

Genome Update: length distributions of sequenced prokaryotic genomes

Genomes of the month

This month, four genomes from two different species will be discussed. The first organism, *Rhodospseudomonas palustris*, is a metabolic 'fox', in that it can do many different things. *R. palustris* is one of the most metabolically versatile bacteria known – it can grow utilizing any one of the four modes of metabolism that support life. In contrast, the three different *Prochlorococcus marinus* isolates are perhaps more like metabolic 'hedgehogs' in that they can do only one thing, but they do it very well. *P. marinus* requires only light, CO₂ and inorganic materials to live, and the three different genomes reflect adaptation to different ecological environments, in terms of wavelength and intensity of available light.

R. palustris is a purple photosynthetic bacterium, belonging to the α -Proteobacteria. *R. palustris* can obtain energy from light, inorganic compounds or organic compounds, allowing survival and growth under a wide range of conditions. The genome of *R. palustris* strain CGA009 consists of a circular chromosome of about 5.46 Mbp in length, which is slightly below the average size (5.6 Mbp) for free-living α -Proteobacteria (see Fig. 1). The genome encodes 4836 predicted genes (see <http://genome.ornl.gov/microbial/rpal/> for a detailed list of genes), including all genes necessary for growth with CO₂ as the sole carbon source; about 15% of the genome is devoted to transport (Larimer *et al.*, 2004). *R. palustris* is ideally suited for use as a biocatalyst and it might be possible to bioengineer this organism to produce large amounts of H₂ from plant biomass (Larimer *et al.*, 2004).

P. marinus is a dominant part of the phytoplankton in oceans and is responsible for a large fraction of photosynthesis globally. The three *P. marinus* genomes sequenced vary in size from 1.66 to

2.41 Mbp, as shown in Table 1. The MED4 strain is adapted to high light intensities and has the smallest genome of any known oxygenic phototroph, whilst the MIT9313 strain is low-light adapted and has a larger genome (Rocap *et al.*, 2003). These two strains represent isolates with the largest evolutionary distance within the *Prochlorococcus* lineage (Rocap *et al.*, 2003). *P. marinus* strain SS120 can grow in very low light levels and as such represents an organism adapted to the extreme edge of an environmental niche (Dufresne *et al.*, 2003). It is interesting to step back and have a look at these genomes from just the point of view of examining their differences in size. *P. marinus* strain MIT9313 is almost half again as large as the other two strains. Furthermore, the *P. marinus* strain MED4 genome is about three times smaller than the *Gloeobacter violaceus* genome (Nakamura *et al.*, 2003), discussed in last month's 'Genome Update' (Ussery, 2004), and is four times smaller than the genome of the cyanobacterium *Anabaena nostoc* (Kaneko *et al.*, 2001). To date, eight cyanobacterial genomes have been sequenced (see <http://www.cbs.dtu.dk/services/GenomeAtlas/show-kingdom.php?kingdom=Bacteria&phyla=Cyanobacteria> for the full list).

Method of the month – genome size

This month, the 'method' of genome comparison is merely to have a look at the distribution of sizes of microbial

genomes. This is the first in a series of articles discussing each of the columns of numbers in Table 1. Sequenced prokaryotic genomes currently range in size from 490 885 bp for the *Nanoarchaeum equitans* genome to 9 105 828 bp for the *Bradyrhizobium japonicum* genome. As can be seen for the three *Prochlorococcus* genomes in Table 1, even for different strains of the same species, there can be considerable size variation. This is true for other bacteria as well; for example, *Escherichia coli* genomes can vary in size by more than 1 000 000 bp (Ochman & Jones, 2000); and the *Bacillus cereus* main chromosome contains a 2.4 Mbp stable core, while the remaining roughly 3 Mbp of the chromosome are highly variable (Carlson & Kolsto, 1994). Even within a given isolate, quite a bit of variation can occur. For example, the same clinical isolate of *Campylobacter jejuni* (NCTC 11168) differs in cell morphology, colonization abilities and microarray expression patterns, depending on the age of the culture from which the bacteria is grown (Gaynor *et al.*, 2004).

Although there are indeed large variations within prokaryotic genomes, in perspective, the 20-fold size difference seen is quite small, compared to the more than 1 000 000-fold size range found in eukaryotic microbial genomes (McGrath & Katz, 2004). Fig. 1 illustrates the genome

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

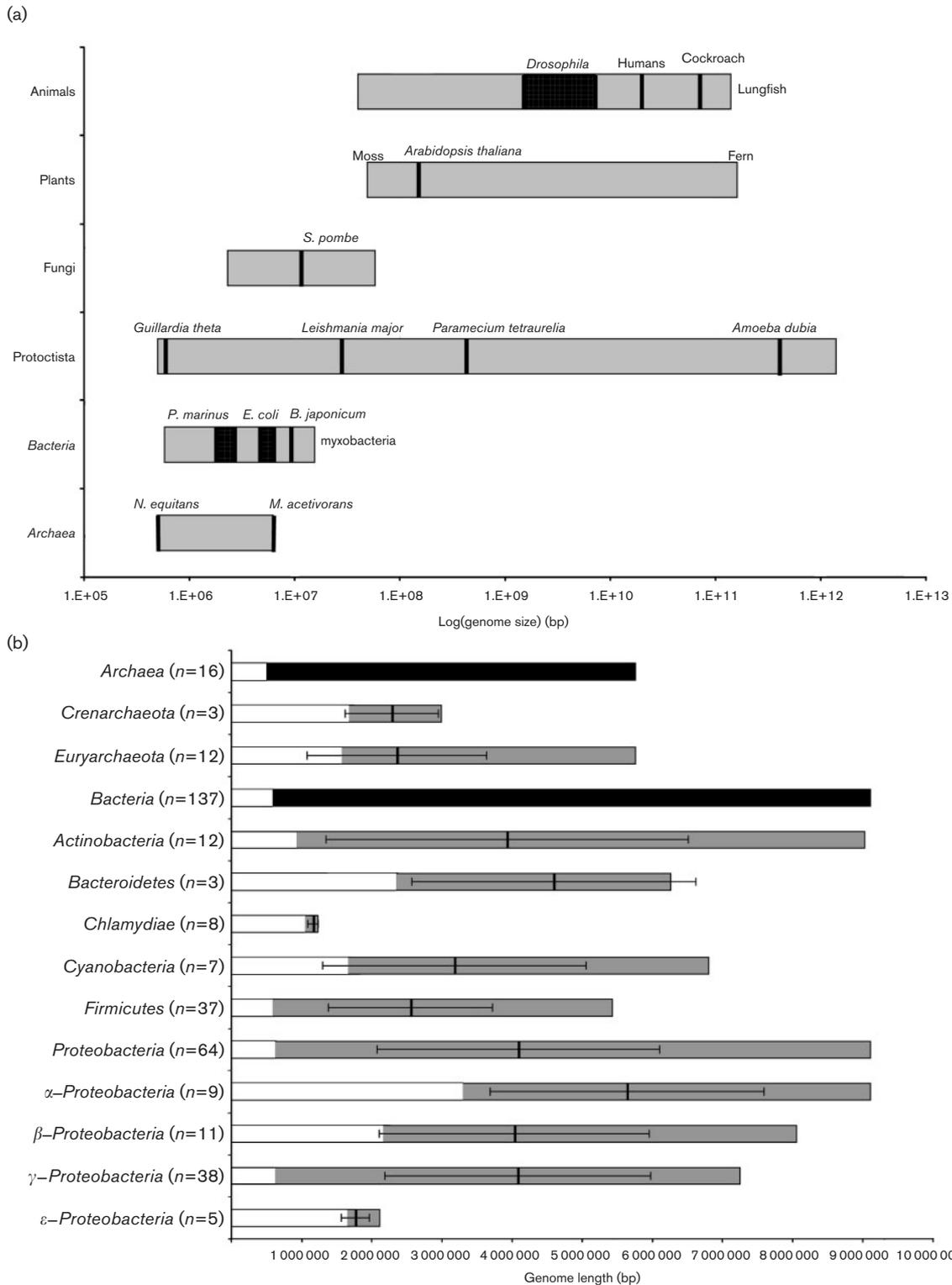


Fig. 1. (a) Estimates of genome sizes, based on data from DOGS – Database of Genome Sizes (<http://www.cbs.dtu.dk/databases/DOGS/>). Ranges for genomes from *Drosophila*, *E. coli* and *P. marinus* are shown, as discussed in the text. For simplicity this figure includes one line for the kingdom Protocista, which is defined as nucleated micro-organisms and their descendants, exclusive of fungi, animals and plants (Margulis & Schwartz, 1998). (b) Size ranges of sequenced prokaryotic genomes; 'n' is the number of sequenced genomes for the given phylum or group. The solid black lines refer to archaeal and bacterial genomes, whilst the shaded areas refer to genomes within a specific phylum. Note that this scale is linear, as opposed to the log scale in (a). *M. acetivorans*, *Methanosarcina acetivorans*.

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 DNA DataBase of Japan (DDBJ).

Genome	Size (bp)	AT content (%)	rRNA operons	tRNAs	CDS	Accession no.
<i>Rhodospseudomonas palustris</i> CGA009	5 459 213	35.0	2	49	4832	BX571963
<i>Prochlorococcus marinus</i> MED4	1 657 990	69.2	1	38	1716	BX548174
<i>Prochlorococcus marinus</i> MIT9313	2 410 873	49.3	2	44	2274	BX548175
<i>Prochlorococcus marinus</i> SS120	1 751 080	63.6	1	40	1882	AE017126

size range for various types of organisms; note that the genomes of microbial eukaryotes vary in size from that of a small bacterial genome (e.g. the 510 000 bp *Guillardia theta* genome) to several hundred times larger than the human genome (e.g. the 670 000 000 000 bp *Amoeba dubia* genome). For comparison, the sizes of plant and animal genomes are also included in Fig. 1. The range of sizes for various *Drosophila* genomes are shown, as well as the above-mentioned ranges for the *P. marinus* and *E. coli* genomes.

Fig. 1(b) shows the sizes of sequenced archaeal and bacterial genomes. Note that the scale for this plot is linear. The results are shown for all phyla containing three or more genomes. Thus, for the archaeal genomes, the (one) *Nanoarchaeota* genome is not shown. Members of the *Actinobacteria* have a wide range of genome sizes, and include two large *Streptomyces* genomes (8.7 and 9.0 Mbp). Surprising (at least in our opinion) is the observation that the *Firmicutes* have smaller genomes (an average of about 2.6 Mbp) than the *Proteobacteria* (average 4.1 Mbp). Although there are 37 *Firmicutes* genomes sequenced, the set chosen is still a bit biased, and thus might not be a good reflection of the genome size distribution actually found in nature. As an example of this, the δ/ϵ -proteobacterial genomes that have been sequenced are fairly small in size (average 1.8 Mbp), although the genomes of some of the myxobacteria (δ -*Proteobacteria*) have been estimated to be more than 12 Mbp long (Pradella *et al.*, 2002). This is a case where a bacterial genome is longer than the eukaryotic *Schizosaccharomyces pombe* genome (Wood *et al.*, 2002), with the bacteria encoding perhaps twice as many genes. Members of the α -*Proteobacteria* tend to have larger genomes (perhaps masked by the common occurrence of multiple chromosomes

and large plasmids in this subdivision of organisms), and include *R. palustris*, mentioned above, as well as the current largest sequenced bacterial genome (*B. japonicum*). On the other hand, the alpha group also contains the reduced *Rickettsia* genomes and is thought to be the progenitor of mitochondria. It is interesting to note that, in terms of the 'minimal set of genes' necessary for life, this experiment has been performed many times in the evolution of these endosymbiotic genomes (Klasson & Andersson, 2004).

In response to last month's column, people have asked about how our list of sequenced genomes is obtained, since we have extra genomes, in addition to the ones available from EMBL or GenBank. Basically, the genomes on our web page (<http://www.cbs.dtu.dk/services/GenomeAtlas/show-kingdom.php?kingdom=Bacteria&sortKey=DATESORT>) reflect those that are publicly available, either from EMBL or from a genome sequencing centre that is willing to allow us to download raw unannotated sequence files. Currently there are three places from which we download information: the Sanger Institute (<http://www.sanger.ac.uk/Projects/Microbes/>), the US DOE Joint Genome Institute (http://www.jgi.doe.gov/JGI_microbial/html/index.html) and the University of Oklahoma's Advanced Center for Genome Technology (<http://www.genome.ou.edu/>). We would, of course, be delighted to include more genomes from other places – suggestions are welcome!

Next month, the 'method' of genome comparison discussed will be the AT content, which currently varies from 27.9% for *Streptomyces coelicolor* to 77.5% for *Wigglesworthia glossinidia*, for the sequenced prokaryotic genomes publicly available. Obviously, the AT content is not homogeneously distributed throughout the

chromosome and it is known, for example, that promoter regions are in general more AT-rich than the genome average (Pedersen *et al.*, 2000).

Acknowledgements

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

David W. Ussery and Peter F. Hallin

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

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

Carlson, C. R. & Kolsto, A. B. (1994). A small (2.4 Mb) *Bacillus cereus* chromosome corresponds to a conserved region of a larger (5.3 Mb) *Bacillus cereus* chromosome. *Mol Microbiol* **13**, 161–169.

Dufresne, A., Salanoubat, M., Partensky, F. & 18 other authors (2003). Genome sequence of the cyanobacterium *Prochlorococcus marinus* SS120, a nearly minimal oxyphototrophic genome. *Proc Natl Acad Sci U S A* **100**, 10020–10025.

Gaynor, E. C., Cawthraw, S., Manning, G., MacKichan, J. K., Falkow, S. & Newell, D. G. (2004). The genome-sequenced variant of *Campylobacter jejuni* NCTC 11168 and the original clonal clinical isolate differ markedly in colonization, gene expression, and virulence-associated phenotypes. *J Bacteriol* **186**, 503–517.

Kaneko, T., Nakamura, Y., Wolk, C. P. & 19 other authors (2001). Complete genomic sequence of the filamentous nitrogen-fixing cyanobacterium *Anabaena* sp. strain PCC 7120. *DNA Res* **8**, 205–213, 227–253.

Klasson, L. & Andersson, S. G. (2004). Evolution of minimal-gene-sets in host-dependent bacteria. *Trends Microbiol* **12**, 37–43.

Larimer, F. W., Chain, P., Hauser, L. & 16 other authors (2004). Complete genome sequence of the metabolically versatile

photosynthetic bacterium *Rhodospseudomonas palustris*. *Nat Biotechnol* **22**, 55–61.

Margulis, L. & Schwartz, K. V. (1998).

FIVE KINGDOMS – An Illustrated Guide to the Phyla of Life on Earth. New York: W. H. Freeman & Co.

McGrath, C. L. & Katz, L. A. (2004). Genome diversity in microbial eukaryotes. *Trends Ecol Evol* **19**, 32–38.

Nakamura, Y., Kaneko, T., Sato, S. & 16 other authors (2003). Complete genome structure of *Gloeobacter violaceus* PCC 7421, a cyanobacterium that lacks thylakoids. *DNA Res* **10**, 137–145.

Ochman, H. & Jones, I. B. (2000). Evolutionary dynamics of full genome content in *Escherichia coli*. *EMBO J* **19**, 6637–6643.

Pedersen, A. G., Jensen, L. J., Brunak, S., Staerfeldt, H. H. & Ussery, D. W. (2000). A DNA structural atlas for *Escherichia coli*. *J Mol Biol* **299**, 907–930.

Pradella, S., Hans, A., Sproer, C., Reichenbach, H., Gerth, K. & Beyer, S. (2002). Characterisation, genome size and genetic manipulation of the myxobacterium *Sorangium cellulosum* So ce56. *Arch Microbiol* **178**, 484–492.

Rocap, G., Larimer, F. W., Lamerdin, J. & 21 other authors (2003). Genome divergence in two *Prochlorococcus* ecotypes reflects oceanic niche differentiation. *Nature* **424**, 1042–1047.

Ussery, D. W. (2004). Genome Update: 161 prokaryotic genomes sequenced, and counting. *Microbiology* **150**, 261–263.

Wood, V., Gwilliam, R., Rajandream, M. A. & 132 other authors (2002). The genome sequence of *Schizosaccharomyces pombe*. *Nature* **415**, 871–880.

DOI 10.1099/mic.0.27032-0