

### Genome Update: proteome comparisons

#### Genomes of the month

Four new genomes will be discussed in this month's Genome Update. The list of organisms given in Table 1 includes an archaeon, two bacteria and an eukaryote. Two genomes will only be mentioned briefly. The protozoan *Cryptosporidium hominis* causes gastroenteritis and diarrhoea around the world. Like the *Cryptosporidium parvum* genome (Abrahamson *et al.*, 2004), that of *C. hominis* has undergone genome reduction to a size of about 9 Mbp, and both genomes have slightly fewer than 4000 genes (Xu *et al.*, 2004). *Photobacterium profundum* is a member of the vibrio group (Thompson *et al.*, 2004), and its genome consists of two chromosomes and one plasmid (Table 1). At the time of writing, the GenBank files have been deposited, but the genome report has not yet been published. One very interesting feature of this genome is the larger number of rRNAs.

The archaeon *Haloarcula marismortui* lives in high salt conditions – around 4.5 M salt in the Dead Sea. Its genome consists of nine replicons (two chromosomes and seven plasmids, Table 1), totalling about 4.3 Mbp in length, with 4242 predicted proteins (Baliga *et al.*, 2004). There seems to be a trend (based on the three genomes sequenced so far) for halophilic archaeal genomes to have multiple replicons, with the largest replicon being a bit more G + C-rich (~65%) than the smaller ones, which have a G + C content of around 55%. *H. marismortui* has the largest halophilic archaeal genome sequenced to date.

*Mycoplasma hyopneumoniae* strain 232 is an important member of the porcine respiratory disease complex and is the causative agent of swine mycoplasmosis. *M. hyopneumoniae* belongs to the class *Mollicutes*, members of which are characterized by small, A + T-rich genomes

and the absence of a permanent cell wall (Waites & Talkington, 2004). The genome of *M. hyopneumoniae* 232 is the eighth of the genus *Mycoplasma* to have been sequenced; it is 892 758 bp long, has an A + T content of 71% and encodes 692 predicted protein coding sequences (Minion *et al.*, 2004). The genome contains a single 16S–23S rRNA operon, one 5S rRNA gene and 30 tRNA coding sequences. Virulence factors in *M. hyopneumoniae* have not been clearly established yet, except the cilium adhesin protein P97. The cilium adhesin gene contains six paralogues in the genome but only the interaction with the surface protein P97 enables *M. hyopneumoniae* to attach to the cilia of the respiratory epithelium. A second gene in the operon is designated P102; this gene also contains six paralogues but their function(s) are unknown. Protein secretion occurs through an abbreviated membrane protein secretory system, consisting of SecA, SecD, SecY, PrsA, DnaK, Tig and LepA. A comparison of the genome of *M. hyopneumoniae* 232 with those of other *Mycoplasma* species will be presented below.

#### Method of the month: visualizing proteome comparisons

This month we will present two ways of comparing bacterial proteomes (i.e. all the

proteins encoded by a genome); both methods are based on the identification of sequence homologies in protein sequences, detected by BLASTP.

We have collected all annotated proteins of all eight *Mycoplasma* genome sequences currently available, and blasted each of the individual sequences against the collection. For each organism, we extracted the number of genes distinct for that organism and the number of genes shared with the other species. These two numbers are shown in Fig. 1(a), and reflect to some extent an evolutionary distance or similarity between the individual species. The large circle in the middle represents a pool of all 6053 proteins from the *Mycoplasma* genomes. The smaller intersecting bubbles represent the individual *Mycoplasma* genomes, and the area (size) is proportional to the total number of proteins for that genome. The intersecting areas illustrate the number of homologues between a given genome and the pool of proteins, excluding hits from the genome itself. (Otherwise, all of the proteins would have a hit!) Only hits having an E-value of  $10^{-15}$  or better were included. One interesting observation is that there seems to be a core set of roughly 500 proteins for each genome that is found in at least one other *Mycoplasma*. We have measured how many species the average

**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.
<i>Haloarcula marismortui</i> ATCC 43049 <sup>T</sup> Chr. 1	3 131 724	37.6	3131	48	2	AY596297
<i>Haloarcula marismortui</i> ATCC 43049 <sup>T</sup> Chr. 2	288 050	42.8	281	1	1	AY596298
<i>Haloarcula marismortui</i> ATCC 43049 <sup>T</sup> pNG700	410 554	40.9	362	0	1	AY596296
<i>Haloarcula marismortui</i> ATCC 43049 <sup>T</sup> pNG600	155 300	41.7	166	1	0	AY596295
<i>Haloarcula marismortui</i> ATCC 43049 <sup>T</sup> pNG500	132 678	45.5	131	0	0	AY596294
<i>Haloarcula marismortui</i> ATCC 43049 <sup>T</sup> pNG400	50 060	42.6	51	0	0	AY596293
<i>Haloarcula marismortui</i> ATCC 43049 <sup>T</sup> pNG300	39 521	40.0	40	0	0	AY596292
<i>Haloarcula marismortui</i> ATCC 43049 <sup>T</sup> pNG200	33 452	44.4	42	0	0	AY596291
<i>Haloarcula marismortui</i> ATCC 43049 <sup>T</sup> pNG100	33 303	45.7	36	0	0	AY596290
<i>Haloarcula marismortui</i> ATCC 43049 <sup>T</sup> Total	4 274 642	~ 38	4240	50	4	AY596290– AY596297
<i>Mycoplasma hyopneumoniae</i> 232 Main	892 758	71.4	691	1	30	AE017332
<i>Photobacterium profundum</i> SS9 Chr. 1	4 085 304	58.0	3416	150	14	CR354531
<i>Photobacterium profundum</i> SS9 Chr. 2	2 237 943	58.8	1997	19	1	CR354532
<i>Photobacterium profundum</i> SS9 plasmid	80 033	56.0	67	0	0	CR377818
<i>Photobacterium profundum</i> SS9 Total	6 403 280	~ 58	5480	169	15	–
<i>Cryptosporidium hominis</i> , 11 chromosomes	~ 9 160 000	68.7	3944	45	6	AAEL01000001– AAEL01001422

gene has orthologues to – the connectivity. For each gene in every genome, we have counted the number of genomes that this gene has homology to and divided it by the species count (7); the resulting number ranges between 0 (0/7, found in no other species) and 1 (7/7, are found in all other species). The average for each genome is located at the bottom of each circle, and ranges from 0.58 to 0.71. On the other hand, the number of organism-specific proteins encoded by the various *Mycoplasma* genomes varies – ranging from around 500 (513 for *Mycoplasma mycoides*, 500 for *Mycoplasma penetrans*) to a mere five (for *Mycoplasma genitalium*).

In Fig. 1(b) we have constructed a BLAST table, again showing protein homology, but this time between all combinations of *Mycoplasma* genomes and additional genomes of *Clostridium perfringens* 13, *Escherichia coli* O157:H7, *Streptococcus agalactiae* NEM316, *Lactobacillus plantarum* WCFS1 and *Bacillus subtilis* 168. On the diagonal the table shows the number of proteins that have homologous hits within the proteome itself as indicated by the red colour, scaled from 10% (white) to 36% (red). The green area on each side of the diagonal shows the number of

proteins that has homologous hits *between* genomes. Only hits having 85% of overlapping alignments and an E-value of  $10^{-5}$  or better are counted. When looking at the two *columns* of *M. mycoides* and *M. penetrans*, it is observed that these organisms share a low percentage of their genes with the other species, simply because they have almost twice as many genes. What should be noted is the fact that both species display many paralogues.

In addition to these results, we have generated a neighbour-joining tree based on 16S rRNA gene sequences of all species, using CLUSTAL\_X. The relatively short evolutionary distance for *M. genitalium* and *Mycoplasma pneumoniae* could be an explanation as to why these two species have not developed or acquired many distinct genes compared with other species. It is our anticipation to develop a web-based platform on our Genome Atlas web pages that will allow for BLASTP proteome comparisons such as these between any chosen set of sequenced microbial genomes. Readers are encouraged to check the supplemental web page for updates and links to this service when it is operational.

### 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/GenUp012/>

### Acknowledgements

This work was supported by a grant from the Danish National Research Foundation.

**Tim T. Binnewies, Peter F. Hallin, Hans-Henrik Staerfeldt and David W. Ussery**

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)

**Abrahamsen, M. S., Templeton, T. J., Enomoto, S. & 17 other authors (2004).** Complete genome sequence of the apicomplexan, *Cryptosporidium parvum*. *Science* **304**, 441–445.

**Baliga, N. S., Bonneau, R., Facciotti, M. T. & 12 other authors (2004).** Genome sequence of *Haloarcula marismortui*: a halophilic archaeon from the Dead Sea. *Genome Res* **14**, 2221–2234.

**Minion, F. C., Lefkowitz, E. J., Madsen, M. L., Cleary, B. J., Swartzell, S. M. & Mahairas, G. G. (2004).** The genome sequence of *Mycoplasma hyopneumoniae* strain 232, the agent of swine mycoplasmosis. *J Bacteriol* **186**, 7123–7133.

**Thompson, F. L., Iida, T. & Swings, J. (2004).** Biodiversity of vibrios. *Microbiol Mol Biol Rev* **68**, 403–431.

**Waites, K. B. & Talkington, D. F. (2004).** *Mycoplasma pneumoniae* and its role as a human pathogen. *Clin Microbiol Rev* **17**, 697–728.

**Xu, P., Widmer, G., Wang, Y. & 15 other authors (2004).** The genome of *Cryptosporidium hominis*. *Nature* **431**, 1107–1112.

DOI 10.1099/mic.0.27760-0