Genomics update

A square archaeon, the smallest eukaryote and the largest bacteria

Michael Y. Galperin*

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

The past 2 months brought significant advances in genome sequencing for all three domains of life. These include the genome sequence of the marine unicellular green alga *Ostreococcus tauri* that has the smallest cell size (usually <1 μm) of all known eukaryotes, the genome sequence of the square-celled archaeon *Haloquadratum walsbyi* and its relatives, and a dozen of new bacterial genomes, including some of the largest ones sequenced so far – *Myxococcus xanthus* and *Rhodococcus* sp. RHA1 whose genome sizes are more than 9 Mb – and the genome of the oil spill-degrading bacterium *Alcanivorax borkumensis* (Table 1).

*Ostreococcus tauri* belongs to the family Prasinophyceae, an early branching group of green algae that retains some primitive features of ancestral green plants. Its single cell contains a nucleus with a single nuclear pore and a greatly reduced cytoplasm with one chloroplast and one mitochondrion. It carries a single copy of each of the core cell cycle genes and produces relatively primitive light-harvesting antenna complexes (Robbens et al., 2005; Six et al., 2005). Despite these traits, *O. tauri* grows very fast and occasionally causes algal blooms in coastal waters. *Ostreococcus tauri* is a common member of marine phytoplankton; in oligotrophic waters of Atlantic, Indian and Pacific oceans it may account for up to 90% of the autotrophic biomass. The genome of *O. tauri* consists of 20 chromosomes, ranging in size from 0.16 to 1.07 Mb (Derelle et al., 2006) with the total size of 12.56 Mb. Thus, *O. tauri* genome is the smallest of any autotrophic eukaryotes sequenced so far and is just a tad larger than the genomes of *Saccharomyces cerevisiae* and several other yeasts. It contains 8166 predicted open reading frames, which is similar to the protein set of the actinobacterium *Streptomyces coelicolor* A3(2) and much less than the protein sets of *Rhodococcus* sp. RHA1 (see below) and of two bacteria with previously sequenced genomes, *Burkholderia xeno- vorans* LB400 and *Bradyrhizobium japonicum* USDA 110. As much as 81% of the *O. tauri* genome consists of coding DNA, which is the highest number of all sequenced eukaryotic genomes. Overall, this genome shows the signs of compaction (streamlining), similar to those seen in *Pelagibacter ubique* and *Prochlorococcus marinus*, other ubiquitous marine autotrophs. It appears that adaptation to the relatively stable marine environment in these entirely different lineages occurred along the same lines and was dominated by massive gene loss. There is no doubt that this sequence will provide a valuable reference point for the comparative analysis of various plant genomes. There is also certain commercial potential: several *O. tauri* genes have been subject of two patents issued in 2005 and 2006 to BASF Plant Science GmbH (see GenBank accession numbers CS020113 and CS351561 for examples).

Two papers published back-to-back in the online journal *BMC Genomics* describe, respectively, the genome sequence of the haloarchaeon *H. walsbyi* and the metagenome of the microbial community of a saturated brine (crystallizer) pond, from which this organism was isolated (Bolhuis et al., 2006; Legault et al., 2006). First described as ‘a square bacterium’ in 1980 by Anthony Walsby and referred ever since as ‘Walsby’s square bacterium’ or ‘Walsby’s square archaeon’; *H. walsbyi* has the shape of a thin rectangle with a side of up to 10 μm but only 0.25 μm thick (Walsby, 2005). This organism has not been cultivated until 2004, when two groups succeeded in growing it in an axenic culture in hypersaline media containing pyruvate (Bolhuis et al., 2004; Burns et al., 2004). Cultivation of *H. walsbyi* paved the way to the complete genome sequence of this very unusual organism that can survive in the brine containing 3.3 M NaCl and 2 M MgCl2. The genome sequence of *H. walsbyi* revealed proton- and chloride-translocating bacteriorhodopsins and a dihydroxyacetone-specific phosphoenolpyruvate-dependent phosphotransterase system (PTS), the first one found in any archaeon. The authors note that dihydroxyacetone is an end-product of glycerol metabolism by *Salinibacter ruber*, a bacterium that inhabits the same
<table>
<thead>
<tr>
<th>Species name</th>
<th>Taxonomy</th>
<th>GenBank accession</th>
<th>Genome size, bp</th>
<th>Proteins</th>
<th>Sequencing centre</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>New organisms</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ostreococcus tauri</td>
<td>Eukaryota, Viridiplantae</td>
<td>CR954201-954220</td>
<td>12.56 Mb</td>
<td>8166</td>
<td>CNRS – Univ. Pierre et Marie Curie-Paris, Banyuls sur Mer, France</td>
<td>Derelle et al. (2006)</td>
</tr>
<tr>
<td>Haloquadratum walsbyi</td>
<td>Archaea</td>
<td>AM180088, AM180089</td>
<td>3 132 494, 46 867</td>
<td>2646</td>
<td>Max Planck Institute of Biochemistry, Martinsried, Germany; Univ. Groningen, Netherlands, and Univ. Miguel Hernandez, Alicante, Spain</td>
<td>Bolhuis et al. (2006)</td>
</tr>
<tr>
<td>Mycobacterium sp.</td>
<td>Actinobacteria</td>
<td>CP000384, CP000385</td>
<td>5 705 448, 215 075</td>
<td>5615</td>
<td>DOE Joint Genome Institute, Walnut Creek, CA, USA</td>
<td>Unpublished</td>
</tr>
<tr>
<td>Rhodococcus sp. RHA1</td>
<td>Actinobacteria</td>
<td>CP000431, CP000432, CP000433, CP000434</td>
<td>7 804 765, 1 123 075, 442 536, 332 361</td>
<td>9145</td>
<td>University of British Columbia, Vancouver, Canada</td>
<td>Unpublished</td>
</tr>
<tr>
<td>Rubrobacter xylanophilus</td>
<td>Actinobacteria</td>
<td>CP000386</td>
<td>3 225 748</td>
<td>3140</td>
<td>DOE Joint Genome Institute</td>
<td>Unpublished</td>
</tr>
<tr>
<td>Cytophaga hutchinsonii</td>
<td>Flexibacteres</td>
<td>CP000383</td>
<td>4 433 218</td>
<td>3785</td>
<td>DOE Joint Genome Institute</td>
<td>Unpublished</td>
</tr>
<tr>
<td>Mesorhizobium sp. BNC1</td>
<td>α-Proteobacteria</td>
<td>CP000390, CP000389, CP000391, CP000392</td>
<td>4 412 446, 343 931, 131 247, 47 561</td>
<td>4543</td>
<td>DOE Joint Genome Institute</td>
<td>Unpublished</td>
</tr>
<tr>
<td>Roseobacter denitrificans</td>
<td>γ-Proteobacteria</td>
<td>CP000362</td>
<td>4 133 097</td>
<td>3946</td>
<td>Translational Genomics Res. Institute, Phoenix, AZ, USA</td>
<td>Schneiker et al. (2006)</td>
</tr>
<tr>
<td>Alcanivorax borkumensis</td>
<td>α-Proteobacteria</td>
<td>AM286690</td>
<td>3 120 143</td>
<td>2755</td>
<td>German Research Center for Biotechnology, Braunschweig, Germany</td>
<td>Unpublished</td>
</tr>
<tr>
<td>Pseudalteromonas atlantica</td>
<td>γ-Proteobacteria</td>
<td>CP000388</td>
<td>5 187 005</td>
<td>4281</td>
<td>DOE Joint Genome Institute</td>
<td>Unpublished</td>
</tr>
<tr>
<td>Myxococcus xanthus</td>
<td>δ-Proteobacteria</td>
<td>CP000113</td>
<td>9 139 763</td>
<td>7331</td>
<td>The Institute for Genomic Research, Rockville, MD, USA</td>
<td>Goldman et al. (2006)</td>
</tr>
<tr>
<td>Helicobacter acinonychis</td>
<td>α-Proteobacteria</td>
<td>AM260522, AM260523</td>
<td>1 553 927, 3 661</td>
<td>1618</td>
<td>Penn State Univ., University Park, PA, USA, and Max Planck Institute for Infection Biology, Berlin, Germany</td>
<td>Eppinger et al. (2006)</td>
</tr>
<tr>
<td>Borrelia afzelii</td>
<td>Spirochaetes</td>
<td>CP000395, CP000400, CP000402</td>
<td>905 394, 30 017, 26 533</td>
<td>927</td>
<td>Fritz Lipmann Institute, Jena, Germany</td>
<td>Gloeckner et al. (2006)</td>
</tr>
<tr>
<td><strong>New strains</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Escherichia coli 536</td>
<td>γ-Proteobacteria</td>
<td>CP000247</td>
<td>4 938 920</td>
<td>4629</td>
<td>Univ. Würzburg and Univ. Göttingen, Germany</td>
<td>Hochhut et al. (2006)</td>
</tr>
<tr>
<td>Shigella flexneri 5 str. 8401</td>
<td>γ-Proteobacteria</td>
<td>CP000266</td>
<td>4 574 284</td>
<td>4116</td>
<td>State Key Lab. for Molecular Virology and Genetic Engineering, Beijing, China</td>
<td>Nie et al. (2006)</td>
</tr>
<tr>
<td>Clostridium perfringens</td>
<td>Firmicutes</td>
<td>CP000246</td>
<td>3 256 683</td>
<td>2876</td>
<td>The Institute for Genomic Research, Rockville, MD, USA</td>
<td>Myers et al. (2006)</td>
</tr>
<tr>
<td>Clostridium perfringens SM101</td>
<td>Firmicutes</td>
<td>CP000312, CP000315, CP000313, CP000314</td>
<td>2 897 393, 38 092, 12 397, 12 206</td>
<td>2631</td>
<td>The Institute for Genomic Research, Rockville, MD, USA</td>
<td>Myers et al. (2006)</td>
</tr>
</tbody>
</table>

a. Encoded in all chromosomes and megaplasmids; includes pseudogenes.

crystallizer ponds, and suggest that dihydroxyacetone utilization by H. walsbyi could promote synthrophic growth of these two organisms (Bolhuis et al., 2006). In the second paper, the authors use a metagenomic approach and analyse a fosmid library collected from the same crystallizer ponds. As expected, the majority of collected sequences could be attributed either to H. walsbyi or to S. ruber. However, H. walsbyi-related sequences revealed a variety of metabolic genes that have not been detected in the sequenced genome. The authors conclude that various strains of H. walsbyi, while inhabiting a common ecological niche and sharing a common core genome, encode a variety of auxiliary metabolic enzymes. As a result, the combined genome (pangenome) of H. walsbyi actually encodes much wider metabolic capability than the one found in any single organism. This conclusion is remarkably similar to the one made from an analysis of the prochlorococcal pangenome (Hess, 2004;
Coleman et al., 2006). These observations suggest the existence of yet another level of microbial diversity, the largely unappreciated variation between individual populations within a single microbial species.

Among the newly sequenced bacterial genomes, two – *Rhodococcus* sp. RHA1 and *M. xanthus* – clearly stand out because of their size. The genome of *Rhodococcus* sp. RHA1 consists of four linear segments, a 7.8 Mb chromosome and three megaplasmids (Table 1) and encodes more predicted proteins than any other actinobacterial genome sequenced to date. *Rhodococcus* sp. RHA1 is capable of degrading a wide variety of polychlorinated biphenyls (PCBs), as well as benzoate and phthalate, which makes it a potentially valuable agent for bioremediation of contaminated environments. The PCB-degrading genes of *Rhodococcus* sp. RHA1 are located primarily on the megaplasmids, which have been characterized previously (Shimizu et al., 2001; Warren et al., 2004). In contrast, benzoate degradation genes appear to be located on the chromosome (Patrauchan et al., 2005). Availability of the complete genome will help in understanding the cellular requirements for PCB degradation and will allow cross-genome comparisons with *B. xenovorans*, a phylogenetically distant bacterium that is also capable of degrading PCBs and also has an unusually large genome size.

Two more actinobacterial genomes were sequenced recently by the US Department of Energy Joint Genome Institute. One of them comes from a fast-growing *Mycobacterium* sp. strain MCS, isolated from Creosote wood preservative-contaminated soil in Libby, Montana, and capable of metabolizing polycyclic aromatic hydrocarbons (PAHs), such as pyrene (Miller et al., 2004). Well-known carcinogenic properties of PAHs make this organism and related mycobacteria important participants in the PAHs bioremediation processes.

*Rubrobacter xylanophilus* is a moderately thermophilic representative of a deeply branching group of actinobacteria that is highly resistant to radiation (Ferreira et al., 1999). It is capable of degrading plant hemicellulose and xylan, which could be important for paper industry. *Rubrobacter* is a widespread group of bacteria, whose representatives have been detected in marine and Australian desert soil samples. Incidentally, *Rubrobacter*-related bacteria were implicated in rosy discoloration of masonry and lime wall paintings of a historically important building in Burggen, Germany (Schabereiter-Gurtner et al., 2001).

*Cytophaga hutchinsonii* is the first sequenced representative of the family Flexibacteraceae (formerly *Cytophaga*-Flexibacter group), which is a member of the phylum *Bacteroidetes*. *Cytophaga hutchinsonii* is the type strain of the genus *Cytophaga*, a heterogeneous group that includes marine, freshwater and terrestrial species. *Cytophaga hutchinsonii* has been a model organism for studying the mechanisms of gliding motility, which does not use flagella or type IV pili (McBride, 2004). It has also attracted interest by its ability to degrade such polymers as chitin, peptin and cellulose.

There are two α-proteobacteria in the current list. One is the EDTA-degrading bacterium *Mesorhizobium* sp. BNC1 (formerly known as *Agrobacterium* sp. BNC1), which can use EDTA as a sole source of carbon and nitrogen (Bohuslavek et al., 2001). This organism could be useful in bioremediation of toxic heavy metals that may remain in solution in the EDTA-bound form. The other one is the marine aerobic photosynthetic bacterium *Roseobacter denitrificans*, a relatively rare phototroph that is capable of growing both aerobically and anaerobically, using nitrate as a terminal electron acceptor. The details of this project are available online at http://genomes.tgen.org/rhodobacter.html.

The two γ-proteobacteria in Table 1, *Alcanivorax borkumensis* and *Pseudoalteromonas atlantica*, are both marine bacteria, widespread in the open ocean. *Pseudoalteromonas atlantica* is known primarily as a producer of agarase and other commercially important saccharolytic and proteolytic enzymes. It is often found attached to eukaryotic hosts, such as seaweed or crab shells. Owing to its capacity to produce large amounts of exopolysaccharide and form biofilms on various surfaces, *P. atlantica* is believed to be one of the primary colonizers of new marine habitats. The genome sequence is expected to shed light on the ability of *P. atlantica* and its relatives to produce a variety of biologically active compounds, including antibacterial, antifungal and antiviral molecules.

*Alcanivorax borkumensis* is also a widespread marine bacterium, found in ocean samples worldwide. However, it is barely detectable in unpolluted environments but becomes the dominant microbe in oil-polluted waters (Harayama et al., 2004; Yakimov et al., 2005). As its name suggests, *A. borkumensis* effectively degrades alkanes and uses oil hydrocarbons as exclusive source of carbon and energy, which puts it into the category of ‘hydrocarbonlastic’ bacteria. *Alcanivorax borkumensis* can metabolize long-chain normal and branched alkanes, as well as certain phytanes and alkylcycloalkanes. Like its recently sequenced relative *Hahella chejuensis*, *A. borkumensis* produces glycolipid biosurfactants. These surfactants participate in emulsification of hydrocarbons, increasing their bioavailability. *Alcanivorax borkumensis* has a relatively compact genome sequence with a variety of alkane-metabolizing genes, some of which are similar to those in the much larger genome of *Pseudomonas putida* (Sabirova et al., 2006; Schneiker et al., 2006). However, in contrast to the streamlined genomes of such marine bacteria as *P. ubique* or *P. marinus*, mentioned
above, A. borkumensis retains a sophisticated system of signal transduction that includes 30 sensory histidine kinases and 30 response regulators. Further studies of A. borkumensis should help in designing the superbugs that would help minimize the ecological damage caused by oil spills. The example of A. borkumensis is a good illustration for the ‘rare biosphere’, recently described by Sogin and colleagues (2006), thousands of low-abundance populations that account for great phylogenetic diversity in the deep sea.

The second largest genome in the current crop is the long-anticipated genome of M. xanthus, a δ-proteobacterium that has long been used as a model organism to study gliding motility, morphogenesis and bacterial cell differentiation (Kaplan, 2003; Kaiser, 2004; Søgaard-Andersen, 2004). As noted in a previous column, various pieces of M. xanthus DNA have been sequenced by Monsanto Corp. and released as supplementary information to the US patent 6833447 for myxococcal nitrite reductase (see http://patft.uspto.gov/netacgi/nph-Parser? applicationId=8000000200&fpType=7). Now, thanks to the involvement of TIGR scientists, the whole 9.14 Mb chromosome of M. xanthus – the longest circular DNA molecule and the longest bacterial chromosome sequenced so far – has become publicly available in GenBank. At the time of this writing, the description of the M. xanthus genome (Goldman et al., 2006) was still unavailable, but even a cursory look at the encoded protein set revealed an impressive variety of functionalities. A preliminary census of signalling proteins in M. xanthus using previously described methodology (Galperin, 2005; 2006) revealed at least 138 sensory histidine kinases, 133 response regulators, 102 Ser/Thr protein kinases, 17 diguanylate cyclases, 7 c-di-GMP phosphodiesterases, 12 adenylate cyclases and 31 methyl-accepting chemotaxis proteins. These and other M. xanthus proteins contain numerous PAS, TPR, FHA and other signalling domains. For example, M. xanthus encodes 23 proteins with the recently described PATAN domain, far more than any other organism with a completely sequenced genome (Makarova et al., 2006).

Besides its role as a model organism, M. xanthus is an important soil bacterium with rich secondary metabolism. There is no doubt that the availability of its genome will give a big boost to a variety of research programmes.

An interesting study on the evolution of pathogenesis has been performed by comparing the genome sequences of Helicobacter acinonychis, gastric pathogen of cheetahs, lions and tigers and its close relative, human pathogen Helicobacter pylori (Eppinger et al., 2006). Genome comparisons have been used to deduce the direction of the pathogen jump between host species or, in the authors’ own words, to decide ‘who ate whom and when did it happen?’ Analysis of the core genes conserved between all the genomes suggests that the host jump was from early humans to cats, has occurred within the last 200 000 years and was accompanied by fragmentation of many genes encoding outer membrane proteins.

Other recently completed sequencing projects include the genomes of the spirochete Borrelia afzelii (Gloeckner et al., 2006) an uropathogenic strain of Escherichia coli (Hochhut et al., 2006), new strains of Shigella flexneri (Nie et al., 2006) and of two strains of Clostridium perfringens (Myers et al., 2006).

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


