microbiology Comment

Genome Update: DNA repeats in bacterial genomes

Genomes of the month – close relatives

Four new bacterial genomes have been deposited with GenBank/EMBL/DDBJ since last month's Genome Update. The list is shown in Table 1 and represents a wide range of organisms, with three of the four organisms having close relatives that have already been sequenced: the saprophytic soil bacterium *Bacillus licheniformis*, which is a close cousin of *Bacillus subtilis*; *Yersinia pseudotuberculosis*, which is thought to represent the species from which *Yersinia pestis* originated; the spirochaete *Borrelia garinii*; and the compost bacterium *Symbiobacterium thermophilum*, which depends on the presence of other *Bacillus* bacteria for growth (Ueda *et al.*, 2001). At the time of writing, reports on the *Bacillus licheniformis* and *Y. pseudotuberculosis* genomes have been published, which will be briefly discussed below.

The *Bacillus licheniformis* ATCC 14580^T genome is 4·2 Mbp long, has an AT content of 54·8 % and encodes 4208 predicted protein-encoding genes (Rey *et al.*, 2004). *Bacillus licheniformis* is a close relative of *Bacillus subtilis*, which is the widely used model organism for Gram-positive bacteria. Roughly 80 % of the predicted coding sequences of *Bacillus licheniformis* have identified orthologues among the genes of *Bacillus subtilis*. The two genomes are approximately the same size and have a similar composition. *Bacillus licheniformis* has been used for many years in the industrial production of various enzymes, antibiotics and other products of commercial interest and, like *Bacillus subtilis*, it is usually non-pathogenic and easy to grow. It lacks, however, the ability to actively take up DNA (natural competence), despite the fact that it harbours orthologues of most of the *Bacillus subtilis* genes known to be involved in the mechanisms of competence.

The strain of *Y. pseudotuberculosis* that has been sequenced is IP32953, an isolate from a human patient (Chain *et al.*, 2004). This specific *Y. pseudotuberculosis* strain was used for comparison with the genomes of two *Y. pestis* strains (KIM 10 and CO92). The difference between the mildly pathogenic progenitor *Y. pseudotuberculosis* and the deadly *Y. pestis*, responsible for wiping out about half of the population of Europe during the Middle Ages, appears to be, on the one hand, '*the addition of a mere 32 new chromosomal genes and two plasmids specific to* Y. pestis' and, on the other hand, the absence of several hundred genes: 317 genes are missing in *Y. pestis* compared with *Y. pseudotuberculosis*, in addition to the 149 pseudogenes in the *Y. pestis* strains (Chain *et al.*, 2004).

A third *Y. pestis* (strain 91001) genome has also been published recently (Song *et al.*, 2004) and deposited with GenBank/EMBL/DDBJ, as discussed in a previous Genome Update column (Ussery & Hallin, 2004). *Y. pestis* strain 91001 was isolated from the rodent Brandt's vole (*Microtus brandti*), and is avirulent to humans. The genome consists of one main 4·6 Mbp chromosome and an additional 200 kbp in four plasmids (pPCP1, pCD1, pMT1 and pCRY). Three of the four plasmids are similar, whereas pCRY is a novel plasmid discovered in this work and encodes a cryptic type IV secretion system. The genome has 4037 predicted genes, with 141 of them being pseudogenes; note that this number is close to that found in the two other *Y. pestis* genomes. The chromosomal organization of the new strain shows many rearrangements compared with the other *Y. pestis* sequenced strains CO92 and KIM 10. The non-pathogenicity and host-specificity of strain 91001 may be explained through the large genome deletion in the chromosome or/and some pseudogenes. We wanted to see which of the 32 genes are found in the other two *Y. pestis* genomes but absent in the *Y. pseudotuberculosis* genome. However, upon closer examination of the report and supplementary material by Chain *et al.* (2004), we could not determine specifically which 32 genes were referred to in the abstract for the article. With the additional sequence of a FIFTH *Yersinia* genome (*Yersinia enterocolitica*) available from the Sanger Centre (see http://www.sanger.ac.uk/Projects/Y_enterocolitica/), there is much work to do in terms of genomic comparisons within *Yersinia* species.

Method of the month – comparison of levels of DNA repeats in bacterial genomes

This month we will discuss comparing the number of DNA repeats in sequenced bacterial genomes. It has been known for more than 30 years that the DNA in eukaryotic cells contains large stretches of

***Microbiology* Comment** provides a platform for readers of *Microbiology* to communicate their personal observations and opinions in a more informal way than through the submission of papers.

Most of us feel, from time to time, that other authors have not acknowledged the work of our own or other groups or have omitted to interpret important aspects of their own data. Perhaps we have observations that, although not sufficient to merit a full paper, add a further dimension to one published by others, or we may have a useful piece of methodology that we would like to share.

Guidelines on how to submit a *Microbiology* Comment article can be found in the Instructions for Authors at http://mic.sgmjournals.org

It should be noted that the Editors of *Microbiology* do not necessarily agree with the views expressed in *Microbiology* Comment.

Chris Thomas, Editor-in-Chief

Table 1. Summary of the published genomes discussed in this Update

Note that the accession number for each chromosome is the same for GenBank, EMBL and the DDBJ.

Name	Length (bp)	AT content (%)	No. of genes	tRNAs	rRNAs	Accession no.
Bacillus licheniformis ATCC 14580^T	4 222 336	53·8	4208	72	7	CP000002
Borrelia garinii PBi	904 246	71·7	832	33	1	CP000013
Yersinia pseudotuberculosis IP32953	4 744 671	52·4	3974	85	7	BX936398
Symbiobacterium thermophilum IAM 14863^T	3 566 135	31·3	3337	98	0	AP006840

short repeats, whilst DNA from bacterial cells contains few repeats. As mentioned in a previous Genome Update, the size of bacterial genomes ranges approximately 20-fold, from about 500 000 bp to around 10 000 000 bp. Many eukaryotic genomes, on the other hand, can be from 1000 times to 1 000 000 times larger (i.e. in the range of a billion base pairs for plants and animals, to about a trillion base pairs for some amoebae), although the number of genes is roughly in the range of only tenfold more for eukaryotes than prokaryotes. The likely explanation is that bacteria have had plenty of time (replication cycles) to streamline their DNA, such that most of the bacterial genome encodes proteins, whilst only a tiny part of many animal and plant genomes contains protein-encoding sequences. So bacteria are different in that they generally have fewer repeats, and these repeats often have some selective advantage, or 'reason' for their existence; one such reason being to enhance diversity, by generating a population of slightly different sequences for a given locus.

We search for global direct repeats by taking a 100 nt window and determining the best DNA sequence match within the entire chromosome. This value is stored at nucleotide position 50, then a new window is chosen, moving over 1 nt, and the processes are repeated until the entire chromosome is searched, with only a gap of 50 nt on either edge not having assigned values (Skovgaard *et al.*, 2002).

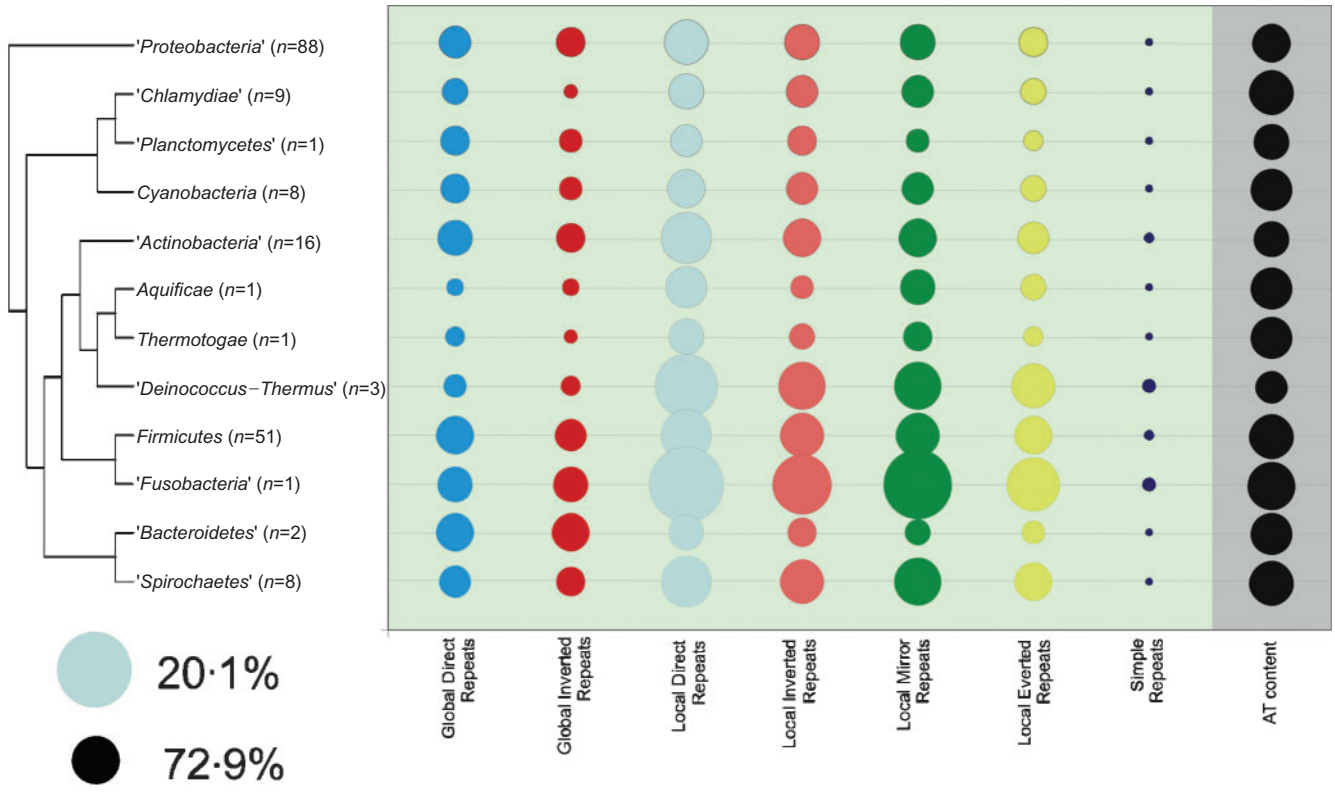

Fig. 1. Comparison of levels of DNA repeats in 189 bacterial genomes. The area of the circle represents the fraction of repeats in the genome, and the largest coloured circle reflects about 20 % repeats – i.e. 20 % of the genome has repeats such that the best match in the genome for a given window size is at least 80 %. The column on the far right, with black circles, gives the AT content, which can affect the level of repeats. See the text for additional details. The number of sequenced genomes in each phylum is given beside the 16S rRNA gene-based phylogenetic tree on the left of the diagram.

Inverted repeats are searched for in a similar manner, except the match is looked for on the opposite strand. Local repeats are determined by finding the best match for a 15 nt region, within a 100 bp window, as described previously (Jensen *et al.*, 1999). There are four possible types of repeats: direct and inverted, as described above; mirror repeats (where the DNA sequence is inverted, on the same strand, in a true palindrome); and everted repeats, where the 5′ to 3′ DNA sequence is repeated on the other strand, in the 3′ to 5′ direction (Jensen *et al.*, 1999). Local direct repeats can form slipped–mispaired structures, resulting in deletions or duplications. Inverted repeats can form cruciforms and stem–loop structures, whilst certain mirror repeats can form triple-stranded DNA and everted repeats can form parallel-stranded helices. Finally, we also measure 'simple repeats', which consist of short tandem repeats, such as homopolymer tracts (van Belkum *et al.*, 1998).

The mean values for the above-mentioned seven different types of repeats are shown in Fig. 1, for 12 different bacterial phyla. Note that the number of genomes per phylum varies considerably, and that although the '*Fusobacteria*' have the largest fraction of repeats, this is from only one genome, whilst on the other hand there are currently 88 proteobacterial genomes in our database. However, in spite of the lack of a good spread of genomes amongst the phyla, some general trends can still be seen. In general, the levels of global repeats are quite low, with only two phyla having more than 5 % of the genome with repeats of 80 % or greater for 100 bp windows. (In many eukaryotic genomes, this number is more than 50 %.) There also seem to be higher levels of direct repeats than inverted repeats, both at the global and local levels. Finally, the level of short 'simple repeats' is quite low, less than 1 % for nearly all the genomes examined ('*Fusobacteria*' and '*Deinococcus–Thermus*' both have 1·1 % – the highest values, but again note that these are only a few genomes). The local and simple repeats are more likely to occur in genomes that are either very GC-rich or AT-rich, since the probability of finding a given unique base goes from 1 in 4 for 50 % AT content to only 1 in 2 for a genome with 100 % AT or 100 % GC. A link to a detailed table giving values for all sequenced bacterial chromosomes can be found on the supplemental web page.

Supplemental web pages

Web pages containing supplemental material related to this article can be accessed from the following url: http://www.cbs.dtu.dk/services/GenomeAtlas/suppl/GenUp010/

Acknowledgements

This work was supported by a grant from the Danish Center for Scientific Computing.

David W. Ussery, Tim T. Binnewies, Rodrigo Gouveia-Oliveira, Hanne Jarmer and Peter F. Hallin

Center for Biological Sequence Analysis, Department of Biotechnology, Building 208, The Technical University of Denmark, DK-2800 Kgs. Lyngby, Denmark

Correspondence: David W. Ussery (dave@cbs.dtu.dk)

Chain, P. S. G., Carniel, E., Larimer, F. W. & 20 other authors (2004). Insights into the evolution of *Yersinia pestis* through whole-genome comparison with *Yersinia pseudotuberculosis*. *Proc Natl Acad Sci U S A* **101**, 13826–13831.

Jensen, L. J., Friis, C. & Ussery, D. W. (1999). Three views of microbial genomes. *Res Microbiol* **150**, 773–777.

Rey, M. W., Ramaiya, P., Nelson, B. A. & 18 other authors (2004). Complete genome sequence of the industrial bacterium *Bacillus licheniformis* and comparisons with closely related *Bacillus* species. *Genome Biol* **5**, r77.

Skovgaard, M., Jensen, L. J., Friis, C., Stærfeldt, H.-H., Worning, P., Brunak, S. & Ussery, D. W. (2002). The atlas visualisation of genome-wide information. *Methods Microbiol* **33**, 49–63.

Song, Y., Tong, Z., Wang, J. & 26 other authors (2004). Complete genome sequence of *Yersinia pestis* strain 91001, an isolate avirulent to humans. *DNA Res* **11**, 179–197.

Ueda, K., Ohno, M., Yamamoto, K. & 8 other authors (2001). Distribution and diversity of symbiotic thermophiles, *Symbiobacterium thermophilum* and related bacteria, in natural environments. *Appl Environ Microbiol* **67**, 3779–3784.

Ussery, D. W. & Hallin, P. F. (2004). Genome Update: annotation quality in sequenced microbial genomes. *Microbiology* **150**, 2015–2017.

van Belkum, A., Scherer, S., van Alphen, L. & Verbrugh, H. (1998). Short-sequence DNA repeats in prokaryotic genomes. *Microbiol Mol Biol Rev* **62**, 275–293.

DOI 10.1099/mic.0.27628-0