Genomics update

The quest for biofuels fuels genome sequencing

Michael Y. Galperin*
National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA.

The list of recently completed microbial genome projects (Table 1) shows further progress in sequencing genomes of poorly studied environmental bacteria. The genome of *Aquifex aeolicus*, sequenced 10 years ago, has been joined by genomes of two more representatives of the phylum *Aquificae*. The genome of *Polaribacter* sp. MED152, a marine member of *Bacteroidetes*, revealed a combination of heterotrophic metabolism with light energy capture by proteorhodopsin. In addition, six genomes from the phylum *Chlorobi* more than doubled the number of sequenced genomes of green sulfur bacteria.

In eukaryotic genomics, important news was the release by the JGI scientists of a draft genome of the soft-rot ascomycete fungus *Trichoderma reesei*, also known as *Hypocrea jecorina* (Martinez et al., 2008). *Trichoderma reesei* is filamentous fungus that is widely used in biotechnology as a producer of various cellulases and hemicellulases for the hydrolysis of plant cell walls. This organism has attracted renewed interest owing to its potential use in the conversion of lignocellulosic residues to biofuel. The GenBank version of the draft genome of *T. reesei* consists of 2236 contigs, assembled into 170 scaffolds and containing ~34 Mbp of DNA, representing ~99% of the whole genome. The current assembly did not assign the scaffolds to any of the seven chromosomes of *T. reesei*, but allowed identification of 9129 predicted protein-coding genes (Martinez et al., 2008). Comparison of *T. reesei* with *Fusarium graminearum* (Gibberella zeae) and *Neurospora crassa* revealed a certain degree of synteny between these three genomes. A surprising finding was the relatively low number of glycoside hydrolases (cellulases, hemicellulases and pectinases) encoded by *T. reesei* genome. The authors suggest that successful utilization by *T. reesei* of its limited set of cellulolytic enzymes to efficiently degrade plant cell walls could be due to (i) clustering of the respective genes that ensures co-expression of the right combination of hydrolytic enzymes, and (ii) secretion of secondary metabolites (Martinez et al., 2008).

Although phylogenetically unrelated to *T. reesei*, the γ-proteobacterium *Cellvibrio japonicus* also encodes an efficient machinery for degrading plant cell walls that includes 130 predicted glycoside hydrolases (DeBoy et al., 2008).

The current list includes two actinobacterial genomes, representing the soil bacterium *Kocuria rhizophila* (Takarada et al., 2008) and a new strain of the human gut symbiont *Bifidobacterium longum* (Lee and O’Sullivan, 2006; Lee et al., 2008). The genus *Kocuria* belongs to the family *Micrococccinea* and was separated from *Micrococcus* just a few years ago (Stackebrandt et al., 1995). Accordingly, *K. rhizophila* ATCC 9341, parental strain of the sequenced *K. rhizophila* DC2201, was until recently classified as *Micrococcus luteus* and used as a standard quality control strain in a number of applications, including testing of antimicrobial compounds (Tang and Gillevet, 2003). The genus name was assigned to honour Miroslav Kocur, Slovakian microbiologist who dedicated many years to studying *M. luteus* (Rosypal and Kocur, 1963; Kocur, 1986). *Kocuria rhizophila* is an environmental actinomycete that is often associated with plant roots. Despite its small (for a soil actinomycete) 2.7 Mbp genome, *K. rhizophila* appears to encode the full set of key metabolic enzymes. However, it encodes fewer proteins participating in secondary metabolism, including single genes for a non-ribosomal peptide synthetase and a polyketide synthase. The relatively high tolerance of *K. rhizophila* to various organic compounds correlates with the presence of a large number of genes encoding various membrane transporters, including drug efflux pumps (Takarada et al., 2008).

The two newly sequenced genomes of *Aquificae* represent two major families in this phylum. *Hydrogenobaculum* sp. Y04AAS1 belongs to the family *Aquificaceae*, which also includes *A. aeolicus*, the best-characterized member of the phylum, whereas *Sulfitohydrogenibium* sp. Y03AOP1 belongs to the family *Hydrogenothermaceae*. Both are thermophilic chemolithoautotrophs, isolated from...
### Table 1. Recently completed microbial genomes (May–July 2008).

| Species name | Taxonomy | GenBank accession | Genome size (bp) | Proteins (total) | Sequencing centre | Reference |
|--------------|----------|-------------------|------------------|------------------|------------------|-----------|
| *Trichoderma reesei* | Eukaryota, Fungi | AAIL00000000 | 34 Mbp | 9129 | JGI | Martinez et al. (2008) |
| *Kocuria rhizophila* | Actinobacteria | AP009152 | 2 697 540 | 2357 | NITE | Takarada et al. (2008) |
| *Hydrogenobacterium* sp. Y04A0S1 | Aquificae | CP001130 | 1 559 514 | 1629 | JGI | Unpublished |
| *Sulfuriflavitum* sp. Y03AO1 | Aquificae | CP000180 | 1 838 442 | 1721 | JGI | Unpublished |
| *Candidatus Amoebobacteriaceus* | Bacteroidetes | CP000112 | 1 894 264 | 1283 | JGI | Unpublished |
| *Polaribacter* sp. MED152 | Bacteroidetes | NZ_AANA00000000 | 2 967 150 | 2646 | JCVI | González et al. (2008) |
| *Chlorobaculum parvum* | Chlorobi | CP001099 | 2 289 249 | 2043 | JGI | Unpublished |
| *Chlorobium limicola* | Chlorobi | CP001097 | 2 763 181 | 2434 | JGI | Unpublished |
| *Chlorovibrio halodurans* | Chlorobi | CP001100 | 3 293 456 | 2710 | JGI | Unpublished |
| *Pelodictyon phaeoalcaliphilum* | Chlorobi | CP001110 | 3 018 238 | 2707 | JGI | Unpublished |
| *Prosthecobacter aestuarii* | Chlorobi | CP001108 | 2 512 923 | 2327 | JGI | Unpublished |
| *Nathanaeobacter thermophilus* | Firmicutes | CP000134 | 3 165 557 | 2906 | JGI | Unpublished |
| *Methylobacterium populi* | α-Proteobacteria | CP000129 | 5 800 441 | 5365 | JGI | Unpublished |
| *Oligotropha carboxidovorans* | α-Proteobacteria | A8KN00000000 | 3 745 772 | 3754 | Mississippi State U. | Paul et al. (2008) |
| *Wolfia australis* | β-Proteobacteria | AM99887 | 1 482 455 | 1275 | Sanger Institute | Kaslon et al. (2008) |
| *Raistonia picothri* | β-Proteobacteria | CP001068 | 3 942 557 | 452 | JGI | Unpublished |
| *Califibrio japonicus* | γ-Proteobacteria | CP000094 | 4 576 573 | 3754 | JCVI | DeBoy et al. (2008) |
| *Erwinia icasamensis* | γ-Proteobacteria | CU468128 | 4 077 (total) | 3622 | MPIMG | Kubo et al. (2008) |
| *Proteus mirabilis* | γ-Proteobacteria | AM942759 | 4 063 606 | 3685 | Sanger Institute | Pearson et al. (2008) |
| *Geobacter lovleyi* | δ-Proteobacteria | CP0001089 | 3 917 761 | 3476 | JGI | Unpublished |
| *Candidatus Phytoplasma mali* | Tenericutes | CU468464 | 601 943 | 479 | MPIMG | Kubo et al. (2008) |
| *Mycoplasma arthritis* | Tenericutes | CP0001047 | 820 453 | 631 | JCVI | Dybvig et al. (2008) |
| *Bifidobacterium longum* DDJO10A | Actinobacteria | CP0000605 | 2 375 792 | 2003 | JGI | Lee et al. (2008) |
| *Chlorobium phaseolicolor* | Chlorobi | CP001101 | 2 736 403 | 2499 | JGI | Unpublished |
| *Lactobacillus casei* BL23 | Firmicutes | FM177140 | 3 079 196 | 3044 | INRA | Unpublished |
| *Streptococcus pneumoniae* | Firmicutes | CP0001015 | 2 078 953 | 2115 | JCVI | Doppazo et al. (2001) |
| *Rhodopseudomonas palustris* TIE-1 | α-Proteobacteria | CP001096 | 5 744 941 | 5246 | JGI | Unpublished |
| *Burkholderia cenocepacia* J2315 | β-Proteobacteria | AM747270–AM747273 | 8.05 (total) | Sanger Institute | Unpublished |
| *Burkholderia multivorans* ATCC 17616 | β-Proteobacteria | AP009385–AP009388 | 6.99 (total) | 6112 | Tohoku U. | Unpublished |
| *Neisseria gonorrhoeae* | γ-Proteobacteria | CP001113 | 4 574 041 | 5246 | JGI | Unpublished |
| *Methylococcus capsulatus* | γ-Proteobacteria | CP000934 | 4 576 573 | 3754 | JCVI | DeBoy et al. (2008) |
| *Neisseria meningitides* | γ-Proteobacteria | CP001090 | 4 827 641 | 4805 | JCVI | Unpublished |
| *Streptococcus thermophilus* | γ-Proteobacteria | CP001091–CP001094 | 2.34 (total) | 2142 | Bielefeld U. | Unpublished |
| *Salmonella enterica subsp. enterica* | γ-Proteobacteria | CP001120 | 4 888 768 | 4779 | JCVI | Unpublished |
| *Salmonella enterica subsp. enterica* | γ-Proteobacteria | CP001112 | 91 374 | 3874 | JCVI | Unpublished |
| *Salmonella enterica subsp. enterica* | γ-Proteobacteria | CP001113 | 4 827 641 | 4805 | JCVI | Unpublished |
| *Salmonella enterica subsp. enterica* | γ-Proteobacteria | CP000604 | 176 473 | 1764 | JCVI | Unpublished |
| *Salmonella enterica subsp. enterica* | γ-Proteobacteria | CP001112 | 3 605 | 293 | JCVI | Unpublished |
| *Stenotrophomonas maltophilia* | γ-Proteobacteria | CP001111 | 4 573 969 | 4039 | JGI | Unpublished |
| *Treponema pallidum subsp. pallidum* | spirochaetes | CP000805 | 1 139 457 | 1028 | Baylor | Matejkova et al. (2008) |

Sequencing centre names are abbreviated as follows: Baylor, Human Genome Sequencing Center, Baylor College of Medicine, Houston, Texas, USA; Bielefeld U., Centrum für Biotechnologie, Universität Bielefeld, Bielefeld, Germany; INRA, Institut National de la Recherche Agronomique, Domaine de Vilvert, Jouy en Josas, France; JCVI, J. Craig Venter Institute, Rockville, Maryland, USA; JGI, US Department of Energy, Joint Genome Institute, Walnut Creek, California, USA; Korea NIH, Center for Infectious Disease and Research, Korea National Institute of Health, Seoul, Korea; Mississippi State U., Mississippi State University, Mississippi State, Mississippi, USA; MPIMG, Max-Planck-Institute for Molecular Genetics, Berlin, Germany; NITE, Genome Analysis Center, Department of Biotechnology, National Institute of Technology and Evaluation, Shibuya-ku, Tokyo, Japan; Sanger Institute, The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridgehire, UK; Tohoku U., Department of Environmental Life Sciences, Graduate School of Life Sciences, Sendai, Miyagi, Japan; UNAM, Centro de Ciencias Genomicas, Universidad Nacional Autonoma de Mexico, Cuernavaca, Mexico.
hot springs at Yellowstone National Park at 60–75°C and capable of growing in microaerophilic conditions by using reduced sulfur compounds and/or hydrogen as electron acceptors and CO₂ as the source of carbon (Stöhr et al., 2001; Reysenbach et al., 2005). However, the former is an acidophile, growing at or below pH 3.0, and the latter grows at neutral pH values. The genome size of *Hydrogenobaculum* sp. YO4AAS1 is very close to that of *A. aeolicus*, whereas *Sulfituhydrogenibium* sp. YO3AOP1 features a 300 kb larger genome and almost a hundred of extra proteins. Availability of these new genomes should provide a much-needed insight into the physiology of *Aquificae*, one of the earliest-branching bacterial lineages.

Of the two members of the highly diverse phylum *Bacteroidetes* in the current list, the first one, *Candidatus Ameobophilus asiaticus*, is an obligate intracellular symbiont of the amoebae *Acanthamoeba* sp. (Hom et al., 2001). However, it has a much larger genome and encodes far more proteins than *Candidatus Sulcia* muelleri, another member of the *Bacteroidetes* that is an endosymbiont of sharpshooters (McCutcheon and Moran, 2007). In addition, JGI scientists plan to sequence the genome of *Candidatus Cardinium hertigii*, a symbiont of *Encarsia* wasps. Comparison of Ca. A. asiaticus with *Ca. S. muelleri* and Ca. C. hertigii on one hand and to free-living *Bacteroidetes* on the other should provide further clues to the mechanisms of bacterial adaptation to the endosymbiotic lifestyle.

The second *Bacteroidetes* member, *Polaribacter* sp. MED152, is a marine bacterium that was isolated from the surface water of north-western Mediterranean Sea off the Catalan coast (González et al., 2008). In the original GenBank submission, it was listed as a strain of *Polaribacter dokdonensis* (Yoon et al., 2006), with which it shares 99.6% similar 16S rRNA sequence. However, because of certain phenotypic differences between the two, the authors have chosen to refer to the sequenced organism simply as ‘strain MED152’. Together with the previously described *Gramella forsetii* (Bauer et al., 2006), *Polaribacter* sp. MED152 represents the marine *Bacteroidetes* that in certain conditions may comprise up to 20% of the bacterioplankton. Physiology of these bacteria is still poorly understood, and the authors use the genome of MED152 to offer a very attractive scheme of a ‘dual lifestyle’ for this organism. Based on the abundance of protease and glycosidase genes, they propose that the normal *modus operandi* for MED152 includes gliding motility in search for suitable polymers and their subsequent degradation for carbon, nutrients and energy (González et al., 2008). However, once suitable polymeric substrates have been exhausted, MED152 must sustain itself in a nutrient-poor environment. In contrast to *G. forsetii*, MED152 encodes proteorhodopsin, an *H*⁺-translocating light-dependent ion pump that can use light energy to charge the membrane, generating the proton-motive force. In fact, exposure to light does not stimulate growth of MED152 but appears to stimulate bicarbonate uptake and, conceivably, assimilation of carbon dioxide (González et al., 2008). Accordingly, MED152 encodes a variety of (predicted) light sensors that have not been seen in other members of *Bacteroidetes*. As noted in the accompanying insightful comment (Kirchman, 2008), the ability of marine bacteria to absorb light and use it to supplement their energy needs has important consequences for the understanding of the global carbon cycle.

In the past 2 months, JGI scientists released six complete genomes of *Chlorobi* (green sulfur bacteria), five of which, *Chlorobaculum parvum*, *Chlorobium limicola*, *Chlororhodopetrum thalassium*, *Pelodictyon phaeo-clathratiforme* and *Prosthecochloris aestuarii*, represent new species and one, *Chlorobium phaeobacteroides* represents a new strain of the species that had its first sequenced genome 2 years earlier (Table 1). Like other green sulfur bacteria, all these strains are anoxygenic phototrophs that live in strictly anaerobic sulfide-rich environments. They gain energy from photosynthesis, which relies on type I reaction centres and uses sulfide, sulfur and/or thiosulfate as electron acceptors, and fix carbon through the reverse TCA cycle (Overmann and Garcia-Pichel, 2000; Frigaard and Bryant, 2004). The species differ in their ecological niches and the relative amounts of carotene pigments and bacteriochlorophylls a, c, d and e. Green sulfur bacteria play a key role in carbon, nitrogen and sulfur turnover in anoxic freshwater aquatic environments and are a potential source of biomass for biofuels. In addition, *Prosthecochloris aestuarii*, which forms multilayered biofilms, has been implicated in microbial infection of coral reefs. Comparative analysis of these genomes should clarify many unanswered questions in physiology of these interesting and important organisms.

*Natranaerobius thermophilus* strain JW/NM-WN-LF is an anaerobic, halophilic alkali-thermophile isolated from sediments of a solar-heated, alkaline, hypersaline soda lake at Wadi An Natrun, Egypt (Mesbah et al., 2007). Its optimum growth conditions are 53°C, pH 9.5 and between 3.3 and 3.9 M Na⁺. It cannot grow at pH lower than 8.3 (or higher than 10.8). This organism belongs to a separate lineage in the class *Clostridia* and is currently assigned to the separate order *Natranaerobiaceae* and family *Natranaerobiales*. A detailed analysis of its genome sequence should clarify the adaptations of *N. thermophilus* to its unique ecological niche but it is already obvious that they include a Na⁺-dependent *Fₖ⁺ₐ*-type ATP synthase, very similar to the ones in the recently sequenced genomes of *Alkaliphilus metalliredigens* and *Alkaliphilus oremlandii*.

---

Journal compilation © 2008 Society for Applied Microbiology and Blackwell Publishing Ltd, *Environmental Microbiology*, 10, 2471–2475
No claim to original US government works
Other organisms with newly sequenced genomes include the chemolithoautotrophic $\alpha$-proteobacterium Oligotropha carboxidovorans (Paul et al., 2008), copper-resistant $\beta$-proteobacterium Ralstonia pickettii 12J, plant epiphyte Erwinia tasmaniensis (a non-pathogenic relative of widespread plant pathogens (Kube et al., 2008b), endophytes of the poplar tree Methylobacterium populi (Van Aken et al., 2004) and Stenotrophomonas maltophilia R551-3, tetrachloroethene-dechlorinating $\delta$-proteobacterium Geobacter lovleyi (Sung et al., 2006; Strycharz et al., 2008), new strains of Rhizobium etli, Treponema pallidum and many others (Table 1).

The current list also includes genomes of two mollicutes, Candidatus Phytoplasma mali and Mycoplasma arthritidis. The first one is a phytopathogen infecting apple, cherry, apricot and plum trees. It was isolated in Heidelberg, Germany, from an apple tree displaying symptoms of apple proliferative disease and is the first mycoplasma to have a linear chromosome (Kube et al., 2008a). The second one causes arthritis in rats and mice and is remarkable for carrying a lysogenic bacteriophage (Dybvig et al., 2008).

However, the greatest surprise in the mycoplasma studies came not from genome sequencing labs but from taxonomists. Although mycoplasmas have long been listed in the Division Tenericutes (International Committee on Systematic Bacteriology-Subcommittee on the Taxonomy of Mollicutes, 1995), this clade was usually considered together with Rickettsia and Chlamydia and not treated as an actual taxonomic unit. Instead, Mollicutes were considered a class in the phylum Firmicutes, which was consistent with the available phylogenetic analyses (Falah and Gupta, 1997; Ciccarelli et al., 2006). However, in the recent edition of Bergey’s Manual of Systematic Bacteriology, class Mollicutes was excluded from the phylum Firmicutes and moved to the new phylum Tenericutes (Ludwig et al., 2008). While there might have been valid reasons for doing that (for example, many mycoplasma use a non-standard genetic code with UGA codon coding for tryptophan instead of terminating translation), the cited reason for that move was comparative analysis of mycoplasmal sequences by Ludwig and Schleifer (2005), published in a book to which many researchers had no access. Given that the goal of Bergey’s Manual is introduction of ‘phylogenetic framework’ (Ludwig et al., 2008), it seems unfortunate that such important changes are being made without a public discussion or at least a publication in a peer-reviewed journal. After all, massive investments in microbial genome sequencing worldwide have moved bacterial taxonomy from a purely academic sphere into the realm of the biotechnological marketplace, and relatively minor changes in classification could have serious effect on the priorities in future genome sequencing projects.

**Acknowledgements**

M.Y.G. is supported by the Intramural Research Program of the NIH, National Library of Medicine. The author’s opinions do not reflect the views of NCBI, NLM or the National Institutes of Health.

**References**

Bauer, M., Kube, M., Teeling, H., Richter, M., Lombardot, T., Allers, E., et al. (2006) Whole genome analysis of the marine Bacteroidetes ‘Gramella forsetii’ reveals adaptations to degradation of polymeric organic matter. *Environ Microbiol* 8: 2201–2213.

Chung, G.T., Yoo, J.S., Oh, H.B., Lee, Y.S., Cha, S.H., Kim, S.J., and Yoo, C.K. (2008) The complete genome sequence of Neisseria gonorrhoeae NCCP11945. *J Bacteriol* 190: 6035–6036.

Ciccarelli, F.D., Doerks, T., von Mering, C., Creevey, C.J., Snel, B., and Bork, P. (2006) Toward automatic reconstruction of a highly resolved tree of life. *Science* 311: 1283–1287.

DeBoy, R.T., Mongodin, E.F., Fouts, D.E., Tailford, L.E., Khouri, H., Emerson, J.B., et al. (2008) Insights into plant cell wall degradation from the genome sequence of the soil bacterium Cellvibrio japonicus. *J Bacteriol* 190: 5455–5463.

Dopazo, J., Mendoza, A., Herrero, J., Caldara, F., Humbert, Y., Friedli, L., et al. (2001) Annotated draft genomic sequence from a Streptococcus pneumoniae type 19F clinical isolate. *Microb Drug Resist* 7: 99–125.

Dybvig, K., Zuhua, C., Lao, P., Jordan, D.S., French, C.T., Tu, A.H., and Loraine, A.E. (2008) The genome of Mycoplasma arthritidis. *Infect Immun* 76: 4000–4008.

Falah, M., and Gupta, R.S. (1997) Phylogenetic analysis of mycoplasmas based on Hsp70 sequences: cloning of the dnaK (hsp70) gene region of Mycoplasma capricolum. *Int J Syst Bacteriol* 47: 38–45.

Frilgaard, N.U., and Bryant, D.A. (2004) Seeing green bacteria in a new light: genomics-enabled studies of the photosynthetic apparatus in green sulfur bacteria and filamentous anoxygenic phototrophic bacteria. *Arch Microbiol* 182: 265–276.

González, J.M., Fernández-Gómez, B., Fernández-Guerra, A., Gómez-Consarnau, L., Sánchez, O., Coll-Lladó, M., et al. (2008) Genome analysis of the proteorhodopsin-containing marine bacterium Polaribacter sp. MED152 (Flavobacteria). *Proc Natl Acad Sci USA* 105: 8724–8729.

Horn, M., Harzenetter, M.D., Linner, T., Schmid, E.N., Muller, K.D., Michel, R., and Wagner, M. (2001) Members of the Cytophaga-Flavobacterium-Bacteroides phylum as intracellular bacteria of acanthamoebae; proposal of ‘Candidatus Amoebobius asiaticus’. *Environ Microbiol* 3: 440–449.

International Committee on Systematic Bacteriology-Subcommittee on the Taxonomy of Mollicutes (1995) Revised minimum standards for description of new species of the class Mollicutes (Division Tenericutes). *Int J Syst Bacteriol* 45: 605–612.

Kirchman, D.L. (2008) New light on an important microbe in the ocean. *Proc Natl Acad Sci USA* 105: 8487–8488.
