microbiology Comment

Genome Update: base skews in 200+ bacterial chromosomes

Nine new microbial genomes have been published since the last Genome Update was written, seven of which are from bacterial organisms and the other two are *Plasmodium* genomes. These are summarized in Table 1 and include two α-*Proteobacteria*, *Anaplasma marginale* and *Zymomonas mobilis*, the latter of which is an ethanologenic bacterium; the deep-sea γ-proteobacterium *Idiomarina loihiensis*; the thermophilic *Bacillus*-related species *Geobacillus kaustophilus*; a PCE-dechlorinating bacterium, *Dehalococcoides ethenogenes*, which is the first sequenced genome from the phylum *Chloroflexi*; *Salmonella enterica*, which causes typhoid in humans; and *Campylobacter jejuni*, which is the major cause of human bacterial gastroenteritis. The two *Plasmodium* genomes, *Plasmodium chabaudi* and *Plasmodium berghei*, have also been sequenced for use as model malaria species. A brief discussion of these is given below.

Genomes of the month

Anaplasma marginale is an intracellular bacterium of the order *Rickettsiales*, and is the most prevalent tick-borne livestock pathogen throughout the world. The St Maries strain of *A. marginale* has a genome length of 1·2 Mb with 949 protein-encoding genes and a high coding density of 86 % (Brayton *et al.*, 2005). The G+C content of 49·8 mol% is unusually high for an obligate, intracellular organism (other sequenced rickettsiae species average 31 mol%). Due to reductive evolution, many *Rickettsiales* bacteria contain a large number of pseudogenes – *A. marginale*, however, contains few inactive copies of functional genes and has only 14 genes defined as functional pseudogenes. The surface coat of *A. marginale* contains two immunodominant surface protein complexes, mainly Msp2 and Msp3, from the msp2 and msp3 superfamily. Simultaneous switching of the msp2 and msp3 variants during infection allows the bacteria to generate antigenic variants, maintaining a persistent infection in the host organism. No vaccine currently exists for *A. marginale*, and a primary focus for ongoing studies must be the immunodominant members of the two previously mentioned superfamilies.

Four strains of *Campylobacter*, including *Campylobacter coli* RM2228, *Campylobacter jejuni* RM1221, *Campylobacter lari* RM2100, and *Campylobacter upsaliensis* RM3105 have been sequenced and compared, although only *C. jejuni* has been fully sequenced as one contiguous piece (the others have been sequenced to at least eightfold coverage and assembled such that most of the chromosome is covered by a few large pieces). *C. jejuni* consists of a single, circular chromosome of 1·8 Mb, has a G+C content of 30·31 mol% (see Table 1) and 94 % of the genome represents coding sequences (Fouts *et al.*, 2005). A comparison of the newly sequenced *Campylobacter* species to *C. jejuni* strain NCTC 11168, sequenced by Parkhill *et al.* (2000), revealed that the main difference between the two *C. jejuni* strains is the presence of four large integrated elements (IE) within strain RM1221. Also, as expected, the two *C. jejuni* strains appear to be closer to each other than to the other *Campylobacter* species.

Dehalococcoides ethenogenes strain 195 was obtained from an anaerobic sewage digestor. It is the only known bacterium with the ability to completely dechlorinate groundwater pollutants such as tetrachloroethane (PCE) and trichloroethane (TCE) to the non-toxic substance ethane, unlike other anaerobic dehalorespirers such as *Dehalobacter restrictus* that perform incomplete dechlorination to the toxic *cis*-dichloroethane (DCE). The genome of *D. ethenogenes* is composed of a 1·4 Mb circular chromosome containing 1591 CDS (see Table 1) as well as large duplicated regions and several integrated elements, which represent 13·6 % of the genome. *D. ethenogenes* contains genes for 17 reductive dehalogenases (RD), 16 of which are found in close proximity to genes for transcription regulators, suggesting stringent regulation of RD activity (Seshadri *et al.*, 2005). Unlike other gene groups in the *D. ethenogenes* genome, RD genes display a strong orientation bias in that all RD operons are oriented in the direction of replication. The discovery of genes encoding nitrogenase and other nitrogenase-essential components indicates that *D. ethenogenes* is able to fix nitrogen, suggesting a nitrogen-fixing autotroph as an ancestor.

Geobacillus kaustophilus strain HTA426 is a thermophilic *Bacillus*-related species

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

The accession number for each chromosome is the same for GenBank, EMBL and DDBJ. Note that of the four *Campylobacter* genomes, only *C. jejuni* has been fully sequenced in one piece. *P. berghei* and *P. chabaudi* have been sequenced to 7479 and 10 690 pieces, respectively, so we are unable to extract data regarding A+T content and number of tRNAs and rRNAs.

Name	Genome length (bp)	A+T (mol%)	No. of genes	tRNAs	rRNAs	Accession no.
Anaplasma marginale St Maries	1 197 687	50·2	949	37	1	CP000030
Dehalococcoides ethenogenes 195	1 469 720	51·1	1591	46	1	CP000027
Campolybacter jejuni RM1221	1 777 831	69·7	1838	44	3	CP000025
Campylobacter lari RM2100	1 500 000	73·1	1554	42	1	AAFK00000000
Campylobacter upsaliensis RM3195	1 660 000	69·8	1782	43	3	AAFJ00000000
Campylobacter coli RM2228	1 680 000	71·4	1764	43	3	AAFL00000000
Geobacillus kaustophilus HTA426	3 544 776	47·9	3498	87	9	BA000043
Geobacillus kaustophilus pHTA426	47 890	55·8	42	0	0	AP006520
Idiomarina loihiensis L2TR	2 839 318	53·0	2628	56	4	AE017340
Salmonella enterica ATCC 9150	4 585 229	47·8	4093	82	7	CP000026
Zymomonas mobilis ZM4	2 056 416	53·7	1998	51	3	AE008692
Plasmodium berghei	17 996 878	–	5864	–	–	CAAI01000000
Plasmodium chabaudi	16 866 661	–	5698	–	–	CAAJ01000000

isolated from sediment taken from the Mariana trench. It can grow in temperatures up to 72 °C with an optimal growth temperature of 60 °C. This is the first reported genome sequence of a thermophilic *Bacillus*-related species (Takami *et al.*, 2004). Therefore, a comparative analysis of the *G. kaustophilus* genome with the genomes of the mesophilic bacilli (*Bacillus anthracis*, *Bacillus cereus*, *Bacillus halodurans*, *Bacillus subtilis* and *Oceanobacillus iheyensis*) may reveal features characteristic of thermophilic adaptation. The *G. kaustophilus* genome is 3·5 Mb in size with a G+C content of 52·1 mol%, the smallest genome of this grouping, but the highest G+C content. It has a single plasmid (pHTA426) of 47 kb with 42 ORFs. The genome encodes 3498 genes of which (1) no orthologues were found for 839 genes (757 groups) in the other *Bacillus* species, while (2) 488 genes (419 groups) showed no significant homology to any other reported gene product. There are 1308 genes common to all six bacilli (1257 orthologous groups): of the 271 genes reported to be essential for growth of *B. subtilis* under non-limiting conditions, 233 were found in the *G. kaustophilus* genome (Kobayashi *et al.*, 2003). Notable for their absence are the genes encoding the teichoic acid biosynthetic enzymes with the exception of *tagE*, suggesting either that teichoic acid is synthesized by an alternative pathway or perhaps that *G. kaustophilus* has a different negatively charged polymer in its cell wall.

A very interesting finding is the absence of both *glyQ* and *glyS* orthologues that encode the α- and β-subunits, respectively, of glycyl tRNA synthetase. Since the bacterium cannot survive without the ability to charge tRNAGly, either this activity resides in a non-orthologous protein or tRNAGly is mischarged by another aminoacyl tRNA synthetase with subsequent modification of the aminoacyl group to glycine, as happens, for example, for glutamine in *B. subtilis*.

Dissection of those genomic features that correlate with thermophily suggest that an increased G+C content of rRNA, amino acid composition and asymmetric substitution of some amino acids contribute to thermophilic adaptation. However, there is no correlation between thermophily and synonymous codon usage in the case of *G. kaustophilus*, contrasting with that found for other thermophilic prokaryotes (Singer & Hickey, 2003). In a search for candidate genes that may contribute to thermophily, of particular note is the presence in only *G. kaustophilus* of a gene encoding a protamine P1-type protein (51 % similarity to protamine P1 of the Koala bear), the first finding of such a protein among prokaryotes. Additionally, among this group of bacilli, *G. kaustophilus* has unique genes encoding proteins for spermine/spermidine biosynthesis and for both rRNA and tRNA methyltransferases. It is likely that there are additional thermophily adaptation activities encoded among the *G. kaustophilus*-specific genes.

Idiomarina loihiensis is a deep-sea γ-proteobacterium that has recently been isolated from a hydrothermal vent on a submarine volcano in Hawaii. Here it lives in the partially oxygenated cold waters at the periphery of the vent. Its genome consists of a single circular chromosome of 2·8 Mb and has a mean G+C content of 47 mol%. As shown in Table 1, 2628 ORFs, four rRNA operons and 56 tRNA genes are predicted. *I. loihiensis* may survive a wide range of growth temperatures and salinities. As seen in many other deep-sea bacteria (Ivanova *et al.*, 2000), it exhibits a limited ability to utilize carbohydrates as its sole source of carbon and energy. Many typical carbohydrate degradation enzymes present in other proteobacteria appear to be missing. Instead, in comparison to other γ-proteobacteria, it shows an abundance of amino acid transport and degradation enzymes, suggesting that the primary source of carbon and energy may be amino acids rather than sugars. Similar to other deep-sea vent microorganisms, *I. loihiensis* produces a highly viscous exopolysaccharide that has been suggested to be used by vent micro-organisms to develop biofilms (Hou *et al.*, 2004).

Salmonella enterica and *Salmonella bongori*. More than 2,000 serovars comprise *S. enterica*, and *S. enterica* serovars often have a broad host range and cause gastrointestinal and systemic diseases. Two serovars, Paratyphi A and Typhi, are restricted to humans and cause only systemic disease. The *S. enterica* serovar Paratyphi A (strain ATCC 9150) genome is 4·6 Mb long, with an A+T content of 47 mol% and has 4263 annotated CDS (McClelland *et al.*, 2004). There are also 82 tRNA genes, 7 rRNA clusters and 36 structural RNAs identified. Comparing the Paratyphi A genome with the Typhi genome by using sequence and microarray analysis has shown that both genomes have independently accumulated many pseudogenes among their 4400 CDS (173 for Paratyphi A and 210 for Thyphi) and only 30 genes are degraded in both serovars. These 30 genes include many of the known virulence genes for gastrointestinal infections.

Zymomonas mobilis ZM4 is another α-proteobacterium, and although its genome is almost twice as large as *A. marginale* (it consists of a single, circular chromosome of 2·06 Mb), they are both at the smaller end of the α-*Proteobacteria*, with species such as *Bradyrhizobium japonicum* being almost 10 times as large as *A. marginale* with a genome size of 9·1 Mb (Seo *et al.*, 2005). *Z. mobilis* has a G+C content of 46·3 mol% (see Table 1) and is an ethanologenic bacterium with great potential for use in the industrial production of ethanol as an alternative fuel. Analysis of its genome revealed that *Z. mobilis* has the ability to produce several hexose-metabolizing enzymes, which enables it to utilize sucrose, fructose and glucose as well as mannose, raffinose and sorbitol. The gene encoding an essential enzyme in the Embden–Meyerhof–Parnas pathway was not found, suggesting that *Z. mobilis*, like other *Zymomonas* species, utilize the Entner–Doudoroff pathway for glucose catabolism. Furthermore, two genes encoding enzymes in the TCA pathway were lacking, indicating the presence of an alternative to this pathway. *Z. mobilis* ZM4 was compared to the *Z. mobilis* ZM1 strain, revealing the presence of 54 ORFs in ZM4 not found in ZM1. These genes peculiar to ZM4 presumably account for the higher rates of growth, glucose uptake and ethanol production seen in ZM4 compared to ZM1. Such genes may prove invaluable when creating recombinant bacterial strains that ferment higher levels of ethanol.

Comparison of two rodent malaria parasite species *Plasmodium chabaudi* and *Plasmodium berghei* with the human-infectious malaria species *Plasmodium falciparum* revealed that the rodent parasites have 4391 orthologous genes in *P. falciparum*, which represent a universal plasmodium gene set (Hall *et al.*, 2005). Proteins could be categorized into four strategies for gene expression during the parasite life cycle: (1) housekeeping, (2) host-related expression, (3) strategy-specific expression and (4) stage-specific expression.

Method of the month – base skews and DNA replication origins

This month we are looking at base skews in the recently sequenced genomes. Based on origin and terminus predictions using ORIGINX software (P. Worning, L. J. Jensen, P. F. Hallin, H.-H. Stærfeldt & D. W. Ussery, unpublished data), available from www.cbs.dtu.dk/services/GenomeAtlas/suppl/origin/, we have included A/T and G/C skews for the leading and lagging strands as well as signal-to-noise ratios (S/N ratio). The S/N value measures the ratio between signal strength ($I_{p,max}-I_{p,min}$) and $I_{p,min}$ where I_p is the summed information content of all oligonucleotides of leading

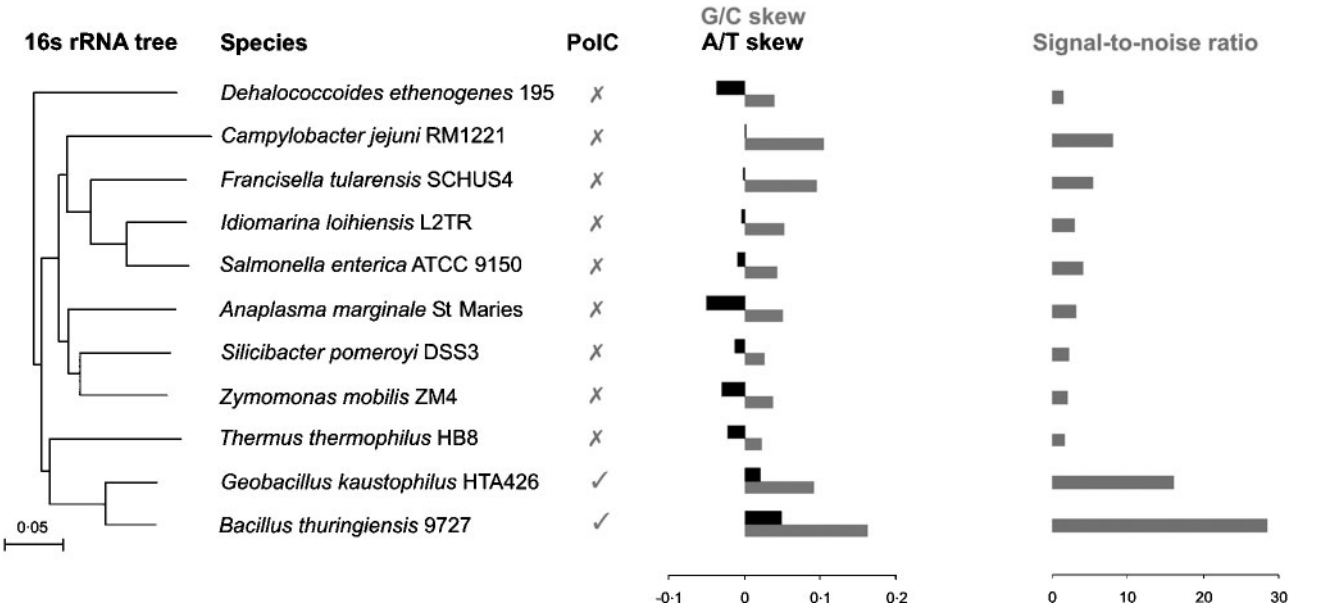

Fig. 1. G/C and A/T skews and S/N ratios in the context of the 16S rRNA tree. Only the Firmicutes have the PolC polymerase subunit causing the positive A/T skew. Also, it can be seen that the new phylum of *Chloroflexi* (*Dehalococcoides ethenogenes*) aligns distantly from the other recently sequenced genomes.

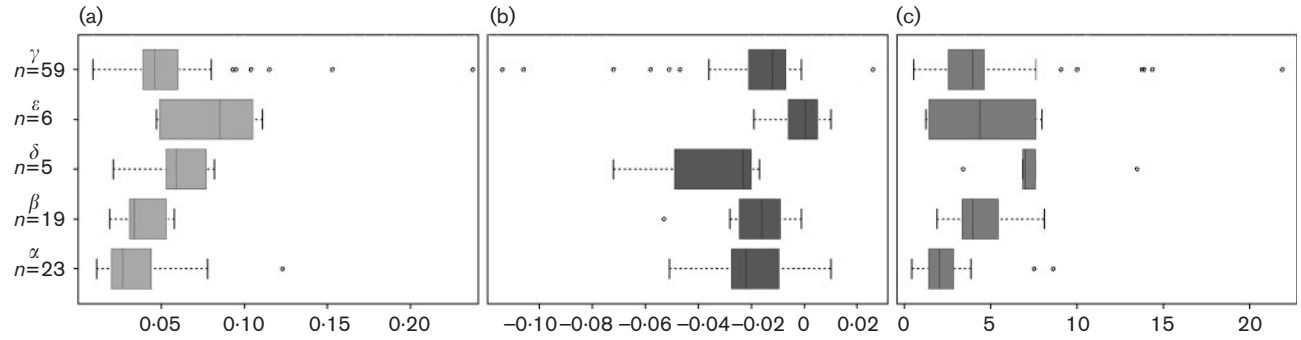

Fig. 2. G/C (a) and A/T skews (b) and S/N ratios (c) for different subdivisions of the *Proteobacteria*.

versus lagging strands, assuming *p* is the origin. Based on programming done by Worning and colleagues (P. Worning, L. J. Jensen, P. F. Hallin, H.-H. Stærfeldt & D. W. Ussery, unpublished data), we have measured this signal and produced origin plots for all circular prokaryotic sequences in our database.

Although this month's genomes are few in number in comparison with the contents of our entire database, a trend among these skews can be derived. We have collected the presence/absence of the polymerase C subunit (PolC), A/T and G/C skews and S/N ratios and listed the newly sequenced genomes according to their location in a neighbour joining tree, drawn from an alignment of 16s rRNA. This comparison is shown in Fig. 1. As described above, the Firmicutes show high S/N ratios as well as high G/C and A/T skews. The positive A/T skews, as suggested by Worning and colleagues (P. Worning, L. J. Jensen, P. F. Hallin, H.-H. Stærfeldt & D. W. Ussery, unpublished data), are explained by the presence of PolC. We also noticed a trend in that the γ- and ε-*Proteobacteria* show high G/C skews and small negative A/T skews, whereas the α-*Proteobacteria* show large negative A/T skews and smaller positive G/C skews.

Are the differences in base skews related to phyla? We extracted G/C and A/T skews and S/N ratios from the Genome Atlas Database. From the Box and Whiskers plot in Fig. 2(a) it can be observed that the G/C skews are higher for γ- and ε-*Proteobacteria* compared to those of α- and β-*Proteobacteria*. A significant difference in G/C skew exists between α- and ε-*Proteobacteria*. In Fig. 2(b) the A/T skew is observed to be significantly higher for ε-*Proteobacteria* than for other *Proteobacteria*. As was the case among the recently sequenced genomes shown in Fig. 1, the γ-*Proteobacteria* display higher S/N ratios compared to those of the α-*Proteobacteria* as shown in Fig. 2(c).

Additionally, bar charts were constructed to visualize base skews and composition in various chromosomal regions, namely (1) 150 bp upstream regions, (2) ORFs, leading/lagging strands and (3) the entire chromosome. These bar charts can be explored for every organism having annotated ORFs at www.cbs.dtu.dk/services/GenomeAtlas/suppl/GenUp014/

Supplemental web pages

Additional web pages containing supplementary material related to this article can be accessed from www.cbs.dtu.dk/services/GenomeAtlas/suppl/GenUp014/

Acknowledgements

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

Peter F. Hallin, Natasja Nielsen, Kevin M. Devine,† Tim T. Binnewies, Hanni Willenbrock and David W. Ussery

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

†Present address: Department of Genetics, Smurfit Institute, Trinity College Dublin, Dublin 2, Ireland

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

Brayton, K. A., Kappmeyer, L. S., Herndon, D. R., Dark, M. J., Tibbals, D. L., Palmer, G. H., McGuire, T. C. & Knowles, D. P., Jr (2005). Complete genome sequence of *Anaplasma marginale* reveals that the surface is skewed by two superfamilies of outer membrane proteins. *Proc Natl Acad Sci U S A* **102**, 844–849.

Fouts, D. E., Mongodin, E. F., Mandrell, R. E. & 18 other authors (2005). Major structural differences and novel potential virulence mechanisms from the genomes of multiple *Campylobacter* species. *PLoS Biol* **3**, doi 10.1371/journal.pbio.0030015.

Hall, N., Karras, M., Raine, J. D. & 27 other authors (2005). A comprehensive survey of the *Plasmodium* life cycle by genomic, transcriptomic, and proteomic analyses. *Science* **307**, 82–86.

Hou, S., Saw, J. H., Lee, K. S. & 19 other authors (2004). Genome sequence of the deep-sea γ-proteobacterium *Idiomarina loihiensis* reveals amino acid fermentation as a source of carbon and energy. *Proc Natl Acad Sci U S A* **101**, 18036–18041.

Ivanova, E. P., Romanenko, L. A., Chun, J., Matte, M. H., Matte, G. R., Mikhailov, V. V., Svetashev, V. I., Huq, A., Maugel, T. & Colwell, R. R. (2000). *Idiomarina* gen. nov., comprising novel indigenous deep-sea bacteria from the Pacific Ocean, including descriptions of two species, *Idiomarina abyssalis* sp. nov. and *Idiomarina zobellii* sp. nov. *Int J Syst Evol Microbiol* **50**, 901–907.

Kobayashi, K., Ehrlich, S. D., Albertini, A. & 96 other authors (2003). Essential *Bacillus subtilis* genes. *Proc Natl Acad Sci U S A* **100**, 4678–4683.

McClelland, M., Sanderson, K. E., Clifton, S. W. & 32 other authors (2004). Comparison of genome degradation in Parathypi A and Typhi, human-restricted serovars of *Salmonella enterica* that cause typhoid. *Nat Genet* **36**, 1268–1274.

Parkhill, J., Wren, B. W., Mungall, K. & 18 other authors (2000). The genome sequence of the food-borne pathogen *Campylobacter jejuni* reveals hypervariable sequences. *Nature* **403**, 665–668.

Seo, J.-S., Chong, H., Park, H. S. & 19 other authors (2005). The genome sequence of the ethanologenic bacterium *Zymomonas mobilis* ZM4. *Nat Biotechnol* **23**, 63–68.

Seshadri, R., Adrian, L., Fouts, D. E. & 22 other authors (2005). Genome sequence of the PCE-dechlorinating bacterium *Dehalococcoides ethenogenes*. *Science* **307**, 105–108.

Singer, G. A. & Hickey, D. A. (2003). Thermophilic prokaryotes have characteristic patterns of codon usage, amino acid composition and nucleotide content. *Gene* **317**, 39–47.

Takami, H., Takaki, Y., Chee, G.-J., Nishi, S., Shimamura, S., Suzuki, H., Matsui, S. & Uchiyama, I. (2004). Thermoadaptation trait revealed by the genome sequence of thermophilic *Geobacilus kaustophilus*. *Nucleic Acids Res* **32**, 6292–6303.

DOI 10.1099/mic.0.27889-0