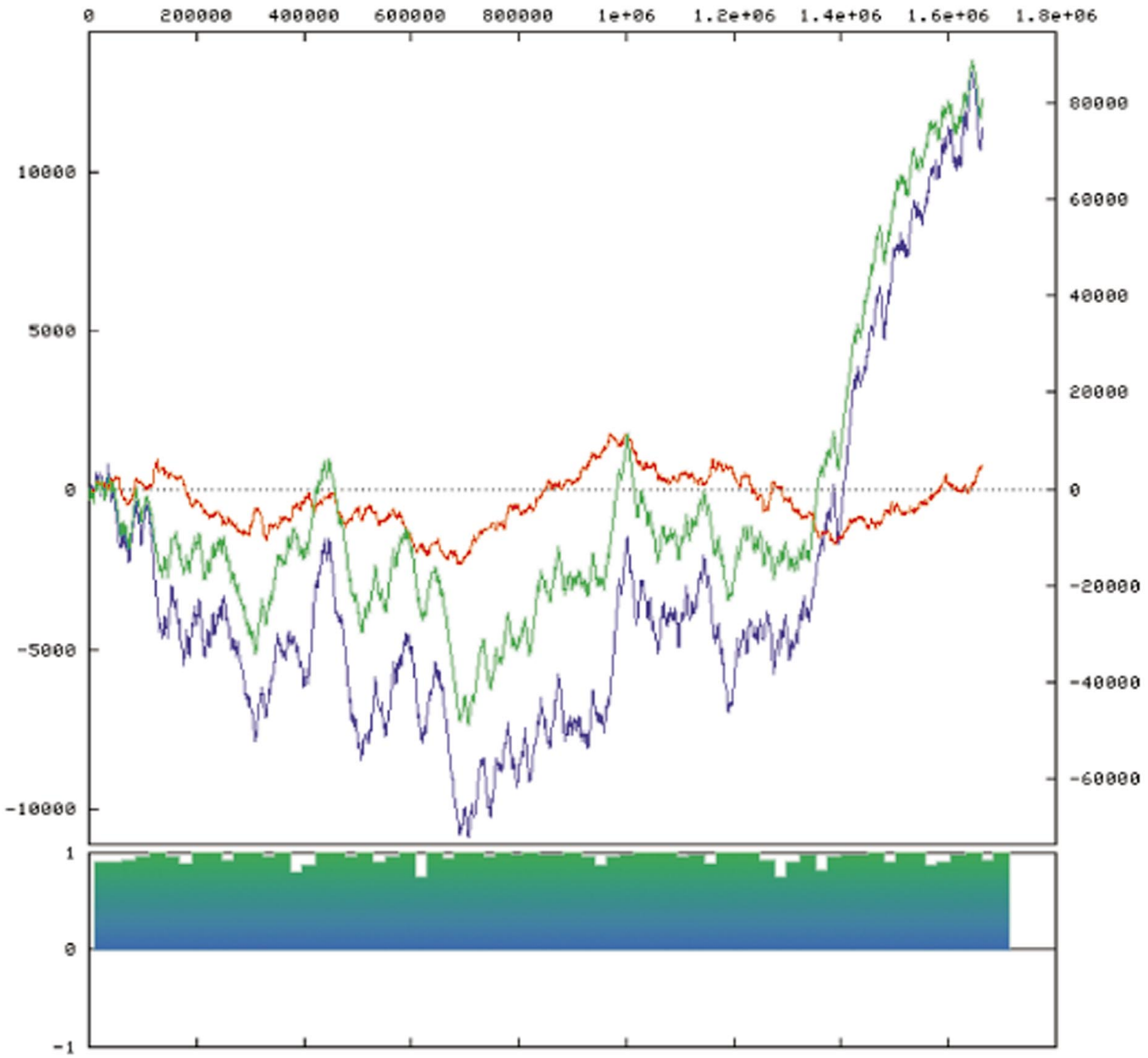
1. F. R. Blattner *et al*., *Science* **277**, 1453 (1997).
2. F. Kunst *et al*., *Nature* **390**, 249 (1997).
3. J. R. Lobry, *Mol. Biol. Evol.* **13**, 650 (1996); *Science* **272**, 745 (1996); *Biochimie* **78**, 323 (1996).
4. Purine excess: $c_{(l)} = S[d_{A,S} + d_{G,S} - D_{T,S} - d_{C,S}]$, where S is the base present at the current sequence position (l), the sum is performed over the range 1 to l, and $d_{X,Y} = 1$ if X = Y; and 0 if X ≠ Y. Interchanging the A and T subscripts in this equation defines the keto excess.

A DNA sequence can be uniquely described as a walk through a three-dimensional vector space, defined by two orthogonal axes for the two types of base pair and a third perpendicular axis, that repre



sents the sequence position (*3*) (Fig. 2). An A in the sequence corresponds to movement in the positive *x* direction and a T to the opposite. G and C are mapped by analogous steps along the *y* axis and sequence position increases along *z*. For example, starting at the origin of such a coordinate system, if the first base encountered is G, then the vector trace generates the point (0, +1, +1), where the indices are the usual Cartesian coordinates. If the second base is A, the trace extends to (+1, +1, +2), and so forth. The trace corresponding to GAATTTC continues on through (+2, +1, +3), (+1, +1, +4), (0, +1, +5), and (−1, +1, +6) to (0, 0, +7). Negative values of sequence position can also be used, which allows the origin to correspond to any convenient point in the sequence. As indicated by Fig. 2, the purine-excess and keto-excess functions that we have graphed for the nine prokaryotic genomes consist of steps along one or the other of two diagonal axes in this sequence space. Alternatively, the functions can be visualized as projections of the vector sequence trace onto one or the other of two vertical planes that cut the base-composition plane along the designated axes.

5. The precise locations of the three known termini (Table 1) actually fall slightly *beyond* the maximum of the purine excess curve and they coincide in every known case with the end of a segment that has a sharply negative slope in the coding-strand excess curve.
6. H. O. Smith, J.-F. Tomb, B. A. Dougherty, R. D. Fleischmann, J. C. Venter, *Science* **269**, 538 (1995).
7. V. A. Dzelzkalns and L. Bogorad, *EMBO J.*, **7**, 333 (1988).
8. J. M. Freeman *et al*., data not shown.
9. C. Freiberg *et al*., *Nature* **387**, 394 (1997).
10. If DNA damage to the lagging-strand *template* dominates over synthesis errors, however, this conclusion would be reversed because a purine-rich strand is less vulnerable to damage.
11. M. P. Francino and H. Ochman, *Trends Genet.* **13**, 240 (1997).
12. In the case of *E. coli* phage I, the purine-excess plot has a minimum at the replication origin and a major dip just previous to it. From the origin, there is a rise that has a continuous run of ORFs coded on the presented strand (61% of total ORFs in the genome) that are thus transcribed along the phage's one-way replication direction. The dip region is coded exclusively on the complementary strand (31% of total ORFs). The other 8% alternate between strands at the start of the dip.
13. B. Liu *et al*., *Nature* **366**, 33 (1993); *Proc. Natl. Acad. Sci. U.S.A.* **91**, 10660 (1994); A. M. Deshpande and C. S. Newlon, *Science* **272**, 1030 (1996).
14. S. C. Mohr *et al*., *Biol. Bull.*, in press (1998).
15. J. Graber *et al*., unpublished experiments.
16. R. D. Fleischmann *et al*., *Science* **269**, 496 (1995).
17. C. J. Bult *et al*., *Science* **273**, 1058 (1996).
18. G. Plunkett, personal communication.
19. R. Himmelreich *et al*., *Nucl. Acids Res.* **24**, 4420 (1996).



20. C. M. Fraser *et al*., *Science* **270**, 397 (1995).
21. J.-F. Tomb *et al*., *Nature* **388**, 539 (1997).
22. T. Kaneko *et al*., *DNA Res.* **3**, 109 (1996); the origin is tentatively assigned to the position of *dnaA*.

23. We thank B. Rogers for helpful discussions regarding visualizations, the Boston University Office of Information Technology and the Scientific Computing and Visualization Group for supercomputing resources, and an anonymous reviewer for helpful suggestions. S.C.M. is partially supported by a Training Grant from the U.S. National Human Genome Research Institute (T32 HG00041-03). Grant DE-FG02-98ER62558 from the U.S. Department of Energy supported this research.

9 December 1997; revised 27 February 1998; accepted 9 March 1998

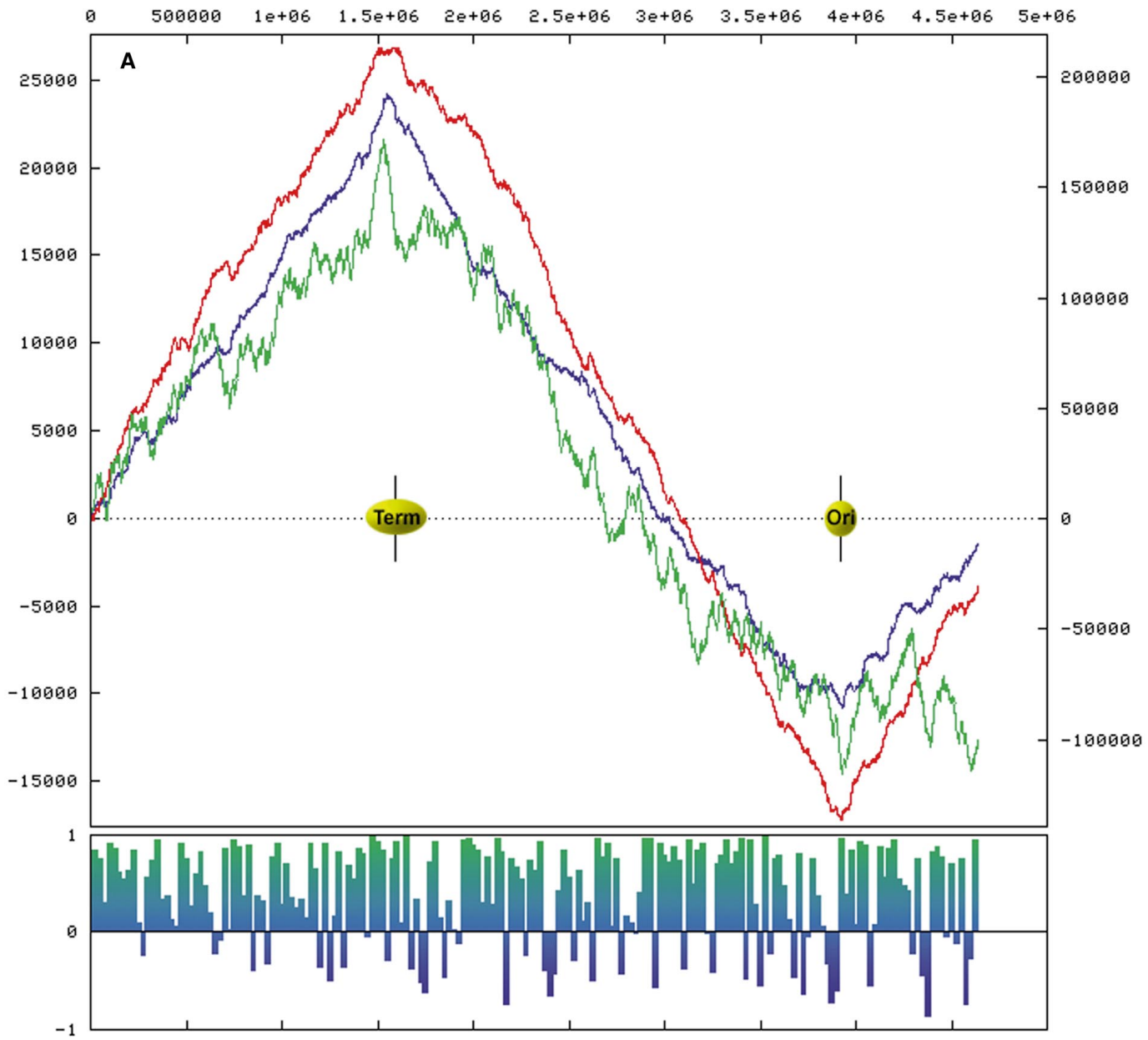


Fig. 1. Purine excess (blue curves), keto excess (red curves), and coding-strand excess (green curves) for the complete genomes of (**A**) *E. coli (1)*, (**B**) *H. influenzae* (*16*), and (**C**) *M. jannaschii* (*17*). Known origins and termini of replication are marked. Abscissa represents the genomic sequence position from the beginning to the end of the genome; left ordinate represents the count of purine and keto excesses; right ordinate represents the Watson coding-strand excess count at a given position. Green histograms across the bottom of each graph display the correlation coefficients between purine excess and coding-strand excess for-25 kb windows. Click on each image to enlarge. Graphs of six additional genomes (Table 1) can be viewed on the Web at http://bmerc-www.bu.edu/genomeplot/

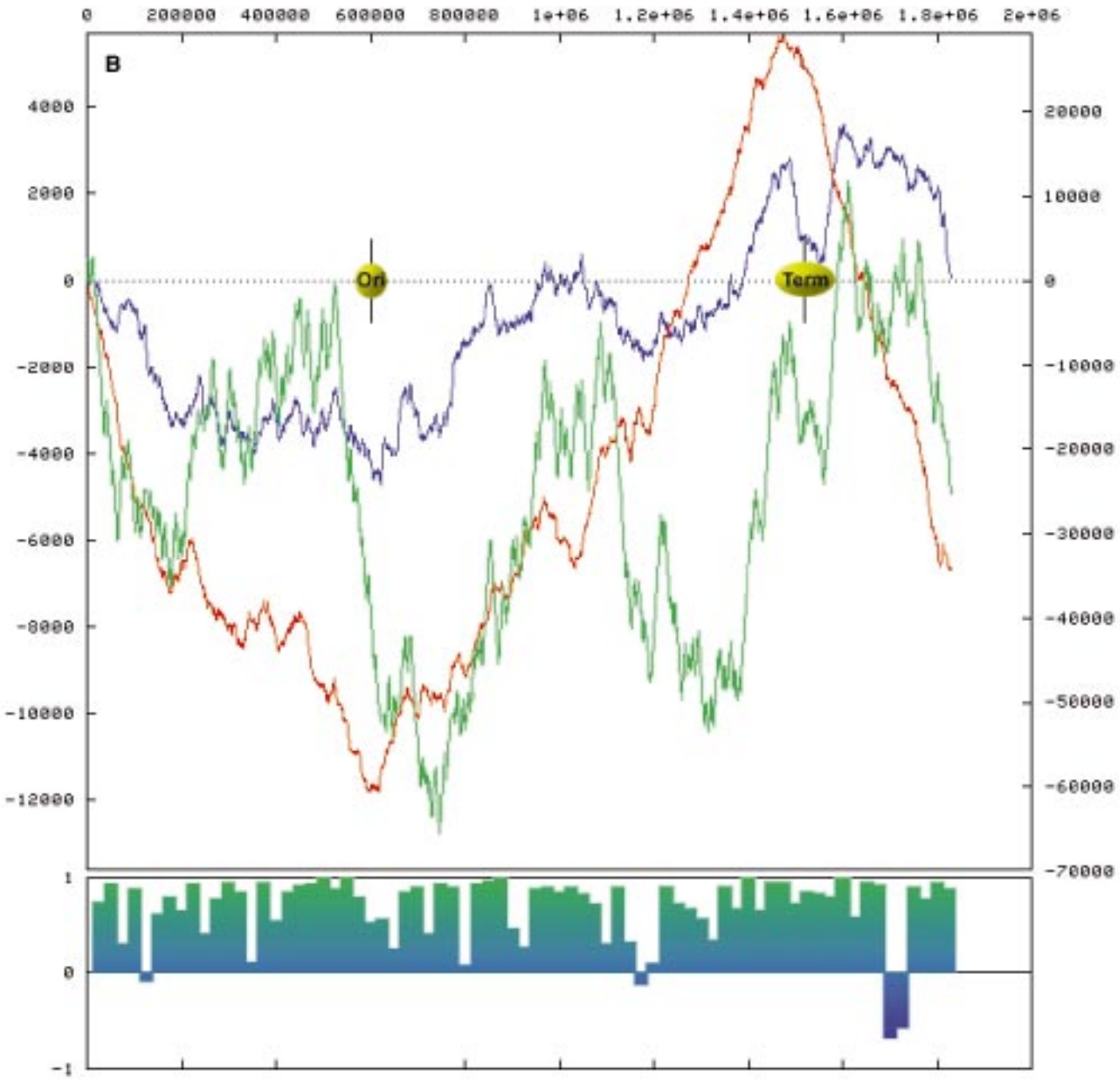
absence of any other information, the global minimum of the purine excess locates the probable origin of replication, and its maximum is the likely terminus for prokaryotic genomes. Similar regularities may emerge from the impending deluge of eukaryotic DNA sequences. We have already shown that the patterns of purine-excess plots correlate well with phylogenetic position for mitochondrial DNAs (*14*), and graphs of coding-strand excess in the *Saccharomyces cerevisiae* genome tend to match the purine-excess curves (*15*).

James M. Freeman
Biomolecular Engineering Research Center, Boston University, Boston, MA 02215, USA
Thomas N. Plasterer
Department of Pharmacology, Boston University
Temple F. Smith
Biomolecular Engineering Research Center, Boston University
Scott C. Mohr
Department of Chemistry, Boston University

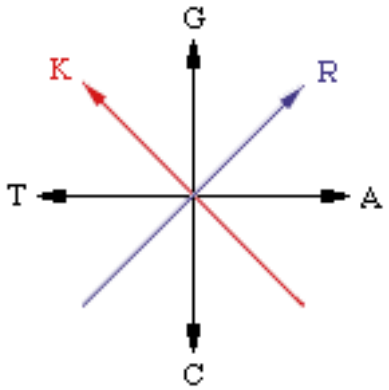
REFERENCES AND NOTES

1. F. R. Blattner *et al*., *Science* **277**, 1453 (1997).
2. F. Kunst *et al*., *Nature* **390**, 249 (1997).
3. J. R. Lobry, *Mol. Biol. Evol.* **13**, 650 (1996); *Science* **272**, 745 (1996); *Biochimie* **78**, 323 (1996).
4. Purine excess: $\chi_{(l)} = \Sigma[\delta_{A,S} + \delta_{G,S} - \delta_{T,S} - \delta_{C,S}]$, where S is the base present at the current sequence position (l), the sum is performed over the range 1 to l, and $\delta_{X,Y} = 1$ if X = Y; and 0 if X ≠ Y. Interchanging the A and T subscripts in this equation defines the keto excess.

A DNA sequence can be uniquely described as a walk through a three-dimensional vector space, defined by two orthogonal axes for the two types of base pair and a third perpendicular axis, that repre-

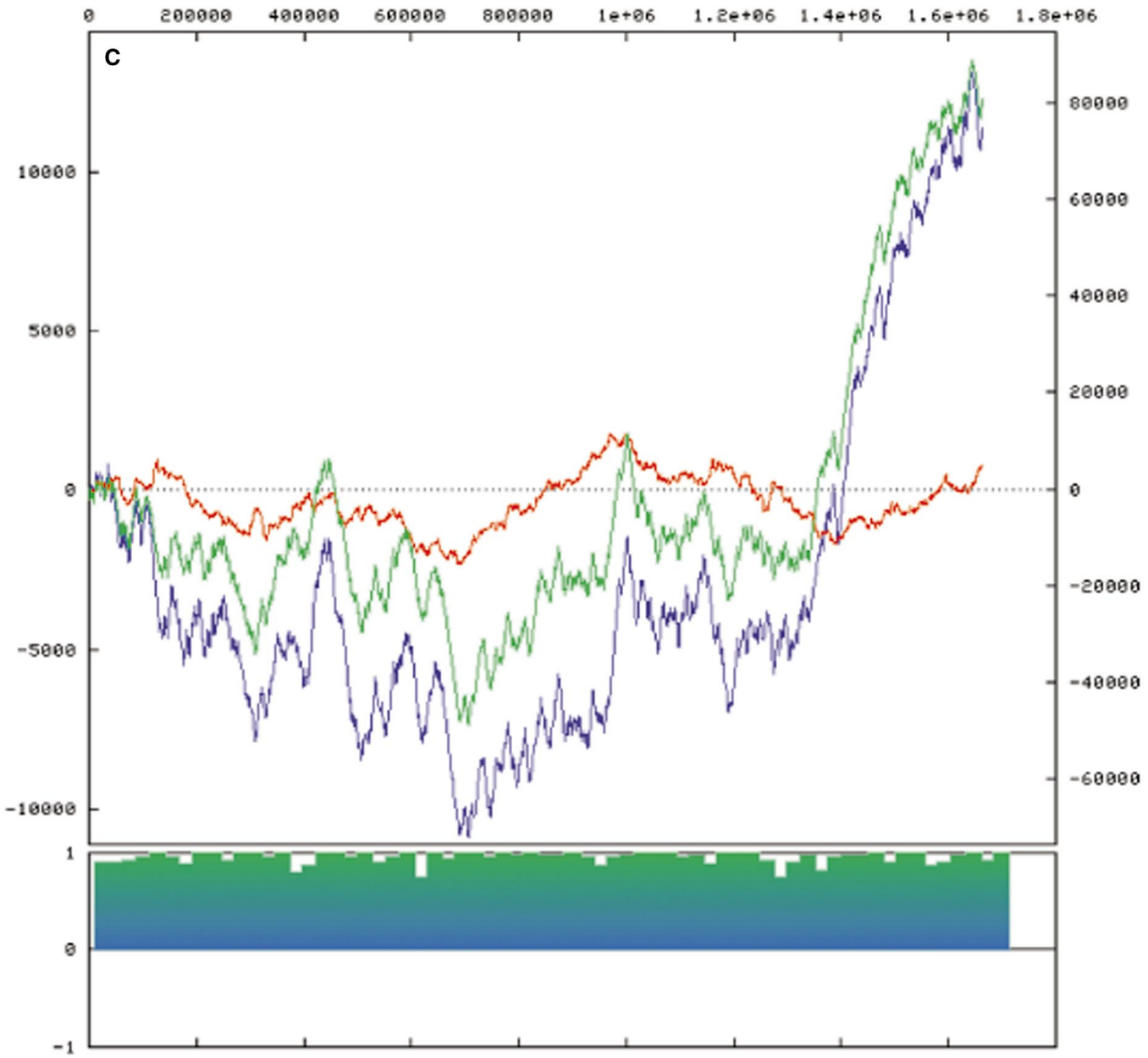


sents the sequence position (*3*) (scheme). An A in the sequence corresponds to movement in the positive *x* direction and a T to the opposite. G and C are mapped by analogous steps along the *y* axis and sequence position increases along *z*. For example, starting at the origin of such a coordinate system, if the first base encountered is G, then the vector trace generates the point (0, +1, +1), where the indices are the usual Cartesian coordinates. If the second base is A, the trace extends to (+1, +1, +2), and so forth. The trace corresponding to GAATTTC continues on through (+2, +1, +3), (+1, +1, +4), (0, +1, +5), and (−1, +1, +6) to (−1, 0, +7). Negative values of sequence position can also be used, which allows the origin to correspond to any convenient point in the sequence. As indicated by the scheme, the purine-excess and keto-excess functions that we have graphed for the nine prokaryotic genomes consist of steps along one or the other of two diagonal axes in this sequence space. Alternatively, the functions can be



visualized as projections of the vector sequence trace onto one or the other of two vertical planes that cut the base-composition plane along the designated axes.

5. The precise locations of the three known termini (Table 1) actually fall slightly *beyond* the maximum of the purine excess curve and they coincide in every known case with the end of a segment that has a sharply negative slope in the coding-strand excess curve.
6. H. O. Smith, J.-F. Tomb, B. A. Dougherty, R. D. Fleischmann, J. C. Venter, *Science* **269**, 538 (1995).
7. V. A. Dzelzkalns and L. Bogorad, *EMBO J.*, **7**, 333 (1988).
8. J. M. Freeman *et al.*, data not shown.
9. C. Freiberg *et al.*, *Nature* **387**, 394 (1997).
10. If DNA damage to the lagging-strand *template* dominates over synthesis errors, however, this conclusion would be reversed because a purine-rich strand is less vulnerable to damage.



11. M. P. Francino and H. Ochman, *Trends Genet.* **13**, 240 (1997).
12. In the case of *E. coli* phage λ, the purine-excess plot has a minimum at the replication origin and a major dip just previous to it. From the origin, there is a rise that has a continuous run of ORFs coded on the presented strand (61% of total ORFs in the genome) that are thus transcribed along the phage's one-way replication direction. The dip region is coded exclusively on the complementary strand (31% of total ORFs). The other 8% alternate between strands at the start of the dip.
13. B. Liu *et al*., *Nature* **366**, 33 (1993); *Proc. Natl. Acad. Sci. U.S.A.* **91**, 10660 (1994); A. M. Deshpande and C. S. Newlon, *Science* **272**, 1030 (1996).
14. S. C. Mohr *et al*., *Biol. Bull.*, in press (1998).
15. J. Graber *et al*., unpublished experiments.
16. R. D. Fleischmann *et al*., *Science* **269**, 496 (1995).
17. C. J. Bult *et al*., *ibid.* **273**, 1058 (1996).
18. G. Plunkett, personal communication.
19. R. Himmelreich *et al*., *Nucleic Acids Res.* **24**, 4420 (1996).
20. C. M. Fraser *et al*., *Science* **270**, 397 (1995).
21. J.-F. Tomb *et al*., *Nature* **388**, 539 (1997).
22. T. Kaneko *et al*., *DNA Res.* **3**, 109 (1996); the origin is tentatively assigned to the position of *dnaA*.

23. We thank B. Rogers for helpful discussions regarding visualizations, the Boston University Office of Information Technology and the Scientific Computing and Visualization Group for supercomputing resources, and an anonymous reviewer for helpful suggestions. S.C.M. is partially supported by a training grant from the U.S. National Human Genome Research Institute (T32 HG00041-03). Grant DE-FG02-98ER62558 from the U.S. Department of Energy supported this research.

9 December 1997; revised 27 February 1998; accepted 9 March 1998