microbiology Comment

Genome Update: correlation of bacterial genomic properties

Genomes of the month – 14 new genomes!

Fourteen new microbial genomes have been published since last month's Genome Update was written. Since a discussion of so many genomes would take too much space, only a few select genomes will be discussed in detail; the others are listed in Table 1 and mentioned only briefly. The genomes range in size from 0·6 Mbp for a polydnavirus [*Cotesia congregata bracovirus* (CcBV); Espagne *et al.*, 2004], about the same size as the smallest bacterial genome, to about 34 Mbp for a diatom (*Thalassiosira pseudonana*). The remaining dozen prokaryotic genomes include that of an archaeon (*Methanococcus maripaludis* strain S2; Hendrickson *et al.*, 2004) and those of 11 bacteria: seven *Proteobacteria* (discussed below), two *Firmicutes* (*Bacillus*), one *Bacteroides* species (*Bacteroides fragilis* strain YCH46, Kuwahara *et al.*, 2004) and an actinobacterium (*Nocardia farcinica* strain IFM 10152, a clinical isolate; Ishikawa *et al.*, 2004). In addition, two spirochaete genomes have been published (Qiu *et al.*, 2004) but, as discussed below, these are not included in Table 1.

Aristotle divided life into three categories: animal, plants and minerals. Where does the diatom *T. pseudonana* belong? The answer is easy – it is 'yes'. Yes, it is a plant since it can do photosynthesis. Yes, it is an animal in that it has locomotion, and yes, it is a mineral due to its fancy silica skeleton. Thus, diatoms have been a bit of an enigma, in terms of classification. *T. pseudonana* actually has three genomes – its 'main' nuclear genome, consisting of 35·5 Mbp, spread out over 24 chromosomes, a plastid genome of about 128 kbp and a mitochondrial genome of 34 kbp (Armbrust *et al.*, 2004). The plastid genome is not a chloroplast (derived from an endocytosed photosynthetic prokaryote), but rather reflects an acquired photosynthetic eukaryotic alga. Diatoms are relatively recent organisms, on a geological scale, dating back to a mere 180 million years ago (Pennisi 2004), compared with the more than 3 billion-year history of prokaryotes. Their skeletons make up part of 'diatomaceous earth', which is used in making industrial filters, and also Qiagen columns that many of the readers of this column might be familiar with, for use of purification of DNA.

Most of this month's bacterial genomes (seven of eleven) are from the class *Proteobacteria*. Two different *Burkholderia* species have been sequenced, as discussed below. There are also two examples in this month's collection of genomes in Table 1 where different strains of the same species have been sequenced by different groups (on different continents) and published, with essentially no reference to each other's work. Three of this month's seven proteobacterial genomes are various *Legionella pneumophila* strains. *L. pneumophila* Philadelphia 1^T has been sequenced by a group based in the USA (Chien *et al.*, 2004), while the genomes of *L. pneumophila* strains Lens and Paris were published about a week later by a French group (Cazalet *et al.*, 2004). Similarly, the genome of *Bacillus licheniformis* strain ATCC 14580^T (Veith *et al.*, 2004) was published within a few weeks of that of *Bacillus licheniformis* strain DSM 13^T (Rey *et al.*, 2004), discussed in last month's Genome Update. While it is again beyond the scope of this column to do a more comprehensive comparison, it is felt that a brief overview of the two genomes side by side is useful. The two remaining proteobacterial genomes are those of *Methylococcus capsulatus* strain Bath (Ward *et al.*, 2004) and '*Mannheimia succiniciproducens*' strain MBEL55E (Hong *et al.*, 2004); another *Mannheimia* species (*Mannheimia haemolytica* strain PHL213) is being sequenced at Baylor University (http://www.hgsc.bcm.tmc.edu/projects/microbial/Mhaemolytica/). In addition to the 11 bacterial genomes in Table 1, two new spirochaete genomes have been published (*Borrelia burgdorferi* strains N40 and JD1; Qiu *et al.*, 2004), although these are not included in our list as the DNA sequences for the two genomes were not available at the time of writing. The *Borrelia burgdorferi* genome consists of a main linear chromosome of around 0·9 Mbp, along with another 0·6 Mbp of DNA from about 20 plasmids. The DNA sequences for the main linear chromosomes of strains N40 and JD1 are not finished; although the published manuscript (Qiu *et al.*, 2004) compares the genomes of the two strains, the 267 GenBank files cited in the report contain mostly plasmid sequences, roughly 20 plasmids each for N40 and JD1, and little genomic DNA. The remaining 220 or so GenBank files are for plasmids from many other *Borrelia burgdorferi* strains, making

***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. Only the main chromosome(s) are shown; plasmids are excluded. The genome of *Bacillus licheniformis* ATCC 14580^T from last month's Update is included for comparison with that of *Bacillus licheniformis* DSM 13^T.

Name/strain	Length (bp)	A+T content (%)	No. of genes	tRNAs	rRNAs	Accession no.
Bacillus cereus ZK	5 300 915	64·7	5 134	96	13	CP000001
Bacillus licheniformis ATCC 14580^T	4 222 336	53·8	4 208	72	7	CP000002
Bacillus licheniformis DSM 13^T	4 222 748	53·8	4 286	72	7	AE017333
Bacteroides fragilis YCH46 Main	5 277 274	56·7	4 578	74	6	AP006841
Burkholderia mallei ATCC 23344^T	5 835 527	31·5	4 764	56	3	
Chromosome 1	3 510 148	31·9	2 996	47	1	CP000010
Chromosome 2	2 325 379	31·0	1 768	9	2	CP000011
Burkholderia pseudomallei K96243	7 247 547	31·9	5 855	60	4	
Chromosome 1	4 074 542	32·3	3 460	53	3	BX571965
Chromosome 2	3 173 005	31·5	2 395	7	1	BX571966
Legionella pneumophila Lens	3 345 687	61·6	2 947	43	3	CR628337
Legionella pneumophila Paris	3 503 610	61·6	3 082	43	3	CR628336
Legionella pneumophila Philadelphia 1^T	3 397 754	61·7	2 942	43	3	AE017354
'*Mannheimia succiniciproducens*' MBEL55E	2 314 078	57·5	2 384	60	6	AE016827
Methanococcus maripaludis S2	1 661 137	66·9	1 722	38	3	BX950229
Methylococcus capsulatus Bath	3 304 697	36·4	3 120	46	2	AE017282
Nocardia farcinica IFM 10152 Main	6 021 225	29·2	5 674	53	3	AP006618
Polydnavirus CcBV	567 670	66·1	156	7	0	AJ632304–33
Thalassiosira pseudonana CCMP1335	34 500 000	53·0	11 242	131	–	AAFD01000000

analysis of the data both more interesting for those seeking information about *Borrelia* diversity and a bit frustrating for those of us who want to have the sequence of each bacterial chromosome as one contiguous piece of DNA.

The genomes of two members of the genus *Burkholderia* (belonging to the *β-Proteobacteria*) have been published recently (Nierman *et al.*, 2004; Holden *et al.*, 2004). *Burkholderia mallei* is the causative agent of equine glanders, an acute infection characterized by either pneumonia and necrosis of the tracheobronchial tree if the organism is inhaled, or pustular skin lesions, multiple abscesses and sepsis if the skin is the portal of entry. Melioidosis, the disease caused by *Burkholderia pseudomallei*, is an endemic disease in northern Australia and eastern Asia. Melioidosis is characterized by a broad spectrum of clinical manifestations, ranging from asymptomatic colonization to fulminant sepsis. In contrast to *Burkholderia mallei*, which is an obligate parasite of horses, mules and donkeys, with no other known natural reservoir, *Burkholderia pseudomallei* is a saprophytic organism broadly distributed in water and soil in its endemic regions. Both organisms are listed as Category B agents on the Centers for Disease Control Bioterrorism Agents/Diseases list (http://www.bt.cdc.gov/agent/agentlist-category.asp). Interestingly, Godoy *et al.* (2003) recently showed that, based on results obtained using multilocus sequence typing, *Burkholderia mallei* should be considered as a clone of *Burkholderia pseudomallei*.

The large genomes of both organisms are organized in two replicons (Table 1). In both genomes, the smaller chromosome contains essential metabolic genes, making it indispensable. The *Burkholderia mallei* genome is characterized by the presence of numerous insertion sequences (IS), which are instrumental in mediating genome alterations (deletions, insertions and inversions). *Burkholderia mallei* also contains a tremendous number of simple sequence repeats, which may play an important role in altered protein expression or structure variation. In contrast to *Burkholderia mallei*, the *Burkholderia pseudomallei* genome contains fewer IS elements, but it contains many genomic islands with properties that suggest that they were recently acquired by horizontal gene transfer. When comparing both genomes, it becomes obvious that 1446 genes [627 on chromosome 1 (16 %) and 819 on chromosome 2 (31 %)] present in *Burkholderia pseudomallei* are absent from *Burkholderia mallei* (in comparison, only about 1 % of the genes on chromosome 1 and 4 % of the genes on choromosome 2 in *Burkholderia mallei* are absent from *Burkholderia pseudomallei*). In addition, the disruption by IS-mediated insertions or frameshift mutations in a few pseudogenes in *Burkholderia mallei* results in marked phenotypic differences between both organisms (e.g. differences in motility and secretion capacity). These observations are consistent with the dual existence of *Burkholderia pseudomallei* (as soil-colonizer and human pathogen) and the highly specialized nature of *Burkholderia mallei* (as an intracellular parasite).

A large selection of genes modulating pathogenicity and host–cell interactions were found in the *Burkholderia pseudomallei* genome. These include flagella, type III secretion systems, surface proteins and drug resistance determinants. In *Burkholderia mallei*, numerous genes for non-ribosomal peptide synthesis and

polyketide synthases were found and these genes could be involved in toxin production. In addition, comparative genome hybridization with multiple *Burkholderia mallei* strains (both virulent and avirulent) identified many more putative virulence genes, including type IV pilus biosynthesis genes and genes involved in capsule biosynthesis.

At present the genomes of many other *Burkholderia* species are being sequenced. These include *Burkholderia thailandensis* (a non-pathogenic close relative of *Burkholderia mallei* and *Burkholderia pseudomallei*), multiple strains (including representatives of the epidemic ET12 and PHDC lineages) of *Burkholderia cenocepacia* (an opportunistic pathogen) and several strains with special biodegradation capabilities (including *Burkholderia xenovorans* LB400^T and *Burkholderia vietnamiensis* G4). Together with the published sequences of *Burkholderia mallei* and *Burkholderia pseudomallei*, these sequences will teach us a lot more about the biology of this interesting group of organisms.

The first discovery of a bacterium of the genus *Legionella* came in 1976 when an outbreak of pneumonia at an American Legion convention in Philadelphia led to 29 deaths. Infection with *Legionella pneumophila* results mainly in sporadic and epidemic cases of Legionnaire's Disease. The genome of this pathogen has been sequenced (Chien *et al.*, 2004). The genome of *L. pneumophila* Philadelphia 1^T consists of a single circular chromosome of 3 397 754 bp and a plasmid-like element of 45 kbp (pLP45) that can exist in both chromosomal and episomal forms. A set of genes which might explain the ability of *Legionella* species to survive in so many different environments was also described (Chien *et al.*, 2004). A comparison of the genome of *L. pneumophila* Philadelphia 1^T with that of *Coxiella burnetii* (belonging to the order '*Legionellales*') shows that *L. pneumophila* Philadelphia 1^T shares ~42 % of its genes with *C. burnetii* even though there are big differences in the genome sizes (3·4 and 1·9 Mbp).

In addition to the *L. pneumophila* Philadelphia 1^T genome sequence, the sequences of two additional *L. pneumophila* strains (Lens and Paris) have been reported by Cazalet *et al.* (2004). *L. pneumophila* Paris and *L. pneumophila* Lens each contain one circular chromosome (3 503 610 bp, 3077 genes and 3 345 687 bp, 2932 genes, respectively) with an A + T content of 62 %. Strains Paris and Lens both contain one plasmid (131 885 bp and 59 832 bp, respectively). By comparison of the two different *Legionella* chromosomes, genome plasticity can readily be seen – one chromosome contains an insertion, and the *L. pneumophila* Paris plasmid is almost twice the size of the plasmid in strain Lens. The two *L. pneumophila* chromosomes exhibit a conserved backbone of 2664 genes, but have around 10 to 15 % strain-specific genes, compared with only 2 % strain-specific genes in *Salmonella typhi* (Cazalet *et al.*, 2004). The complete *dot* and *icm* loci, which together direct assembly of a type IV secretion apparatus and a second type IV secretion system, is encoded by the *lvh* region (Tat, type I and type II secretion systems are also present). In addition to this, only *L. pneumophila* Paris contains a type V secretion system. The conjugative transfer, mediated by the type IV secretion system, of plasmids and chromosomal DNA is, for example, one mechanism in *L. pneumophila* that contributes to the genome plasticity. Analysis of the two *Legionella* chromosomes shows extensive genome plasticity and diversity. In addition, we have grouped all similar proteins of the three different strains into clusters of homologues. The number of clusters having proteins shared by any combination of the three strains is shown in Fig. 1.

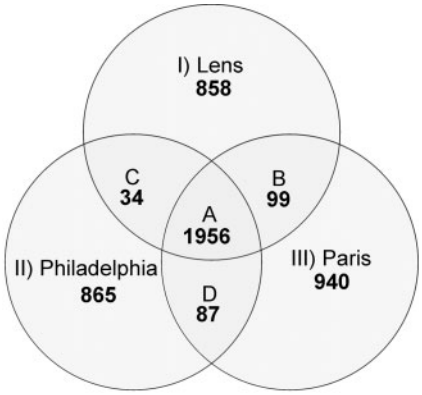

Fig. 1. Comparison of shared protein clusters in *L. pneumophila* strains between (A) all three strains, (B) Lens and Paris strains, (C) Lens and Philadelphia 1^T strains and (D) Philadelphia 1^T and Paris strains. The remaining proteins that are not clustered are placed in groups I, II and III for Lens, Philadelphia 1^T and Paris, respectively.

This month two genomes of the Gram-positive, spore-forming bacteria belonging to the genus *Bacillus* have been released: the third sequenced genome of *Bacillus cereus* (*Bacillus cereus* ZK, GenBank/EMBL/DDBJ accession no. CP000001) and the second sequenced genome of *Bacillus licheniformis* (*Bacillus licheniformis* DSM 13^T; Veith *et al.*, 2004).

The genome sequence of *Bacillus cereus* ZK is 5 300 915 bp long, and by size this places the isolate in the middle of the two previously sequenced genomes with approximately 75 more kilobases than strain ATCC 10987 and 110 fewer kilobases than strain ATCC 14579^T. The AT content of 64·7 % is close to the same level for all three isolates. *Bacillus cereus* ZK has 96 predicted tRNA genes, while the two other strains have 98 (ATCC 10987) and 108 (ATCC 14579^T). Both ATCC 14579^T and ZK have 13 predicted rRNA encoding operons, whereas ATCC 10987 only has 12. With a total of 5134 predicted genes, the ZK strain has fewer predicted genes than the two other strains – 100 genes fewer than ATCC 14579^T and as many as 469 fewer than ATCC 10987. *Bacillus cereus* is a close relative of the pathogenic species *Bacillus anthracis* and *Bacillus thuringiensis*, and its spores are widespread in soil and air, often leading to the contamination of cereals. *Bacillus cereus* is frequently observed multiplying in foods such as cooked rice and may lead to food poisoning (Kotiranta *et al.*, 2000). It is also highly motile and produces a variety of different toxins and antibiotics.

The genome of *Bacillus licheniformis* DSM 13^T consists of 4 222 748 bp, 412 bp more than that of the recently sequenced strain *Bacillus licheniformis* ATCC 14580^T (Rey *et al.*, 2004). Both strains are predicted to have 72 tRNA genes, seven rRNA operons and an AT content of 53·8 %. Strain DSM 13^T is predicted to contain 4286 genes, 78 more than predicted for strain ATCC 14580^T. The two research groups have used different software and strategies to make

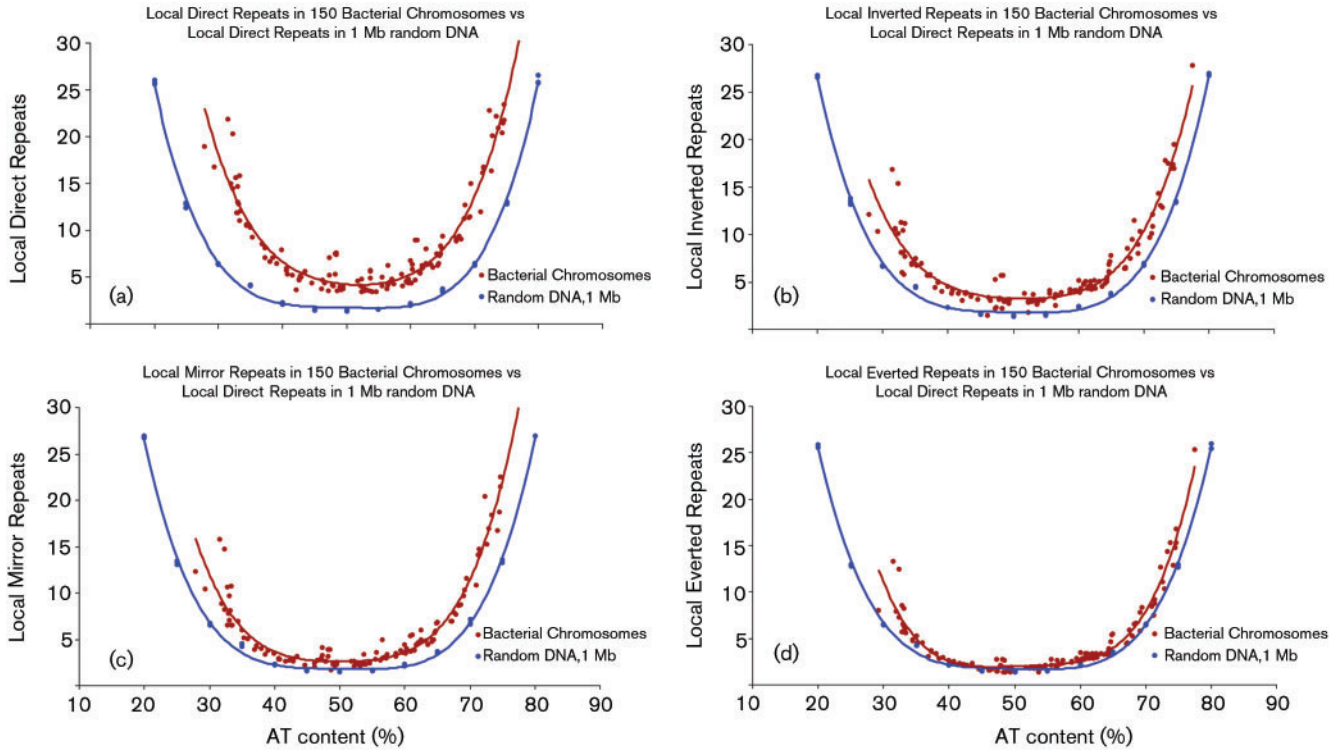

Fig. 2. Comparison of A+T content and levels of local DNA repeats in 150 bacterial genomes. The blue line represents the repeat levels found in 1 Mbp of generated random DNA of a fixed A+T content, and the red line represents the best line through the experimentally observed data (each dot represents the value for a single chromosome). Note that the lowest levels of repeats are, in general, found for genomes of around 50 % A+T content, whilst genomes with either very low or very high A+T content levels tend to exhibit maximal levels of repeats.

their predictions, which may explain at least part of this difference. *Bacillus licheniformis*, unlike *Bacillus cereus*, belongs to the non-pathogenic branch of the genus *Bacillus*. It is closely related to *Bacillus subtilis* and *Bacillus halodurans* and is widely used in industrial processes due to its remarkable ability to produce and secrete proteins at high levels.

Method of the month – correlation of bacterial genomic properties

With so many genomes being published, there is a need for methods of looking at relationships between hundreds of genomes, in addition to comparisons of two or three genomes at a time. This month we will discuss how one can make their own scatter plot to compare chosen parameters of nearly 200 bacterial genomes against each other. As an example, in last month's Genome Update (Ussery *et al.*, 2004) we introduced a bubble diagram to visualize seven different kinds of repeats and their average fraction in the different phyla of bacterial chromosomes. The observed variations can be explained, to a certain extent, by differences in the A+T content of the various genomes. As the A+T content shifts away from 50 %, the nucleotide alphabet changes from four letters towards two letters (100 % A+T or G+C), thereby increasing the probability of repeated sequences. This month we will discuss methods to visualize how changes in A+T content affect the different repeat levels. In Fig. 2, we have plotted A+T content on the *x*-axis and the different types of repeats along the *y*-axis. Local direct repeats (a), local inverted repeats (b), local mirror repeats (c) and local everted repeats (d) are shown for chromosomes of the Genome Atlas Database and for 1 Mb fully randomized DNA sequences with different A+T contents.

For the different kinds of local repeats it is evident that there is a strong correlation between the A+T content and the percentage of repeats. For local direct repeats this is significantly stronger in the whole range of A+T content compared with randomized DNA. One biological explanation is that many genes or operons display similar short domains for similar components of the final protein. This, naturally, does not occur in randomized DNA. Such correlations are likely to affect many other bioinformatic results and we have developed an online tool available on our genome database web pages. This tool allows users to study over 40 different numerical values extracted from the CBS Genome Atlas Database (Hallin & Ussery, 2004) by drawing scatter plots, coloured according to the phyla of the organisms (http://www.cbs.dtu.dk/services/GenomeAtlas/show-compare.php?kingdom=Bacteria).

Supplemental web pages

Web pages containing material related to this article can be accessed from the following url: http://www.cbs.dtu.dk/services/GenomeAtlas/suppl/GenUp011/

Acknowledgements

This work was supported by a grant from the Danish Center for Scientific Computing. T. C. is indebted to the Fund for Scientific Research – Flanders (Belgium) for a position as postdoctoral fellow.

Peter F. Hallin,[1] Tom Coenye,[2] Tim T. Binnewies,[1] Hanne Jarmer,[1] Hans-Henrik Stærfeldt[1] and David W. Ussery[1]

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

[2]Laboratorium voor Microbiologie, Universiteit Gent, Ledeganckstraat 35, B-9000 Gent, Belgium

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

Armbrust, E. V., Berges, J. A., Bowler, C. & 42 other authors (2004). The genome of the diatom *Thalassiosira pseudonana*: ecology, evolution, and metabolism. *Science* **306**, 79–86.

Cazalet, C., Rusniok, C., Bruggemann, H. & 11 other authors (2004). Evidence in the *Legionella pneumophila* genome for exploitation of host cell functions and high genome plasticity. *Nat Genet* Epub ahead of print, doi:10.1038/ng1447.

Chien, M., Morozova, I., Shi, S. & 34 other authors (2004). The genomic sequence of the accidental pathogen *Legionella pneumophila*. *Science* **305**, 1966–1968.

Espagne, E., Dupuy, C., Huguet, E., Cattolico, L., Provost, B., Martins, N., Poirie, M., Periquet, G. & Drezen, J. M. (2004). Genome sequence of a polydnavirus: insights into symbiotic virus evolution. *Science* **306**, 286–289.

Godoy, D., Randle, G., Simpson, A. J., Aanensen, D. M., Pitt, T. L., Kinoshita, R. & Spratt, B. G. (2003). Multilocus sequence typing and evolutionary relationships among the causative agents of melioidosis and glanders, *Burkholderia pseudomallei* and *Burkholderia mallei*. *J Clin Microbiol* **41**, 2068–2079.

Hallin, P. F. & Ussery, D. (2004). CBS genome atlas database: a dynamic storage for bioinformatic results and sequence data. *Bioinformatics* Epub ahead of print, doi:10.1093/bioinformatics/bth423.

Hendrickson, E. L., Kaul, R., Zhou, Y. & 28 other authors (2004). Complete genome sequence of the genetically tractable hydrogenotrophic methanogen *Methanococcus maripaludis*. *J Bacteriol* **186**, 6956–6969.

Holden, M. T., Titball, R. W., Peacock, S. J. & 45 other authors (2004). Genomic plasticity of the causative agent of melioidosis, *Burkholderia pseudomallei*. *Proc Natl Acad Sci U S A* **101**, 14240–14245.

Hong, S. H., Kim, J. S., Lee, S. Y. & 7 other authors (2004). The genome sequence of the capnophilic rumen bacterium *Mannheimia succiniciproducens*. *Nat Biotechnol* **22**, 1275–1281.

Ishikawa, J., Yamashita, A., Mikami, Y., Hoshino, Y., Kurita, H., Hotta, K., Shiba, T. & Hattori, M. (2004). The complete genomic sequence of *Nocardia farcinica* IFM 10152. *Proc Natl Acad Sci U S A* **101**, 14925–14930.

Kotiranta, A., Lounatmaa, K. & Haapasalo, M. (2000). Epidemiology and pathogenesis of *Bacillus cereus* infections. *Microbes Infect* **2**, 189–198.

Kuwahara, T., Yamashita, A., Hirakawa, H. & 7 other authors (2004). Genomic analysis of *Bacteroides fragilis* reveals extensive DNA inversions regulating cell surface adaptation. *Proc Natl Acad Sci U S A* **101**, 14919–14924.

Nierman, W. C., DeShazer, D., Kim, H. S. & 30 other authors (2004). Structural flexibility in the *Burkholderia mallei* genome. *Proc Natl Acad Sci U S A* **101**, 14246–14251.

Pennisi, E. (2004). Genetics. DNA reveals diatom's complexity. *Science* **306**, 31.

Qiu, W. G., Schutzer, S. E., Bruno, J. F., Attie, O., Xu, Y., Dunn, J. J., Fraser, C. M., Casjens, S. R. & Luft, B. J. (2004). Genetic exchange and plasmid transfers in *Borrelia burgdorferi sensu stricto* revealed by three-way genome comparisons and multilocus sequence typing. *Proc Natl Acad Sci U S A* **101**, 14150–14155.

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.

Ussery, D. W., Binnewies, T. T., Gouveia-Oliveira, R., Jarmer, H. & Hallin, P. F. (2004). Genome Update: DNA repeats in bacterial genomes. *Microbiology* **150**, 3519–3521.

Veith, B., Herzberg, C., Steckel, S. & 9 other authors (2004). The complete genome sequence of *Bacillus licheniformis* DSM13, an organism with great industrial potential. *J Mol Microbiol Biotechnol* **7**, 204–211.

Ward, N., Larsen, A., Sakwa, J. & 35 other authors (2004). Genomic insights into methanotrophy: the complete genome sequence of *Methylococcus capsulatus* (Bath). *PLoS Biol* **2**, e303.

DOI 10.1099/mic.0.27720-0