microbiology
Genome Update

Genome update: the 1000th genome – a cautionary tale

There are now more than 1000 sequenced prokaryotic genomes deposited in public databases and available for analysis. Currently, although the sequence databases GenBank, DNA Database of Japan and EMBL are synchronized continually, there are slight differences in content at the genomes level for a variety of logistical reasons, including differences in format and loading errors, such as those caused by file transfer protocol interruptions. This means that the 1000th genome will be different in the various databases. Some of the data on the highly accessed web pages are inaccurate, leading to false conclusions for example about the largest bacterial genome sequenced. Biological diversity is far greater than many have thought. For example, analysis of multiple *Escherichia coli* genomes has led to an estimate of around 45 000 gene families – more genes than are recognized in the human genome. Moreover, of the 1000 genomes available, not a single protein is conserved across all genomes. Excluding the members of the *Archaea*, only a total of four genes are conserved in all bacteria: two protein genes and two RNA genes.

Introduction

Sometime in October or November 2009, depending on which database is consulted, the 1000th prokaryotic genome sequence was completed. This landmark followed within 15 years of the sequencing of the first bacterial genomes, *Haemophilus influenzae* and *Mycoplasma genitalium* (Fleischmann *et al.*, 1995; Fraser *et al.*, 1995). The generation of these and other early genome sequences were costly in terms of both time and money, and the efforts were justly rewarded by patents and publications in the journals *Science* and *Nature*. Much has changed since then, not least by breathtaking technical innovations in sequencing procedures (the end of which is not yet in sight) supported by steadily increasing computer power and constantly improving software (Ansorge, 2009; Kyrpides, 2009). A bacterial genome can now be completely sequenced, assembled and annotated in less than 24 h (Flicek & Birney, 2009; Reeves *et al.*, 2009). The necessary final steps required for full closure of all the gaps and quality assessment of the annotation are, however, still time consuming, although it is possible, even with the current technology, to completely assemble a bacterial genome based on a single run from a 'next generation' machine (Tauch *et al.*, 2008). The imminent 'third generation' sequencing machines are promising vast improvements and one can envision a time in the not-too-distant future when it will be routine to sequence and annotate several bacterial genomes before the morning tea break.

As every microbiologist knows, there has been a dramatic increase in the number of sequenced microbial genomes over the past decade; this is illustrated for prokaryotic genomes in Fig. 1. The exponential growth in the number of finished genomes per year seemed to reach a peak of approximately 180 per year by 2007, and then declined slightly for 2008 and 2009. Have we reached a 'stationary phase' of bacterial genome sequencing? There are possible explanations for this decline, and one key observation is that the number of unfinished genomes deposited in GenBank now is larger than the number of complete genomes, as the production of a 'rough draft' has become relatively inexpensive. As of early January 2010, there are 1024 complete genomes listed at NCBI and more than twice as many genomes (2307) listed as 'in progress'; the Genomes Online Database (GOLD) web pages (see below and link in Table 1) boast more than 6400 microbial genome sequencing projects.

The 1000th genome(s)

In principle, there should be a list where one could go to find the 1000th genome; however, as several genomes are processed and submitted on an almost daily basis to databases, determining the 1000th genome is not as easy as might appear at first. Table 1 lists the set of genomes for the various databases. According to GOLD (which is perhaps one of the best centralized locations for keeping track of this), the 1000th genome is an archaeon: *Methanocaldococcus vulcanius* strain M7 (accession CP001787), sequenced by the Joint Genomes Institute (JGI). The honour of being number 1000 is not as clear for the other databases. For the DNA Database of Japan (DDBJ), there are three candidate genomes which were all added on the same day; these have also been sequenced by the JGI. One of these is another *Methanocaldococcus* (this time *M. fervens*), which has not yet been published (CP001696). The other two genomes are part of the Genomic Encyclopedia of Bacteria and Archaea (GEBA) project (Wu *et al.*, 2009). *Kangiella koreensis* is a member of the phylum *Gammaproteobacteria*, belonging to the order *Oceanospirillales*; this genome is from the type strain (DSM 16069), isolated from a beach in South Korea (Han *et al.*, 2009). *Slackia heliotrinireducens* DSM 20476 is a member of the phylum *Actinobacteria*, belonging to the uncharacterized family *Coriobacteriaceae* (Pukall *et al.*, 2009).

The '1000th genome' from EMBL is a set of *Acetobacter pasteurianus* gen-

 Printed in Great Britain

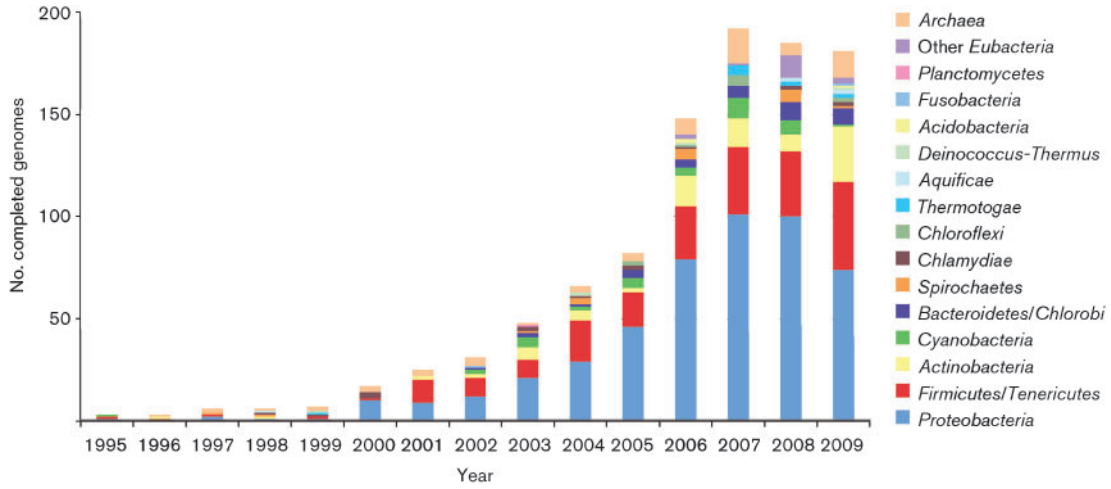

Fig. 1. Increase in the number of genomes completed per year separated by bacterial phylum. Data source: NCBI, complete genomes (http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi).

omes from eight different strains. *Acetobacter* species have been used historically for making vinegar and are known to be quite variable in terms of their genome content, due to a large number of transposons (Azuma *et al.*, 2009). Although the eight *A. pasteurianus* strains are also present in DDBJ and GenBank, only one sequence is found in the 'complete genomes' list on the NCBI web pages, and the other seven genomes, although finished to one contiguous piece, are listed as 'in progress' more than 6 months after being deposited in the International Nucleotide Sequence Database.

Finally, we have listed (Table 1) a set of seven genomes deposited in GenBank around the same time. At one stage, *Vibrio* sp. isolate Ex25, was listed as the 1000th genome on the NCBI pages, although with time, as some older genomes have been removed and the lists have been updated, the 1000th position has moved a bit, so a range of genomes is given. Of the seven genomes in the list, five have at least one publication associated with them at the time of writing; *Vibrio* sp. Ex25 (CP001805) and *Staphylococcus aureus* ED98 (CP001781) genomes have not been described in a publication. Three of the genomes are from the GEBA project (Wu *et al.*, 2009): *Rhodothermus marinus* strain R-10T is a member of the phylum *Bacteroidetes* and was isolated from hot springs off the coast of Iceland (Nolan *et al.*, 2009); *Gordonia bronchialis* DSM 43247 is a member of the phylum *Actinobacteria*, isolated from the sputum of a woman with diseased lungs; and *Haliangium ochraceum* DSM 14365 is a halophilic deltaproteobacterium isolated from coastal sand in Japan. The *Blattabacterium* genome sequenced, which is from a cockroach endosymbiotic strain, belongs to the class *Flavobacteria* in the phylum *Bacteroidetes*; the newly sequenced genome shows evolutionary convergence with gammaproteobacterium endosymbionts (López-Sánchez *et al.*, 2009). Finally, *Comamonas testosteroni* strain CNB-2 was isolated from soil contaminated with 4-chloronitrobenzene and can grow on this pollutant using it as its sole carbon and nitrogen source (Ma *et al.*, 2009). This betaproteobacterium is a member of the order *Burkholderiales*, and was given its name from its ability to metabolize testosterone.

Of the 1000 prokaryotic genomes sequenced so far, only 7 % are archaea, with the rest (93 %) being bacteria. Whether this ratio is truly reflective of the relative proportions in the environment is doubtful. Within the bacteria, members of the phyla *Proteobacteria* and *Firmicutes* make up the majority of the completed genome sequences, and many of the new genomes sequenced each year are found within these two phyla. Together, these account for 72 % (of 930) of bacterial genomes, with 489 (52 %) proteobacteria and 187 (20 %) firmicute genomes (GOLD data). A similar skewed distribution is observed for the archaeal genomes, where the majority are from the phylum *Euryarchaeota* (63 %); with nearly all the rest coming from the phylum *Crenarchaeota* (31 %). The phylum *Nanoarchaeota* has only a single sequenced genome. Such an uneven distribution in available sequences has consequences for the chance that a particular query sequence will identify similarities in the database: when a query sequence from a proteobacterial genome is used to search the microbial database using BLAST, the chance is much higher that a hit will be found than, for instance, searching with a nanoarchaeota sequence. Since the E-value reported by BLAST is based on the expected background noise, and is based on the assumption that the sequences within the database are random (which obviously is not true in this case), the results should be interpreted with caution when the 'best hit' of a

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

The accession number for each chromosome used in the International Nucleotide Sequence Database (INSD; www.insdc.org) is the same for the DDBJ, EMBL and GenBank databases.

Strain	Taxonomic phylum	Length (bp)	AT content (%)	No. genes	No. rRNAs	No. tRNAs	Accession no.
GOLD							
Methanocaldococcus vulcanius M7	*Euryarchaeota*	1 761 737	68.4	1742	2	37	CP001787
DDBJ*							
Slackia heliotrinireducens DSM 20476	*Actinobacteria*	3 165 038	39.8	2765	2	48	CP001684
Methanocaldococcus fervens AG86	*Euryarchaeota*	1 507 251	67.8	1581	2	37	CP001696
Kangiella koreensis DSM 16069	*Alphaproteobacteria*	2 852 073	56.3	2632	2	41	CP001707
EMBL†							
Acetobacter pasteurianus IFO 3283-01	*Alphaproteobacteria*	3 340 249	46.9	3050	5	57	GPID 31129‡
A. pasteurianus IFO 3283-01	*Alphaproteobacteria*	2 907 495	50.0	2628	5	57	AP011121
A. pasteurianus IFO 3283-01-42C	*Alphaproteobacteria*	2 815 241	46.9	2562	4	54	AP011163
A. pasteurianus IFO 3283-03	*Alphaproteobacteria*	2 907 287	50.0	2627	5	57	AP011128
A. pasteurianus IFO 3283-07	*Alphaproteobacteria*	2 906 044	50.0	2626	5	57	AP011135
A. pasteurianus IFO 3283-22	*Alphaproteobacteria*	2 907 267	50.0	2627	5	57	AP011142
A. pasteurianus IFO 3283-26	*Alphaproteobacteria*	2 907 309	50.0	2627	5	57	AP011149
A. pasteurianus IFO 3283-32	*Alphaproteobacteria*	2 904 642	50.0	2625	5	57	AP011156
A. pasteurianus IFO 3283-12	*Alphaproteobacteria*	2 904 624	50.0	2625	5	57	AP011170
GenBank§							
Rhodothermus marinus DSM 4252	*Bacteroidetes*	3 386 737	35.7	2863	1	45	CP001807
Vibrio sp. Ex25	*Gammaproteobacteria*	5 089 025	55.1	4518	11	124	CP001805
Blattabacterium sp.	*Bacteroidetes*	636 850	72.9	587	1	34	CP001487
Comamonas testosteroni CNB-2	*Betaproteobacteria*	5 373 643	38.6	4803	3	79	CP001220
Gordonia bronchialis DSM 43247	*Actinobacteria*	5 290 012	33.0	4696	2	49	CP001802
Haliangium ochraceum DSM 14365	*Deltaproteobacteria*	9 446 314	30.5	6719	2	46	CP001804
Staphylococcus aureus ED98	*Firmicutes*	2 847 542	67.2	2689	5	61	CP001781

*DDBJ, http://gib.genes.nig.ac.jp

†EMBL, http://www.ebi.ac.uk/genomes/bacteria.html

‡GPID is NCBI's Genome Project ID, a genome identifier unique to NCBI and not part of the INSD. The GPID contains a combination of the main chromosome plus six plasmids. The data shown in the table for the other *A. pasteurianus* genomes are for the main chromosome only, although each of these genomes also contains six plasmids in addition to the main chromosome.

§GenBank (also referred to as 'NCBI web pages'), http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi

nanoarchaoeta gene is to a proteobacteria gene.

There are 612 unique species represented within the 1000 genomes. Thus, for many species there are multiple genomes sequenced, usually from multiple strains, and if this redundancy is removed from the data, the coverage of genomes per phylum does not significantly change the distribution; this is true for data from both the GOLD and the NCBI databases. The species for which the most genomes are completed is currently *Escherichia coli*, for which 32 genomes have been completely sequenced. Much of the emphasis so far in bacterial genomics is on (human) pathogenic bacteria, as reflected in the species for which multiple strains have been sequenced, although it is likely that these represent only a fraction of the true bacterial diversity on Earth.

Numerical characteristics of sequenced genomes

One thousand bacterial genomes represent a few billion nucleotides and a few million genes. To analyse such a wealth of information would require immense computing time, so that the 2000th genome would probably be sequenced prior to any in-depth analysis of the first 1000 being completed. Therefore, we have focussed our efforts on a few basic analyses, details of which are described elsewhere (Ussery *et al.*, 2009). A striking difference is noted in the variation of genome size within phyla: whereas members of the phyla *Proteobacteria*, *Cyanobacteria*, *Actinobacteria*, *Chloroflexi* and *Bacteroidetes/Chlorobi* display a wide variation in genome size, those of the phyla *Thermotoga*, *Fusobacteria* and *Aquificae* show little differences in size within the same phylum. (A box and whiskers plot is provided in Supplementary Figure S1, available in Microbiology Online, which is an updated version of Fig. 1 from Ussery & Hallin, 2004a.) Some of this conservation in genome size could be

due to the restricted number of genomes available, although three acidobacteria genomes already display a large variation in genome size variation. Compared with the wide variety in genome size within bacterial phyla, especially when outliers are considered, the variation between phyla is less dramatic. As expected, intracellular parasites, such as members of the phyla *Chlamydiae* and *Thermotogae*, typically have small genomes (1–2 Mbp), whereas those of the phyla *Proteobacteria* and *Chloroflexi* usually have larger genomes (~4 Mb on average). The record holders are currently *Sorangium cellulosum* strain 'So ce 56', a deltaproteobacterium that can produce several bioactive compounds, with a genome of 13 Mbp, and *Candidatus Hodgkinia cicadicola*, an alphaproteobacterium, with a genome of only 143 kbp. Note: the NCBI pages list two larger bacterial genomes, but closer inspection shows that the sizes of these genomes are much smaller than indicated. The GOLD web pages list the largest genome as *Solibacter usitatus* Ellin6076, at 9.965 Mbp, and the second largest as *Mycoplasma gallisepticum* R, which is only 0.996 Mbp (apparently this field is sorted by text rather than numbers). Currently, both EMBL and DDBJ cannot be sorted by genome size on their genome web pages. Thus, the size range between the largest and smallest bacterial genome is approaching 100-fold.

It is also interesting to observe how the average length of genomes being sequenced has increased over the years. Until 2001, the average genome being sequenced was 2.4 Mbp, after that, the average increased to 3.7 Mbp – a shift that probably reflects the improved technology in generating, handling and assembling large genomes, and the decreased costs so that size is now less of a limiting factor.

A second characteristic given by a numerical value is GC content, which again varies between phyla, as discussed in a previous Genome Update article (Ussery & Hallin, 2004b; see Supplementary Figure S2 for an updated version of the box-and-whiskers plot). The phylum *Proteobacteria* contains genomes with widely varying GC content whereas members of the phyla *Actinobacteria* and *Firmicutes* have a much narrower GC content distribution. Members of other phyla commonly have a GC content somewhere between 40 and 50 %, with the exception of phyla for which only a few genomes have so far been sequenced, in which case a skew due to limited data cannot be ruled out.

Superficially, there seems to be a correlation between GC content and genome size. For instance, using the NCBI data, an increase in GC content of roughly 3.7 % GC per Mbp can be observed. Examining the phyla for which we have the most genomes, this tendency was also seen among the proteobacteria (an increase of 5 % GC per Mbp) and the actinobacteria (an increase of 2 % per Mbp). However, a significant relationship was not found among the firmicutes or the archaea. The observed relationship between GC content and genome size, within and between some phyla, could be the result of differences in lifestyle. Many intracellular bacteria with small genomes also have a low GC content (Moran & Baumann, 2000). Since organisms that encounter multiple variable conditions would also require more genes to cope with this variability, genome size could also correlate with ecological niche, so that genome size correlates to some extent with gene content, assuming a more or less invariable gene density.

Fig. 2 shows the variation of gene density per genome among the phyla. As is evident, the gene density does not vary much for the majority of genomes, with a mean of 918 genes per Mbp. The gene density is also quite uniform within each phylum; the only one that deviates from this to any extent is that of the phylum *Chloroflexi*. Most of the genomes within this phylum have a density of around 770 genes per Mbp; however, there are three genomes in the phylum *Dehalococcoides* that have densities around 1050 genes per Mbp. These genomes are also quite short, with an average length of 1.4 Mbp.

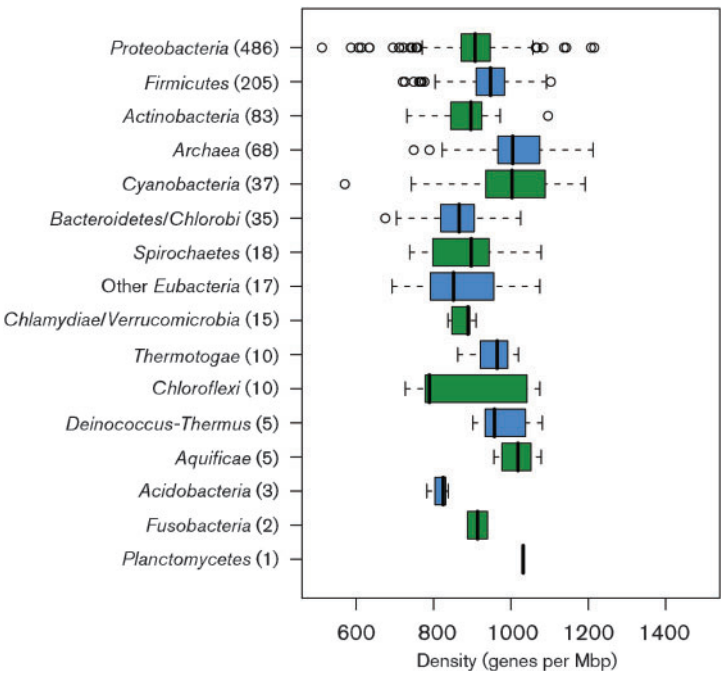

Fig. 2. Box-and-whiskers plot of gene density of prokaryotes per phylum separated by bacterial phylum. Data are from NCBI.

Comparison of gene content

With the release of the complete genome of *M. genitalium*, the quest to identify the prokaryotic genome with the smallest possible number of genes began (Fraser *et al.*, 1995). Some endosymbiotic bacteria have lost a large fraction of their genes as their host provides them with essential biomolecules (Moya *et al.*, 2009). This particularly applies to intracellular bacteria living in symbiosis with an insect host which can have severely reduced genomes specialized to produce nutrients that are not included in the host's diet (McCutcheon *et al.*, 2009). The publication of the extremely small *Carsinella rudii* genome, with only 182 predicted protein-coding genes (Nakabachi *et al.*, 2006), started a discussion on whether or not to accept this as the genome of a true living cell: it was argued that genes coding for a complete machinery for DNA replication, transcription and translation, and a (simplified) metabolic network for energy production would have to be present, but a number of essential genes for these processes are missing in *C. rudii* (Tamames *et al.*, 2007). However, there is no fundamental biological difference between a symbiont depending on its host, for instance for a particular amino acid or an enzyme needed to supercoil its DNA; the requirement to independently produce its own DNA, RNA and proteins would not be met either way. We therefore consider that the genome of *C. rudii* represents that of a prokaryotic cell as validly as any other. The even smaller genome of *Hodgkinia cicadicola* (which, as an exception to the rule, has a very small genome but a high GC content) has specialized to produce vitamin B12 for its insect host, for which it reserves 7 % of its proteome that in total only encodes 169 proteins (McCutcheon *et al.*, 2009).

Another goal of comparative genomics has been to recognize those protein-coding genes that are conserved in all prokaryotic genomes, as these would represent the ultimate core genome of bacteria. The outcome of such analyses obviously depends on the criteria applied to gene definition, to orthologous inclusion as well as on the decision to include reduced symbiont genomes. When the genomes of *H. influenzae* and *M. genitalium* were compared in 1996, a total of 256 genes were recognized as conserved (Mushegian & Koonin, 1996) and this number was reduced to 179 genes when five endosymbiont genomes were added (Gil *et al.*, 2003). Adding *Rickettsia prowazekii* and *Chlamydia trachomatis* genomes reduced the number to 156 conserved orthologous genes (Klasson & Andersson, 2004). This number is certainly expected to further decrease if all 1000 genomes are included (it is acknowledged here that this analysis ignores functionally equivalent genes that lack sequence similarity, as noted by Gil *et al.*, 2004). Only 31 proteins were found to be conserved across 191 species (Ciccarelli *et al.*, 2006). How many proteins do we find conserved across all bacteria? The results might be surprising to some.

For reasons of computational simplicity, we took the described genes of the two smallest symbiont genomes (*C. rudii* and *H. cicadicola*) as a starting point, and compared these with the two largest genomes currently available per bacterial phylum, including the *Archaea*; this resulted in a set of 32 genomes from 16 different phyla. From this artificial subset of 32 genomes, a core genome analysis was performed using a similarity cutoff of at least 50 % identity at the protein level, over at least 50 % of the longest gene in reciprocal pairwise analysis (as described by Tettelin *et al.*, 2005). For the entire set of 1000 genomes, not a single protein-coding gene was found to be conserved. When we exclude members of the *Archaea*, two protein-coding genes were conserved, these were the translation elongation factor EF-Tu and the ribosomal protein S12. Furthermore, the 16S and 23S rRNAs were found to be conserved.

Finally, we compared the first two genomes sequenced, those of *M. genitalium* and *H. influenzae*, with the genomes listed in Table 1. As expected, all genomes of strains in the genus *Methanocaldococcus* were extremely similar, sharing >98 % of their genes in any pairwise comparison (see Supplementary Figure S3). The three gammaproteobacteria (*Vibrio* sp., *H. influenzae* and *K. koreensis*) share 10–15 % of their genes, but there are only nine genes shared between *Methanocaldococcus* and *H. influenzae* Rd, and only two genes are shared between *Methanocaldococcus* and *M. genitalium*. All other combinations of genomes share between 0.2 and 6 % of their genes. The conclusion is that a minimal gene set required for life depends, as does everything else in biology, on the context. There is certainly a conserved set of gene functions necessary for life, but this does not necessarily correlate with a set of conserved gene sequences.

Conclusion

The number of sequenced genomes is quickly expanding, and although there is an International Nucleotide Sequence Database, it appears to not be working so well for genome sequences. Currently, the number of genomes available is a bit more than 1000, but none of the three databases (DDBJ, EMBL or GenBank) contain all of the genomes. Further, some of the web pages contain inaccurate information. There is a real need for a clearly visible place for microbiologists to go to obtain up-to-date information about genome sequences. Currently, there are three different formats for the three different databases, and unfortunately these three databases are not synchronized on a regular basis. Further, rules about when a genome is 'fully sequenced' should be further explored (Kyrpides, 2009; Chain *et al.*, 2009). This is essential for a full and reliable comparison of the vast genomic resources available. What is already obvious, though, is an incredible diversity within these genomes, with essentially no genes being fully conserved across all prokaryotes, using the same standard criteria as used by others (Tettelin *et al.*, 2005). Perhaps it is best to consider using alternative methods of defining gene families, in some way based more on conserved function.

Supplementary data

Supplementary figures showing box-and-whiskers plots of the genome size of prokaryotes per phylum and of the GC content of prokaryotes per phylum, and of the BLAST matrix of the genomes from Table 1 are available with the online version of this paper.

Karin Lagesen,[1,2] Dave W. Ussery[1] and Trudy M. Wassenaar[1,3]

[1]Center for Biological Sequence Analysis, Department of Systems Biology, The Technical University of Denmark, 2800 Lyngby, Denmark

[2]Centre for Molecular Biology and Neuroscience, Institute of Medical Microbiology, Oslo University Hospital, Rikshospitalet, NO-0027, Oslo, Norway, and Department of Informatics, University of Oslo, PO Box 1080 Blindern, NO-0316, Oslo, Norway

[3]Molecular Microbiology and Genomics Consultants, Zotzenheim, Germany

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

Ansorge, W. J. (2009). Next-generation DNA sequencing techniques. *Nat Biotechnol* **25**, 195–203.

Azuma, Y., Hosoyama, A., Matsutani, M., Furuya, N., Horikawa, H., Harada, T., Hirakawa, H., Kuhara, S., Matsushita, K. & other authors (2009). Whole-genome analyses reveal genetic instability of *Acetobacter pasteurianus*. *Nucleic Acids Res* **37**, 5768–5783.

Chain, P. S., Grafham, D. V., Fulton, R. S., Fitzgerald, M. G., Hostetler, J., Muzny, D., Ali, J., Birren, B., Bruce, D. C. & other authors (2009). Genomics. Genome project standards in a new era of sequencing. *Science* **326**, 236–267.

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

Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., Kerlavage, A. R., Bult, C. J., Tomb, J. F., Dougherty, B. A. & other authors (1995). Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* **269**, 496–512.

Flicek, P. & Birney, E. (2009). Sense from sequence reads: methods for alignment and assembly. *Nat Methods* **6** (Suppl. 11), S6–S12.

Fraser, C. M., Gocayne, J. D., White, O., Adams, M. D., Clayton, R. A., Fleischmann, R. D., Bult, C. J., Kerlavage, A. R., Sutton, G. & other authors (1995). The minimal gene complement of *Mycoplasma genitalium*. *Science* **270**, 397–403.

Gil, R., Silva, F. J., Zientz, E., Delmotte, F., González-Candelas, F., Latorre, A., Rausell, C., Kamerbeek, J., Gadau, J. & other authors (2003). The genome sequence of *Blochmannia floridanus*: comparative analysis of reduced genomes. *Proc Natl Acad Sci U S A* **100**, 9388–9393.

Gil, R., Silva, F. J., Peretó, J. & Moya, A. (2004). Determination of the core of a minimal bacterial gene set. *Microbiol Mol Biol Rev* **68**, 518–537.

Han, C., Sikorski, J., Lapidus, A., Nolan, M., Del Rio, T. G., Tice, H., Cheng, J.-F., Lucas, S., Chen, F. & other authors (2009). Complete genome sequence of *Kangiella koreensis* type strain (SW-125^T). *Stand Genomic Sci* **1**, 3. doi:10.4056/sigs.36635.

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

Kyrpides, N. C. (2009). Fifteen years of microbial genomics: meeting the challenges and fulfilling the dream. *Nat Biotechnol* **27**, 627–632.

López-Sánchez, M. J., Neef, A., Peretó, J., Patiño-Navarrete, R., Pignatelli, M., Latorre, A. & Moya, A. (2009). Evolutionary convergence and nitrogen metabolism in *Blattabacterium* strain Bge, primary endosymbiont of the cockroach *Blattella germanica*. *PLoS Genet* **5**, e1000721.

Ma, Y. F., Zhang, Y., Zhang, J. Y., Chen, D. W., Zhu, Y., Zheng, H., Wang, S. Y., Jiang, C. Y., Zhao, G. P. & Liu, S. J. (2009). The complete genome of *Comamonas testosteroni* reveals its genetic adaptations to changing environments *Appl Environ Microbiol* **75**, 6812–6819.

McCutcheon, J. P., McDonald, B. R. & Moran, N. A. (2009). Convergent evolution of metabolic roles in bacterial co-symbionts of insects. *Proc Natl Acad Sci U S A* **106**, 15394–15399.

Moran, N. A. & Baumann, P. (2000). Bacterial endosymbionts in animals. *Curr Opin Microbiol* **3**, 270–275.

Moya, A., Gil, R. & Latorre, A. (2009). The evolutionary history of symbiotic associations among bacteria and their animal hosts: a model. *Clin Microbiol Infect* **15** (Suppl. 1), 11–13.

Mushegian, A. R. & Koonin, E. V. (1996). A minimal gene set for cellular life derived by comparison of complete bacterial genomes. *Proc Natl Acad Sci U S A* **93**, 10268–10273.

Nakabachi, A., Yamashita, A., Toh, H., Ishikawa, H., Dunbar, H. E., Moran, N. A. & Hattori, M. (2006). The 160-kilobase genome of the bacterial endosymbiont *Carsonella*. *Science* **314**, 267.

Nolan, M., Tindall, B. J., Pomrenke, H., Lapidus, A., Copeland, A., Del Rio, T. G., Lucas, S., Chen, F., Tice, H. & other authors (2009). Complete genome sequence of *Rhodothermus marinus* type strain (R-10^T). *Stand Genomic Sci* **1**, 3. doi:10.4056/sigs.42644.

Pukall, R., Lapidus, A., Nolan, M., Copeland, A., Del Rio, T. G., Lucas, S., Chen, F., Tice, H., Cheng, J.-F. & other authors (2009). Complete genome sequence of *Slackia heliotrinireducens* type strain (RSH 1^T). *Stand Genomic Sci* **1**, 3. doi:10.4056/sigs.37633.

Reeves, G. A., Talavera, D. & Thornton, J. M. (2009). Genome and proteome annotation: organization, interpretation and integration. *J R Soc Interface* **6**, 129–147.

Tamames, J., Gil, R., Latorre, A., Peretó, J., Silva, F. J. & Moya, A. (2007). The frontier between cell and organelle: genome analysis of *Candidatus Carsonella ruddii*. *BMC Evol Biol* **7**, 181.

Tauch, A., Schneider, J., Szczepanowski, R., Tilker, A., Viehoever, P., Gartemann, K. H., Arnold, W., Blom, J., Brinkrolf, K. & other authors (2008). Ultrafast pyrosequencing of *Corynebacterium kroppenstedtii* DSM44385 revealed insights into the physiology of a lipophilic corynebacterium that lacks mycolic acids. *J Biotechnol* **136**, 22–30.

Tettelin, H., Masignani, V., Cieslewicz, M. J., Donati, C., Medini, D., Ward, N. L., Angiuoli, S. V., Crabtree, J., Jones, A. L. & other authors (2005). Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial "pan-genome". *Proc Natl Acad Sci U S A* **102**, 13950–13955.

Ussery, D. W. & Hallin, P. F. (2004a). Genome update: length distributions of sequenced prokaryotic genomes. *Microbiology* **150**, 513–516.

Ussery, D. W. & Hallin, P. F. (2004b). Genome update: AT content in sequenced prokaryotic genomes. *Microbiology* **150**, 749–752.

Ussery, D. W., Wassenaar, T. M. & Borini, S. (2009). *Computing for Comparative Microbial Genomics: Bioinformatics for Microbiologists*. London, UK: Springer.

Wu, D., Hugenholtz, P., Mavromatis, K., Pukall, R., Dalin, E., Ivanova, N. N., Kunin, V., Goodwin, L., Wu, M. & other authors (2009). A phylogeny-driven genomic encyclopaedia of Bacteria and Archaea. *Nature* **462**, 1056–1060.

DOI 10.1099/mic.0.038257-0