Eyeing bacterial genomes

Howard Ochman* and Scott R Santos

The density of information in a bacterial genome allows its history, organization and encoded functions to be distilled into a single graphical representation. These features have made it possible to discern the forces acting in and on bacterial genomes at levels not attainable in eukaryotes.

Addresses
Department of Biochemistry and Molecular Biophysics, University of Arizona, Tucson, AZ 85721, USA
*e-mail: hochman@email.arizona.edu

Current Opinion in Microbiology 2003, **6**:109–113

This review comes from a themed issue on
Cell regulation
Edited by Andrée Lazdunski and Carol Gross

1369-5274/03/$ – see front matter

DOI 10.1016/S1369-5274(03)00031-6

Abbreviations
ORF open reading frame
IS insertion sequence

Introduction

Hemophilus influenzae set the standard: it was the first fully sequenced microbe and the first whose genome was portrayed in the now familiar concentric-ring arrangement [1]. Not only are such depictions structurally correct, they can also easily accommodate the growing number of genomic features in an economy of space, so much so that even linear chromosomes have been portrayed in this manner [2]. These circular configurations are presented in nearly half of all primary papers on fully sequenced bacterial genomes. Whereas authors might seem obliged to include them, readers are as apt to overlook them. This is unfortunate because these figures constitute some of the more elegant renderings in the current literature and often allow the reader to understand more about bacterial genomes than is possible from reading the paper itself. In this review, we describe how each of the concentric rings in these features can provide insights into the structure, function or evolution of bacterial genomes.

On the edge

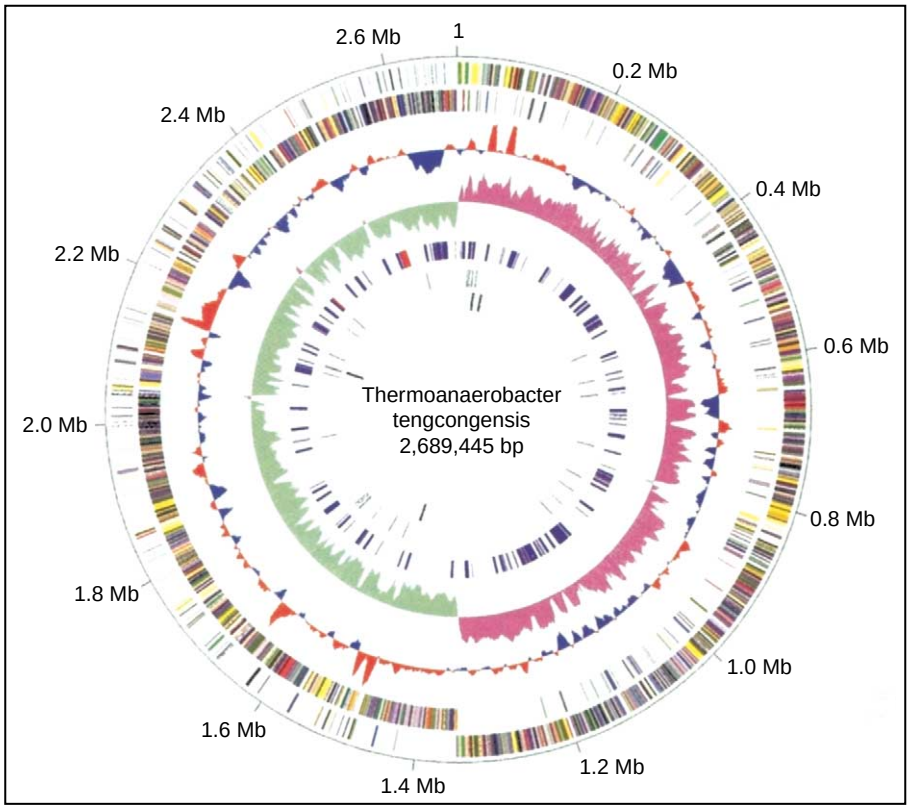
Despite a number of permutations caused by genomic or experimental anomalies, these representations of bacterial genomes have settled on a more-or-less common format (Figure 1). At the periphery are the predicted open reading frames (ORFs), colour-coded with respect to function and assorted by location on the leading or lagging strand. ORFs are recognized by a variety of methods, ranging from uninterrupted coding sequences of prescribed lengths or coding potential, to sequences known to confer an established function or phenotype. Given the expanding number and phylogenetic distribution of microbial sequences in databases, it is becoming easier to recover homologs and to assign functions to the majority of ORFs in any newly sequenced bacterial genome. Even so, in all but the most reduced and derived genomes [3••], novel ORFs (i.e. those having no known homolog in current databases) have been discovered. These regions are likely to represent rapidly evolving genes, DNA derived from an as yet uncharacterized source, or perhaps even non-coding regions erroneously designated as ORFs [4–7]; however, this variation in gene content attests to the very dynamic nature of bacterial chromosomes.

Bacterial genomes vary in the proportion of genes on their leading and lagging strands; and in every genome sequenced to date, most genes, and particularly those that are highly expressed, reside on the leading strand [8,9]. This arrangement is thought to avert head-on collisions between the replication and transcription machineries. Collision of DNA and RNA polymerization complexes can interrupt replication and abort transcription, and such events increase with the expression of genes on the lagging strand [10,11]. Even as the first few genome sequences became available, it was clear that there was substantial variation among bacteria in the degree of coding asymmetry between the strands, and that the low G + C-positive bacteria (*Bacillus subtilis* and the mycoplasmas) displayed the greatest strand bias. *Thermoanaerobacter tengcongensis*, the genome of which is depicted in Figure 1, has almost 87% of its ORFs situated on the leading strand [12]. It now appears that this bias might be associated with the dedicated use of PolC and DnaE for leading- and lagging- strand synthesis, respectively. Species that lack PolC, but rely on DnaE as a subunit of the polymerisation complex during both leading- and lagging- strand synthesis, show less disparity in the coding contents of each strand [13•].

Near the rim

The next circle plots base composition, calculated either as the percentage of G+C bases for each gene, for fixed or overlapping intervals, and is generally displayed as the deviations in G + C content of the prescribed intervals from the average. Bacterial base compositions can vary widely, with sequenced genomes ranging from *22%* G + C in the tsetse fly endosymbiont *Wigglesworthia*

Figure 1



Concentric rings representing features of the *Thermoanaerobacter tengcongensis* genome. *T. tengcongensis* is a Gram-negative anaerobe with a genome size of 2.7 megabases and low G + C content. Proceeding from the periphery, rings one and two contain the predicted ORFs transcribed on the clockwise and anticlockwise strands, respectively, and color-coded according to different functional categories. There is a switch in the density of coding sequences at the replication origin and terminus owing to the differential assortment of genes onto the leading and lagging strands. Ring three depicts base composition, showing regions above and below the genomic mean of 37.6% G + C in red and blue, respectively. Ring 4 shows GC-skew, with values above 0 in purple and those below 0 in green, highlighting the change in the direction of skew at the origin and terminus. Ring five shows repeat sequences, color-coded according to size; and rings six and seven denote the locations of tRNAs and rRNAs, respectively. Reproduced with permission ([12] and additional details found therein).

glossinidia, to 67% G + C in *Pseudomonas aeruginosa*. As G•C basepairs are more thermally stabile than A•T basepairs, it has often been proposed that bacterial base composition is in some way related to growth temperature. However, no such relationship exists [14,15]. Bacterial base compositions principally result from biases in the underlying patterns of mutation, which can differ among species [16,17]. Not surprisingly, there is a significant correlation between growth temperature and the G + C contents in structural RNAs, which must maintain secondary structures through complementary nucleotides; and in such molecules, the increase in the percentage of G + C with temperature is limited to the double-stranded stem regions [14,18].

Although mutational biases account for much of the variation among bacterial base compositions, there is a physiological condition that seems to correspond to differences in G + C content. Aerobic genera tend to have higher G + C content than related anaerobic lineages, a difference observed in several prokaryotic phyla [19•]. However, the reason for this is unclear, and the trend runs opposite to that predicted: oxidative damage to DNA acts primarily at guanine residues with a prevalence of G•C to A•T and G•C to T•A mutations [20], which would cause a decrease in G + C content with aerobiosis.

The wide variation in base composition across bacterial taxa contrasts the fact that base composition is relatively homogeneous among genes within a bacterial genome. Because of this, base composition is used as a guide to the ancestry of a gene, such that gene sequences having atypically high or low G + C contents are likely to have been introduced by lateral gene transfer.

Scanning the circles that display G + C contents, the inflections in base composition have been variously

assigned as phage and other mobile elements, species-specific genes, pathogenicity islands and genes with no known homologs — all features that corroborate their acquisition by lateral gene transfer. The availability of complete sequences of two Salmonellae [21,22] along with genome sequences of closely related *Escherichia coli* [23,24], allows evaluation of the relationship between base composition, phylogenetic distribution and gene function. Among the 1000 or so genes found in *Salmonella* but not in *E. coli*, there is a preponderance of phage genes and known virulence determinants, many of which have relatively low G + C contents [21,22]. Therefore, several features of these genes indicate that lateral transfer is responsible for most of the genomic differences between these enteric species.

There is however, not always a direct correlation between base composition and gene ancestry: genes of common base compositions can be acquired; and similarly, very anomalous sequences can be ancestral. In some low G + C genomes, such as that of *C. perfringens*, the few regions of atypically high G + C content correspond to the rRNA operons, which are virtually never transferred. The problem of determining the ancestry of every gene in a genome is greatly simplified in taxa where the availability of genome sequences for several strains allows indepth phylogenetic analysis of each gene. However, there are still very few bacterial groups for which such information on multiple genomes is available, and the examination of base composition and other sequence features provides an alternative approach for detecting potentially transferred genes, without relying on comparisons to other organisms.

Moving towards the middle

Aside from biases in the number of genes residing on the leading and lagging strands, there are also asymmetries in the nucleotide contents of each DNA strand [25]. In general, there is an excess of guanines (and usually thymines) on the leading strand, a characteristic of genomes usually referred to as 'GC-skew'. The degree of skew is expressed as the difference in the number of guanines and cytosines on a given strand relative to the total number of these residues (G − C/G + C); and computational analyses of the earliest bacterial genome sequences revealed that the direction of skew changed abruptly at the replication origin and terminus [26].

Although a strand-biased distribution of genes could potentially produce differences in nucleotide contents (owing to the selection for certain codons, or to mutational biases introduced by transcription), the degree of GC-skew is not associated with the coding capacity of each strand [8,9]. Instead, GC-skew arises during replication, because the discontinuously synthesized lagging strand undergoes particular errors at higher rates. The distribution of GC-skew around a chromosome now serves as a device to rapidly pinpoint the locations of the replication origin and terminus, without experimental verification. This property is prevalent in bacterial but not archaeal genomes, however.

New insights into the evolutionary processes underlying these asymmetries have come from recent comparisons of homologous genes across bacterial genomes. In such comparisons, it is possible for orthologs to be oriented on the same or different strand, and it appears that the movement of orthologs between strands influences gene distribution and nucleotide asymmetries. When genes switch from one strand to the other (owing to inversions or translocations), they acquire specific mutations and adopt the compositional skew of their new resident strand [27•,28,29•]. Surprisingly, the accumulation of strand-specific mutations actually increases the likelihood that genes moving from the leading to the lagging strand will incur lethal mutations and be eliminated from the population. By reconstructing the ancestral orientation of orthologs in both *Chlamydia* and spirochaetes, Mackiewicz *et al.* [30•] found that in each genome, the relative number of orthologs that switched from the leading to the lagging strand was significantly lower than the number inverted in the opposite direction. Thus, over evolutionary time, the preferential elimination of such genes will produce some of the asymmetries observed in bacterial genomes.

Inner circles

Bacterial genomes can contain several classes of repeated sequences, ranging from short polynucleotide tracts (microsatellites) to large dispersed elements. It is now fairly common for genome papers to note the position(s) of rRNA operons, which range from one to 15 copies in bacteria. There is a loose relationship between the rRNA operon copy number and genome size: intracellular pathogens and symbionts with genomes in the 0.5–1 Mb range, such as *Rickettsia*, *Mycoplasma* and *Buchnera*, each harbor one copy, whereas both *E. coli* and *B. subtilis* have genome sizes of approximately 4.5 Mb and contain seven and 10 copies, respectively. It is not simply that larger genomes can accommodate additional rRNA operons: the copy number appears to be modulated by environmental conditions and resource availability [31,32].

There is also considerable variation in the number of translocatable elements, such as insertion sequences (IS), within bacterial genomes. Although IS copy numbers can vary within species, the genomes of recent pathogens (as opposed to the chronic intracellular pathogens with reduced genomes) have accumulated exceptionally high numbers of translocatable elements. Whereas strains of *E. coli* might harbor up to 50 IS elements, *Shigella flexneri*, a pathogenic derivative of *E. coli*, contains over 300 IS elements, constituting nearly 10% of its 4.6 Mb genome [33].

The presence, maintenance and non-random distribution of repetitive elements in bacterial genomes might imply a functional role to their existence. For example, the Correia repeats (CR) and CR-enclosed elements in pathogenic *Neisseria* occur predominately within the intergenic regions near virulence, metabolic and transporter genes and might impact their regulation and expression [34]. In *E. coli*, stress response genes contain a significantly higher number of short close repeats, which are thought to act as sites of recombination via slipped-strand mispairing during DNA, RNA or protein synthesis thereby increasing phenotypic diversity during stressful conditions [35•]. Similarly, in *Mycoplasma*, phenotypic diversity in adhesin and lipoprotein genes is generated by recombination between repeats [36••].

Chromosomes that contain a high proportion of repetitive elements typically exhibit numerous duplications [37]. In a recent study, Frank *et al.* [38•] tabulated the density of repeats >200 bp in length in many of the fully sequenced genomes. Although they detected little association between genome size and the proportion of repeats, all of the intracellular microbes considered had very low numbers of repetitive elements. They conjectured that the initial reductions in genome size in these organisms occurs through deletions mediated by homologous recombination between repeats, which are eventually eliminated from the genome along with the intervening sequences.

Conclusions

At first glance, sequenced bacterial chromosomes appear to merely illustrate aspects of genome structure and content. But they also provide indepth information about the way that bacterial genomes are organized and have evolved. Whereas it is obvious that we should ascertain the functions of all of the genes encoded by an organism, fresh insights into the inner workings of a genome have also been gained by examining the strand location of genes, the patterns of GC-skew, the distribution of repetitive, transposable and duplicated sequences, and regions of deviant base composition. As the number of sequenced genomes and the use of this information increases, additional rings will certainly be added to the picture. Comparative genomics and microarrays have already supplied new ways of depicting gene repertoires and the relationships among genomes, and genome-wide experimental studies will necessitate the inclusion of information beyond that derived computationally.

References and recommended reading

Papers of particular interest, published within the annual period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, Bult CJ, Tomb JF, Dougherty BA, Merrick JM *et al.*: **Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd**. *Science* 1995, **269**:496-512.

2. Bentley SD, Chater KF, Cerdeno-Tarraga AM, Challis GL, Thomson NR, James KD, Harris DE, Quail MA, Kieser H, Harperv D *et al.*: **Complete genome sequence of the model actinomycete *Streptomyces coelicolor* A3(2)**. *Nature* 2002, **417**:141-147.

3. •• Tamas I, Klasson L, Canback B, Naslund AK, Eriksson AS, Wernegreen JJ, Sandstrom JP, Moran NA, Andersson SG: **50 million years of genomic stasis in endosymbiotic bacteria**. *Science* 2002, **296**:2376-2379.

This comparison of two endosymbiont genomes showed that bacteria sequestered in hosts for millions of years display virtually almost no differences in gene repertoire or genome organization. These bacteria have vastly reduced genomes and have not innovated any new genes as they co-evolved with their insect hosts.

4. Fischer D, Eisenberg D: **Finding families for genomic ORFans**. *Bioinformatics* 1999, **15**:759-762.

5. Skovgaard M, Jensen LJ, Brunak S, Ussery D, Krogh A: **On the total number of genes and their length distribution in complete microbial genomes**. *Trends Genet* 2001, **17**:425-428.

6. Ochman H: **Distinguishing the ORFs from the ELFs: short bacterial genes and the annotation of genomes**. *Trends Genet* 2002, **18**:335-337.

7. Mira A, Klasson L, Andersson SGE: **Microbial genome evolution: sources of variability**. *Curr Opin Microbiol* 2002, **5**:506-512.

8. McLean MJ, Wolfe KH, Devine KM: **Base composition skews, replication orientation, and gene orientation in 12 prokaryote genomes**. *J Mol Evol* 1998, **47**:691-696.

9. Francino MP, Ochman H: **A comparative genomics approach to DNA asymmetry**. *Ann NY Acad Sci* 1999, **870**:428-431.

10. French S: **Consequences of replication fork movement through transcription units *in vivo***. *Science* 1992, **258**:1362-1365.

11. Liu B, Alberts BM: **Head-on collision between a DNA replication apparatus and RNA polymerase transcription complex**. *Science* 1995, **267**:1131-1137.

12. Bao Q, Tian Y, Li W, Xu Z, Xuan Z, Hu S, Dong W, Yang J, Chen Y, Xue Y *et al.*: **A complete sequence of the *T. tengcongensis* genome**. *Genome Res* 2002, **12**:689-700.

13. • Rocha E: **Is there a role for replication fork asymmetry in the distribution of genes in bacterial genomes?** *Trends Microbiol* 2002, **10**:393-395.

The author provides evidence that the possession of PolC correlates with significantly more genes being encoded in the leading strand of bacterial genomes than in the lagging strand. However, there were no significant associations between GC-skew and possession of PolC and only a weak correlation between gene strand biases and GC-skew, suggesting that the different types of replication biases result from different types of structural asymmetry.

14. Galtier N, Lobry JR: **Relationships between genomic G + C content, RNA secondary structures, and optimal growth temperature in prokaryotes**. *J Mol Evol* 1997, **44**:632-636.

15. Hurst LD, Merchant AR: **High guanine-cytosine content is not an adaptation to high temperature: a comparative analysis amongst prokaryotes**. *Proc R Soc Lond Ser B* 2001, **268**:493-497.

16. Sueoka N: **Directional mutation pressure and neutral molecular evolution**. *Proc Natl Acad Sci USA* 1988, **85**:2653-2657.

17. Muto A, Osawa S: **The guanine and cytosine content of genomic DNA and bacterial evolution**. *Proc Natl Acad Sci USA* 1987, **84**:166-169.

18. Wang HC, Hickey DA: **Evidence for strong selective constraint acting on the nucleotide composition of 16S ribosomal RNA genes**. *Nucleic Acids Res* 2002, **30**:2501-2507.

19. • Naya H, Romero H, Zavala A, Alvarez B, Musto H: **Aerobiosis increases the genomic guanine plus cytosine content (GC%) in prokaryotes**. *J Mol Evol* 2002, **55**:260-264.

Although there have been numerous attempts to implicate environmental factors as the cause for differences in the base composition among bacterial genomes, there have been few selective forces appear to operate across distantly related bacteria. The association between an aerobic lifestyle and G + C content reported here is apparent in divergent taxa. But, as the authors note, the reason for this trend remains obscure.

20. Wang D, Kreutzer DA, Essigmann JM: **Mutagenicity and repair of oxidative DNA damage**. *Mut Res* 1998, **400**:99-115.

21. Parkhill J, Dougan G, James KD, Thomson NR, Pickard D, Wain J, Churcher C, Mungall KL, Bentley SD, Holden MT *et al.*: **Complete genome sequence of a multiple drug resistant *Salmonella enterica* serovar Typhi CT18**. *Nature* 2001, **413**:848-852.

22. McClelland M, Sanderson KE, Spieth J, Clifton SW, Latreille P, Courtney L, Porwollik S, Ali J, Dante M, Du F *et al.*: **Complete genome sequence of *Salmonella enterica* serovar Typhimurium LT2**. *Nature* 2001, **413**:852-856.

23. Blattner FR, Plunkett G III, Bloch CA, Perna NT, Burland V, Riley M, Collado-Vides J, Glasner JD, Rode CK, Mayhew GF *et al.*: **The complete genome sequence of *Escherichia coli* K-12**. *Science* 1997, **277**:1453-1474.

24. Perna NT, Plunkett G iii, Burland V, Mau B, Glasner JD, Rose DJ, Mayhew GF, Evans PS, Gregor J, Kirkpatrick HA *et al.*: **Genome sequence of enterohaemorrhagic *Escherichia coli* O157:H7**. *Nature* 2001, **409**:529-533.

25. Karlin S, Brocchieri L, Trent J, Blaisdell BE, Mrazek J: **Heterogeneity of genome and proteome content in bacteria, archaea, and eukaryotes**. *Theor Pop Biol* 2002, **61**:367-390.

26. Lobry JR: **Asymmetric substitution patterns in the two DNA strands of bacteria**. *Mol Biol Evol* 1996, **13**:660-665.

27. • Szczepanik D, Mackiewicz P, Kowalczuk M, Gierlik A, Nowicka A, Dudek MR, Cebrat S: **Evolution rates of genes on leading and lagging DNA strands**. *J Mol Evol* 2001, **52**:426-433.

The authors demonstrate that genes situated on the leading strand have a lower divergence rate than those situated on the lagging strand and that sequences that have recently changed strands are most prone to mutational change. They also discuss the possible effects that this phenomenon might have on phylogenetic analysis.

28. Rocha EPC, Danchin A: **Ongoing evolution of strand composition in bacterial genomes**. *Mol Biol Evol* 2001, **18**:1789-1799.

29. • Dalevi DA, Eriksen N, Eriksson K, Andersson SGE: **Measuring genome divergence in bacteria: a case study using *Chlamydian* data**. *J Mol Evol* 2002, **55**:24-36.

These authors demonstrate that gene order differences between *Chlamydia* spp. is predominately caused by rearrangement events within the genome. Additionally, the authors find no evidence for horizontal gene transfer events in generating the observed genome divergence.

30. • Mackiewicz P, Mackiewicz D, Gierlik A, Kowalczuk M, Nowicka A, Dudkiewicz M, Dudek MR, Cebrat S: **The differential killing of genes by inversions in prokaryotic genomes**. *J Mol Evol* 2001, **53**:615-621.

An interesting finding that the asymmetry in mutation rates on the two strands of DNA could affect the maintenance of certain genomic rearrangements.

31. Condon C, Liveris D, Squires C, Schwartz I, Squires CL: **rRNA operon multiplicity in *Escherichia coli* and the physiological implications of *rrn* inactivation**. *J Bacteriol* 1995, **177**:4152-4156.

32. Klappenbach JA, Dunbar JM, Schmidt TM: **rRNA operon copy number reflects ecological strategies of bacteria**. *Appl Environ Microbiol* 2000, **66**:1328-1333.

33. Jin Q, Yuan Z, Xu J, Wang Y, Shen Y, Lu W, Wang J, Liu H, Yang J, Yang F: **Genome sequence of *Shigella flexneri* 2a: insights into pathogenicity through comparison with genomes of *Escherichia coli* K12 and O157**. *Nucleic Acids Res* 2002, **30**:4432-4441.

34. Liu SV, Saunders NJ, Jeffries A, Rest RF: **Genome analysis and strain comparison of Correia repeats and Correia repeat-enclosed elements in pathogenic *Neisseria***. *J Bacteriol* 2002, **184**:6163-6173.

35. • Rocha EPC, Matic I, Taddei F: **Over-representation of repeats in stress response genes: a strategy to increase versatility under stressful conditions?** *Nucleic Acids Res* 2002, **30**:1886-1894.

The authors demonstrate that a high number of short nearby repeats are present in the stress response genes of *Escherichia coli* and that these sites might induce phenotypic variability in these genes. Other types of repeats, such as long repeats that would be capable of homologous recombination, are almost absent from these genes, suggesting that homologous recombination between *E. coli* genes might not play a significant role in generating genomic variability.

36. •• Rocha EPC, Blanchard A: **Genomic repeats, genome plasticity and the dynamics of *Mycoplasma* evolution**. *Nucleic Acids Res* 2002, **30**:2031-2042.

The importance of repeats in the evolution of bacterial genomes is considered by examining *Mycoplasma*, which possess some of the smallest genomes examined to date. The authors find that repeats are a common feature of *Mycoplasma* genomes and conclude that repeats drive genome evolution in *Mycoplasma* by serving as recombinational hot spots.

37. Achaz G, Rocha EPC, Netter P, Coissac E: **Origin and fate of repeats in bacteria**. *Nucleic Acids Res* 2002, **30**:2987-2994.

38. • Frank AC, Amiri H, Andersson SGE: **Genome deterioration: loss of repeated sequences and accumulation of junk DNA**. *Genetica* 2002, **115**:1-12.

By comparing the number and distribution of repeat sequences in numerous sequenced genomes, the authors search for genetic and ecological correlates, and potentially the function of repeats, in bacterial genomes.